



**THE** SPECIAL ISSUE  
**IIOAB**  
**JOURNAL**

**VOLUME 7 : NO 9 : NOVEMBER 2016 : ISSN 0976-3104**

**Institute of Integrative Omics and  
Applied Biotechnology Journal**

*Dear Colleagues,*

*I am thrilled to extend a warm welcome to each of you to our distinguished scientific journal dedicated to the exploration of Emerging Technologies in Networking and Security.*

*As the Editor of this special IIOAB Journal issue, I am excited to witness the rapid evolution and convergence of advanced technologies in the domains of networking and security. Our journal stands as a platform for disseminating cutting-edge research, innovative methodologies, and transformative applications that redefine the paradigms of networking architecture and cybersecurity.*

*Your expertise, dedication, and scholarly contributions play a pivotal role in shaping the future landscape of networking and security technologies. Your groundbreaking research, novel algorithms, and visionary insights contribute significantly to the development of robust and secure networking infrastructures, addressing the ever-growing challenges posed by cyber threats and ensuring the integrity of information systems.*

*The dynamic nature of this field presents both challenges and opportunities for innovation. Your commitment to exploring emerging technologies not only drives technological advancements but also holds immense potential to safeguard critical infrastructures, preserve data privacy, and shape the future of digital connectivity.*

*I encourage each of you to share your visionary insights, submit your pioneering research, and actively engage in the vibrant discussions within our journal. Let us collaboratively nurture an environment where ideas flourish, collaborations thrive, and knowledge paves the way for the next generation of networking and security solutions.*

*Thank you for your unwavering dedication to advancing the frontiers of Emerging Technologies in Networking and Security. I eagerly anticipate the transformative discoveries and breakthroughs that will emerge from your invaluable contributions.*

*Warm regards,*

*Prof. B. Madhshudanan  
Editor-in-Chief*

## ACCIDENT PREVENTION IN VEHICLE WITH EFFECTIVE RESCUE OPERATION

K. Karthick<sup>1</sup>, T. Praveen<sup>2</sup>, R. Yogalakshmi<sup>3</sup>, R. Anushya<sup>4</sup>, Mehajabeen<sup>5</sup><sup>1</sup>Dept. of Computer Science and Engineering, Kongunadu College of Engineering and technology, Trichy. INDIA<sup>2</sup>Dept. of Computer Science and Engineering, Vel Tech High Tech Dr.RangarajanDr.Sakunthala Engineering College, Avadi, Chennai, INDIA<sup>3,4,5</sup>Dept. of Computer Science and Engineering, Vel Tech High Tech Dr.RangarajanDr.Sakunthala Engineering College, Avadi, Chennai, INDIA

## ABSTRACT

**Aims:** In recent times, drowsiness & sudden heart attack is one of the major cause for highway accidents. Accidents due to drowsiness and sudden heart attack are prevented and controlled when the vehicle is out of control. In this project, in addition to prevention mechanism we presented some immediate rescue operation by transmitting the driver status to nearby hospital or hotel using GPS technology (i.e. WSN inbuilt in vehicle). **Materials and methods:** The term used here for the recognition that the driver is drowsy or physically not good (i.e. low heart beat level) is by using eye blink sensor, heart beat sensor. These types of accidents occurred due to driver can't able to control the vehicle, when he/she wakes. When either heart beat sensor or eye blink sensor activates, then the vehicle is slow down by auto steering mechanism. In-addition to it, the WSN built-in the vehicle transmits the driver status (i.e.) drowsy or sudden stroke to connected server end (i.e. nearby hospital with the help of GPS technology). **Results:** In server side a simple VB application is created which displays the eye position and heart beat level along with latitudinal and longitudinal value of the vehicle. **Conclusion:** It is also possible to control the vehicle from server end that is by start and stop signal, this is done with the help of WSN. This method used to drive car safely and effectively by rescue operation. EALs are capable of determining the working length of the HRF and that Root ZX showed a higher accuracy rate in detection of simulated horizontal root fractures. It should be emphasised that the results obtained in this in vitro study cannot be directly extrapolated to the clinical situation, but can provide an objective examination of a number of variables that are not practical to test clinically.

Published on: 08<sup>th</sup>- August-2016

## KEY WORDS

Drowsiness detection, heart beat level detection, WSN in vehicle, Automatic movement control, GPS technology.

\*Corresponding author: Email: [karthivel.me@gmail.com](mailto:karthivel.me@gmail.com)

## INTRODUCTION

“Driving to save lives, time, and money in spite of the conditions around you and the actions others.” This is the slogan for Defensive Driving. Vehicle accidents are most common if the driving is inadequate. These happen on most factors if the driver is drowsy or if he is alcoholic. Driver drowsiness is recognized as an important factor in the vehicle accidents. It was demonstrated that driving performance deteriorates with increased drowsiness with resulting crashes constituting more than 20% of all vehicle accidents. But the life lost once cannot be re-winded. Advanced technology offers some hope avoid these up to some extent. This project involves measure and controls the eye blink using IR sensor and heart beat level using PULSE OXIMETERSENSOR. TheIR transmitter is used to transmit the infrared rays in our eye. The IR receiver is used to receive the reflected infrared rays of eye. If the eye is closed means the output of IR receiver is high otherwise the IR receiver output is low. This to know the eye is closing or opening position. PULSE OXIMETER SENSOR is designed to give digital output of heart beat when a finger is placed on it. When the heart beat detector is working, the beat LED flashes in unison with each heartbeat. This digital output can be connected to microcontroller directly to measure the Beats per Minute (BPM) rate. It works on the principle of light modulation by blood flow through finger at each pulse. Output from both sensors are summed up and given to the logic circuit to indicate the final output (i.e.) to transmit driver status to connected nearby hospital or hotel (server side). By using wire-less technology if the driver gets a heart attack or he is drunk it will send signals to nearby hospital to indicate the driver status [1].

This project involves controlling accident due to unconscious through Eye blink and heart beat level. A car simulator study was designed to collect physiological data for validation of this technology. Methodology for analysis of physiological data, independent assessment of driver drowsiness and development of drowsiness detection algorithm by means of sequential fitting and selection of regression models is presented. We can automatically park the vehicle by first using Automatic braking system, which will slow down the vehicle and simultaneously will turn on the parking lights of the vehicle and will detect the parking space and will automatically park the car preventing from accident. By using wire-less technology such as Car Talk2000 if the driver gets a heart attack or he is drunk it will send signals to vehicles nearby about this so driver become alert. [2]

Sleep related accidents tend to be more severe, possibly because of the higher speeds involved and because the driver is unable to take any avoiding action, or even brake, prior to the collision. Horne describes typical sleep related accidents as ones where the driver runs off the road or collides with another vehicle or an object, without any sign of hard braking before the impact. Accidents are also caused when street lights are out especially on highways, long distance routes [3]. Here, usually the upper dipper lights are in upper mode. So, when the driver fails to change the mode of the light and at the same time when the car comes from the opposite side. it causes the opposite driver to miss the judgment and gives rise to accident. Accidents are also caused due to the intruders coming suddenly in either side of the vehicle i.e. front, left or right. Due to which the driver misses the judgment and meets with an accident. The Objective of this project is to develop a system to keep the vehicle and driver lifeseure and protect it by the occupation of the intruders [4].

**Contributions:** This paper makes the following contributions:

**Formal model of Vehicle section:** We develop a formal model of vehicle section which shows the overall model of the system. It contains heart beat sensor and eye blink sensor to monitor the humans.

**Modular analysis:** We show how to calculate the heart beat and to measure the eye movement of the human being.

**Implementation:** We develop a prototype model and if any abnormal movement is captured by the sensors and information are signals are transmitted to warn the human beings in order to avoid accidents.

**Experiments:** We present results from experiments from the overall hardware representation and if any accident is occurred the information if transmitted to the nearest hospital through GPS system.

## RELATED WORK

Driver drowsiness resulting in reduced vehicle control is one of the major causes of road accidents. Driving performance deteriorates with increased drowsiness with resulting crashes constituting 48% of all vehicle accidents. The vehicle crashes result in more than 1500 fatalities, 71 000 injuries, and an estimated \$12.5 billion in diminished productivity and property loss. Many efforts have been made recently to develop on-board detection of driver drowsiness. A number of approaches have been investigated and applied to characterize driver drowsiness using physiological

## RELATED WORKS ON VEHICLE CONTROLLING

A driver state of drowsiness can also be characterized by the resulting vehicle behavior such as the latera position, steering wheel movements, and time-to-line crossings whom correspondence should be addressed not intrusive, they are subject to several limitations related to the vehicle type, driver experience, and geometric characteristics and condition of the road. Among these various possibilities, the monitoring of a driver's eye state by a camera is considered to be the most promising application due to its accuracy and Non-intrusiveness. The driver's symptoms can be monitored to determine the driver's drowsiness early enough to take preventive actions to avoid an accident [5].

Though many studies have developed image-based driver alertness recognition systems using computer vision techniques, many problems still remain. First, eye detection remains a challenging problem with no inexpensive or commercial solutions. For some applications, eye feature detection can be satisfactory, but these only used frontal

face images taken with controlled lighting conditions. In a car, the constantly changing lighting conditions cause dark shadows and illumination changes, such that effective techniques in stable lighting often do not work in this challenging environment. The performance of current algorithms degrades significantly when tested across different postures and illumination conditions, as documented in a number of evaluations. A second problem is that current systems do not use identification and correlation analysis of various visual measures. Typical visual characteristics of a driver with a reduced alertness level include longer blink duration, slow eyelid movement [6].

**DROWSINESS FEATURES**

The drowsiness features are characterized by the blinking frequency of the eye by the driver.

- \*Awake-conscious-normal
- \*Drowsy-less conscious-risky
- \*sleep-out of conscious-at extreme risk

**PROPOSED SYSTEM**

The total essence and the functioning of the vehicle section is represented in the following block diagrams [7]. These block diagrams mainly consist of 9 parts. They include

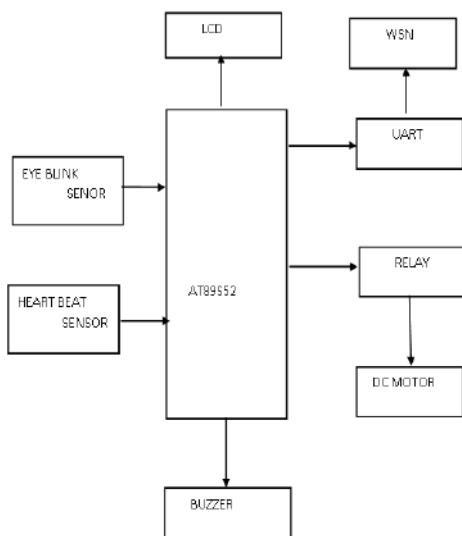
- Eye Blink Sensor
- Heart beat sensor
- LCD
- AT89S52 Microcontroller
- Buzzer
- Relay
- DC Motor
- UART
- WSN
- GPS

In block diagram there is two section, namely

- Vehicle section
- Monitoring and controlling unit

Hence a circuit diagram for both the units are illustrated with clear interfacing unit.

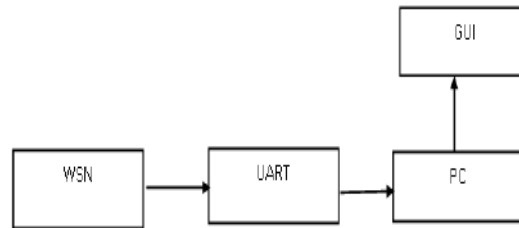
**VEHICLE SECTION**



**Fig:1. Vehicle section**

In vehicle diagram, it contains heart beat sensor, eye blink sensor which transmit signal to microcontroller to process further information. The microcontroller reads the sensor value and produce buzzer alarm along with auto parking of the vehicle.[8]

### MONITORING AND CONTROLLING



**Fig: 2. Monitoring and Controlling section**

In controlling side, a simple visual basic is created which displays the driver status i.e. driver's heart beat level and eye blink position (opened or closed) [9]. It not only indicates sensor value instead it is also possible to indicate the vehicle location i.e. Latitudinal and longitudinal value to the server end with the help of GPS technology.

Screenshot of visual basic application is shown below.



**Fig: 3.Result Analysis**

### CONCLUSION

The vehicle is at a very high speed on highways due to which handling is tough and getting the vehicle to halt in such a condition is difficult. Due to this many automobile companies are trying to research onto how an accident which occurs due to driver fatigue can be prevented.

In this project we will generate a model which can prevent such an incident. The Purpose of such a model is to advance a system to detect fatigue symptoms in drivers and control the speed of vehicle to avoid accidents. The main components of the system consist of an eye blink sensor and heart beat sensor for driver blink acquisition. It also holds main part of GPS used to link Google server to transmit the vehicle location. Advanced technology offers some hope avoid these up to some extent.

### FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

### ACKNOWLEDGEMENT

None

### CONFLICT OF INTERESTS

There is no conflict of interest amongst the authors.

## REFERENCES

- [1] Y Wang, J Yang, H Liu, Y Chen, M Gruteser, RP Martin. [2013] Sensing Vehicle Dynamics for Determining Driver Phone Use, *MobiSys*.
- [2] H Yan, EC McLaughlin, J Hanna. [2013] Loss of 19 firefighters in Arizona blaze 'unbearable,' governor says
- [3] J Yang, S. Sidhom, G. Chandrasekaran, T Vu, H Liu, N Cecan, Y Chen, M Gruteser, RP Martin. [2011] Detecting Driver Phone Use Leveraging Car Speakers. *MobiCom*
- [4] J. Lindquist, J Hong. [2011] Undistracted Driving: A Mobile Phone That Doesn't Distract. *HotMobile*.
- [5] TEXIEVE: Detecting Drivers Using Personal Smartphones by Leveraging Inertial Sensors. C Bo, X Jian, XY Li. [2010]
- [6] J Palmer. [2009] Robot Firefighting team debuts,
- [7] JBAMuintz, MA Viergever. [1998] An overview of medical Image Registration Methods, 9-12
- [8] LGBrown. [1991] A survey of Image Registration Methods. 29-37
- [9] Prevention of Accident Due to Drowsy By Using Eye Blink, B. praveenkumar, K. Mahendranan, ISSN: 2319-8753
- [10] Design Of Accident Prevention System Using QRD 1114 and CNY 70 Sensor, p. kadam, kavita. ISSN: 2250-2459
- [11] Yang Jie et al, Sensing Driver Phone Use With Acoustic Ranging through car speaker. *IEEE Transaction on mobile computing*, (volume 11, issue: 9) 2012
- [12] DF RC. Ltd, "Firefighting robot," brochure.
- [13] MVF-5, "Dok-ing," website, 2013. [Online]. Available: <http://www.rylandresearch.co.uk/remotely-operated-vehicles/firemote-4800>
- [14] [www.osun.org/pdf/microcontroller](http://www.osun.org/pdf/microcontroller)
- [15] [www.electronicsforu.com/voice ic.pdf](http://www.electronicsforu.com/voice%20ic.pdf)
- [16] <http://www.eng.chula.ac.th/files/langearforum/download/langearforum2553/CarTalkITS.pdf>

## WEB LOG MINING - A STUDY

Geetha Krishnagandhi <sup>1</sup>and Suresh Gnana Dhas<sup>2</sup>

<sup>1</sup> Bharathiar University, Coimbatore, Department of BCA, G.T.N. Arts College (SSC), Dindigul, TN, INDIA

<sup>2</sup> Dept. of Computer Science and Engineering, Vivekanandha College of engineering for women (Aut.) Tiruchengode, Namakkal Dt. TN, INDIA

### ABSTRACT

The analysis of web log files may give information that are useful for improving the services offered by web portals and information access and retrieval tools, giving information on problems occurred to the users and gain knowledge from the web. A particular useful kind of knowledge, which can be applied to improve the performance of web service. Web log file analysis began with the purpose to offer to web site administrators a way to ensure adequate bandwidth and server capacity to their organization. It is way to evaluate the effectiveness of a Web site and its information access tools is through the mining of web log files. This study reports on initial findings on a specific aspect that is highly relevant to web mining and extracting useful patterns from the web.

Published on: 08<sup>th</sup>– August-2016

#### KEY WORDS

Web Mining; Web Log Mining;  
Web Structure Mining; Web  
Content Mining; Web Usage  
Mining

\*Corresponding author: Email: [geethachouthri@gmail.com](mailto:geethachouthri@gmail.com); Tel.: +91 8760821422

## INTRODUCTION

### Data Mining

Data Mining is an important research area as there is a huge amount of data available in most of the applications. To extract useful information and Knowledge from that large amount of data. It is an interdisciplinary research field to database systems, statistics, machine learning, information retrieval etc. Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge presentation are the important data mining processes [1].

The World Wide Web has become one of the most important media to store, share and distribute information. Web Mining is the data mining technique that automatically discovers or extracts the information from web. Mining the web is discovering knowledge from hypertext and WWW. The www is one of the longest rising areas of intelligence accessible internet and the number is still increasing. Every websites attract millions of users and visitors. These visitors behind vast amount of website.

Web log mining is a promising tool to study user behaviours, benefit web-site designers with a better organization and services. Effective web log mining system consists of data processing, sequential pattern mining and visualization [2].

Many existing systems that can be used to analyse the traversal path of web-site visitors, their performance is still far from satisfactory. In often unclear where a specific document is located. And usually a great portion of time is needed to look for and find the appropriate information. [3]

When user accesses websites are recorded in web log file, web server log file is a simple plain text file. Log file contain noisy and ambiguous data which may affect results of mining process [4].



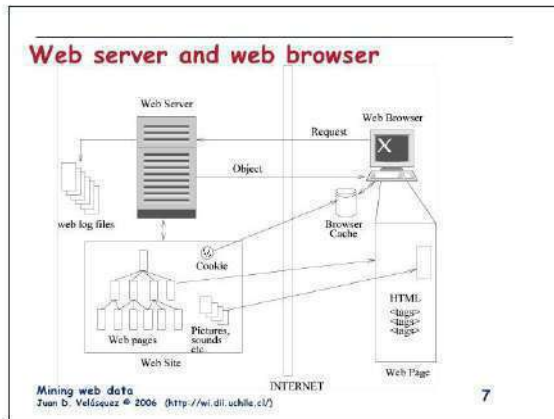


Fig: 1 - Web Server and Web Browser – Mining Web Data

### WEB MINING CATEGORIES

Web mining can be broadly classified into three domains

- ◆ Web Structure Mining
- ◆ Web Content Mining
- ◆ Web Usage Mining



Fig: 2. Web Mining Classification

### Web Structure Mining

Web Structure Mining is to deal with the structure of the hyperlinks within the web itself. It is the process to discover the model of link structure of the web pages. To catalogue the links, generate information such as the similarity and relations among them by taking the advantage of hyperlink topology. The goal of structure mining is to generate structured summary about the website and web page.

It is the technique to analysis and explain the links between different web pages and web sites. It mainly focuses on developing web crawlers. It works on hyperlinks and mines the topology of their arrangement. Web structure mining can classify the web pages and produce results such as the similarity and relationship between different web sites [5].

### Web Content Mining

Web Content Mining is the process of retrieving the information from web into more structured forms and indexing the information to retrieve quickly. It focuses mainly on the structure within a document i.e. inner document level. It is also related with text mining because much of the web contents are text, but is also quite different from these because web data is mainly semi structured in nature and text mining focuses on unstructured text [6].

It focuses on extracting knowledge from the contents or their description of the web documents. It involves techniques the summarising, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behaviours.

### Web Usage Mining

Web Usage Mining is the process of discovering a meaningful patterns from data generated by client server transactions on one or more web servers [7].

It focuses on digging the usage of web contents from the logs maintained on web servers, cookies logs, applications server logs etc. It works on how and when user moves from one type of content to others. Thus, it can provide association between different contents.

## OVERVIEW OF WEB MINING

With the rapid and explosive growth of information available over the internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of web data, web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc., As a result, web users are always drowning in an ocean of information and facing the problem of information overload when interacting with the web.

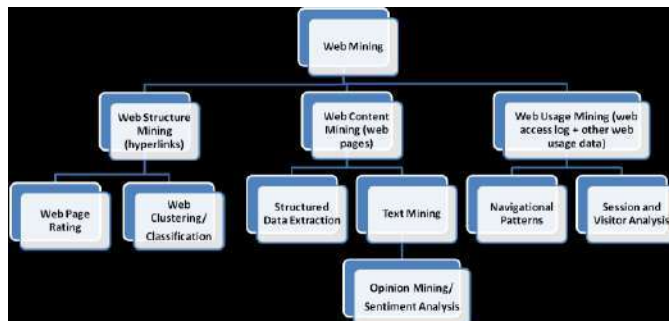


Fig. 3. Overview of Web Mining

Web site is a collection number of web pages grouped under the same domain name. In web mining, data is distributed on various places and web pages may contain data is not only text form it could text, audio, images video and images and navigate between them via hyperlinks. When user accesses website, log files are created. Log file recorded information about the each user. Tremendous uses of web, Web log files are growing at faster rate.

### Problems in Web related Search

- ◆ Finding relevant information
- ◆ Finding needed information
- ◆ Learning useful knowledge
- ◆ Recommendation of information

Web involves three types of data, the actual data from WWW, the web log data obtained by the users who browsed the web pages and the web structure data.

Web Server maintains the web log file. Log files are located in different locations like web server, web proxy server, and client browser.

## Web Log File

Web server log file is a simple plain text file which record information about each user. Log file contain information about user name, IP address, date, time, bytes transferred, access request. A time a user requests a resource from the particular site. When user submit request to a web server that activity are recorded in web log file. Log file range 1KB to 100MB [8].

Log file gives significant information to web server.log file information about:

1. Which pages were requested in website?
2. How many bytes sent to user from server?
3. What type of error occurs?

When user submit request to a web server that activity are recorded in web log file. Log file used for debugging purpose. Analysing log file are used to detecting attacks on web.

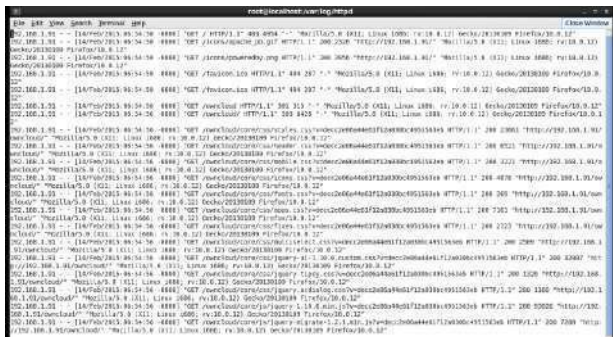


Fig: 4. Sample Web Log File

## Location of Log File

Web log file is located in three different locations.

### Web server logs

Web log files provide most accurate and complete usage of data to web server. The log file do not record cached pages visited. Data of log files are sensitive, personal information so web server keeps them closed.

### Web proxy server

Web proxy server takes HTTP request from user, gives them to web server, then result passed to web server and return to user. Client send request to web server via proxy server. The two disadvantages are: Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction. The request interception is limited.

### Client Browser

Log file can reside in clients' browser window itself. HTTP cookies are pieces of information generated by a web server and stored in user's computer, ready for future access.

## Types Of Web Server Logs

Generally four types of server logs.

### Access log file

Data of all incoming request and information about client of server. Access log records all requests that are processed by server.

### **Error log file**

List of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log [joshila]. Access and error logs are mostly used, but agent and referrer log may or may not enable at server.

### **Agent log file**

Information about user's browser, browser version. Referrer log file: This file provides information about link and redirects visitor to site.

### **Log File Format**

Web log file is a simple plain text file which record information about each user. Display of log files data in three different format

- ◆ W3C Extended log file format
- ◆ NCSA common log file format
- ◆ IIS log file format

NCSA and IIS log file format the data logged for each request is fixed. W3C format allows user to choose properties, user want to log for each request [9].

### **W3C Extended Log File Format**

W3C log format is default log file format on IIS server. Field are separated by space, time is recorded as GMT (Greenwich Mean Time). It can be customized that is administrators can add or remove fields depending on what information want to record. In W3C format of year is YYYY-MM-DD. Omitting unwanted attributes field when log file size is limited [W3C].

### **NCSA Common Log File Format**

National Centre for Supercomputing Application format. NCSA is recorded basic information about user request such as user name and remote host name, date, time, request type, HTTP status code and numbers of bytes send by server. NCSA is fixed format, it cannot customized. It is available for website but not for FTP site. Format of year is DD/MM/YYY. Fields are separated by space, time is local time.

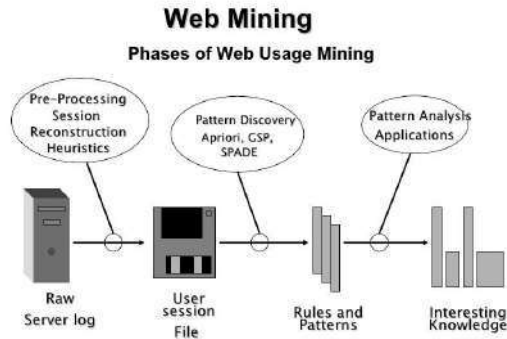
### **IIS Log File Format**

IIS format is not customized, it is fixed ASCII format. Fields are separated by comma, easy to read. Time recorded in local time. Its records more information than NCSA format. Fields in IIS are client IP address, user name, date, time, service and instance, server name, server IP address, time taken, client bytes sent, server bytes sent, service status code, windows status code, request type, target of operation, parameters.

## **PRE-PROCESSING**

The aim of data pre-processing is to select essential features, clean data from irrelevant records and finally transform raw data into sessions. In the pre-processing, the data cleaning process includes removal of records, graphics, videos and the formal information. The failed records with the failed HTTP status code and robots cleaning. The Web usage Mining phases are

- 1) Data Pre-processing.
- 2) Pattern Discovery.
- 3) Pattern Analysis Stage



**Fig: 5. Phases of Web Usage Mining**

First, the data pre-processing phase includes the web log data cleaning, user identification, session identification and data transformation [10]. Condensed and transformed in order to retrieve and analyse significant and useful information.

So the main steps of this phase are:

- 1) Extract the web logs that collect the data in the web server.
- 2) Clean the web logs and remove the redundant information.
- 3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in pattern analysis to create summary reports.

Second Pattern Mining can be performed on log records to find association patterns, sequential patterns and trend of web accessing.

The pattern discovery phase involves the discovery of frequent sequences.

The pattern analysis phase involves the analysis of the frequent patterns.

#### **Data Preprocessing**

Before using data for pattern discovery it need to be cleaned to get data in specific format. Pre-processing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data pre-processing is to improve data quality and increase mining accuracy. Pre-processing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process. The main task is to “clean” the raw web data such as web logs. Data pre-processing involves data cleaning, user identification and data transformation [11].

In data cleaning phase, the web log is examined and irrelevant on redundant items such as image, sound, video files and executable gif files and HTML files, removal of HTTP errors recorded created by crawlers, removed from the log.

The user’s identification phase involves identification of users from the log data.

1. A new IP identifies a new user.
2. If the same IP is used, but different operating system in terms of type and version is being used, then that is considered as new user.

Pre-processing is necessary, because log file contain noisy & ambiguous data which may affect result of mining process. Some of web log file data are unnecessary for analysis process and could affect detection of web attack.

Data pre-processing is an important steps to filter and organize only appropriate information before applying any web mining algorithm. Pre-processing reduce log file size also increase quality of available data. The purpose of data pre-processing is to improve data quality and increase mining accuracy. Pre-processing consists of field extraction, data cleansing, user identification, session identification. In this paper main task is to “clean” the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process. So the main steps of this phase are:

- ◆ Extract the web logs that collect the data in the web server.

- ◆ Clean the web logs and remove the redundant information.

## MATERIAL AND METHODS

Web log data pre-processing is a complex process and takes 80% of total mining process. Log data is pre-treated (cleaning) to get reliable data. There are four steps in pre-processing of the log data.

### **Field Extraction**

The log entry contains various fields which need to be separate out for the processing. The process of separating field from the single line of the log file is known as field extraction. The server used different characters which work as separators. The most used separator character is ',' or 'space' character.

### **Data Cleaning**

Data Cleaning is the removal of outliers or irrelevant data. It is the process to remove noisy and unnecessary data. Remove log entry nodes contain extension like jpg, gif means remove request such as multimedia files, image, page style file [12].

Data cleaning is usually site specific, and involves extraneous references to embedded objects that may not be important for purpose of analysis, including references to style files, graphics or sound files. Therefore some of entries are useless for analysis process that is cleaned from the log files. By data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data.

Analysing the huge amounts of records in server logs is a cumbersome activity. So initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, are also downloaded which are not useful for further analysis are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be removed from log files [9]. Thus removal process in the experiment includes the records of graphics, videos and the format information: The records have filename extension of GIF, JPEG, CSS and so on, which can be found in the URI field of the every record, can be removed. This extension files are not actually the user interested web page, rather it is just the documents embedded in the web page. So it is not necessary to include in identifying the user interested web pages. This cleaning process helps in discarding unnecessary evaluation and also helps in fast identification of user interested patterns.

The records with the failed HTTP status code: The HTTP status code is then considered in the next process for cleaning. By examining the status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. This cleaning process will further reduce the evaluation time for determining the used interested patterns. Method fields, Robots cleaning are the software tool to extract from the website.

This helps in accurate detection of user interested patterns by providing only the relevant web logs. Only the patterns that are much interested by the user will be resulted in the final phase of identification if this cleaning process is performed before start identifying the user interested patterns.

## USER IDENTIFICATION

This step identify individual user by using their IP address. If new IP address, there is a new user. If IP address is same but browser version or operating system is different then it represents different user.

The log file after cleaning is considered as Web Usage Log Set – WULS. The next important and complex step is unique user identification. The complexity is due to the local cache and proxy servers. To overcome this cookies are used. But users may disable cookies. Another Solution is to collect registration data from users. But users neglect to give their information due to privacy concerns. So majority of records does not contain any information in the user-id and authentication fields. The fields which are useful to find unique users and sessions are :

- 1) A new IP Address indicates a new user.

- 2) User Agent - The same IP but different web browsers or different operating systems in terms of type and version means a new user.
- 3) Referrer URL – Suppose a the topology of a site is available, if a request for a page originates from the same IP Address as other already visited pages, and no direct hyperlink exists between these pages, it indicates a new user.

Users and sessions are identified by using these fields as follows. If two records has same IP address check for browser information. If user agent value is same for both records then they are identified as from same user.

### **Session Identification**

Each user spends total time in each web page. Session means time duration spent in web pages. A referrer-based method is used for identifying sessions. If IP address, browsers and operating systems are same, the referrer information should be taken. The CS referrer is checked, a new user session is identified if the URL in the refer URL-field is a large interval usually more than 30 minutes between the accessing time of this record.

The goal of session identification is to divide the page accesses of each user into individual sessions. These sessions are used as data vectors in various classification, prediction, clustering into groups and other tasks. If URL, in the referrer URL, field in current record is not accessed previously or if referrer URL field is empty then it is considered as a new user session. Reconstruction of accurate user sessions from server access logs is a challenge task and time oriented heuristics with a time limit of 30 min is followed.

From WULS, the set of user sessions are extracted as referrer based method and time oriented heuristics. Every record in WULS must belong to a session and every record in WULS can being to one user session only. After grouping the records into sessions the path completion step follows.

### **Path Completion**

Path completion step is carried out to identify missing pages due to cache and 'Back' Path set is the incomplete accesses pages in a user session. It is extracted from every user session set.

Path Combination and Completion : Path Set (PS) is access path of every USID identified from USS. It is defined as:  $PS = \{USID,(URI),Date, RLength)\dots(URI, Date, RLength)\}$

Where RLength is computed for every record in data cleaning stage. After identify path for each USID path combination is done if two consecutive pages are same. In the user session if any of the URL specified in the Referrer URL is not equal to the URL. In the previous record then that URL in the Referrer URL field of current record is inserted into this session and thus path completion is obtained.

### **Pattern Discovery**

The Pattern Discovery Phase is the key component of the Web mining. Pattern discovery converge the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition, etc. applied to the Web domain and to the available data.

### **Content Clustering**

At the final stage of pre-processing of content data have m web pages that have used to mining content of these pages and integrating them with usage and structure patterns. The symbols m, n, and k to denote the number of documents, the number of terms, and the number of clusters, respectively. The symbol S to denote the set of n documents that we want to cluster,  $C_1, C_2, \dots, C_k$  to denote each one of the k clusters, and  $n_1, n_2, \dots, n_k$  to denote the sizes of the corresponding clusters. The K-means clustering algorithm is one of the clustering algorithm that is used the vector-space model to represent each document. The best known approach that is based on partitioning is k-means clustering, a simple and efficient algorithm used by statisticians for decades. The idea is to represent the cluster by the centroid of the documents that belong to that cluster (the centroid of cluster C is defined as). The cluster membership is determined by finding the most similar cluster centroid for each document. After clustering done, similar pages are assigned to same cluster that can be used in recommendation process.

## Page Ranking

Finally, by employing the HITS algorithm on structure data system generate ranked pages. In HITS concept, Kleinberg identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to Kleinberg, "Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs".[13]

## Pattern Analysis

**Pattern recognition** is a branch of machine learning that focuses on the recognition of patterns and regularities in data, although is in some cases considered to be nearly synonymous with machine learning. Pattern recognition systems are in many cases trained from labeled "training" data (supervised learning), but when no labeled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning).

The terms pattern recognition, machine learning, data mining and knowledge discovery in databases (KDD) are hard to separate, as they largely overlap in their scope. Machine learning is the common term for supervised learning methods and originates from artificial intelligence, whereas KDD and data mining have a larger focus on unsupervised methods and stronger connection to business use. In pattern recognition, there may be a higher interest to formalize, explain and visualize the pattern; whereas machine learning traditionally focuses on maximizing the recognition rates. Yet, all of these domains have evolved substantially from their roots in artificial intelligence, engineering and statistics; and have become increasingly similar by integrating developments and ideas from each other.

In machine learning, **pattern recognition** is the assignment of a label to a given input value. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of *classes* (for example, determine whether a given email is "spam" or "non-spam"). However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values and parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence.

Pattern recognition algorithms generally aim to provide a reasonable answer for all possible inputs and to perform "most likely" matching of the inputs, taking into account their statistical variation. This is opposed to pattern matching algorithms, which look for exact matches in the input with pre-existing patterns.

## CONCLUSION

In this paper, I presented a preliminary analysis of the web log mining according to the methodology for gathering information from log files, mining information to extract knowledge from them, analyse the useful patterns and how to presented. The aim of this work was to report on initial findings about the study of web content mining, web structure mining and web usage mining and how to extract patterns from the data from web log files and discover the knowledge

## FUTURE WORK

A new web page recommendation framework is to be proposed to make efficient Web Log Mining. First, users' navigational patterns are extracted from web usage data simultaneously web content data and web structure data is also taken after pre-processing and pattern discovery is performed on a server data's and based on the pattern discovery the recommendations are generated. Proposed framework is to combine the special features of web mining algorithms and to create a new algorithm to maintain efficient web log mining, so it is advantageous as compare to the previous hybrid recommendation frameworks.



**CONFLICT OF INTEREST**

None

**ACKNOWLEDGEMENT**

None

**FINANCIAL DISCLOSURE**

None

**REFERENCES**

- [1] J Han, M Kamber.[2000] Data Mining: Concepts and Techniques Morgan Kaufmann.
- [2] S.Veeramalai,N.Jaisankar and A.Kannan.[August - 2010] Efficient web log mining using enhanced Apriori algorithm with Hash Tree and Fuzzy.
- [3] Ramakrishnan Srikant.[2011] Mining web logs to improve website organization. , IBM Almaden Research Center.
- [4] B.Madasamy, J.Jebamalar Tamilselvi.General web knowledge mining framework (IJCSE).
- [5] Priyanka Patil and Ujwala Patil, Preprocessing of web server log file for web mining. (NCETCT 2012).
- [6] Renata Ivancsy, Istvan Vajk.[ 2006] Frequent Pattern Mining in Web Log Data, Acta Polytechnica Hungarica .3(1)
- [7] Qiang Yang,Charles X.Ling and JianFeng Gao. [2013]Mining web logs for actionable knowledge.
- [8] Zhengtu Yang,Yitong Wang,Masaru Kitsuregawa.[2011 ] An effective system for mining web log.
- [9] J Vellingiri and Chethur Pandian. A Novel technique for web log mining with better data cleaning and transaction identification [ISSN 1549-3636].
- [10] Rajashree Shetlar.[2006] Sequential Pattern Mining from web log data. [ISSN 2250-3676].
- [11] Jiawei Han and Micheline Kamber.[2015] Data Mining concepts and technologies - Morgan Kaufmann Publishers, 2nd edition-Delhi.
- [12] Margaret H.Danham, S.Sridhar [2007] Data Mining Introductory and Advanced Topics --Dorling Kindersley India PVT Ltd. -New Delhi.
- [13] Bing Liu.[2007] Web Data Mining Exploring Hyperlinks, contents and usage data --Springer --New York.

## EFFICIENT AND SECURE DATA TRANSMISSION IN AIR-CRAFTS

Anto Rose, T. Praveen, Karthick, Marudhu Pandiyan\*

Department of Computer Science and Engineering, Vel Tech High Tech Dr.RangarajanDr.Sakunthala Engineering College, Avadi, Chennai, INDIA

### ABSTRACT

**Aims:** Android is the majority popular platform for mobile devices. It facilitates distribution of data and services between applications using an affluent inter-app communication system. While access to resources can be restricted by the Android permission system, enforcing permissions is not satisfactory to prevent security violations, as permissions may be mismanaged, purposely or accidentally. **Materials and methods:** Security is the major issue in aircrafts. So in this paper we present some new security schemes for securing the data communication in aircrafts. We used a new technique that is used to replace conventional radar. For tracking the Aircrafts we are using ADS-B system. It is used for finding the current position of the aircrafts. It uses satellites for updating the current location of the aircrafts. The ADS-B system communicates with the satellite and continuously broadcast their position and other information such as the current velocity of the aircrafts and the movement of other nearby aircrafts. **Results:** If we communicate the aircrafts without the ADS-B it will create many problems like any active attacker can modify our confidential data. There are several kinds of aircraft attacks are available such as Ghost Aircraft Injection and Virtual Trajectory Modification attack. An active attacker can inject, modify and delete the messages. For avoiding all these things and for providing security we are using ADS-B authentication systems. **Conclusion:** It mainly focusing on data integrity. It uses cryptographic concepts for data encryption and Decryption. And also the signature matching and verification is used for security purposes.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

Authentication; ADS-B system; Integrity; Tracking; Batch Verification.

\*Corresponding author: Email: [drsowmyab1@gmail.com](mailto:drsowmyab1@gmail.com) Tel: +91- 8884546649;

### INTRODUCTION

Now a days the usage of aircrafts has been increasing day by day. According to the statistics in Europe the number of registered aircrafts is around 26500 per day. Most of the persons are preferring to travel by air. Because of this Traffic and security problem arisen. The Airways traffic control is mainly based on the radars. For controlling the traffic in airways it uses two types of radars Primary Surveillance Radar(PSR) and Secondary Surveillance Radar(SSR). The Primary Surveillance Radar are fully independent and also it is non-cooperative [10]. The Primary Surveillance Radar is used for transmitting the signals with very high frequency and also it receives the echoes which is reflecting from other aircrafts. By the use of this echo from other aircrafts it will identify the position of the aircrafts. Without the usage or the participation of the particular aircraft we can find the location of aircrafts.

The Secondary Surveillance Radar(SSR) gathers information from the aircrafts all the aircrafts uses the onboard system which is in built in all kinds of aircrafts. Based on the onboard we can transmit our message is delivered to the Secondary Surveillance Radar. The information contains many useful information about the radars like the identification code of the particular aircraft, at what height it is flying and the altitude of the aircrafts.

Both the Primary Surveillance Radar and the Secondary Surveillance Radar has some disadvantages both are high cost and difficult to manage.

And also both the radars are not providing high security. The features of these two radars will not sets for Military based application. It does not maintains integrity, availability and confidentiality. Because of these drawbacks[2] we cannot use Primary Surveillance Radar and Secondary Surveillance Radar. The new technique is used for replacing the drawbacks which is available in the Primary Surveillance Radar and the Secondary Surveillance Radar. The Automatic Surveillance Broadcast System is used [5]. It replaces the Primary Surveillance Radar and Secondary Surveillance Radar. Most of the countries like Australia and Canada is using Automatic Surveillance Broadcast System. It is standardized by the Federal Aviation Administration. In the traditional radar system the

aircrafts only respond to the ground station. The Automatic Surveillance Broadcast System does not use Radars. It uses only the satellite [4] based on this it communicates to the ground station. Mainly it is based on the GPS. Based on the GPs it continuously broadcast their position to other aircrafts and also it broadcast the velocity and the height.

By the use of these information from the ground the ground controller can control and find out the current location of the aircraft [3] [8]. Because of this the pilot can take decision in an efficient manner and also it helps to control the traffic. The ADS-B system plays vital role in air ways traffic control and also in communication system. The mashup technology is to be used in the ADS-B system. The ADS-B combined with the mashup.

The ground controller can track or monitor the status of the aircrafts in their website. The ADS-B system broadcast their message through wireless channel. It transmits their message without the use of any cryptographic technique. The ground controller and the aircraft uses only the single low cost ADS-B receiver. The Active attacker can attack the data from the ADS-B system. So the ADS-B[9] system uses data integrity concept. In Data Integrity the data cannot be modified by anyone and also it uses source integrity. Both the data integrity and the source integrity gives more security to the ADS-B system based data transmission. The source integrity is also otherwise known as authenticity. In source integrity while sending the message to the receiver before sending the message [13] itself it should get permission from the receiver. If the receiver or the ground controller has given permission then only it will start sending message. The batch verification is the additional security scheme which is to be used in the [12] aircrafts. It uses signatures if multiple signature is receiving the ground station the Batch verification scheme verifies all the signatures. The batch verification does not allow the partial verification of signature.

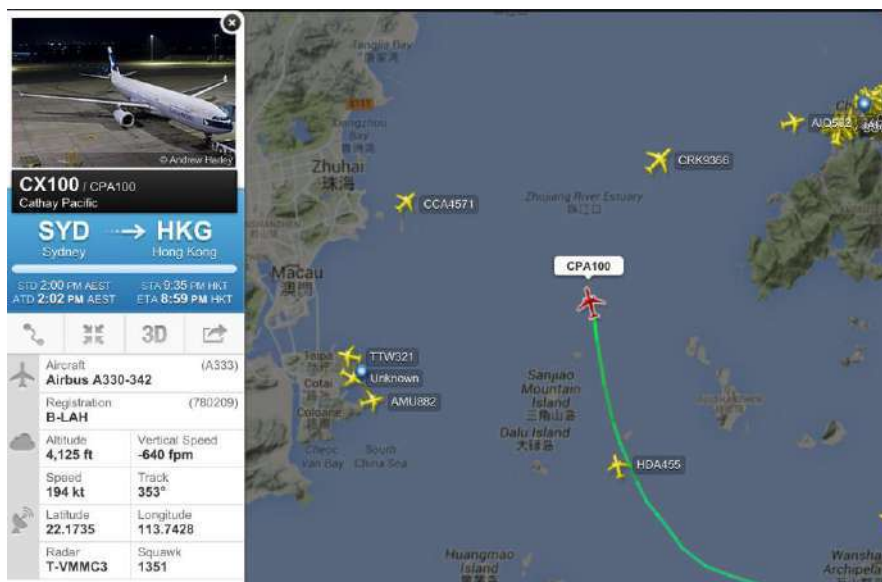


Fig: 1.ADS- B Mashup

The ADS-B is viewed from the ground by using some smart objects which gives the neat graphical representation on the web page of the ground controller.

### Contributions

1. In this paper we proposed and used the efficient authentication method that is ADS-B authentication System.
2. Used Batch Verification Scheme in this each and the every airline has the responsibility to generate the private keys.
3. It uses two types of verification scheme one is partial verification scheme and the another one is batch Verification

## RELATED WORK

In this Security is the main issue so we used some integrity concepts for authentication. The Symmetric key encryption technique is used for encrypting or sharing the same key between the sender and the receiver. But it is very difficult to deploy the key values.

The Digital Signature is the good Method for Batch Verification and also it provides more authenticity than the Symmetric Key Encryption.

### Tracking aircrafts

The aircrafts can be tracked based on the ADS-B system. The ADS-B system uses GPS. It does not use Radar. The Radar communication is not that much effective when comparing with the GPS Communication. It receives the signals from the other or nearby aircrafts based on that we can track the location of the aircrafts and also we can track the Velocity of the aircrafts. [4][8]

The aircraft tracking is very difficult because of the traffic. Now a days all are using the airways to travel so the traffic is more. Because of this much traffic the security becomes one of the major issues in aircrafts.



Fig. 2. Tracking Aircrafts.

The ADS-S uses a special tool by the use of that tool the aircrafts are to be monitored in their web site.

### Use of GPS in ADS-B

The Radars are very costly and also it does not provides efficient and effective communication. The Gps is very easy to transmit the signal. The Aircrafts uses GPS for communicating to the ground station. The GPS periodically broadcast the signal.

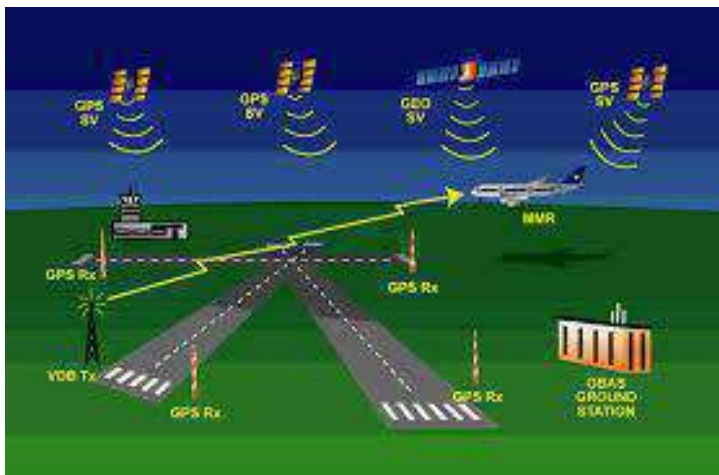


Fig. 3. Aircraftsuses GPS for Data Transmission

By the use of GPS the current location and the Velocity of the aircrafts can be tracked easily. Because of this the pilot can track the aircrafts and the decision making an we can control the traffic easily.

The GPS is very cheap comparing with the Radars.

## DEFINITIONS

### Mathematical definition

Let R and S are to be two cyclic graphs of prime order P. Let M be a generator of R. Then the bilinear map is to be represented by  $R \times R \rightarrow S$ .

1. Bilinearity:  $a \cdot b = Z$ ;
2. Non- degeneracy: If G is a generator of the graph G then the value of the G is not equals 1.
3. Computable: It uses some efficient algorithm to execute. The algorithm is  $e(u,v)$ .

### Hierarchical based signature

There are three levels of Hierarchical identity Based signature. It is a tree based structure .In Hierarchical based signature the Level-0 is the root of the tree. In this all the signatures which are all storing in the database forms a tree based structure.

The airlines are the second levels of the hierarchical data structure. The ADS-B scheme uses Batch verification.  $ID_i, Root^A$  which is used for generating secret keys. The value  $RT_i$  belongs the secret key which is to be used in the  $Root^A$ .

## SECURITY MODELS

It uses three phases they are setup phase, Query phase and output phase. In setup phase it sends the query to the LEVEL -1 identity.  $T^*$ . The value of C is used to generate the keys.

In the Query Phase the Query for the aircrafts  $IG_c$  is to be generated.

In the output phase the message M are to be verified by using the signature.

- 1) Verify(M, IDA X M IDF x z) = 1;
- 2) A Extract A Query on IDA.
- 3) A Extract F Query on (IDA, IDF).
- 4) Signing Query on (IDA, IDF, M)

## CONTROLLING TRAFFIC

The usage of the aircrafts is increasing day by day so in aircrafts the security and the traffic control is the major issue.

For Controlling the traffics it uses the ADS-B authenticates System. It uses the GPS and sends or broadcast the signal to the ground station.

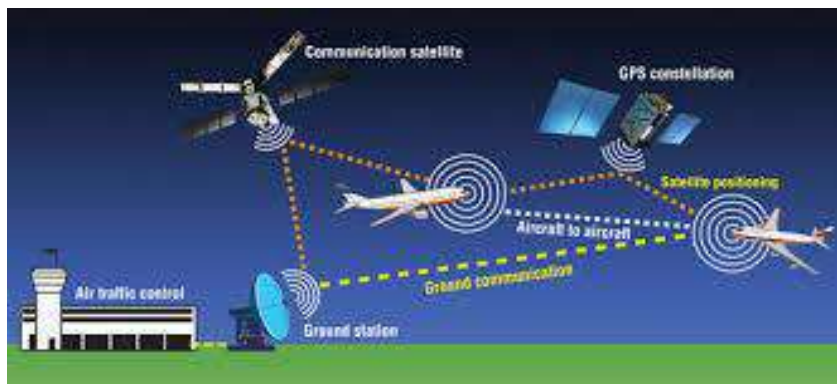


Fig: 4. Traffic Control in Aircrafts

The above image describes the traffic control system. Based on the GPS it tracks the signal coming from the Aircrafts and the Velocity of the aircraft. All the aircrafts contains onboard based on the onboard it transmits the signal to the ground station.

The Ground Station Controller uses mash up and they monitors all the movement of the aircrafts from the ground station if any traffic occurs they will sends message to the pilot. The pilot immediately changes the direction of the aircraft and also it tracks the velocity by uses all these we can track the flight easily.

The ADS-B system is used for tracking the aircrafts. Earlier the radars are use but it does not track the signal properly and also it is very difficult to maintain and also it is very costly. When comparing the ADS-B system with the Radar the Radars causes many aircraft accidents because it does signal properly and also it is very difficult to maintain and also it is very costly.

## PERFORMANCE

The usage of the ADS-B system reduce the traffic control and also it helps the ground station to track the aircrafts and also it guides the pilot to change the location if any problems occurs. In addition it uses batch Verification which is used for the security purpose and also it uses the partial verification of the signatures. It uses hierarchical tree based for storing the signature in the database

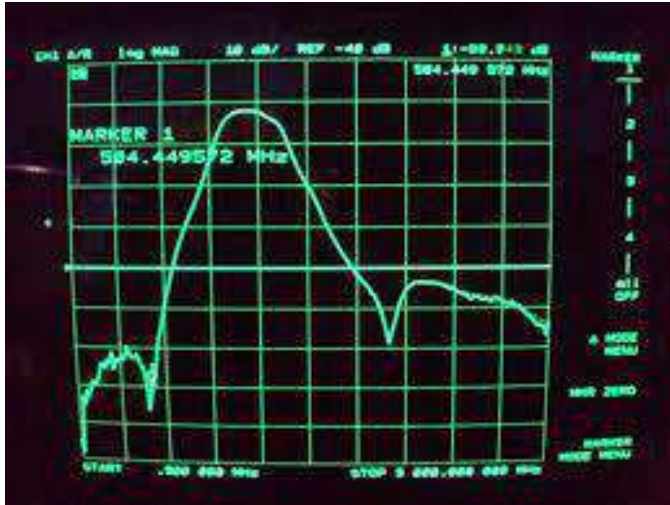


Fig: 5. Performance Measure Of ADS-B System.

The above graphical diagram represents the performance of the ADS-B system compares with the radar. It Shows the performance comparison of ADS-B and the Performance of Radar comparing with the radar communication the ADS-B is much more better. Now a days most of the developed countries are using this system only. Because of high cost and the less performance automatically the usage of the radar is decreasing and also it does not provide that much security.

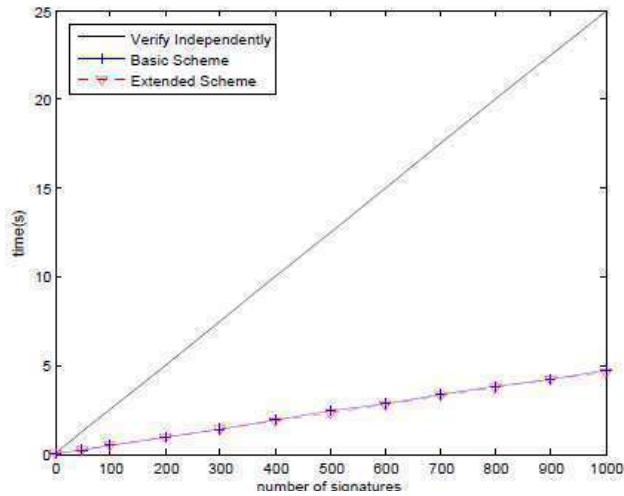


Fig: 6. Performance Measure Of Batch Verification

The Batch Verification Scheme uses Signatures for the authentication Purpose. Before transmitting the signal to the aircraft both the sender and the receiver.

## DISCUSSION

The ADS-B authentication scheme and the batch verification scheme is used for the security purpose. It uses Source integrity and destination integrity. Both the sender and the receiver has to verify the signature so it provides more security than other system.

## CONCLUSION

In this paper we propose the new and efficient scheme for the authentication scheme ADS-B and also it uses the three levels of the hierarchical structure for Batch verification it is used to reduce the verification cost. Based on the ADS-B system we can track the aircraft and also we can control the traffic. By the use of the Batch Verification we can verify the signature so that the security is high comparing with the previous one. The ADS-B uses GPS so the cost is low and also it uses source integrity and signature verification. Comparing with previous one this scheme is more effective and efficient.

## CONFLICT OF INTEREST

None declared.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

No financial support was received for this work.

## REFERENCES

- [1] Sampigethaya K, Poovendra R. [2010] Visualization & assessment of ADS-B security for green ATM. In: 29th Digital Avionics Systems Conference, DASC 2010. pp. 3.A.3-1 – 3.A.3-16. *IEEE*
- [2] Schäfer M, Lender V, Martinovic I. [2013] Experimental analysis of attacks on next generation air traffic communication. In: Proceedings of 11th International Conference on Applied Cryptography and Network Security, ACNS 2013. LNCS, 7954: 253–271. Springer
- [3] Krozel J, Andrisani D, Ayoubi MA, Hoshizaki T, Schwalm C. [2004] Aircraft ADS-B data integrity check. In: 4<sup>th</sup> Aviation Technology, *Integration and Operations Forum*, 1–11
- [4] Baek J, Byon YJ, Hableel E., Al-Qutayri M. [2014] Making air traffic surveillance more reliable: a new authentication framework for automatic dependent surveillance-broadcast (ads-b) based on online/offline identity-based signature. *Security and Communication Networks*
- [5] Yoon, H., Cheon, JH, Kim, Y Batch . [ (2004)] verifications with id-based signatures. In: 7th International Conference on Information Security and Cryptology, ICISC 2004. LNCS, 3506: 233–248 Springer
- [6] McCallie D, Butts J, Mills R. [2011] Security analysis of the ADS-B implementation in the next generation air transportation system. *International Journal of Critical Infrastructure Protection*, 4(2):78–87
- [7] Purton, L., Abbass, H., Alam, S. [2010] Identification of ADS-B system vulnerabilities and threats. In: Australian Transport Research Forum. 1–16
- [8] Samuelson K, Valovage E, Hall D. [2006] Enhanced ADS-B research. In: *IEEE Aerospace Conference*, 1–7. *IEEE*
- [9] Schäfer M, Lenders V, Martinovic I. [2013] Experimental analysis of attacks on next generation air traffic communication. In: Proceedings of 11th International Conference on Applied Cryptography and Network Security, ACNS 2013. LNCS, 7954: 253–271
- [10] Strohmeier, M., Lenders, V., Martinovic, I.: Security of ADS-B: State of the art and beyond. CoRR abs/1307.3664 (2013)
- [11] Strohmeier, M., Lenders, V., Martinovic, I.: Lightweight location verification in air traffic surveillance networks. In: Proceedings of the 1st ACM Workshop on Cyber-Physical System Security. pp. 49–60. CPSS '15, ACM, New York, NY, USA (2015), <http://doi.acm.org/10.1145/2732198.2732202>
- [12] Strohmeier M, Schäfer M, Lenders V, Martinovic I. [2014] Realities and challenges of nextgen air traffic management: The case of ADS-B. *IEEE Communications Magazine* 52(5):111–118
- [13] Wesson KD, Humphreys TE, Evans BL. [2004] Can cryptography secure next generation air traffic surveillance? *IEEE Security & Privacy Magazine*
- [14] Yoon H, Cheon JH, Kim Y. [2004] Batch verifications with id-based signatures. In: 7th International Conference on Information Security and Cryptology, ICISC 2004. LNCS, 3506: 233–248

# AN EFFICIENT ALGORITHM FOR DETECTING OUTLIERS IN A DISTRIBUTED ENVIRONMENT USING MINIMAL IN-FREQUENT ITEM SET PATTERN MINING

Chandra Ravi Chandran<sup>1</sup> and Ajitha Padmanabhan<sup>2</sup>

<sup>1</sup>Dept. Of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, INDIA

<sup>2</sup>Dept.of M.Sc [SS&CS], K. G College of Arts and Science, Coimbatore, Tamil Nadu, INDIA

## ABSTRACT

Published on: 08<sup>th</sup>– August-2016

Outlier's detection in a distributed data surges the classification and prediction of the tasks easier and accurate. Outliers are unusual patterns that occurs rarely and has less incidence in the data Distance based and density based algorithms exists in literature. The In-Frequent item set mining can be utilized to detect outliers which may increase the performance in terms of accuracy. An efficient algorithm is proposed in this paper to detect outliers by defining a certain minimum support threshold for identifying outliers in distributed data by mining minimal in-frequent patterns in the data.

### KEY WORDS

Outliers; In-Frequent Pattern;  
distributed data mining;

\*Corresponding author: Email: [ajitha.mca@gmail.com](mailto:ajitha.mca@gmail.com); Tel.: +91-9843862331

## INTRODUCTION

Outliers are an unusual pattern that occurs rarely and has fewer incidences in the data and normally have lesser support [1]. When the data are from various sources and distributed, there are chances for existence of outliers. Detecting of outliers in the disseminated environment is highly challenging and has very few explorations in this arena [6]. Outliers are dissimilar or inconsistent data that deviates from the normal with smallest measurement [8]. Distributed Data Mining is mining data from different and various sources. Perceiving outliers from datasets has many applications like credit card fraud detection, medical diagnosis, market segmentation and e-commerce[4].

## RELATED WORK

Existing methodologies exist for detecting outliers in a centralised environment frequent pattern mining is used for outlier's detection a distributed data without candidate generation [3]. There are different type of approaches for outliers detection like distance based, density based, clustering based and distribution based. Artificial Intelligence Based approaches to outlier detection like Support Vector methods, fuzzy logic based methods, Genetic algorithm based methods are also available in the literature [14].

Frequent pattern item set detects outliers and assign outlier score to each data point based on the frequent item set it contains. Most of the existing literature shows only frequent item set mining, which may be easier to eliminate outliers. Basically, discovering infrequent patterns in the data sets are considered as outliers. Outliers itself is the attributes that are minimally consistent with the pattern of the data [9].

Mining in-frequent items is proposed in a algorithm called AfRIM [11]. The in-frequent items are searched in top-down manner but with minimum or zero support. MSApriori algorithm is proposed [7] to identify the in-frequent item set based on the high confidence rules and multiple support thresholds which decreases the efficiency. Multiple support thresholds considers data sets of individual nature and to be provided for each and every data sets separately which may in-turn reduces the efficiency [9,10].



Mining frequent item sets are identified in the association rules for fast discovering the frequent item sets so that occurrence of data are considered [12] other in-frequent are not considered and discarded as outliers. Confabulation –inspired Association Rule Mining (CARM) [3] discussed mining both frequent and infrequent pattern set mining inspired on cogency based approach. In-frequent item set discovery by single pass through the association rule datasets. But this approach is based on the conditional probability that exists. Outlier detection using distance based, density based, frequent patterns, density based, distance based, Artificial neural networks, information theoretic based approaches [13][5].

## METHODOLOGY

Integrating In-frequent pattern mining for outlier detection is of a novel approach as it interestingly offers high accuracy of outlier's discovery in vast amounts of data. This paper discusses the outlier detection in distributed sources. Current literature shows detecting outliers in distance based and density based outlier's detection. A new methodology is discussed here to detect outliers. In-Frequent item set pattern discovery in outliers by having automatically assigning a parameter to the mini-support. Secondly finding closed frequent item sets to reduce the memory if the data sets are of large nature with a minimum support threshold.

### Algorithm : CiFPMDiscover

1. Input the data from various sources
2. Identify all frequent itemsets and generate individual candidates that are not discovered.
3. Frequent Pattern Support is calculated to check whether superset for the same support as frequent patterns exists or not.
4. If  $FPSupp = SuperSupp$  then  
 $iFP = D = \{i_1, i_2, \dots, i_n\}$   
 else  
 $iFP = \{NULL\}$
5. Till all the  $iFP = NULL$
6. Iterate till all the possibilities of super set checked with other  $MinSupp$
7. CiFPM is generated when no supersets of same support count
8. Terminate all the item set generation
9.  $MinSupp = \{\alpha\}$
10. If  $MinSupp$  then OutDet
11. Terminate the process

The algorithm Closed in-Frequent Pattern Mining Discover is used to discover outliers and discard it when there is no superset that has same support count as the original itemset. It increases accuracy in finding outliers with single pass so outliers can be easily found.

## RESULTS

CiFPMDiscover algorithm, finds in-frequent pattern mining with closed itemsets so that it provides minimal space to find outliers. The datasets considered are BreastCancer Winscoin datasets.

**Table: 1. Class Distribution of Wisconsin Cancer Breast Cancer Dataset**

Case	Class codes	Number of instances
Commonly occurring classes	2	65.5%
Rare class	4	34.5

**Table-1** shows the class distribution of Wisconsin breast cancer datasets. Commonly occurring classes shows the normal classes and rare class shows the outliers in the datasets. When comparing the CiFPMDiscover with other algorithms of FPOF, CBLOF the detecting of outliers is shown below.

**Table-2** shows the minimum support threshold for identifying outliers in having minimum support threshold for breast cancer datasets in benchmarked UCI machine repository datasets. If the minimum threshold of  $\alpha$  is reached the dataset is considered as outliers and they are discarded.

Table: 2. comparison of proposed CiFPMDiscover to FPOF, CBLOF

Number of Records	Number of Outliers Detected		
	FPOF	CBLOF	CiFPMDiscover
0	0	0	0
4	3	4	3
8	7	7	6
16	14	14	11
24	21	21	18
40	31	32	30
48	35	35	35
56	39	38	36
64	39	39	36
72	39	39	38
80	39	39	38
100	39	39	38
112	39	39	39

Table: 3 .Execution times in respect to the centralized algorithm

Dataset/l	5	10	15
Breast Cancer	230.1	126.4	96.5
Poker	210.1	112.3	83.3
Cov Type	230.1	126.4	96.5

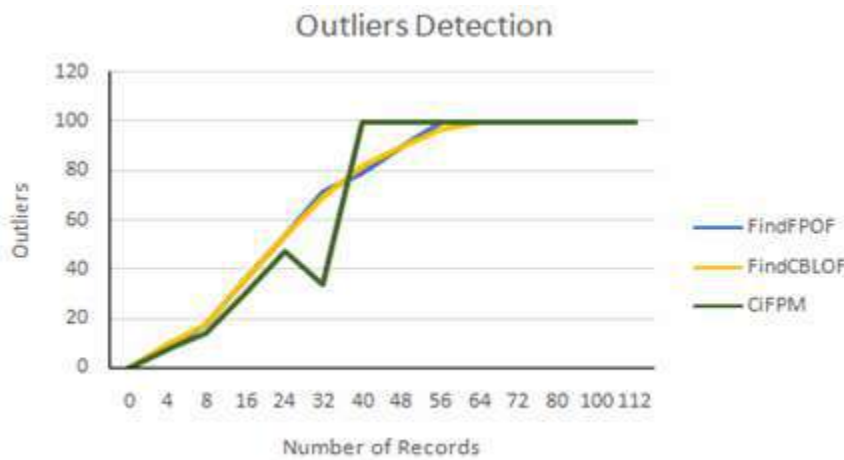


Fig: 1.Comparison of Number of Outliers detected

The **Figure- 1**, shows the comparison of number of outliers detected from the proposed cifpm to the other algorithms for the datasets, breast cancer, poker and ecotype data sets available in uci machine repository.

### CONCLUSION AND FUTURE WORKS

CiFPMDiscover algorithm detects outliers with using minimum support threshold. Using the closed infrequent pattern detection by discarding the attributes that does not support with minimum threshold limit. The proposed algorithm deals with single pass in datasets and saves in memory limitage. Accuracy and memory requirements that considered for discovering outliers is comparatively efficient then the existing methods. Detection of outliers in distributed data sources can be further extended to domain based outlier detection. Automatic detection of outliers based on the dataset may be also explored further.

## CONFLICT OF INTEREST

None declared.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

No financial support was received for this work.

## REFERENCES

- [1] Adda M, Wu L, Feng Y. [2007] Rare itemset mining. In Proceedings of the 6th International conference on machine learning and applications (ICMLA '07) (pp. 73–80). Washington, DC: IEEE Computer Society.
- [2] Agrawal.R, Srikant.[1994] Fast Algorithms for Mining Association Rules. In Proceedings of VLDB'94, 478-499,
- [3] AzadehSoltaniandM.-R.Akbarzadeh-T.[2014]onfabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets, *IEEE transactions on neural networks and learning systems*, Cateni
- [4] SV Colla,[2013] Data Processing for Outliers Detection, Pattern Recognition: Methods and Application, iConcept Press Ltd, ISBN: 978-1-922227-08-9.:1-21.
- [5] Chandola V, Banerjee A, Kumar V. [2009]. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.
- [6] Chandra.E, P Ajitha,[2015] An Algorithm For Detecting Outliers In Distributed Environment” *International Journal of Applied Engineering Research* ISSN 0973-4562 10(1): 1519-1523 © *Research India Publications*.
- [7] Han JJ, Pei Y Yin and R Mao.[2004] Mining frequent patterns without candidate generation: a frequent-pattern tree approach, *Data Mining and Knowledge Discovery*, 8(1): 53–87.
- [8] Hawkins.D.[ 1980] Identification of Outliers. Chapman and Hall, Reading, London, ISBN 978-0412219009
- [9] He Z, Xu X, Huang, JZ, Deng S.[2005] FP-Outlier: Frequent Pattern Based Outlier Detection. *Computer Science and Information Systems*, 2( 1): 103-118.
- [10] Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. [2014] *Mining of Massive Datasets*. Cambridge University Press. ISBN 978-1107015357.
- [11] Liu B, Hsu W, Ma Y. [1999] Mining association rules with multiple minimum supports. In Proceedings of 5th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '99) (pp 337–341). New York
- [12] Moens.S, Aksehirli E, Goethalsn.B. Frequent Itemset Mining for Big Data”, *IEEE International Conference on Big Data*, 111-118.
- [13] Pimentel MA, Clifton DA, Clifton L, Tarassenko LA.[ 2014] review of novelty detection. *Signal Processing*, 99:215–249.
- [14] Rasheed, F, Alhajj, R. [2014] A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences, *IEEE Transactions on Cybernetics*, 44( 4): 569-582.

## ABOUT AUTHORS



*Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University ,Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007. She has totally 15 yrs of experience in teaching including 6 months in the industry. Presently she is working as Professor, Department of Computer Science in Bharathiar University, Coimbatore. She has published more than 50 research papers in National, International Journals and Conferences in India and abroad. She has guided more than 20 M.Phil., Research Scholars. She is Life member of CSI and editor in various International Journals.*



*P.Ajitha., received her B.Com from Bharathiar University, Coimbatore in 1998 and received MCA from Bharathiadsan University, Trichy in 2001. She obtained her M.Phil in the area of Data Mining in 2004. She has 9 years of experience in teaching and 3 months of industrial experience. Currently, she is working as Assistant Professor, Department Of M.Sc Software Systems and Computer Science, K.G College of arts and science, Coimbatore and pursuing her Ph.D in Bharathiar University, Coimbatore. She has presented more than 10 research papers in National and International Conferences and published mre than 5 papers in an various International Journals. Her Research Interest lies in Distributed Data Mining, Machine Learning and Artificial Intelligence. She is Life a member of CSI, life member of Institute of Advanced Scientific Research and also a member IASCIT.*

## EFFICIENT IDENTIFICATION OF DESICCATED LUMBAR IVD FROM MRI

Leena Silvester<sup>1</sup> and Mathusoothana S Kumar<sup>2</sup><sup>1</sup>Computer Science Department, College Of Engineering, Attingal, Thiruvananthapuram, INDIA<sup>2</sup>Information Technology Department, Noorul Islam University, INDIA

## ABSTRACT

Segmentation of Intervertebral disc (IVD) plays an important role in clinical diagnosis of lower back pain. Automatically segmenting the IVD in Magnetic Resonance Imaging (MRI) image is extremely challenging as variations in soft tissue contrast and radio-frequency (RF) in homogeneities cause image intensity variations. This paper proposes connected component analysis on intervertebral discs of mid-sagittal MRI data and detection of the abnormal disc. This approach uses union-find algorithm to determine the final label for each pixel. In the initial phase, each pixel is assigned a temporary label. During the final phase, it scans the image and converts the entire provisional label into final. Label associated with the root is having minimum label and it is finalized as the final label. This is the simplest approach and it takes less time to detect a disc as degenerative or not. This proposed algorithm is able to identify the number of lumbar disc which is degenerative. Also by means of visual inspection, we are able to identify the disc is degenerative or not. Here, the degenerative disc is L5-S1 which can be determined visually since it is broken. Localization of intervertebral discs is to locate the intervertebral discs by a point within a bounding box around the discs. The width, height, diameter of the degenerative disc is less than all other disc. Making use of these various statistics and the intensity profile we show that L5-S1 is degenerative. Experimental results show that the proposed method is very efficient and robust with respect to image slices.

Published on: 08<sup>th</sup>– August-2016

## KEY WORDS

Intervertebral Disc  
Degeneration; Lumbar  
Segmentation; Connected  
Component, Localisation of IVD

\*Corresponding author: Email: [leenasilvester@gmail.com](mailto:leenasilvester@gmail.com) Tel.: +91-9486856119

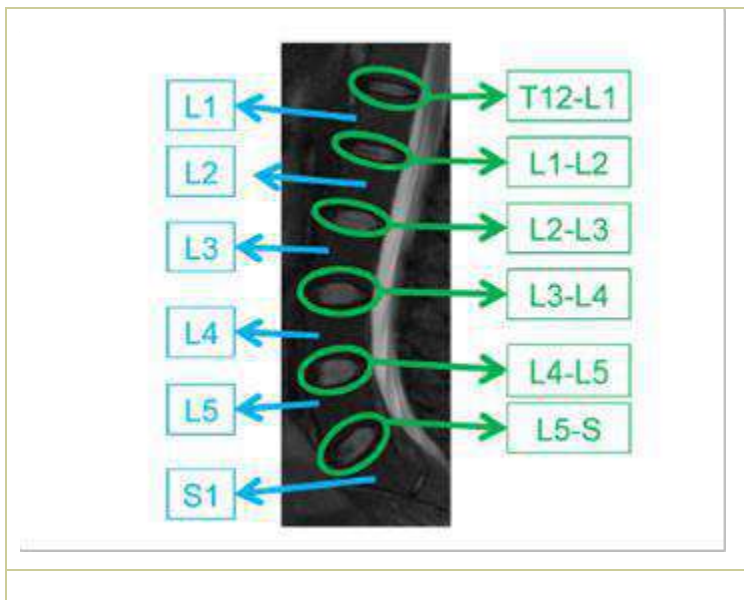
## INTRODUCTION

The most common health problem, low back pain is having high prevalence increases with age. Also, most studies have found that both the mean and median prevalence of low back pain was higher in women. From some studies it is also found a higher prevalence amongst older women as compared with older men [1]. Two studies observed that the occurrence of the low back pain is due to obesity or high body mass index (BMI) [2,3]. The dominant factors influence the onset and course of back pain includes low educational status smoking, obesity, older age, female gender, physically strenuous work, sedentary work, a stressful job, job dissatisfaction and psychological factors such as anxiety or depression, weight, and structural defects of the spinal column[4]. Congenital defects of the vertebrae as well as unequal leg length can cause abnormal loading in the spine. Degenerative spine conditions involve the gradual wear and tear to the discs, joints and bones of the spine over time, usually caused by aging. However, some studies reveal that, there is a correlation between the low back pain between the trauma and nontrauma patients. Low back pain patients with a history of trauma had more severe facet arthrosis than do nontrauma patients[4]. Also there is a scarcity of radiologist [5].

Most widely used imaging modality for the evaluation of intervertebral disc degeneration is the Magnetic resonance imaging (MRI). MRI extracts arbitrary slice orientation including coronal and sagittal views. The type of image “weighting” used during image acquisition determines MR contrast. The most commonly used images for lumbar spine applications are T1-weighted, T2-weighted, intermediate-weighted (proton density), and short tau inversion recovery (STIR). Over a last decade of years, diagnosis of disc degeneration is based on 2D analysis of MR image. T1-weighted images are useful for detection of fat, and fat acts as a natural contrast agent for detection of epidural or paraspinal lesions, marrow infiltration or replacement, focal bone lesions, and also the diagnosis of lipid-containing lesions, especially hemangiomas. In T2-weighted images identifies the bright structures, fluid-containing structures, such as CSF or urinary bladder, and “bright” hyperintense fluid signal. Also, it is useful for the detection of areas of bone marrow edema and is critical for disc related diagnosis. T2-weighted images are useful to show disc desiccation, hyperintensity zones, and Modic end plate findings.

Automatic localization of intervertebral discs from lumbar MRI is the prerequisite to the diagnosis of lower back pain. There has been much progress in research in determining a point within each lumbar disc. The human spine consists of 24 spinal bones, called vertebrae. Lower portion of the spine is known as the lumbar spine. An intervertebral disc, acts like a shock absorber, sits between each pair of vertebrae. Discs are made up of about 80% water, and the water content get decreased as the age progresses, and lose their ability to act as shock absorbers and these discs are known as degenerative. Discs have a tough outer wall (the annulus, a series of strong ligament rings) and a soft center (the nucleus). The nucleus is spongy and provides most of the disc's ability to absorb shock.

Localization of intervertebral discs is to locate the intervertebral discs by a point within or a bounding box around the discs as in [6], while the segmentation task is to provide a fine contour that accurately delineates a contour around the vertebra. Labeling, on the other hand, is to identify the anatomical nomenclature of each structure (e.g. labeling each of the five lumbar vertebrae as L1, L2, L3, L4 and L5). Localization and labelling of a sagittal lumbar T2-weighted MRI is shown in **Figure- 1**. Lumbar area is the second last area of the vertebral column; which are the weight-bearing portion of the spine. The lowest lumbar vertebra is L5 and the highest is L1. Intervertebral discs are labelled based on the enclosing vertebrae [14].



**Fig: 1. T2-Weighted MR sagittal view of lumbar region showing lumbar vertebrae (L1-L5) and the six discs connected to them (T12-L1 down to L5-S). The disc L5-S is sometimes called L5-S1 [10].**

Normal IVD looks like a saturated oval while the degenerative IVD appears to be dry and flat oval shaped. The rest of the paper is organized as follows: Section II, briefly review of segmentation of intervertebral disc and the localization and labeling of lumbar disc. Section III explains the proposed method. Some of the experimental results are shown in Section IV. Results are discussed in section V. Finally, in Section VI we give our conclusion remarks.

## STATE OF THE ART AND CONTRIBUTION

Peng et al. [7] proposed an automatic segmentation of the whole spine column. The procedure relies on intensity profile of the intervertebral disc by convolving it with the image. After fitting all discs on the template, disc centers are searched and then vertebral centers and it is extracted. They tested their technique only on 5 subjects, but the performance of the method for low quality scout data, or when local disc detection fails, is unclear. Weiss et al. [8] propose a semi-automatic approach for localizing and labeling the intervertebral discs. This algorithm is applied separately on upper and lower halves of the spine. The user will manually select a single seed point in the C2-3 disc and detects the remaining discs using intensity threshold values, filters and noise suppression operators. The success of this method is dependent on imaging quality and thresholding values. Zheng et al. [9] develop an approach to segment lumbar vertebrae from digital video fluoroscopic images and validated on synthetic data and a single in vivo sequence. These images are noisier than the standard MR radiographs but have a time component. Fourier descriptors were used to describe the vertebral body shape. This description was incorporated within Hough Transform algorithm. Masaki et al. [10] presents a procedure for automated localization the spine and the discs

based on intensity and a line detection filter to find the straight line between the spinal cord and the ID with the hough transform. With the help of five radiologists they validated 9 out of 10 cases that the automatic image plane is better than manual. A more recent study has been done by Pekar et al. [11] for the automatic detection of intervertebral discs in 3D MR scout scans as part of scan geometry planning (a step beyond selecting a particular slice from an existing scan). They used a 2D image filter for disc candidate detection are located by a filter using eigenvalues analysis of the image Hessian matrix, followed by 3D connected components to find disc centers. They uses a clever search is to locate the next disc through the candidates by distance constraint. If a disc is absent, new point is introduced and it is interpreted as missing disc due to abnormality. Chwialkowski et al. [12] introduce a heuristic-based procedure for automated localization of discs, vertebrae, and spinal cord analysis of the pattern of changes in grey intensities along the disc, and they also compared the variation of gray level intensities in healthy and damaged discs. Schmidt et al. [13] presents a probabilistic inference method using a part-based model to localize and label intervertebral discs in 3-D full back MR images. It incorporates appearance and shape information and uses the A\* algorithm for efficiently pruning the search space. This estimate is computed by exact inference on a tree-structured subgraph and guarantees global optimum. Alomari et al. [14, 15] proposed a graphical model for the lumbar disc localization that captures both pixel- and object-level features [14]. Using a Gibbs distribution, their model assumes local and global levels. Spatial information at the pixel level (global level latent variables) and at the object level (local latent variable), they model the spatial distribution of the discs and the relative distances. They use generalized expectation-maximization for optimization, which achieves efficient convergence of disc labels. They use a probabilistic model for automatic localization and labeling of the discs and outputs in a point inside each disc [15]. They orient each disc horizontally and then they use an Active Shape Model [16] to get a bounding box of each disc and extract intensity and texture features from each disc. Then they construct five classifiers and a voting scheme by running heterogeneous learning algorithms (SVM, PCA+LDA, PCA+Naive Bayes, PCA+QDA and PCA+SVM) [17] to detect a herniated disc. Ayse Betul Okay et al. [18] labels the lumbar vertebrae and discs together simultaneously employ a second-order MRF chain in our current work. The local image features are extracted from the image by employing Pyramidal Histogram of Oriented Gradients (PHOG) and a novel descriptor image projection descriptor (IPD). These features are trained with Support Vector Machines (SVM) and each pixel in the target image is locally assigned a score. These local scores are combined with the semi-global geometrical information like the distance ratio and angle between the neighboring structures under the Markov Random Field (MRF) framework. Their work can handle the missing structures in the MR images. Michopoulou et al. [19] demonstrated the classification of intervertebral discs into normal or degenerated using a Fuzzy C means algorithm in conjunction with atlas approach for spine MRI images which combines prior anatomical knowledge by means of a rigidly registered probabilistic disc atlas with fuzzy clustering techniques incorporating smoothness constraints. Combined algorithm minimizes the severe leakage of disc border due to the overlapping grey-level values between disc and surrounding tissues. Alomari et al. [20] does a survey of the localization, labeling, and segmentation problems for the various vertebral column structures from the available medical imaging modalities. Subarna Ghosh et al. [21] propose a tight bounding box for each disc after localization method. They computed HOG (Histogram of Oriented Gradients) features along with SVM (Support Vector Machine) as classifier to achieve 99% disc localization accuracy on 53 clinical cases. Bhole et al. [22] proposed an automatic detection of lumbar vertebrae and extract a rough ROI and finally the tight bounding box for each disc. They achieve 98.8% accuracy for disc labeling on 67 sagittal images.

In most of the previous research, researchers have been focused on finding a point inside the disc, which is a prerequisite of challenging segmentation step in order to diagnose a disc abnormality. In our work we provide tight bounding boxes for each disc in the lumbar region so that we automatically provide diagnostic results after complicated segmentation.

## METHODS

We present a fully automated method of accessing quantitatively from mid-sagittal MRI.

### Image preprocessing

Preprocessing of the image starts with enhancing the image. The noise and low contrast responsible for these failures are the fundamental obstacles to successful automatic segmentation of MRI images. In order to reduce the random noise we apply 3\*3 median spatial filters. Median filters will not blur the edges as much as a comparable linear low pass filters

### Image Normalization

Image normalization tries to reduce the effect of variation in the input images. Two common methods of normalization are contrast stretching and histogram equalization. Contrast stretching applies a linear transformation to the input image so that the intensity histogram is stretched across the full range of possible pixel intensity values. Image normalization also tries to resize the input image using either nearest-neighbor interpolation or bilinear interpolation or bicubic interpolation. When the specified output size is smaller than the size of the input image and method is 'bilinear' or 'bicubic', resizing applies a lowpass filter before interpolation to

reduce aliasing. Resizing reduces the processing time. Contrast-limited adaptive histogram equalization (CLAHE) was performed as one preprocessing step.

## Gradient computation and edge detection

Preprocessing step includes the computation of gradient of the image. Change of image occurs at the boundary and gradient is used to find the boundary. Magnitude of the gradient describes how quickly the image changes in either x or y direction. Therefore, edge detection with Sobel operator is applied in  $-x$  and  $-y$  directions.

## Connected Component analysis

Connected Component analysis is done in 3 phases; pre-scanning phase, analysis phase and scanning phase. In the first phase, pre-scanning phase, provisional labels are assigned and building of an equivalence array is maintained. This phase scans pixels and its neighbours (in a distance  $r$  are considered) and represent the equivalence information as a rooted tree using union-find algorithm. Union-find algorithm maintains the equivalence information of the provisional labels assigned as a rooted tree. Three operations are needed to implement union-find algorithm: makeset, find and union. The second phase, analysis phase, does not access the image directly. It analyses the union-find algorithm to determine the final label for each pixel. During this phase, it scans the image and converts the entire provisional label into final. Label associated with the root is having minimum label and it is finalized as the final label.

## Algorithm

To achieve fully automatic intervertebral disc location from a set of sagittal image slices, the following procedure is to be performed.

### Algorithm 1: Connected Component analysis-Algorithm

1. Scan the image, left to right, top to bottom
2. If the pixel is 1, then
  - (a) If only one of its upper and left neighbours has a label then copy the label
  - (b) If both have the same label, then copy the label
  - (c) If both have different label, build an equivalence array and form the union of those two pixels by making the root node of one of the pixels point to the root node of the other.
  - (d) Otherwise, assign a new label to this pixel and enter this label in equivalence table
3. Repeat Step 2 until no nonzero pixels are left and the equivalence classes are completely determined

During the second pass, scan the image, left to right, top to bottom and find the lowest label for each equivalence set in the equivalence array. Scan the image for the second time and replace each label by the root i.e., Assign each pixel the label of the equivalence class

To achieve fully automatic intervertebral disc location from a set of sagittal image slices, the following procedure is to be performed.

### Algorithm 2: Segmentation of IVD

Step 1: Let us consider a set of  $N$  sample spine images  $X \in \mathfrak{R}$ , where  $X = \{x_1, x_2, \dots, x_N\}$  taking values in an  $N$ -dimensional image space and  $\mathfrak{R}$  represents universe. Assume each image  $x_i \in \mathfrak{R}$ , have a sequence of sagittal image  $I_1, I_2, \dots, I_k$ .

Step 2: Select a best MR image slice  $I_i$  from a sequence of sagittal image  $\zeta = \{I_1, I_2, \dots, I_k\}$  in order to save processing time and achieve better detection results.

Step 3: Preprocessing can be carried out on this slice.

Step 4: Normalization enhances the contrast of the image by transforming the values using contrast-limited adaptive histogram equalization (CLAHE) and then resize the image

Step 5: Compute the gradient of the image.

Step 6: Remove all connected components from the binary image.

Step 7: Perform connected component analysis on the binarized preprocessed image to locate and labeling all the visible intervertebral discs (IVD) in the best slice  $I_i$ .

Step 8: Search for missing discs in other slices  $I_i \in \zeta$ .

Step 9: Calculate the length and width of minimum bounding rectangle of each intervertebral disc.

Step 10: Calculate  $H_{avg} = T_p / t$  where  $t$  is the height of the IVD and  $T_p$  is the sum of the pixels of the height of each IVD

Step 11: Calculate minimum height  $H_{min}$  and check with threshold value for abnormal condition.

## RESULTS

This database is freely available from spine web [23] and it was used to detect vertebrae and intervertebral disc (DataSet 7 in the webpage) [24]. The dataset contains 15 T2-weighted turbo spin echo MR images and the reference manual segmentation.

### Quality analysis

Before quantitative analyses, the quality of the segmentations was visually within the first 2 seconds of the visual inspection. Input image is represented in **Figure-2**. Boundaries of the disc are identified and plotted in **Figure-3**. It is noted that the last disc L5-S1 is broken into two regions and it is identified as degenerative. Localisation of disc is shown in **Figure-4** after identifying the disc and the application of the algorithm.

### Quantitative analysis

**Table-1** shows various statistic of IVD. The height, width, standard deviation and area of the discs are calculated and plotted in **Figure-5 to Figure-8**.



Fig: 2. Input Image



Fig: 3. Boundary of Disc

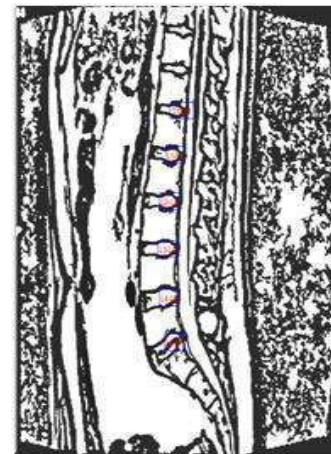


Fig: 4. Localisation of IVD

Standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A standard deviation close to 0 indicates that the data points tend to be very close to the mean of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. So we can conclude that L5-S1 is degenerative since it is having high standard deviation. The width, height, diameter of the disc is also decreased. The width of the degenerative disc is 4.78 which is less than all other disc. The minimum value of height is 4.35, which is the degenerative one. Diameter of the degenerative is less than all other disc and it is 21.56. Statistics of IVD is shown in **Table-1**. Thus; we can conclude that L5-S1 is degenerative



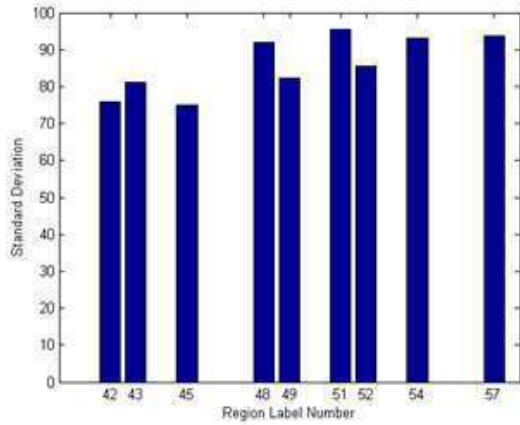


Fig. 5. Standard Deviation

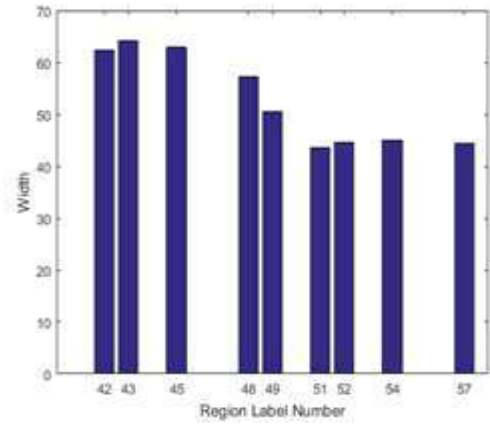


Fig. 6. Width of disc

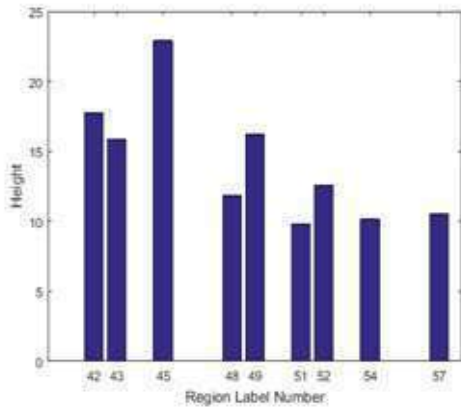


Fig. 7. Height of disc

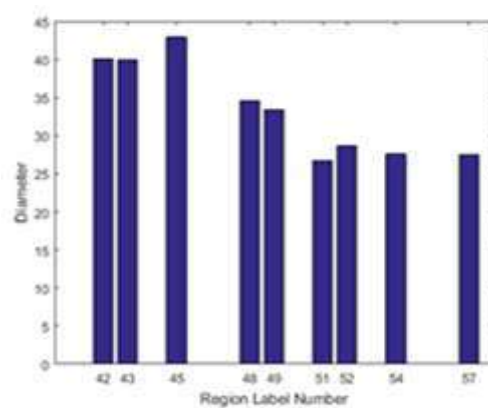


Fig. 8. Diameter of disc

### Evaluation of Tissue Properties within the Disc

For a normal healthy disc, the degree of hydration and resultant signal intensity diminishes as one proceeds outward from the geometric center to the peripheral annular fibers. Hence, if we plot the normal distribution of image intensities, it forms an axially symmetrical, bell-shaped curve. With the degenerative disc, the symmetry of the grey intensities is not a bell shaped curve and it is totally disturbed, and the center of the peak is flattened. The peak of IVD is flattened in the [Figure-9](#).

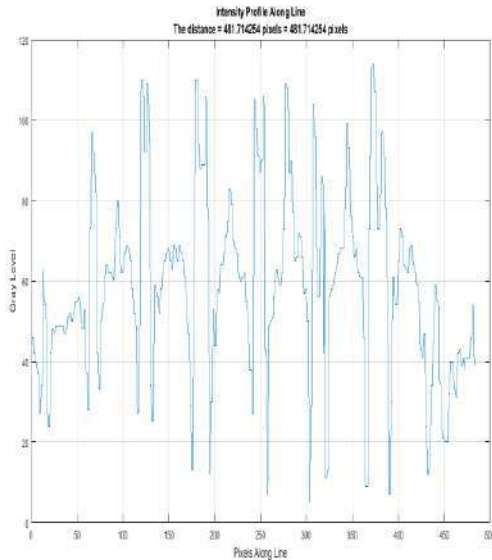


Fig: 9. Intensity profile

Table: 1. Statistics of IVD

Height	Width	Diameter	Standard Deviation
4.78	4.35	21.56	46.39

## DISCUSSION

The results are shown in **Figure-3** to **Figure-8**. From this it is clear that the width, height, diameter of the degenerative disc is decreased. The width of the degenerative disc is 4.78 which is less than all other disc. The minimum value of height is 4.35, which is the degenerative one. Diameter of the degenerative is less than all other disc and it is 21.56. Statistics of IVD is shown in Table 1. The intensity profile of the disc is shown in figure 9 which shows that L5-S1 is degenerative Thus; from all the results we can conclude that L5-S1 is degenerative.

## CONCLUSION

This proposed work focuses on the 2D segmentation of intervertebral disc. Here, T2 images are the input of the algorithm since recent studies are concentrated on T2 images. This method is a fully automatic and no manual seed selection is required for the working of this procedure. Application of connected component analysis on a preprocessed binary image will result in reliable classification of disc as normal or abnormal while maintaining robustness. Computation time is also less. The result will depend upon the preprocessed image. If the image is a smooth one we may lose a connected component and thus it may lead to missing a disc or algorithm will result in incorrect classification of disc as degenerative or not.

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

None.

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Hoy D, Brooks P, Blyth F, Buchbinder R. [2010] The epidemiology of low back pain. *Best practice & research Clinical rheumatology* 24(6):769-781.
- [2] Vogt MT, Lauerma WC, Chirumbole M et.al. [2002] A community- based study of postmenopausal white women with back and leg pain: health status and limitations in physical activity. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 57(8):M544–550.
- [3] Webb R, Brammah T, Lunt M. [2003] Prevalence and predictors of intense, chronic, and disabling neck and back pain in the UK general population *Spine* 28(11):1195–202.
- [4] Peterson CK, Bolton JE, Wood AR. [2000] A cross-sectional study correlating lumbar spine degeneration with disability and pain, *Spine* 25(2):218.
- [5] Bhargavan M, Sunshine JH, and Schepps B. [2002] Too few radiologists?. *American Journal of Roentgenology*, 178(5): 1075–1082.
- [6] Ghosh S, Malgireddy MR, Chaudhary V, Dhillon. [2012] G A new approach to automatic disc localization in clinical lumbar MRI: combining machine learning with heuristics. In: *Proceedings of IEEE international symposium on biomedical imaging*: 114–117.
- [7] Peng Z, Zhong J, Wee W, Lee Jh. [2005] Automated vertebra detection and segmentation from the whole spine MR images. In: *Proc. Engineering in Medicine and Biology, IEEE-EMBS 2005*: 2527–2530.
- [8] Weiss KL, Storrs JM, and Banto RB. [2006] Automated spine survey iterative scan technique, *Radiology*.239(1):. 255–262.
- [9] Zheng Y, Nixon MS, Allen R. [2004] Automatic segmentation of lumbar vertebrae in digital videofluoroscopic imaging, 23(1): 45–52.
- [10] Masaki T, Lee Y, Tsai DY, Sekiya M, and Kazam K. [2006 ] Automatic determination of the imaging plane in lumbar MRI, in *Proc.SPIE Med. Imag* 1252–1259.
- [11] Pekar V, Bystrov D, Heese HS, et al. [2007] Automated planning of scan geometries in spine MRI scans. in *Proc. MICCAI*. New York: Springer, *Lecture Notes Computer Science*,4791:601–608.
- [12] Chwialkowski, MP., Shile, PE., Peshock, RM., Pfeifer, D., Parkey, RW. [1989]: Automated detection and evaluation of lumbar discs in mr images. In: *Proc. of IEEE EMBS*.
- [13] Schmidt S, Kappes J, Bergtholdt M, Pekar V, Dries S, Bystrov D, and Schnörr C. [2007] Spine detection and labeling using a parts-based graphical model. in *Proc. 20th Int Conf Inf Process Med Imag* 4584: 122–133.
- [14] Alomari RS, Corso JJ, and Chaudhary V. [2011] Labeling of lumbar discs using both pixel- and object-level features with a two-level probabilistic model. *IEEE Trans. Med. Imag* 30(1):1–1.
- [15] Corso JJ, Alomari RS, and Chaudhary V. [2008] Lumbar disc localization and labeling with a probabilistic model on both pixel and object features. in *Proc. Med. Image Comput. Comput.-Assisted Intervention Conf vol 5241*: 202–210.
- [16] Cootes T. [1995] Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1): 38–59.
- [17] MathuSoothana S Kumar, R. and Muneeswaran, K. [2011] An Improved Face Recognition Technique Based on Modular LPCA Approach. *Journal of Computer Science (JCS)*, 7(12):1900-1907.
- [18] Oktay AB and Akgul YS. [2013] Simultaneous localization of lumbar vertebrae and intervertebral discs with svm based mrf. *IEEE TMI*
- [19] Michopoulou S, Costaridou L, Panagiotopoulos E, Speller R, Panayiotakis G, Todd-Pokropek A. [2009] Atlas-based segmentation of degenerated lumbar intervertebral discs from mr images of the spine. *IEEE Trans Biomed Eng.*
- [20] Alomari RS, Ghosh S, Koh ,Chaudhary V. [2015] Vertebral Column Localization,Labeling, and Segmentation. *Spinal Imaging and Image Analysis* 18:193.
- [21] Ghosh S, Malgireddy MR, Chaudhary V, Dhillon G. [2012] A new approach to automatic disc localization in clinical lumbar MRI: Combining machine learning with heuristics. 9th IEEE International Symposium on InBiomedical Imaging (ISBI): 114-117.
- [22] Bhole C, Kompalli S, and Chaudhary V. [2009] Context-sensitive labeling of spinal structures in mri images, in *The Proceedings of SPIE medical imaging*.
- [23] Spineweb. [2013] <http://spineweb.digitalimaginggroup.ca/>, Accessed: 2016-01-29.
- [24] Yao J, Burns JE, Munoz H, Summers RM. [2012] Detection of vertebral body fractures based on cortical shell unwrapping. *Springer in Medical Image Computing and Computer-Assisted Intervention–MICCAI* : 509–516.

# DESIGNING AN INTEREST SEARCH MODEL USING THE KEYWORD FROM THE CLUSTERED DATASETS

E. Ajitha, A. Nirmal Kumar\*, A. Jayanthi, D. Daya Florance

Assistant Professor, Dept. of Information Technology, Veltech Hightech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai. INDIA

## ABSTRACT

A social blogging service is one of the most abundant resources for data collection about one's personal interest on various things. Collection of these data will result in the explosion of short text messages. The analyzed and collected data contains enormous amount of noise and redundancy. The small scale data set along with a database is designed to collect and handle the text streams. These datasets are clustered based on time (Day wise). Keyword filtering is used to remove the noisy and outlier datasets. A user purchase model is designed which holds two search criteria Such as General Purchase and Profile Based Purchase. The interest is monitored in a continuous manner in social blog through a Hadoop Server.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

Clusters, Keyword filtering, Outlier, Recommendation, Hadoop

\*Corresponding author: Email: [sa.nirmalkumar@gmail.com](mailto:sa.nirmalkumar@gmail.com); Tel.: + 91 9443998771

## INTRODUCTION

Social blogging has been so popular since its evolution in the world of internet and the amount of short text messages in the form of posts, likes, and chats range up to several millions per day. Inquiring social blog for a specific person's interest will be helpful in saving time over searching for his/her favorites over internet. Existing systems does not hold this property of displaying one's interest. Recommender systems, so far designed, deals only with the past browsed history of the user. Our proposed model discusses to deliver the user interest by monitoring one's personal interests over a social blog.

Hadoop server is used for continuous monitoring. Hadoop splits files into large blocks and saves them across cluster's nodes. To process the stored data, it transfers confined code to the nodes for parallel processing of data. This approach takes advantage of locality of the data nodes manipulating the data they have accessed to allow the dataset to be processed faster and more efficiently.

The increasing popularity has resulted in overwhelming of data even though they are informative. Retrieving those data is a tedious process even though filtering is used. Summarization along with clustering can be introduced to overcome these difficulties in retrieving a particular data.

## RELATED WORKS

A clustering technique is used for handling real time data sets [1]. The simple one-pass algorithms are too inefficient for the incoming data streams. From the application point of view, these algorithms are not enough to process the clusters. Since, they are large in size and these algorithms do not address the size. It is suggested to have micro cluster which refers to the data locality that points to the location of data inside a cluster.

A semi supervised co-clustering with side information [2] is used to process these clusters. This technique

describes to carry additional information about the main text such as author name, publication details and so on. Due to which the overhead in searching the cluster and its related co-clusters is increased. Side information is only necessary when the main context is damaged or corrupted due to error in the cluster. In those situations it is better to avoid collecting the side information since data corruption occurs very less in a cluster. Rootdatasets can only be concentrated for specific information to be mined.

BIRCH (Balanced Iteration Reducing and Clustering Using Hierarchies) is an improved way to cluster the datasets of related entities and related datasets of similar entities [3]. Balanced iteration additively and variably clusters the related datasets from various data localities and produces the best cluster based on the recent time frame. Best clusters are produced considering the available memory space. It will be more suitable for large clusters to hold on a certain threshold value in retrieving for particular dataset which can possibly eliminate outlier datasets.

Bursty feature representation along with the text mining technique is used on a classical static datasets. Bursty features are nothing but a part of text which is commonly repeated over and over in several other posts, the text may refer to a particular product or an event that are generated on a relatively short period of time [4]. It identifies the feature for clustering the datasets which in turn causes noisy datasets to be retrieved along with the queried information. This also results in exponential increase in cluster size.

Periodical time frame (say three sessions / event) and calculating the frequency of bursty texts over those time frames or session may improvise the system. The cluster with the highly populated bursty texts is selected for mining the useful information.

A different approach to overcome the difficulties faced by [5] in terms of previous time period cluster which combines online grouping algorithm with the existing scalable clustering techniques to produce fast and adaptive clusters of text streams [6]. Yet it fails to retain too old clusters due to memory constraints. It is suggested that sufficient weightage to the old clusters can be allotted and based upon which they can be dropped or kept for usage. The more historical data with less weightage can be neglected if more weightage clusters of the same period have evolved.

A framework for clustering hefty text and definite text streams [7]. So far [4] [5] [6] has only met with problem of time and space. But a real time handling and clustering along with segmentation for organizing documents in applications has been done. It uses statistical summarization methodology to cluster the data to individual text streams or as a categorical data streams that represents those respective topics. The drawback of this proposal is that the temporal locality of the clusters is not taken into account. Moreover repeated querying of different kinds cannot be answered as quickly as the incoming data streams.

It can be easily overcome by having a continuous monitoring of social profiles so as to get the instant updates of each activity. Hadoop server can be deployed for such monitoring of data streams. Hadoop uses a special MapReduce function to count on each repeated activity of an individual user. It also implements a programming model for processing the huge amount of incoming text streams.

To identify the keyword in a sentence (post or status) a new concept which constructs a lexical chain for text summarization [8] is introduced. Summarization by this technique causes only strong chains to be identified. This paper does not address the sentence granularity which extracts the central constituents from the text. It does not implement necessary keyword filtering techniques to control the length of the posts.

## DISADVANTAGES OF EXISTING SYSTEM

1. More complex datasets cannot be handled.
2. Waiting time is increased.
3. Less accuracy.

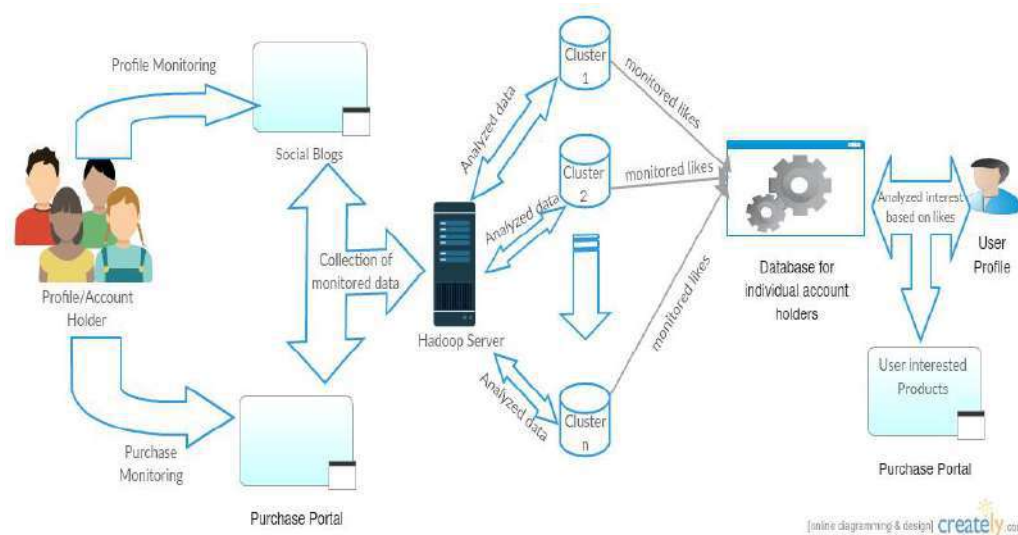
4. Low data transmission rate.
5. Replication or requests due to inefficient offline grouping algorithm.
6. Datasets containing noise are also retrieved when mining important posts.

## PROPOSED WORK

Designing a purchase model on user interest with the filtered keyword from the user posts and the frequency of posts on a particular product or a particular topic is implemented. Purchase Portal will have two options like General Purchase & Profile based Purchase. In General Purchase, usual items of a shopping site is recommended for purchase. In Profile based Purchase, Items are displayed based on the User's Interest. Related Items and Items which are purchased more often are also displayed to the user based on the User Interest.

Continuous monitoring of user profile maintained in social blog is carried. Following the user posts and likes, a Hadoop Server collects the data from user's profile and records it in a database. This database has bi-directional connectivity with both the user account on the blog and the purchase model designed for shopping. Updation of posts which user makes in the blogging site with minimal time delay is implied for the ease of shopping user's favorite product.

Summarization of users posts are produced by the means of continuous monitoring of user profile over days spanning to several weeks. This helps in predicting the user interest very closely with less error rate and achieving high accuracy as the time extends. Recommendation based on history is also available for inactive blog users.



## ADVANTAGES OF PROPOSED SYSTEM

1. Reducing sampling rate consumes less time over computation of online clustering algorithm.
2. Accuracy is improved by mining sub clusters for additional filter.
3. Reliability over time frames.
4. Replica of requests is avoided.
5. Supports a range of data analysis tasks such as reports or historical survey.

## OPERATIONS PERFORMED

1. DATA INSERTION & UPDATE.
2. CONSTRUCTION OF UP-TREE.
3. PURCHASE THE ITEM.
4. DISCARD UNPROMISING ITEMS.
5. FINDING THE PATTERN FROM UP TREE.

## DATA INSERTION & UPDATE

In administrator side, one can add particular data and also update the data. Here, data refers to products with its associated values. Each item contains the profit and quantity. We can add and update the profit and quantity values for product list. All inserted values are added in the database. Also updated values are added in the database. We can also view the added and updated values. These values along with the items are will be displayed in the product list.

## CONSTRUCTION OF UP-TREE

A compact tree structure is used for discovering immense utility item sets and maintaining the information about patterns within databases. Immense utility item sets can be generated from UP-Tree efficiently with only two scans of original databases.

In the first scan, transaction unit of each transaction is computed. At the same time, transaction utility weight of each single item is also accumulated. An item and its supersets are unpromising to be immense utility if its weight is less than the defined threshold. An item is called a promising item if weight is greater than the minimum utility threshold. Otherwise it is called an unpromising item. Generally, an item is also called a promising item if its overestimated utility is no less than minimum utility. Otherwise it is called an unpromising item.

New transaction unit after discarding unpromising items is called reorganized transaction. By reorganizing the transactions not only less information is needed to be recorded in UP-Tree. Since the utilities of unpromising items are excluded, reorganized transaction must be no larger than utility weight. In the second scan, reorganized transactions are inserted into the UP Tree. Hence, the high potential utility item sets can be efficiently generated from the UP-Tree.

## ELEMENTS IN UP-TREE

In a UP-Tree, each node  $N$  consists of name, count,  $nu$ , parent, hyperlink and a set of child nodes where name is the node's item name; count is the node's support count;  $nu$  is the node's node utility, i.e., falsehood utility of the node; parent records the parent node of  $N$ ; hyperlink is a node link which points to a node whose item name is the same as node's name.

A table named header table is employed to enable the traversal of UP-Tree. In header table, each entry records an item name, a falsehood utility, and a hyperlink. The hyperlink points to the last referral of the node which has the same item as the entry in the tree. By following the hyperlinks in header table, the nodes having the same name can be traversed efficiently.

## PURCHASE THE ITEM

In client side user can enter all details. Then user can login using particular username and password. All the inserted also updated items are added into the product list. Then select user wanted items then add all items into cart products with count of the each item. A warning message will display in dialogue box when the customer type the quantity above the constraint value mentioned in the database. All selected items are displayed in the cart product list. Then purchase the required items.

## DISCARD UNPROMISING ITEMS

In Frequent Pattern growth mining, the minimum utility is entered. Based on the value of minimum utility we can find out the promising items and unpromising items. Transactional-weighted utility of an item set is the sum of the transaction utilities of all the transactions. If it's utility is less than a user specified minimum utility threshold. An items set is called a low utility item set. To select specific rules from the set of all rules, constraints on significance and interest can be used. The best known constraints are fixing the threshold based on support & confidence provided by trusted source. Then the low utility item sets are discarded at end of the transaction.

In Data Mining the task of finding the most hit pattern in large databases is very important as it computationally more expensive, when a more number of patterns exist. These patterns which are mined during the various approaches make the user very difficult to identify the patterns which are very interesting for them. The goal of most hit itemset mining is to identify all frequent itemsets. Once the frequent itemsets are identified, association rules are generated for the identified itemsets.

In the real world, however, each item in the purchase portal has a different importance/price and single customer will be interested in buying a number of same products. Therefore, finding only classic frequent patterns in a database cannot fulfill the need of searching the valuable itemsets that contribute the most to the profit in a retail business.

## FINDING THE PATTERN FROM UP-TREE

Searching process for immense utility item set mining is difficult because a product or a item of a low utility item set may be a immense utility item set. If transactional-weighted utility is no less than a user specified threshold value. An item set is called a high utility item set. Based on the TWU we can find out promising items and unpromising items. Based on the threshold value we discard the unpromising items. Then find out the promising items. Candidate item sets are generated with the previously discussed database scans. Mining high utility item sets from database refers to the discovery of item sets with high utility like profit.

## CONCLUSION

The proposed work is a prototype which supports continuous text stream summarization for enhanced blogging site with many facilities. It employs a text stream clustering algorithm to compress posts and texts into TCVs and maintains them online. Then, it uses a summarization algorithm for generating summaries contained in online as well as in history with random time durations. The topic progression can be observed automatically to produce varying timelines for text streams. In future work, development of a multi-topic version of clustering algorithm in a distributed system can be introduced and to evaluate it on more complete and large-scale data sets. Monitoring of more than one account can be considered for future implementations

### CONFLICT OF INTEREST

Authors declare no conflict of interest.

### ACKNOWLEDGEMENT

None.

### FINANCIAL DISCLOSURE

No financial support was received to carry out this project.



## REFERENCES

- [1] CC Aggarwal, J Han, J Wang, and PS.Yu. [2003] A framework for clustering evolving data streams, in *Proc.29th Int Conf Very Large Data Bases*, 81–92.
- [2] RuchaBhutada and DA Borikar.[2015] An Approach to Semi-Supervised Co-Clustering With Side Information in TextMining, *International Journal of Engineering Trends and Technology (IJETT)* –19 ( 6)
- [3] T Zhang, R Ramakrishnan, and M Livny.[ 1996] BIRCH: An efficient data clustering method for very large databases, in *Proc ACM SIGMOD Int Conf Manage Data*, 103–114
- [4] Q He, K Chang, EP Lim, J Zhang.[ 2007] Bursty feature representation for clustering text streams, in *Proc SIAM Int. Conf Data Mining*, 491–496.
- [5] J Zhang, Z Ghahramani, and Y Yang.[ 2014] A probabilistic model for online document clustering with application to novelty detection, in *Proc Adv Neural Inf Process Syst*, 1617–1624.
- [6] Shi Zhong.. Efficient Streaming Text Clustering” Department of Computer Science and EngineeringFlorida Atlantic University, Boca Raton, FL 33431
- [7] CC Aggarwal and PS Yu.[ 2010] On clustering massive text and categorical data streams, *Knowl Inf Syst*, 24(2): 171–196.
- [8] R Barzilay and M Elhadad.[ 1997] Using lexical chains for text summarization, in *Proc. ACL Workshop Intell. Scalable Text Summarization*, 10–17.

# MULTILAYER PERCEPTRON WEIGHT OPTIMIZATION USING BEE SWARM ALGORITHM FOR MOBILITY PREDICTION

J. Ananthi<sup>1\*</sup> and V. Ranganathan<sup>2</sup>

<sup>1</sup>Dept. of Electronics and Communication Engineering, Dhanalakshmi College of Engineering, Chennai, TN, INDIA

<sup>2</sup>Dept. of Electrical and Instrumentation, Dr. Mahalingam College of Engineering & Technology, Pollachi, TN, INDIA

## ABSTRACT

The progress in wireless networks has led to the rising demand of quality in service, reduced delay, seamless network, communication anytime/anywhere, and a lot more. The wireless network provides global services/communication through integrated networks. The wireless networks and beyond (4G) makes people free from cable and guarantee a fully distributed communication with promising Quality of Service (QoS). Hence, Mobility prediction or precise and competent forecast of mobile users trail is of prevailing significance for entire network performance. Mobility prediction along with wireless communication protocols helps in better energy, resource management in a network scenario and provides improved quality to the wireless users. This paper proposes mobility prediction based on a Multi-Layer Perceptron (MLP) network optimized with Bee Swarm Algorithm (BA). The proposed model evaluates mobility prediction using mobility traces from wide production wireless network. The Swarm Intelligence (SI) is used in many complex optimization problems in continuous search. The BA is the foraging behaviour of bees in searching food sources. The BA algorithm integrates the network for optimization of weights.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

Wireless Networks, Mobility Prediction, Multi-Layer Perceptron (MLP), Bee Swarm Algorithm (BA)

\*Corresponding author: Email: [ananthi\\_research@rediffmail.com](mailto:ananthi_research@rediffmail.com)

## INTRODUCTION

Wireless networks serve as the transport mechanism among devices and between devices and the traditional wired networks (enterprise networks and the Internet). Wireless-networks make people free from predetermined locations and introduce mobility in most of the aspects in the human life. The wireless networks have bandwidth and resource usage constraints due to increases in the number of users, which lead to termination in communication with reduction in success rate. Since the users are mobile most of the time mobility must be introduced in the systems. The mobility management schemes should be efficient enough to ensure seamless communication.

Wireless network, are classified into two divisions: With and without infrastructure. A mobile node which has free access to move around, even while conducting communication among other nodes, also known as an infrastructureless wireless networks, can allow free nodal access whilst the base stations remain fixed. Whereas Ad hoc networks or infrastructureless networks, have no fixed stations although the nodes are few to move about while communicating, and all the nodes act as routers. Ad Hoc Network mobile networks try to dynamically establish themselves to form routes and form their own 'on the fly' networks [1].

Wireless technology communications have a vast range of capacities which concentrate on a unique needs, and offers multiple benefits like flexibility, portability, increased productivity, and lowered installation costs. For example, WLAN or wireless local area networks allow users to shift their laptops within the surroundings of their office space, without the need for additional items like cables, wires or the worry of losing network connectivity. The lack of wires enables greater flexibility, increased efficiency rates and reduction in wiring costs [2].

Data synchronization in network systems are visible in Ad Hoc networks, enabled through Bluetooth technologies and it allows sharing application among devices. Bluetooth technology reduces cable connectivity in any peripheral devices, such as printers. Personal digital assistants (PDAs) and mobile devices, such as cell phones, enable synchronization to personal databases and give access to wireless e-mail features, browsing the web and general internet access. The above technologies save costs and enable diverse applications to fulfill their goals, such as manufacturing shop floors to first responders.

Wireless networks provide optimal services to mobile terminals and it is a fundamental function of the above networks to be aware of where the points of attachment exist at any given point of time. Predictions made by mobile networks are commonly used to aid many network management tasks and at a network level, such mechanisms reduce congestions and enhance service provision qualities. But at an application level, mobile networks are exploited to enable multiple location services. Thus, mobility predictions are an important function of the above techniques and it can determine the location of mobile terminals by manipulating available information carefully. And the accuracy of prediction devices highly depends on movement of user models and the prediction algorithm adopted [3].

The mobility prediction is helpful in allocating resources since it predicts the movement of the user priorly which depends on the past movement of the user. Mobility prediction may be defined as finding the next Access Point (AP) that a user will attach to in a network. The mobility of the user is determined from mathematical models that depict human mobility. The proactive methods like mobility prediction guarantee QoS to the users. Mobility predictions of wireless devices helps the users in smart access and useful to the service provider in planning of the infrastructure and to provide better QoS. GSM based mobile prediction networks also contain multiple routers and APs, like other mobile networks. Mobility prediction essentially tries to predict the movement of mobile users based on prior mobility models. The user's next location as the user is traveling in the network is helpful for infrastructure planning, resource allocation and future network requirement prediction [4].

All connections to its wireless network have been stored as trace files which is a valuable mine of information. The trace files contain the wireless Network Interface Card (NIC), Medium Access Control (MAC) addresses and the time of connection/disconnection for each access point. Since MAC addresses are unique, can safely assume that each connection/disconnection associated with a MAC ID belongs to a unique user and can be treated as a mobile node.

The wireless trace dataset contains user histories for thousands of wireless users. This data to perform simulation to verify the accuracy of the model versus actual movement, study user movement, prediction benefits, etc. [5]. A large portion of recent research still assumes that user mobility and the connection trace for a MT are strongly dependent.

There are a large number of prediction systems that have been proposed which attempt to measure or capture some regularity of the user's mobility in order to extrapolate from this knowledge about the future behaviour of the user's MT. Real life mobility traces have Mobility Prediction in Wireless Networks Using NN 323 shown that this assumption of user mobility and connection trace of the MT is not as valid as most researchers believe [6].

User movement trajectories are generally logged in at the time when a mobile device is connected to an Access Point (AP), which represents a specific AP of the nth user location has moved from defined time [7-9]. Mobile devices actual movement is called User Actual Path (UAP) which has the form

$$u_i = \langle ap_1, ap_2, \dots, ap_n \rangle$$

The User Mobility Pattern (UMP) is traced from the logs obtained from all APs which show the frequent path used by a mobile device [10-12].

The UMP is used to form the mobility rules. The user's mobility pattern from its actual movement is obtained called as user's real path,

$$U_j = \{AP_1, AP_2, AP_3, \dots, AP_n\}$$

The mobility rules obtained from the users real path are

$$l_1 \rightarrow l_2$$

$$l_1, l_2 \rightarrow l_3$$

$$l_1, l_2, l_3 \rightarrow l_4, \dots, l_n$$

Each rule has a confidence 'c' and support 's'. The rule that generate highest confidence is selected.

For a given set of patterns for mobility, Head class label determine the next location prediction and is given by

$$a_1, a_2, \dots, a_n \rightarrow b_i$$

Where  $a_i \in A$  and  $b_i \in B$

Optimization can be defined as the process of finding best solution or result under given circumstances. Generally optimization is used for maximizing or minimizing the value of a function, it may be local optimum or global optimum. In optimization there are different types of problems are being utilized like, Liner optimization, Non liner optimization, Dynamic optimization etc. and all have different techniques for solving. Efficiency in system optimizations and processes is sssential for the proper function and economics of mant engineering and science domains. This process is key for the proper functioning of many domains and many problems are solved by adopting approximate and rigorous mathematical search techniques. These intensive approaches have used linear, integer and dynamic programming to arrive at optimal solutions for moderately sized problems. But real life optimmmization problems have encountered engineering problems because of its huge size and complex solution space. Thus, finding exact solutions to these problems is quite difficult and requires an emponential amount of computing time and power to increase the number of decision variables. In order for researchers to overcome such problems. Approximate evolutionary-based decisions and proposals have been created to overcome the search for near-optimum solutions [13].

This paper suggests MLP weight optimization using BA. Section 2 reviews related literature. Section 3 describes the methods employed in this work. Section 4 describes experiment results and Section 5 concludes the work.

## RELATED WORKS

A new geo-statistical unsupervised learning technique to identify useful information on mobile phone using hidden patterns was presented by Manfredini et al., [14]. These are regarding different use of the city in time and space related to individual mobility, outlining the technology's potential for the urban planning community. The methodology ensured a reference base that reports the specific effect of activities on recorded Erlang data and a set of maps showing each activity's contribution to local Erlang signal. This technique chose results as significant to explain specific mobility/city use patterns and tested their significance and interpretation from an urban analysis and planning perspective at a Milan urban region scale.

Prediction-based replication methods which achieved service coverage through replication of a central server proposed by Surobhi&Jamalipour [15] were unable to accurately predict future topological changes and maintain service coverage in a post-emergency network. The realistic mobility model including users' post-emergency complex behavioral changes were proposed. A Machine-To-Machine (M2M) networking-based service coverage framework for post-emergency environments accurately predicted new user mobility and optimal replication. It used these predictions to achieve continuous service coverage. Simulation verified the proposals effectiveness.

A complete framework that proactively defined QoS/QoE-aware policies for Long-Term Evolution (LTE)-connected vehicles to select most adequate radio access from available access technologies that maximized QoE throughout mobility path was introduced by Taleb&Ksentini [16]. The policies were communicated to users following 3GPP standards and enforced by user equipment devices. Two different models to model the network selection process were proposed. Network selection process was modeled using a time-continuous Markov chain, and its performance was evaluated through NS2-based simulations considering two wireless access technologies like WiFi and cellular networks.

A framework, with schemes which integrate user mobility prediction models with bandwidth availability prediction models to support mobile multimedia services requirements was proposed by Nadembega et al., [17]. It specifically proposed schemes that predict paths to destinations, times when users enter/exit cells along predicted paths, and available bandwidth in cells on predicted paths. A request for mobile streaming service was accepted with these predictions, only when there was enough (predicted) bandwidth, along the destination path, to support the service. Simulation showed that the new approach outperformed current bandwidth management schemes in supporting mobile multimedia services better.

Topology design problem in a predictable Delay Tolerant Networks (DTN) where time-evolving topology was known a priori or is predicted was studied by Li et al., [18]. The purpose of reliable topology design problem was to build a sparse structure from original space-time graph so that (1) for any pair of devices, there was a space-time path connecting them with reliability higher than required thresholds; (2) total structure cost is minimized. Finally, simulations on random DTNs, a synthetic space DTN, and a real-world DTN tracing data proved the efficiency of the new method.

Location-based adaptive video quality planning approaches, in-network caching, content prefetching, and long-term radio resource management were discussed and proposed by Abou-zeid&Hassanein [19]. Insights on energy savings were provided. Then a cross-layer framework that jointly optimized resource allocation and multi-user video quality using location predictions was presented. Finally, some future research directions for location-aware media delivery in conclusion were highlighted.

A scalable hybrid bandwidth-efficient Adaptive Service Discovery Protocol (ASDP) for Vehicular Networks presented by Abrougui et al., [20] finds service provider and routing information simultaneously resulting in overall bandwidth savings. The new service discovery protocol adapted the service provider's advertisement zone size based on an adaptation mechanism. Results showed the protocol's scalability. They indicate that the techniques can achieve significant success (more than 90 percent), while guaranteeing low response time (in milliseconds) and low bandwidth use compared to current service discovery techniques.

A novel physical layer authentication scheme exploiting time-varying a Carrier Frequency Offset (CFO) associated with pairs of wireless communications devices was proposed by Hou et al., [21]. Combining these biases and mobility-induced Doppler shift, characterized as a time-varying CFO, is used as a radiometric signature for wireless device authentication. In the new authentication scheme, variable CFO values at different communication times were estimated. Kalman filtering predicted current value by tracking past CFO variations, modeled as an autoregressive random process. Simulation confirmed the effectiveness of the new scheme in multipath fading channels.

A new, fast location-based handoff scheme designed for vehicular environments was presented by Almulla et al., [22]. The protocol was able to accurately predict several APs that a vehicle may visit in the future with the position/movement direction of the vehicle and location information of surrounding APs. It assigned the APs to different priority levels. APs on higher priority levels are scanned first. Simulation showed that the new scheme attained lower prediction error rate and lower link layer handoff latency with limited influence on jitter/throughput.

DTN-Meteo, a new unified analytical model that maps an important class of DTN optimization problems over heterogeneous mobility/contact models to a Markov chain traversal over relevant solution space was proposed by Picu&Spyropoulos [23]. Local optimization algorithms accept/reject candidate transitions (deterministically/randomly), thereby "modulating" transition probabilities. Performance of state-of-the-art algorithms in various real/synthetic mobility scenarios showed that surprising precision is possible against simulations, despite problems complexity and settings diversity.

Ganguly et al., [24] proposed a location based mobility prediction scheme that helped in selecting the appropriate forwarder by predicting the mobility pattern of nodes. DTN specific user mobility involved both periodic and slightly chaotic patterns; chaotic behavior being attributed to the sudden causal events triggering instantaneous node mobility. In this approach, the authors approximate the periodicity of the DTN node mobility and use that knowledge to facilitate forwarding. The authors compared the results, thus obtained and with real location of the nodes in future mentioned time instances and simulation results showed that scheme provided satisfactory results in predicting mobility of nodes to a great extent.

## METHODOLOGY

In this section, datasets, MLP methods and Bee swarm algorithm are described.

### Dataset

The mobility traces used by researchers is provided by Dartmouth College as a community service. In this work one month syslog data is used however mobility trace collected over three years in Dartmouth College is available.

There were 5500 students and 1200 faculty housing the college over the three years during the data collection period. At the outset 476 APs were available and over a point in time it increased to 566. The users were able to use the network across the campus seamlessly as all the APs shared the same SSID. 115 subnets enclosed 188 buildings and hence the wireless gadgets required to acquire new IP addresses at times.

A syslog server log had the AP name, address of the MAC card and message type. It included the timestamp to every message. The messages used by the devices are authenticated, associated, reassociated, roamed and disassociated.

While a mobile gadget selects a network it is primarily authenticated and it associates itself with an AP to enable all traffic linking the device and the network. The mobile gadget reassociated when another AP with better signal strength is available. The device is in roaming when it reassociated with a new access point. When the device moves out of the network coverage or needs the network no more, disassociated message is sent.

In this work trace from users in the Dartmouth College of a single day is used. It is proposed to take into account four attributes with three attributes providing the prior location of the user and the fourth attribute considering the time.

Sample syslog data is given in **Table- 1**.

**Table: 1. Sample syslog data**

Unix Time Stamp	Specific Access Point Associated with the User
1035100785	AdmBldg19AP3
1035100842	AdmBldg20AP3
1035100851	AdmBldg24AP1
1035100908	AdmBldg20AP3
1035100963	AdmBldg24AP1
1035101020	AdmBldg20AP3
1035101022	AdmBldg24AP1
1035101080	AdmBldg20AP1
1035101082	AdmBldg24AP1
1035101139	AdmBldg20AP3

### Establishment of mobility rules

The UMP is

$$l = \langle l_1, l_2, l_3, \dots, l_n \rangle$$

Mobility rules established from this pattern are

$$l_1, l_2, l_3, l_4 \rightarrow l_{c5}$$

$$l_2, l_3, l_4, l_5 \rightarrow l_{c6}$$

$$l_3, l_4, l_5, l_6 \rightarrow l_{c7}$$

....

$$l_{k-4}, l_{k-3}, l_{k-2}, l_{k-1} \rightarrow l_{ck}$$

Where  $l_{ck}$  represents clustered value of AP in the head and represented by all APs close to each other in the network.

**Figure- 1** shows frequent item set for minimum support of 35% and confidence of 10%.

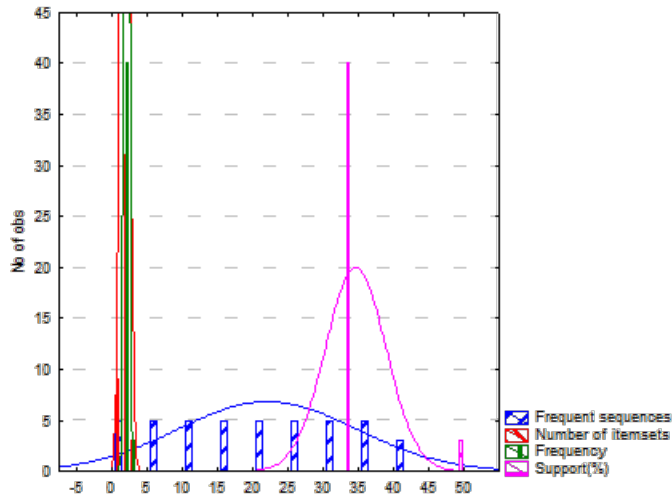


Fig:1 .Frequent Item Set for One Month Data

### Multi-Layer Perceptron (MLP)

A MLP constructs on the architecture of the single layer perceptron. The single layer perceptron is not very practical for the reason that it has limited mapping capability. It is merely pertinent to linearly separable inputs. The MLP yet, can be used as a building block for larger, a lot more practical structures. The restrictions of a simple perceptron may be overcome by using multiple layer architectures, difficult training algorithms and activation functions which are non-binary. A classic MLP arrangement consists of source nodes in the input layer, one or more hidden layers which compute the inputs by applying activation function, and an output layer of nodes illustrated in Figure- 2. The network has an input layer, single hidden layer with non-linear activation and an output layer with linear function. The input signal flows through the hidden layer from the input layer to output layer. The computations performed by this feed forward network can be written mathematically as

$$t = f ( s ) = B \varphi ( As + a ) + b$$

s = inputs

t = outputs

A = first layer weight matrix

a = first layer bias vector

B = second layer weight matrix

b = second layer bias vector

$\varphi$  = non-linearity function.

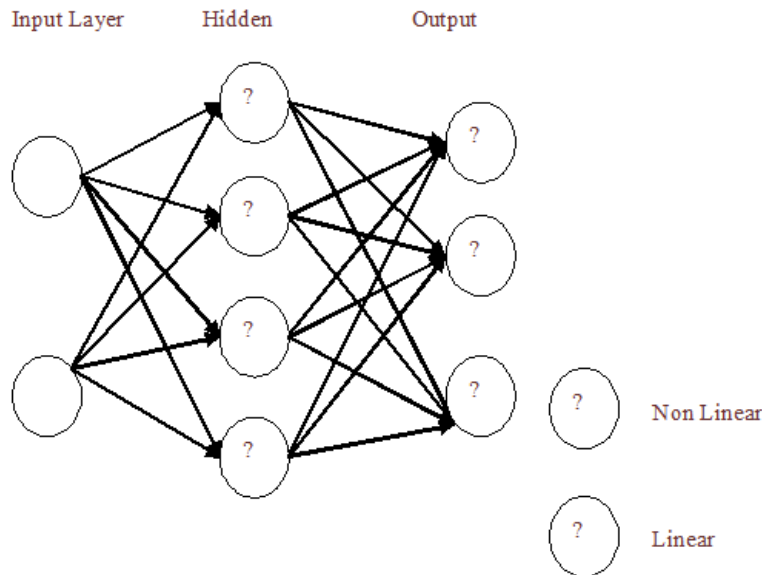


Fig: 2. Multi-Layer Perceptron Architecture

MLP can estimate any continuous function to any level of accuracy of a compact set. On the other hand, the number of hidden layers and weight matrix that ensure optimum network convergence is never known. The solutions to each NN for the input and output data applied is unique. An MLP can categorize non-linear problems successfully.

The network is governed with equations that steer the network to provide precise result with minimum training error [25]. Commonly the training algorithms concentrate on synaptic weights, number of hidden layers, activation function etc. for assuring optimal result. The familiar supervised learning technique that trains the NN is BP. On the other hand, BP gets trapped in local minima and has slow convergence at some time. However to overcome this trapping of local minima some evolutionary algorithm can be used.

Multi-layered feed forward NNs are appropriate for complex pattern classification since it has various characteristics that provide solution for the same. However, the lack of a suitable training algorithm restricts its application for some of the real world environment. Finding a training algorithm which provides a near global optimal set of parameters in a comparatively short interval of time is a complicated task. Evolutionary Algorithms (EA) like Particle swarm intelligence and Genetic Algorithms (GA) explore a large and composite space in an intellectual way to locate parameters nearer to the global optimum [26]. Therefore, they are appropriate to train feed forward networks. MLP-NN have been generally used for forecasting. The common algorithm used for training this network is BP [27, 28].

Another popular NN used for prediction is the partially recurrent networks. They have a special neurons group in the input layer, called context neurons/neurons of state. Thus, in an input layer of partially recurrent networks two neuron types are seen, those which act like the input, receiving outside signals and context neurons receiving output values of a layer delayed by a step. They are useful for time series prediction problem [29-33]. Jordan in 1986 proposed Jordan NN, characterized as context neurons receive a copy from output neurons and themselves. The Jordan network has as many context neurons as output neurons. Recurrent output layer connections to the context neurons have an associated parameter,  $m$ , that, usually take a constant value positive smaller than 1. For time series prediction, a network will have an output neuron representing predicted time series value at futures instances. The network will thus have only a context neuron and its activations at instant  $t$  is given by the following expression:

$$c(t) = mc(t-1) + x(t-1)$$

Where  $x(t-1)$  is output network at instant  $t-1$ .



Remaining network activities are calculated as in multilayer perceptron, where it is enough to consider as input vector a concatenation of the external input activations and context neurons activations:

$$u(t) = (x(t), \dots, x(t-d), c(t))$$

Taking into account the expression of the context neuron activation, it is possible to write:

$$c(t) = \sum_{j=1}^{t-1} \mu^{j-1} x(t-j)$$

Therefore, parameter  $m$  equips Jordan network with certain inertia for this network's context neurons. It was previously seen that context neuron accumulates network output at all previous instants and parameter value  $m$  determines context neuron's sensitivity to retain information.

Popular Jordan NN training algorithm include BPTT and its operation can be summarized by:

- Context neurons activations initialized as zero at the initial instant.
- External input  $(x(t), \dots, x(t-d))$  at instant  $t$  and context neurons activations at instant  $t$  are concatenated to determine input vector  $u(t)$  to network and propagated towards network output obtaining prediction at instant  $t+1$ .
- BP algorithm modifies network weights
- Time variable time increases in one unit with procedure goes to step 2
- Weight adjustment between BP processing elements is carried out based on the difference between NN's target and output values. Error difference in BP is measured by mean square error, as exposed below:

$$E = \sum_{k=1}^m \sum_{j=1}^q (t_{kj} - z_{kj})^2$$

Where  $t_{kj}$  is the  $j$ th target value of the  $k$ th compound, and  $z_{kj}$  is the output. Weights are adjusted to a gradient direction with better fitness [34] as shown in the equation:

$$w_{ji}^{new} = w_{ji}^{old} + \alpha \sum_k \delta_{kj} y_{ki} + \beta \Delta w_{ji}^{old}$$

Where  $j, i$  are adjacent layer indices,  $w_{ji}$  is weight from the previous layer  $i$ th neuron to the  $j$ th neuron in the current layer and  $\Delta w_{ji}$  is the preceding weight change. The variable  $y_{ki}$  represents the  $i$ th output for the  $k$ th pattern. Parameters  $\alpha$  and  $\beta$  are positive constants called learning rate and momentum rate which controls weight adjustments amount during weight update.

BP algorithm, a gradient based method, is the most commonly used in NN training. BP algorithm's inherent problems are encountered when this algorithm is used. First, BP algorithm is easily trapped in local minima for non-linearly separable pattern classification problems/complex function approximation problem [35], leading to back-propagation failure to locate a global optimal solution. Second, BP Optimization algorithm's convergent speed is too slow even if learning goal, a given termination error, is achieved. What is to be emphasized is that BP algorithm's convergent behavior depends on initial values choices in network connection weights as also algorithm parameters like learning rate and momentum. To improve original BP algorithm performance, researcher's concentrated on two factors:

- Better energy function selection [36 & 37];
- Dynamic learning rate and momentum selection [38 & 39].

But these have not removed BP algorithms disadvantages of being trapped in local optima. Specifically, convergent speed will be slower as NN's structure is more complex. The BP has a very high likelihood to be trapped in local minima during the training process. Hence some variation in the PSO is proposed in this chapter.

### Proposed Bee Swarm Optimization Algorithm

Optimization techniques are significant in practice mostly in soft computing. EAs have been used extensively to solve complex optimization problems. They are powerful class of stochastic optimization algorithms which provide solution to problems which cannot be solved analytically.

GA and PSO are some of the EA that have been used in optimization problems. The Bees Algorithm (BA) is among the newer optimization techniques [40] developed upon bees foraging behavior. Formerly proposed the Bee Colony algorithm motivated by the behavior of bees with enhanced performance in optimization problems in contrast to GA, Differential Evolution (DE), and PSO [41].

It is a population-based algorithm that mimics the food foraging behavior of bee swarms. The basic version of the algorithm performs a neighborhood and random search combined. A bee hive simultaneously explores plenty of food sources by stretching itself in many directions and over long distances [42]. Hypothetically, more bees will visit the flower patch with plenty of nectar/pollen that is gathered with less work, while the flower patch with less amount of nectar/pollen is visited by smaller number of bees [43]. Exploring commence in a colony by scout bees inspecting for potential flower patches. They progress unsystematically patch to patch. Some of the bees in the population called as scout bees continue exploration during searching.

The threshold is defined as a mixture of ingredient such as sugar content. The scout bees locating a patch that are rated higher than a threshold dump nectar/pollen and advance to the “dance floor” to show cast its “waggle dance” after returning to hive. The colony communicates through this strange dance. This dance communicates three information’s concerned with the flower patch: Location of the flower patch, distance of the patch from the hive and its status of the quality (i.e. fitness). Based on this information, the bee hive sends bees to the flower patches accurately. The colony assesses the different patches for its relative merit in respect to food quality and energy needed to yield it. The colony understands every bee’s knowledge of outside environment from the waggle dance [44]. The scout bees (i.e. dancer) along with the follower bees goes to the flower patch after the completion of the waggle dance. Furthermore follower bees are sent to the potential flower patches to collect food rapidly and efficiently.

To decide the next waggle dance, the bees observe for the quality of food when yielding from the patch. If the quality of the flower patch is good and is still a potential food source, then it is publicized in the waggle dance and hence further more bees are employed to that patch.

Figure- 3 shows the flowchart of the BA in its simplest form which is dependent to some parameters described in Table- 2.

Table: 2. Parameters for Bee algorithm

Number of Scout bees	N
Sites selected	m
Best sites	e
Bees required for best e sites	n e p
Bees required for other selected (m-e) sites	n s p
Initial patch size	n g h
Iteration	i

The process begins as the scout bees starts harvesting randomly (random search). The fitness of the visited site is evaluated in step 2

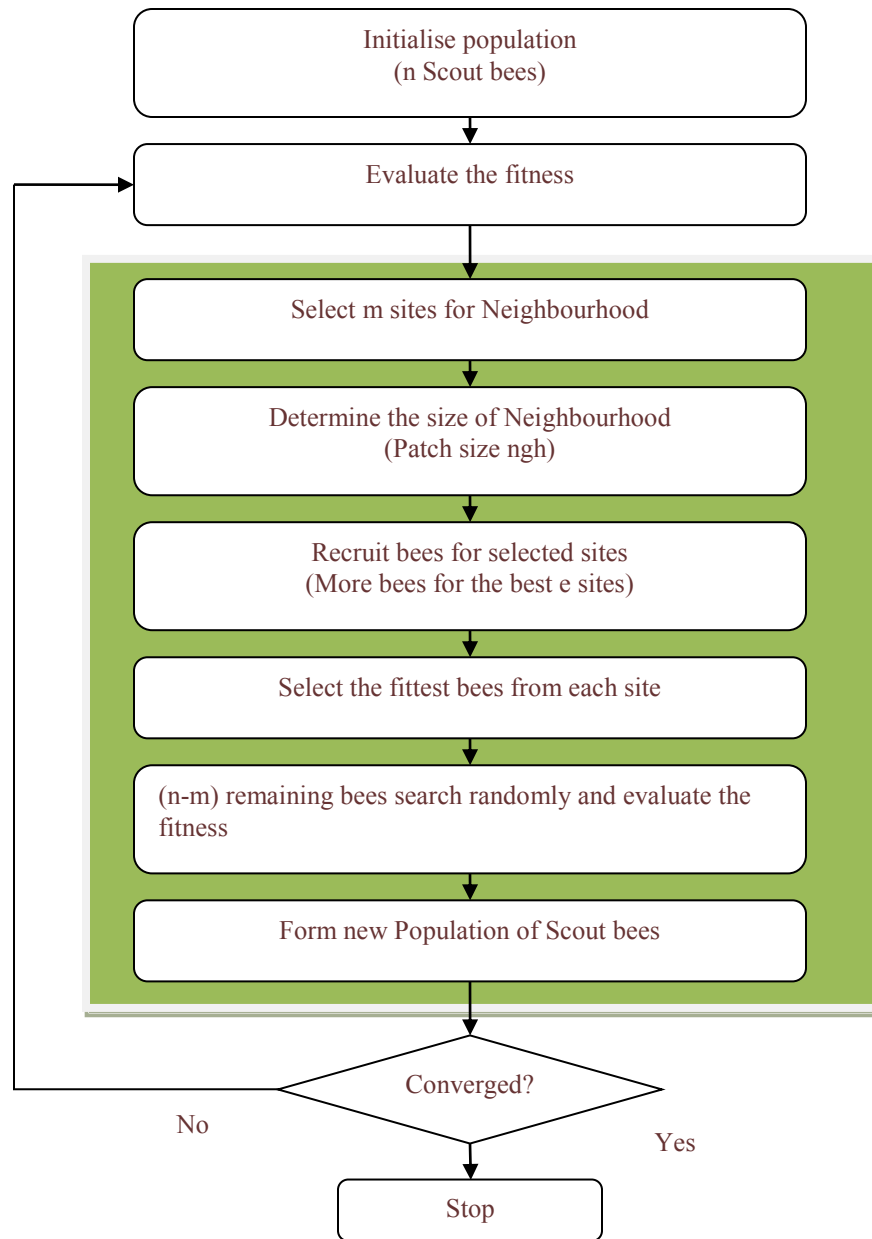


Fig. 3. Pseudo Algorithm for Bee swarm

The Optimization algorithm is used for searching the optimum values of weights, it speeds up the training.

## RESULTS

The mobility prediction accuracy is evaluated using Multilayer perceptron network. The proposed Multilayer perceptron network is optimized with Artificial Bee Colony (ABC) algorithm. The parameters of the proposed Multilayer perceptron are given in [Table- 3](#).

Table: 3.The Parameters of the Proposed Multilayer Perceptron Network

Number of inputs	8
------------------	---

Number of hidden layer	2
Number of outputs	5
Optimization	Weights ,Learning rate and Momentum using bee swarm algorithm
Activation function	Tanh
Algorithm	Bee swarm

In the proposed Bee swarm optimization, an initial population is randomly chosen. The population is run through the BA (as shown in Figure- 4.4). The value used for evaluating the fitness is the best value. The algorithm would have calculated the most favorable solution on convergence. The proposed Bee swarm optimization is used to optimize the MLP network. The weights, learning rate and momentum of the network are optimized using the Bee swarm algorithm. The following Figure- 4 shows the optimization for a number of iterations.

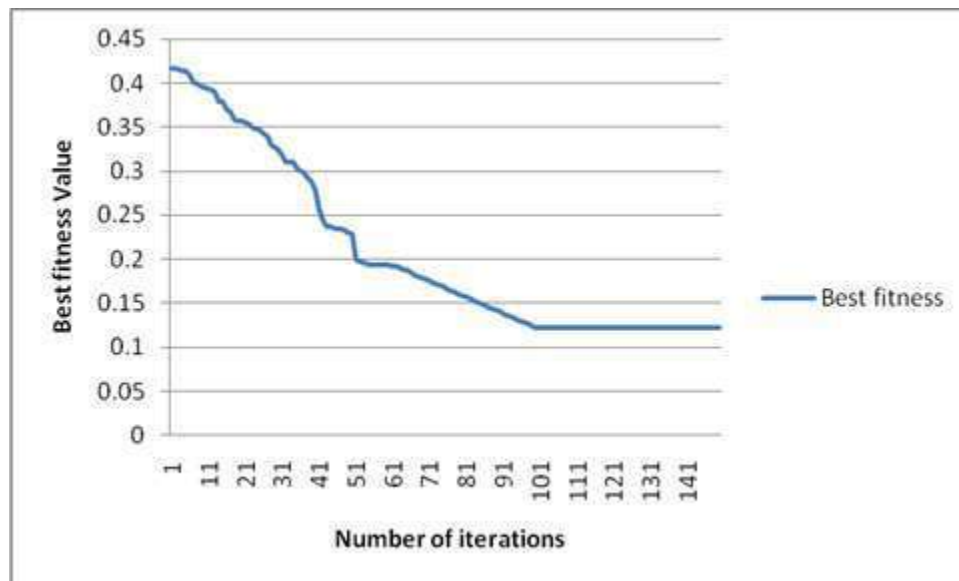


Fig: 4. Best fitness value

The proposed MLP with proposed bee swarm algorithm is compared with MLP without optimization and with Radial basis network. Table- 4 shows the classification accuracy and the RMSE achieved by using different techniques. Table- 5 tabulates the precision and recall. Figure- 5 and 6 shows the classification accuracy and precision and recall respectively.

Table :4. Classification Accuracy Achieved

Technique Used	Classification accuracy	RMSE
Radial basis function network	0.662295	0.3434
Multi-layer perceptron (two hidden layer)	0.744098	0.2776
Multi-layer perceptron (two hidden layer) with optimization	0.913462	0.1372

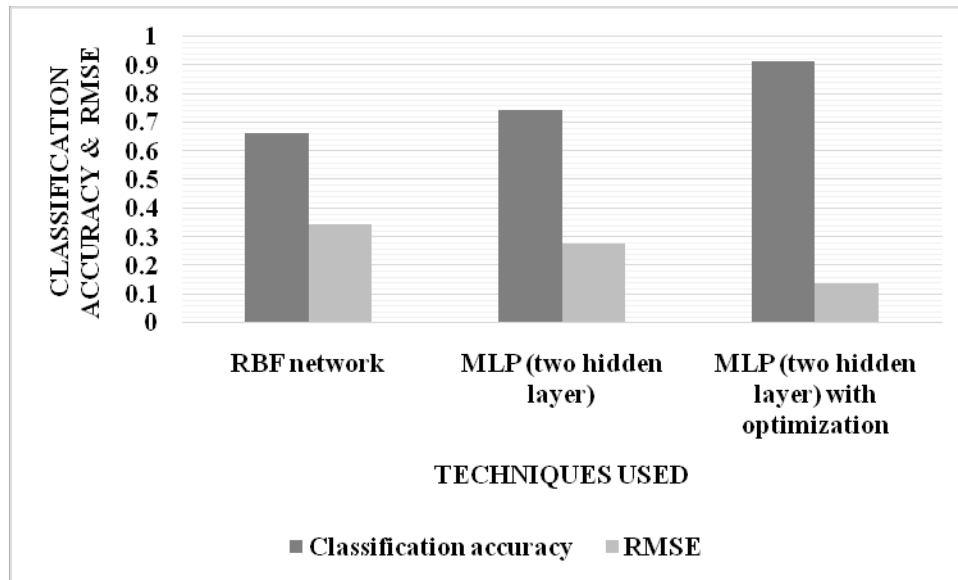


Fig: 5 .Classification Accuracy and RMSE

From the [figure-5](#), it can be observed that the MLP (two hidden layer) with optimization method increased Classification accuracy by 31.87% & 20.43% compared for RBF network and MLP (two hidden layer). The MLP (two hidden layer) with optimization method RMSE decreased by 85.8% & 67.69% compared for RBF network and MLP (two hidden layer).

Table: 5. Precision and Recall

Technique used	Precision	recall
Radial basis function network	0.672	0.662
Multi-layer perceptron (two hidden layer)	0.746	0.741
Multi-layer perceptron (two hidden layer) with optimization	0.929	0.913

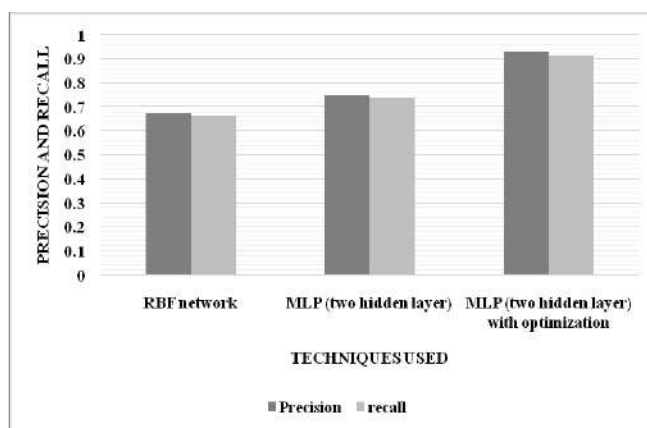


Fig:6. Precision and Recall

From the [Figure- 6](#), it can be observed that the MLP (two hidden layer) with optimization method increased precision by 32.1% & 21.85% compared for RBF network and MLP (two hidden layer). The MLP (two hidden layer) with optimization method recall decreased by 31.87% & 20.79% compared for RBF network and MLP (two hidden layer).

The MLP and RBF was simulated using 10 fold cross validation. The **Figure- 7 to 15** shows the actual and predicted value for sample cross validation for RBF and MLP.

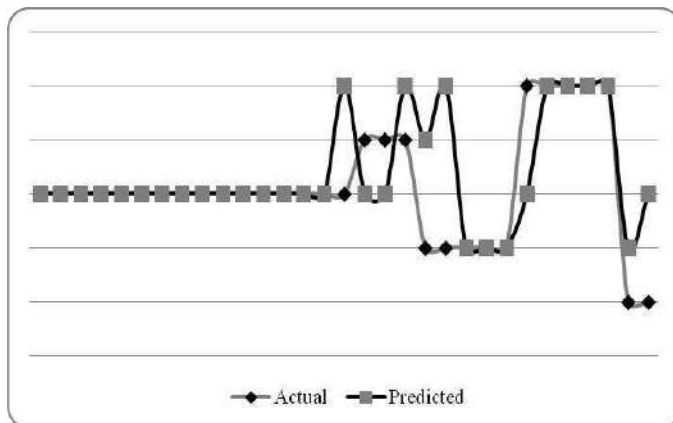


Fig:7. Iteration 4, Multilayer Perceptron

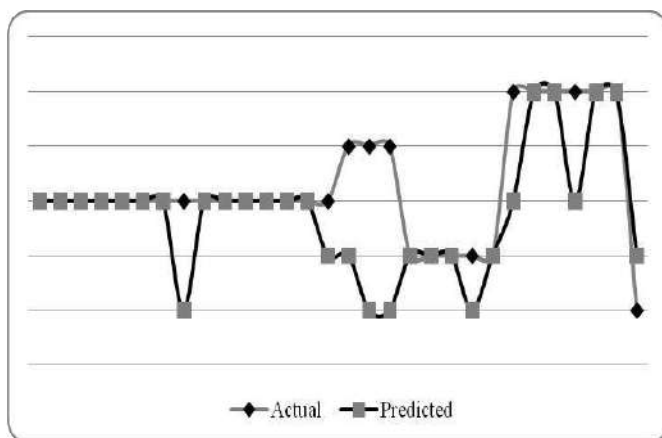


Fig: 8. Iteration 9, Multilayer Perceptron

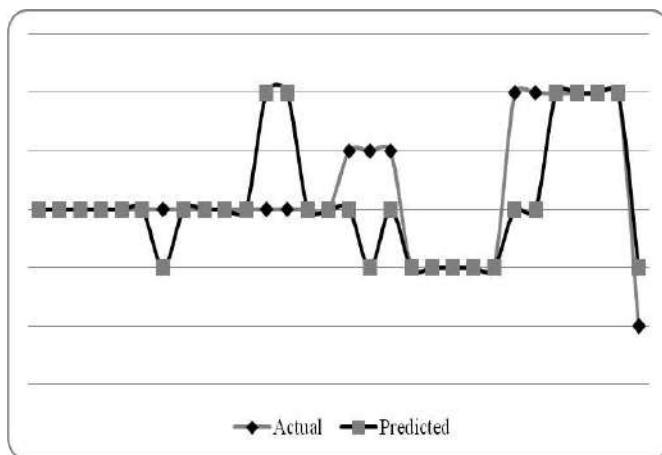


Fig: 9. Iteration 10, Multilayer Perceptron

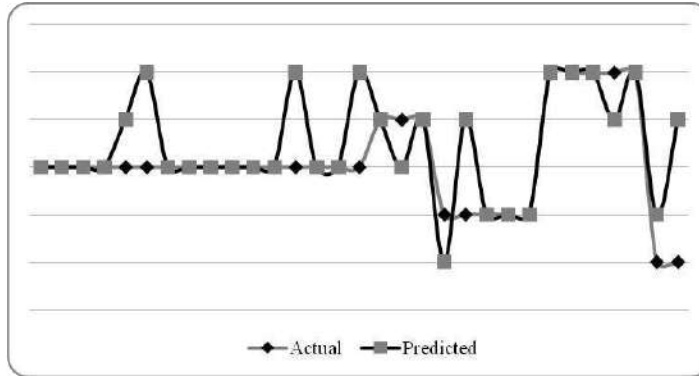


Fig: 10. Iteration 4, Radial Basis function

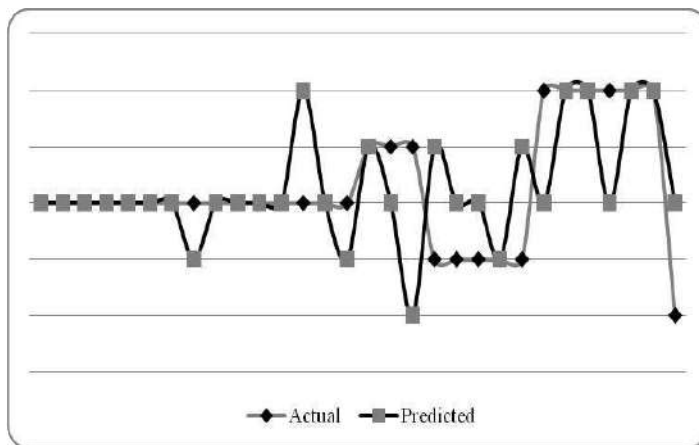


Fig:11. Iteration 9, Radial basis function

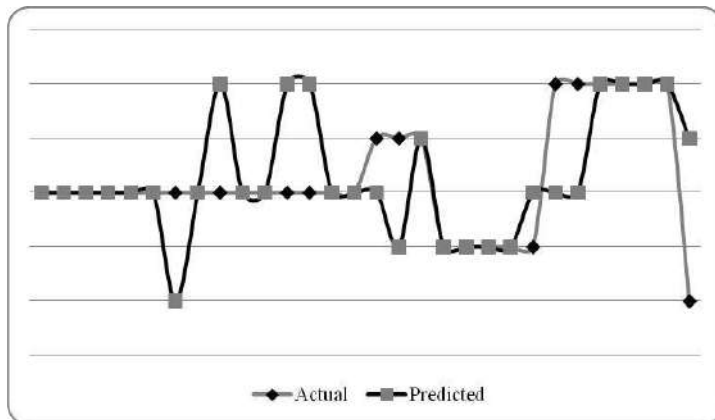


Fig: 12 .Iteration 10, Radial Basis function

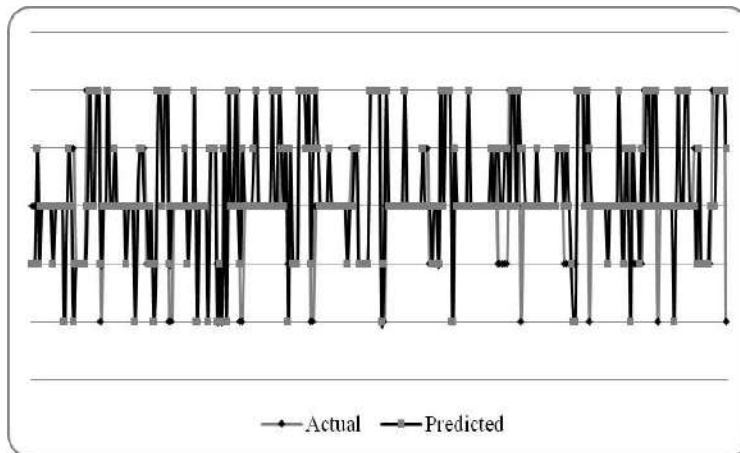


Fig: 13. Radial basis network output: Actual Vs Predicted

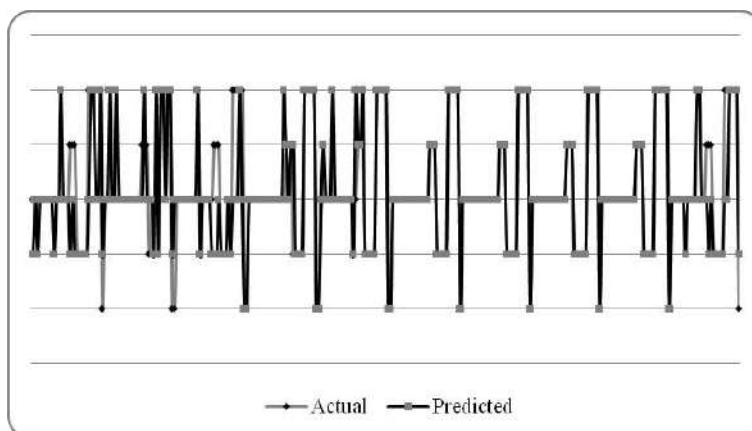


Fig: 14. Multiplayer Perceptron network output: Actual Vs Predicted

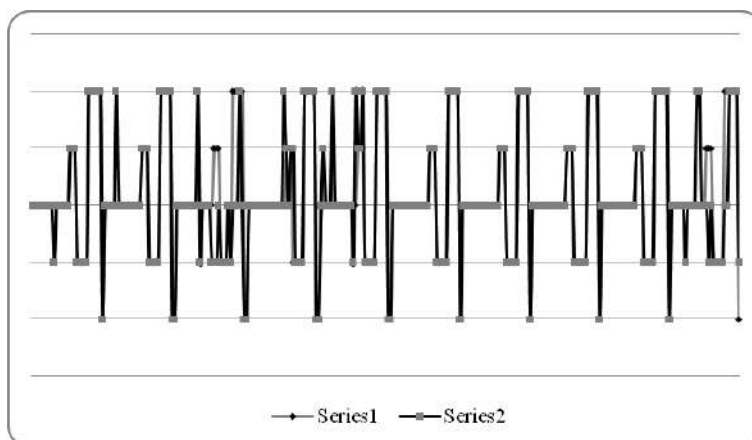


Fig: 15. Multiplayer Perceptron network with bee swarm optimization output: Actual Vs Predicted

It is evident from the Figures and Tables that the Proposed MLP network optimized with bee swarm optimization algorithm performs well compared to MLP without optimization and RBF.



## CONCLUSION

Mobile users expect services to be provided with less latency and better quality. Seamless handoff with least delay in wireless networks is an important criterion to improve QoS. Mobility prediction helps in seamless handoff by allocating resources beforehand. In recent years lot of researches have been done on mobility prediction schemes and models. This work provides a MLP network optimized with bee swarm algorithm in a wireless campus environment for the prediction of mobile user movement. The proposed algorithm predicts the next movement of the user with the help of the mobility rules established from the mobility traces. The data available in the Dartmouth college public domain is used as mobile wireless traces. This one month trace data is used for evaluation in this work. The results of the experiment conducted shows that the MLP network optimized with bee swarm algorithm act upon suitably.

### CONFLICT OF INTEREST

No conflict of interest

### ACKNOWLEDGEMENT

None.

### FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Taneja S, Kush A. [2010] A survey of routing protocols in mobile ad hoc networks. *International Journal of Innovation, Management and Technology*, 1(3):279.
- [2] Karygiannis T, Owens L. [2009] Wireless network security: 802.11, bluetooth and handheld devices. 2002. *National Institute of Standards and Technology (NIST) special publication*.
- [3] Mathivaruni RV, &Vaidehi V.[2008]An activity based mobility prediction strategy using markov modeling for wireless networks. In *Proceedings of the World Congress on Engineering and Computer Science (pp. 22-24).of Computer Science and Telecommunications*, 3(7).
- [4] Prasad PS. [2010] Mobility Modeling, Prediction and Resource Allocation in Wireless Networks (Doctoral dissertation, Auburn University).
- [5] Capka J, Boutab R.[2004]Mobility prediction in wireless networks using neural networks. In *Management of Multimedia Networks and Services:320-333*. Springer Berlin Heidelberg.
- [6] S Lu, R Srikant and V Bhagavan.[1996] Adaptive Resource Reservation for Indoor Wireless LANs”, *Proceedings of IEEE GLOBECOM’96*, London,
- [7] F Akyildiz and W Wang.[2004] The Predictive User Mobility Profile Framework for Wireless Multimedia Networks, *IEEE/ACM Transactions on Networking*, 12(6)
- [8] T Liu, P Bahl and I Chlamtac.[1998]Mobility Modeling, Location Tracking, and Trajectory Prediction in Wireless ATM Networks”, *IEEE Journal on Selected Areas in Communications*, 16(6)
- [9] J Chan, S Zhou and A Seneviratne.[1998] A QoS Adaptive Mobility Prediction Scheme for Wireless Networks”, *IEEE GLOBECOM*,
- [10] S Kwon, H Park and K Lee.[ 2004] A Novel Mobility Prediction Algorithm Based on User Movement History in Wireless Networks, In *Systems Modeling and Simulation: Theory and Applications: Third Asian Simulation Conference, AsianSim 2004*, October 2004.
- [11] S Michaelis and C Wietfeld.[2006] Comparison of User Mobility Pattern Prediction Algorithms to Increase Handover Trigger Accuracy, *IEEE 63rd Vehicular Technology Conference*, Spring 2006.
- [12] Wang G, Guo L. [2013] A novel hybrid bat algorithm with harmony search for global numerical optimization. *Journal of Applied Mathematics*, 2013.
- [13] Manfredini F, Pucci P, Secchi P, Tagliolato P, Vantini S, Vitelli V. [2015] Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region. In *Advances in Complex Data Modeling and Computational Methods in Statistics (pp. 133-147)*. Springer International Publishing.
- [14] Surobhi NA, Jamalipour A. [2014] M2M-Based Service Coverage for Mobile Users in Post-Emergency Environments, *Vehicular Technology, IEEE Transactions on* , 63(7):3294-3303.
- [15] Taleb T, Ksentini A. [2015] VECOS: A Vehicular Connection Steering Protocol, *Vehicular Technology, IEEE Transactions on* , 64, (3):1171-1187
- [16] Nadembega A, Hafid A, Taleb T. (2014). An Integrated Predictive Mobile-Oriented Bandwidth-Reservation Framework to Support Mobile Multimedia Streaming, *Wireless Communications, IEEE Transactions on* , 13(12):.6863-6875
- [17] Li F, Chen S, Huang N, Yin Z, Zhang C, Wang Y. [2015] Reliable Topology Design in Time-Evolving Delay-Tolerant Networks with Unreliable Links, *Mobile Computing, IEEE Transactions on* , 14(6):1301-1314
- [18] Abou-zeid, H., Hassanein, H. [2014] Toward green media delivery: location-aware opportunities and approaches, *Wireless Communications, IEEE -21(4):.38-46*
- [19] Abrougui, K, Boukerche, A., Pazzi RW, Almulla M. [2014] A Scalable Bandwidth-Efficient Hybrid Adaptive Service Discovery Protocol for Vehicular Networks with Infrastructure Support, *Mobile Computing, IEEE Transactions on* , 13(7):1424- , *Vehicular Technology, IEEE Transactions on* ,63(8):3853-3866

- [20] Picu A, & Spyropoulos, T., (2015) DTN-Meteo: Forecasting the Performance of DTN Protocols Under Heterogeneous Mobility, Networking, *IEEE/ACM Transactions on*, 23(2):.587-602
- [21] Ganguly S, Basu S, Roy S, Mitra S. [2015] A location based mobility prediction scheme for post disaster communication network using DTN. In Applications and Innovations in Mobile Computing (AIMoC), 2015 (pp. 25-28). *IEEE*.
- [22] Habib Shah, Rozaida Ghazali, Nazri Mohd Nawi and Nawsher Khan. [2012] Boolean Function Classification using Hybrid Ant Bee Colony Algorithm, *Journal of Computer Science & Computational Mathematics*, 2( 11) .
- [23] Habib, S., Ghazali, R., & Nazri, M, Nawi. [2013] Hybrid Global Artificial Bee Colony Algorithm for Classification and Prediction Tasks, *Journal of Applied Sciences Research*, 5(9).
- [24] Pham, D.T., Koc, E., Ghanbarzadeh, A, Otri, S. [2006] Optimisation of the weights of multi-layered perceptrons using bees algorithm. In: *Proceedings of 5th International Symposium on Intelligent Manufacturing Systems*. 38–46
- [25] H. Yin et al. (Eds.) [2012] Comparing Particle Swarm Optimization Approaches for Training Multi-Layer Perceptron Neural Networks for Forecasting. In: *Intelligent Data Engineering and Automated Learning - IDEAL 2012 Lecture Notes in Computer Science Volume 7435*, 2012, pp 344-351 ©Springer-Verlag Berlin Heidelberg
- [26] P Stagge and B. Senho. [1997]. An extended elman net for modelling time series. In *International Conference on Artificial Neural Networks*,
- [27] Tomasz J Cholewo and Jacek . Zurada. [1997] Neural network tools for stellar light prediction. In *IEEE Aerospace Conference*..
- [28] VM Mladenov, AC Tsakoumis, SS Vladov. [2002] Electric load forecasting with multilayer perceptron and elman neural network. In *IEEE NEUREL*..
- [29] S LAWRENCE CL GILES and AH C TSOI. [2001] Noisy time series prediction using recurrent neural networks and grammatical inference. *Machine Learning*, (43):161-183..
- [30] C SITTE and J SITTE. [2002] Neural networks approach to the randomwalk dilemma of nancial time series. *Applied Intelligence*, (16):163171
- [31] MI Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proc. of the Eighth Annual Conference of the Cognitive Science Society*, pages 531-546. NJ: Erlbaum, 1986.
- [32] RC Eberhart, and Y Shi. *Computational Intelligence: Concepts to Implementations*, Morgan Kaufmann Publishers. San Francisco, (in press).
- [33] Marco Gori, Alberto Tesi. [1992] On the problem of local minima in back-propagation, *IEEE Trans Pattern Anal Mach Intell* 14 (1) 76–86.
- [34] van Ooyen, B Nienhuis. [1992] Improving the convergence of the back-propagation algorithm, *Neural Network* 5 (4) : 465–471.
- [35] M Ahmad, FMA. Salam, supervised learning using the Cauchy energy function, in: *Proc. of International Conference ON Fuzzy logic and Neural Networks*.
- [36] RA Jacobs. [1988] Increased rates of convergence through learning rate adaptation, *Neural Networks* 1 295–307.
- [37] DT Pham, A Ghanbarzadeh, E Koc, S Otri, S Rahim, and M Zaidi. [2005] The bees algorithm, Technical Note, Cardiff University,
- [38] UK, Karaboga, D and B Basturk. [2008] \On the performance of artificial bee colony (ABC) algorithm, *Applied Soft Computing*, 8(1): 687- 697
- [39] Von Frisch K. *Bees: Their Vision, Chemical Senses and Language*. (Revised edn) Cornell University Press, NY., Ithaca, 1976.
- [40] Seeley TD. [1996] *The Wisdom of the Hive: The Social Physiology of Honey Bee Colonies*. Massachusetts: Harvard University Press, Cambridge
- [41] Bonabeau E, Dorigo M, and Theraulaz G. [1999] *Swarm Intelligence: from Natural to Artificial Systems*. Oxford University Press, New York,

# PROTECTING PACKETS AGAINST MALICIOUS NODES IN MOBILE AD HOC NETWORK

Devi \* and Jayakumar

Department of computer science and engineering, R.M.K Engineering College, TN, INDIA

## ABSTRACT

Malicious drops are one of the attacks to drop the packets. It affects the transfer of data from source to destination and sends a fake acknowledgement as received. The acknowledgement that is forwarded will reach the source and wait for the response. Till the timeout it waits and starts to transmit the packet again. Malicious nodes do not have any intension to drop the packets. We proposed a system to reduce malicious nodes prevailing in the transmission path. To avoid these malicious nodes we also propose on-demand routing protocol and digital signature acknowledgement. After the detection of the malicious nodes packet blocking is implemented and secure routing is done.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

attack detection, on-demand routing protocol, digital signature, secure routing, packet dropping

\*Corresponding author: Email: [devisivaranjani29@gmail.com](mailto:devisivaranjani29@gmail.com); Tel.: +919791062604

## INTRODUCTION

Mobile ad hoc network is the collection of two or more devices or nodes with wireless communication and networking capacity that communicate with each other without the aid of any centralized administrator also the wireless nodes that can dynamically from a network to exchange information without using any existing fixed network infrastructure[1-2]. A wireless mobile node can functions both as a router for routing packets from other nodes and as a network host for transmitting and receiving packets[1]. The network consists of peer-to-peer, self-forming and self-healing. The working of ad hoc network is to find a path or route between source node and destination node [3]. Sometimes, the destination node may not have any path to receive packets[4]. The ad hoc network has to be implemented to find a new path for the packet transmission [5].fig1.1

The characteristic of an ad hoc network has no background network for the central control of the Network operations .The network is distributed among the nodes[6]. These nodes in a network should co-operate each other among themselves. When a node tries to communicate to other nodes which are out of its communication radio range, the packets should be forwarded with one or more intermediate nodes [7], [10] fig1.1. The nodes in the ad hoc network dynamically establish routing among themselves, establishing their own network. Mobile ad hoc network is more vulnerable to malicious nodes having a chance of dropping the packets [8, 9]. There are several vulnerabilities in dropping the packets as all the drops are intended to denial the service (DOS) [10, 11].

## PROBLEM STATEMENT

Selectively detecting packet-dropping attacks is extremely challenging in a highly dynamic wireless environment. The difficulty lies in the requirement that we need to detect the place where the packet is dropped but also identify whether the hop is intentional or unintentional. The existing system consists of public auditing for storage. However, this is not suitable for application of homomorphic linear authentication (HLA) because there can be more than one malicious node along the route. Public auditing does not reduce the malicious packet drop instead they can provide a valid proof for the dropping of packets.

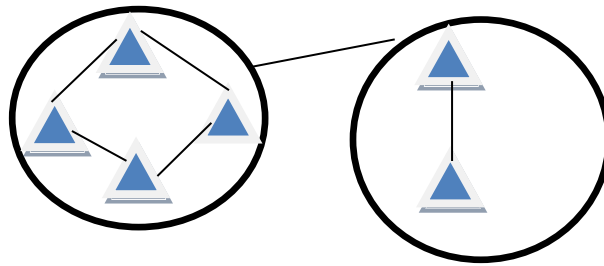


FIG: 1.1. DESCRIBES ABOUT THE AD HOC NETWORK COMMUNICATION

### PROPOSED SYSTEM

We have taken this problem in mobile ad hoc network and proposed a method to overcome the problem. To reduce the malicious node invasion we have proposed on-demand multicast routing protocol and digital signatures. The nodes that send the packets are need to be acknowledged to conform that the nodes reach the destination. Acknowledgements are like a token that the packets are sent safely. If there is no response then we use digital signatures and on-demand routing protocol for security purpose. This protocol helps in providing the security against the attackers.

To improve the security in sending the packets we propose a method of, on-demand routing protocol for transmitting the packets. This will reduce the malicious nodes to drop the packets. Simulation results have been demonstrates the effectiveness of proposed scheme with improved performance as compared to the existing protocol and verifications of the packets.

### SYSTEM ARCHITECTURE

The system architecture shows the overall design of the modules like packet generation, packet dropping, attack detection, digital signature verification, malicious node identification, packet blocking and secure routing. The nodes describes about the activities undergo in the proposed system.

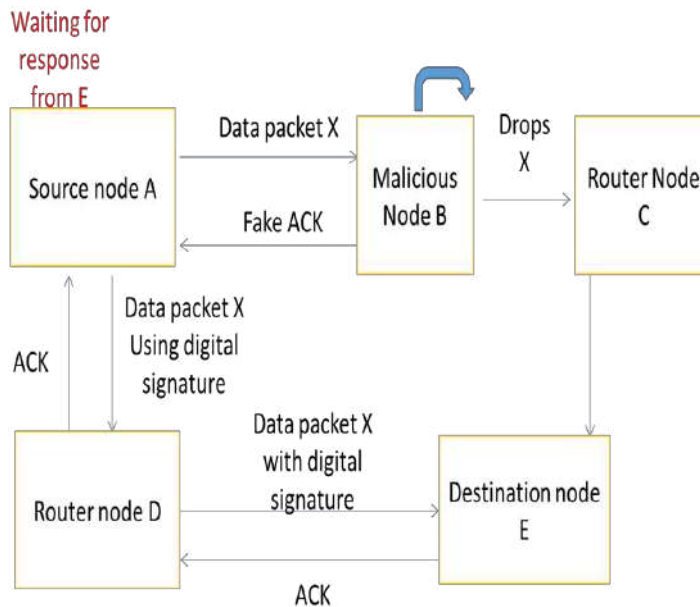


FIG: 4(1) ARCHITECTURE DIAGRAM

## MODULE DESCRIPTION

### PACKET GENERATION

Packets are the files that are to be send from source to destination. A source node has to send these packets through intermediate nodes to reach destination. The intermediate nodes will send the generated packets to the destination node or it will pass to the next nearby node. The source will have a timeout for these packets to be delivered. According to the timeout and delay the packet will be detected as delivered or dropped. If it does not receive the acknowledgement then it proceeds with the verification of where the packets gone.

### PACKET DROPPING

#### 1) Malicious packet drop

On sending a packet it may drop due to a malicious node invasion. It may act like a legitimate node and drop the packets intentionally or to denial the service. Link error is caused due to any internal failure or the failure may occur on any of the link that is connected. Due to the failure of configuration also link errors may occur.

#### 2) Attack detection

Attack detection is detecting at which stage the attack is done. It may happen at the state where the intermediate nodes pass the packets to the destination nodes. It may occur in two ways. 1. Denial of service attack. 2. Suspicious packet drop.

### DIGITAL SIGNATURE VERIFICATION

A digital code is attached to verify its contents and the sender's identity. Digital signatures can be used to certify or to approve documents. Certifying signatures verify the documents creator and show that the document has to been altered since it was signed. Therefore, only the original creator of a document can add a certifying signature. Approved signatures can be added to anyone with a digital id and are used to approve documents, track changes, and accept terms stated with a document. This is applied only for the acknowledgement to verify whether it is from the legitimate node or from a malicious node[11].

### MALICIOUS NODE IDENTIFICATION

Malicious node identification is that it detects the malicious node that are dropping the packets and causing the disruptions in the network. This can be identified by the activity of blocking the packets that are coming its way. A malicious node will only drop the packets and simply sends the acknowledgement as valid that indicates as a legitimate node. A timer is set for the acknowledgement and the response to the sender by this activity malicious node can be found. Once it is identified, it is blocked from further receiving or sending the packets or acknowledgement.

### PACKET BLOCKING

This is the last stage of the implementation. Packet blocking is done to which the node acts as a malicious node and that node is identified and reported. The reported node will be broadcasts as an infected node and that node will be blocked by sending or receiving any packets furthermore. The blocked node will not be used by any of the nodes for their packets transaction. The malicious node will remain idle for the rest of the time[12].

### SECURE ROUTING

For secure routing, we are using an on-demand routing protocol along with the digital signatures to find the malicious nodes that are dropping the packets. Digital signatures are used for secure routing as it ensures the sender and the receiver about the original node. These digital signatures verify the original sender by generating certificates. On-demand routing protocol helps to determine the malicious node in an efficient way such that the nodes can be identified.

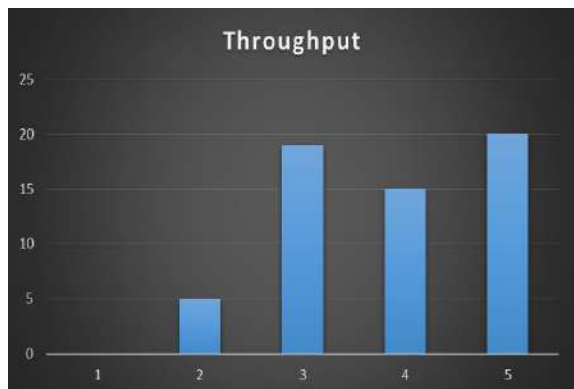


Fig:4(2) Efficient in reducing the malicious nodes

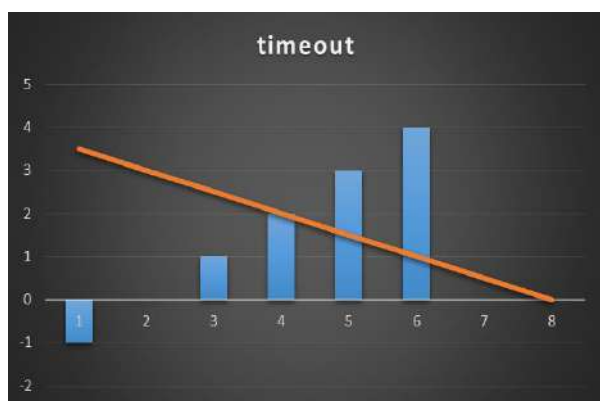


Fig:4(3) Timeout taken to reach the destination is also reduced

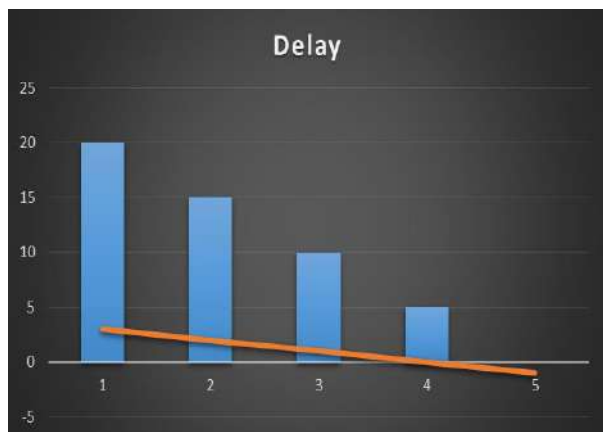


Fig:4(4) Delay time to reach the destination is also reduced

These graphs describe the throughput, timeout and delay to reach the destination from the source.

**THROUGHPUT:** Fig:4(2): The graph shows that how efficient the packets are delivered successful from source to destination.

**TIMEOUT:** Fig:4(3): The graph shows within the time to reach the destination. This may vary but does not change the effectiveness in reaching the destination.

DELAY: Fig4(4):The graph shows the time delay. The time it had taken to reach the destination form the source. These graphs may vary according to the files we are sending.

These graphs show the deviation after there is a malicious attack in the packet transmission. There will be variation in all the graphs showing that a malicious node has dropped the packets and stopped the transmission of packets temporarily.

## CONCLUSION

In this project, protecting packets against malicious nodes is investigated, with the secure routing protocols. The project describes about the problem of compromising nodes and security in mobile ad hoc network. Protecting packets against malicious dropped down because of the existence of malicious nodes. By this protecting of packets, existence of malicious nodes will be reduced and a secure way of routing the packets can be achieved. In future work, it can be implemented with cost effective and can be implemented in other domains.

## ACKNOWLEDGEMENT

None.

## CONFLICT OF INTEREST

No conflict of interest

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Zhiming Xu, Yu Wang, Jingguo Zhu. [2009] A reliable multicast routing protocol for high speed mobile ad hoc network in R-ODMRP, in *IEEE journal of software*.5(1)
- [2] Rajan, C and Shanthi, N.[2013], "Misbehaving attack mitigation technique for multicast security in mobile ad hoc networks (MANET)", *Journal of Theoretical and Applied Information Technology*, vol. 48, no. 3, pp.1349–1357.
- [3] Ganying Ru, Robert J.Kerczewski, Lingjialiu, Samee U.Khan [2013] secure wireless multicast for delay-sensitive data via network coding", on *IEEE transaction on wireless communication* ,1536-1276.
- [4] Zhiguo Wan, Kui Ren, And Ming Gu.[ 2012] USOR: An Unobservable Secure On-demand Routing Protocol For Mobile Ad Hoc Network, *IEEE Transaction On Wireless Communication*, 11(5):1536-1276
- [5] Trust T Mapoka, Simon J Shepherd, Raed A Abdalameed.[2015] A new Multiple Key Management Scheme For Secure Wireless Mobile Multicast, *IEEE Transaction On Mobile Computing*, 14(8):1536-1233.
- [6] Amol Bhosle, Yogadhar Pandey.[2013] Review of authentication and digital signature methods in Mobile ad hoc network on IEEE conference ISSN: 2278 – 1323 , 2(3).
- [7] Muthumanickam Gunasekaran, Kandhasamy Premdatha, "Teap:truse –Enhanced Anonymous On-demand Routing Protocol For Mobile Adhoc Networks", *IET On Information Security* Vol:7,issue:3,pp.203-211,on 2013.
- [8] Yavuz,A.A , Robert Bosch LLC.[2014] An efficient real-time broadcast authentication scheme for command and control messages, *IEEE transaction on information forensics and security*, 9(10).
- [9] Denh Sy, Rex Chen and Lichun Bao.[ 2012] ODA: On-Demand Anonymous Routing in Ad Hoc Networks, on *IEEE international Conference on*
- [10] Hui Xia, Jia Yu1, Zhi-yong Zhang, Xiang-guo Cheng, Zhen-kuan Pan.[ 2014] Trust-enhanced multicast routing protocol based on node's behavior Assessment for MANETs, on 13th International Conference on Trust, Security and Privacy in Computing and Communications,.
- [11] C. Siva Ram Murthy and B. S. Manoj.[2006]Ad hoc Wireless Networks: Architecture and Protocol, First Edition, Pearson Education India
- [12] Tao Shu And Marwan Kurnz, "Privacy-preserving And Truthful Detection Of Packet Dropping Attacks In Wireless Ad Hoc Networks", On *Mobile Computing*, Vol:14, No.4, April 2015.

# ENHANCEMENT OF BLOWFISH ENCRYPTION IN TERMS OF SECURITY USING MIXED STRATEGY TECHNIQUE

Joseph Raj<sup>1</sup> and Shamina Ross<sup>2</sup>

<sup>1</sup>Dept. of Computer Science, Kamaraj College, Thoothukudi-628003, INDIA

<sup>2</sup>Dept. of Computer Applications, Scott Christian College, Nagercoil-629001, INDIA

## ABSTRACT

**Abstract**—Encryption is the process of transforming plain text data into cipher text in order to conceal its meaning and preventing any unauthorized recipient from retrieving the original data. Cryptography has been around for several thousands of years. During this time, different forms of cryptosystems have been developed. Cryptographic algorithms can be divided into symmetric key algorithms and public key algorithms. In the symmetric cryptosystem, the encryption and the decryption keys are the same. Among symmetric cryptosystems, ciphers of different security levels have been developed, ranging from the substitution and transposition ciphers to block ciphers, such as the Blowfish. As of today, the Blowfish has no cryptanalysis. This paper proposed a new algorithm combining Blowfish and the Mixed strategy (MS), named as MS-Blowfish to improve the performance of the Blowfish cryptography algorithm by making modifications to the Feistel (F) function. The outcome of the Blowfish and MS-Blowfish algorithms are compared using Avalanche Effect to show the security enhancement of MS-Blowfish.

Published on: 08<sup>th</sup>– August-2016

## KEY WORDS

Avalanche Effect; Blowfish;  
Cryptanalysis; Feistel  
Network; Mixed Strategy

\*Corresponding author: Email: [v.jose08@gmail.com](mailto:v.jose08@gmail.com); Tel.: 9443151625

## INTRODUCTION

Cryptography plays an important role for protecting data from destructive forces and the unwanted actions of unauthorized users. Cryptographic algorithms have mathematically become more and more complex with time due to the ever increasing need for data security. However, the increase in the complexity of such algorithms incurs more computation overhead, which in turn leads to more execution time and high energy consumption recent years; successful studies have been made to speedup the execution of cryptographic algorithms. The Blowfish algorithm was designed by Bruce Schneier to replace Data Encryption Standard, which was the Federal Information processing Standard Cryptography [1]. It is a symmetrical block cipher [2] having the advantages of secure, fast, easy to implement etc. The operation part of Blowfish consists of XORs and additions on 32-bit words, and only 4KB or even less memory is needed when it runs. The key length of Blowfish is anywhere from 32 bits to 448 bits, which makes datum safe enough. The proposed MS-Blowfish algorithm enhances the performance over Blowfish by modifying the function F of the existing Blowfish. There are a lot of benefits from parallel computing. The advantage of this system is its ability to handle large and extremely complex computations. The basic idea of this research is to simplify complicated cryptographic algorithms by splitting up their tasks to run in parallel successfully so that they execute fast and consume less energy. Amdahl's law states that possible speed gains are limited by the fraction of the software that can't be parallelized to run on multiple cores simultaneously [3]. The Parallel processing, Blowfish and Mixed Strategy concept in Game Theory are combined so that the security is increased. The Avalanche effect is used to show that the proposed MS-Blowfish algorithm possess good diffusion characteristics as that of original Blowfish algorithm [2] [4]. The objective of this research paper is to study the Blowfish algorithm and enhance its performance using Parallel Processing and Mixed Strategy technique.

## RELATED WORK

### System Specification

For this research a Laptop with Intel Pentium T4500 @ 2.30GHz CPU, 4.00GB Dual-Channel DDR3 and Linux Mint 17.1 is used in which the performance data are collected. In this the software encrypts the text file size that ranges from 50 bytes to 208942 bytes.



Their implementation is thoroughly tested and is optimized to give the maximum performance for the algorithm. The performance matrices are the encryption speed, decryption speed, execution time, encryption throughput, decryption throughput, execution throughput and avalanche effect. The Blowfish cryptosystem was implemented using the C programming language in gcc compiler.

## Game Theory

The field of game theory and cryptographic protocol design are both concerned with the study of interactions among mutually distrustful parties. These two subjects have, historically, developed almost entirely independently within different research communities and, indeed, they tend to have a very different behavior. In Game Theoretic settings players are assumed to be rational. A great deal of effort was invested in trying to capture the nature of rational behavior, resulting in a long line of stability concepts. Cryptographic protocols are designed under the assumption that some parties are honest and faithfully follow the protocol, while some parties are malicious and behave in an arbitrary fashion. The game-theoretic perspective, however, is that all parties are simply rational and behave in their own best interests. This viewpoint is incomparable to the cryptographic one, although no one can be trusted to follow the protocol unless it is in their own best interests, the protocol need not prevent irrational behavior [5].

## Mixed Strategy

In the theory of games a player is said to use a mixed strategy whenever he or she chooses to randomize over the set of available actions. Formally, a mixed strategy is a probability distribution that assigns to each available action a likelihood of being selected. If only one action has a positive probability of being selected, the player is said to use a pure strategy. A mixed strategy profile is a list of strategies, one for each player in the game. A mixed strategy profile induces a probability distribution or lottery over the possible outcomes of the game.

One feature of mixed strategy equilibrium is that given the strategies chosen by the other players, each player is indifferent among all the actions that he or she selects with positive probability. In an interpretation advanced in 1973 by John Harsanyi, mixed strategy equilibrium of a game with perfect information is viewed as the limit point of a sequence of pure strategy equilibria of games with imperfect information. Specifically, starting from a game with perfect information, one can obtain a family of games with imperfect information by allowing for the possibility that there are small random variations in payoffs and that each player is not fully informed of the payoff functions of the other players. Harsanyi showed that the frequency with which the various pure strategies are chosen in these perturbed games approaches the frequency with which they are chosen in the mixed strategy equilibrium of the original game as the magnitude of the perturbation becomes vanishingly small. A very different interpretation of mixed strategy equilibria comes from evolutionary biology. To illustrate this, consider a large population in which each individual is programmed to play a particular pure strategy. Individuals are drawn at random from that population and are matched in pairs to play the game. The payoff that results from the adoption of any specific pure strategy will depend on the frequencies with which the various strategies are represented in the population. Suppose that those frequencies change over time in response to payoff differentials, with the population share of more highly rewarded strategies increasing at the expense of strategies that yield lower payoffs. Any rest point of this dynamic process must be Nash equilibrium. The long-run population share of each strategy corresponds exactly to the likelihood with which it is played in the mixed strategy equilibrium [6].

## Parallel processing

In parallel processing, each individual processor works the same as any other microprocessor. The processors act on instructions written in assembly language. Based on these instructions, the processors perform mathematical operations on data pulled from computer memory. The processors can also move data to a different memory location.

Processors rely on software to send and receive messages. The software allows a processor to communicate information to other processors. By exchanging messages, processors can adjust data values and stay in sync with one another. This is important because once all processors finish their tasks, the CPU must reassemble all the individual solutions into an overall solution for the original computational problem. There are two major factors that can impact system performance: latency and bandwidth. Latency refers to the amount of time it takes for a processor to transmit results back to the system. It is not good if it takes the processor takes less time to run an algorithm than it does to transmit the resulting information back to the overall system. In such cases, a sequential computer system would be more appropriate. Bandwidth refers to how much data the processor can transmit in a specific amount of time. A good parallel processing system will have both low latency and high bandwidth.

## BLOWFISH ALGORITHM

The Blowfish algorithm inputs a 64-bit plaintext and then outputs a 64-bit cipher text. It takes a variable-length key, from 32 bits to 448 bits [7], making it ideal for both domestic and exportable use. The algorithm consists of two parts: a key-expansion part and a data-encryption part. Key expansion converts a key of at most 448 bits into several sub key arrays totaling 4168 bytes. The original sub key p-box and s-box are fixed. They are initialized in order with a fixed string that consists of hexadecimal digits of Pi (less the initial 3). Data encryption occurs via a 16-round Feistel network [8] after key expansion. Each round consists of a key-dependent permutation, and a key- and data-dependent substitution. The algorithm uses two boxes: key p-box [18] and key s-box [4] [256], and a core

Feistel function: The two boxes take up  $18 \times 32 + 256 \times 32 = 4186$  bytes memory. The sub keys must be pre-computed before any data encryption or decryption.  
Function F is: Divide XL into four eight-bit quarters: a, b, c, and d

$$F(XL) = F(a, b, c, d) = ((S1, a + S2, b \text{ mod } 2^{32}) \text{ XOR } S3, c) + S4, d \text{ mod } 2^{32}$$

Herein, “+” is addition on 32-bit words, and XOR represents Exclusive OR; S1, a represents key s-box [1] [a], and similar of others.

Figure - 1 shows all operations are XORs and additions on 32-bit words. The only additional operations are four indexed array data lookups per round. The process of decryption is the same as encryption, except that key p-box is used in the reverse order.

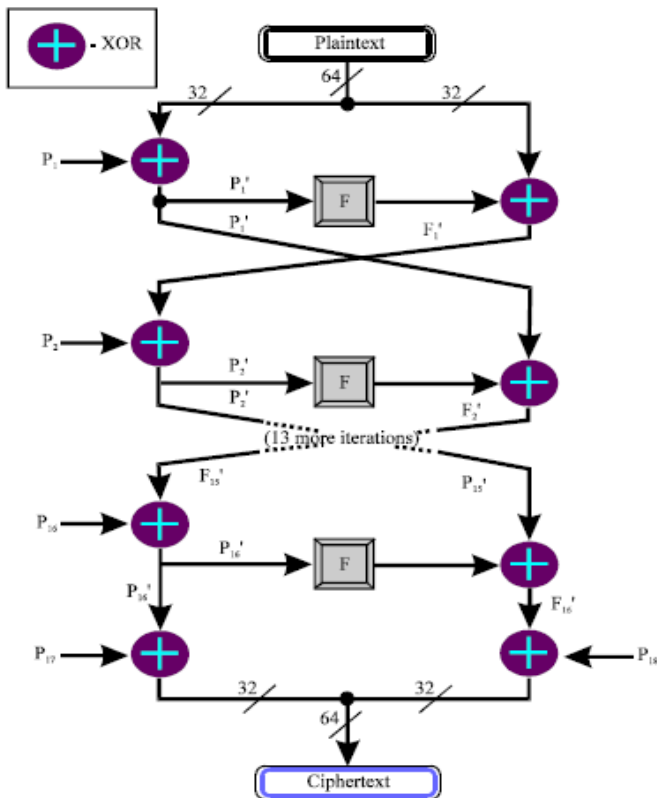


Fig. 1: XORs and additions on 32-bit words

The principle of Blowfish algorithm is both easy to understand and easy to implement. Different with other ciphers, all sub keys of Blowfish are influenced by every bit of the key, that makes the key and the data mingled together completely, which makes it quite difficult to analyze the key [9]. The function F gives the Feistel network a great avalanche effect.

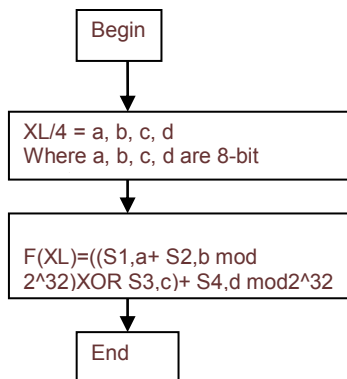
Blowfish cipher is not only secure, but also fast, and suitable for different platforms, therefore, it has a high value of application in the field of information security. Blowfish is among the fastest block ciphers available [10]. Blowfish is used in wide range of applications such as bulk encryption of data files, remote backup of hard disk. Also multimedia applications use blowfish for encryption of voice and media files. It is now being used in biometric identification and authentication, using voice, facial or fingerprint recognition. Geographical information system uses blowfish for cryptographic protection of sensitive data. These applications run in high-end servers, workstations, process bulk amount of data and demand high speed encryption and higher throughput [11]. A study was conducted for different popular key algorithms such as DES, 3DES, AES and Blowfish. They were implemented, and their performance was compared by encrypting input files of varying contents and sizes. The

algorithms were tested on two different hardware platforms, to compare their performance. The results showed that Blowfish had a very good performance compared to the other algorithms [12]. Bruce Schneier made a block cipher speed comparison among Blowfish, RC5, DES, IDEA, 3DES algorithms [13]. The results showed the advantage of Blowfish among block ciphers in terms of speed. The results are shown in Table-1. From the Table-1 it is clear that, the future of Blowfish as a secure algorithm is very promising. Blowfish algorithm is not only secure, but also fast, and suitable for different platforms. So it is widely used in the field of information security [14].

Table: 1. Block Cipher Speed Comparison

Algorithm	Clocks/round	No. of Rounds	Clocks/byte of output
Blowfish	9	16	18
RC5	12	16	23
DES	18	16	45
IDEA	50	8	50
3DES	18	48	108

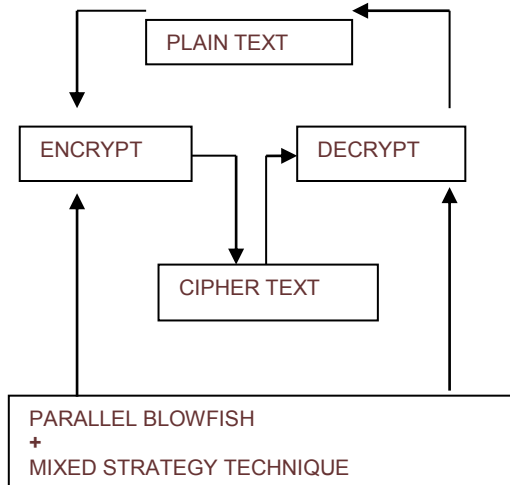
The following figure shows the calculation of the function F(XL) using Blowfish algorithm.



## PROPOSED MIXED STRATEGY-BLOWFISH ALGORITHM AND ANALYSIS

### MS-Blowfish

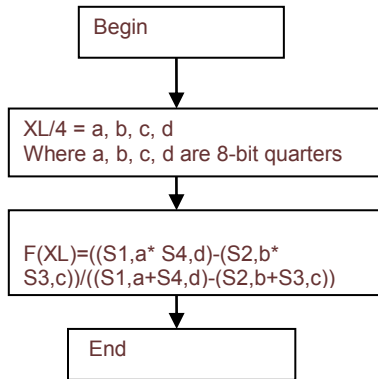
The block diagram shown below represents the structure of MS-Blowfish algorithm.



The proposed MS-Blowfish algorithm is similar to the Blowfish algorithm with a modification in the F function. The modification shows parallel evaluation of different operations within the function. Without violating the security requirements, the Blowfish function F can be modified as follows:-

$$F(XL) = ((S1,a * S4,d) - (S2,b * S3,c)) / ((S1,a + S4,d) - (S2,b + S3,c))$$

This modification supports the parallel evaluation of two multiplication operations and two addition operations. Then parallel evaluation of two subtraction operations. Finally, a division operation. All these operations take place in 3 steps. The following figure shows the calculation of F function using MS-Blowfish.



### Performance Comparisons

In this paper the performance metrics execution time, encryption time, decryption time, throughput, avalanche effect, power consumption are used to evaluate the blowfish algorithm and MS-Blowfish algorithm. The encryption time, the decryption time, the Execution time, is low for Blowfish algorithm than MS-Blowfish algorithm. But the Avalanche Effect is high for MS-Blowfish than Blowfish. MS-Blowfish is the best in terms of security. Blowfish algorithm by itself is highly secure. But above all MS-Blowfish is unbreakable in any circumstances.

## EXPERIMENTAL RESULTS

### Encryption Time

Encryption Time is one of the performance metrics which is defined as the amount of time required for converting plaintext message to cipher text at the time of encryption [15]. Tabulation of results of encryption time with different packet size for Blowfish algorithm and MS-Blowfish algorithm are shown in Table- 2. and Table-3. The encryption time of MS-Blowfish algorithm is slightly more than Blowfish algorithm.

### Decryption Time

Decryption Time is one of the performance metrics which is defined as the amount of time required for converting the cipher text into the plain text at the time of decryption. Tabulation of results of decryption time with different packet size for Blowfish algorithm and MS-Blowfish algorithm are shown in Table- 2. and Table-3. The decryption time for MS-Blowfish algorithm is slightly more than Blowfish algorithm.

### Execution Time

Execution time of an algorithm directly depends on the functionality of the algorithm and it clearly defines that more complex structure originates poor execution time. Higher the key length provides higher security but increases execution time. The speed of the algorithm is determined by the execution time of the algorithm. Tabulation of results of execution time with different packet size for Blowfish algorithm and MS-Blowfish algorithm are shown in Table- 2. and Table- 3. The execution time for MS-Blowfish algorithm is slightly more than Blowfish algorithm.

## Throughput

The throughput of the encryption scheme is calculated by dividing the total plaintext in Megabytes encrypted on the total encryption time for each algorithm in.

$$\text{Throughput} = \text{Total Plaintext in MegaBytes} / \text{Encryption Time}$$

The higher the value of throughput more is the efficiency of encrypting any text with an encryption algorithm. Tabulation of results of throughput with different packet size for Blowfish algorithm and MS-Blowfish algorithm are shown in **Table-2** and **Table-3** for Blowfish and MS-Blowfish algorithms respectively.

**Table: 2. Speed Analysis of Blowfish Algorithm**

Data size	Encryption	Decryption	Execution
50	0.7586	0.7602	0.8875
60	0.7709	0.7722	0.9058
100	0.7919	0.7934	0.9543
250	0.8962	0.8978	1.1592
325	0.9486	0.9497	1.2615
700	1.2776	1.2005	1.7646
900	1.3354	1.3364	2.0352
965	1.3741	1.549	2.29
5350	4.5246	4.4654	8.3683
7400	5.9181	5.8499	11.146
9000	6.9128	5.1447	11.4318
51202	20.9473	16.2216	36.5376
61442	23.8123	19.2313	42.4173
102402	37.7555	31.5148	68.651
208942	63.2736	63.159	126.085
Average Time(millisecond)	11.41983333	10.25639333	21.05967333
Throughput(MB/sec)	2.500233162	2.783848579	1.3557782

**Table: 3. Speed Analysis of MS-Blowfish Algorithm**

Data size	Encryption	Decryption	Execution
50	1.0458	1.0482	1.1875
60	1.0562	1.0586	1.2071
100	1.0908	1.0932	1.2794
250	1.2321	1.2342	1.5609
325	1.3014	1.304	1.7002
700	1.6517	1.7891	2.5419
900	1.8431	1.9559	2.8977
965	1.9028	1.8738	2.8709
5350	6.1375	4.9518	10.1801
7400	7.2244	5.4114	11.7265
9000	8.2606	5.172	12.5251
51202	26.807	23.4668	48.2796
61442	30.7288	26.6701	56.5023
102402	45.6741	43.9294	88.5926
208942	87.9248	88.0193	175.5779
Average Time(millisecond)	14.92540667	13.93185333	27.90864667
Throughput(MB/sec)	1.912996184	2.049422439	1.023060807

### Avalanche Effect

A change in one bit of the plain text or one bit of the key schedule will produce a change in many bits of the cipher text. This change in number of bits in the cipher text whenever there is a change in one bit of the plain text or one bit of the key is called Avalanche Effect [16]. A desirable feature of any encryption algorithm is that a small change in either the plain text or the key should produce a significant change in the cipher text. If the changes are small, this might provide a way to reduce the size of the plain text or key space to be searched and hence makes the cryptanalysis very easy. For a cryptographic algorithm to be secure it should exhibit strong Avalanche effect. Tabulation of results observed by changing one bit of plain text in the sample is shown in **Table-4**. **Figure- 2**. represents the Avalanche effect of Blowfish algorithm and MS-Blowfish algorithm. In the bar chart Blowfish is represented as BF and MS-Blowfish as MSBF.

Algorithm	BF	MSBF
Avalanche	57.1	62.1

The Blowfish algorithm has the lowest Avalanche effect when compared to the MS-Blowfish algorithm discussed here. So it is clear that MS-Blowfish algorithm is more secure than Blowfish algorithm.

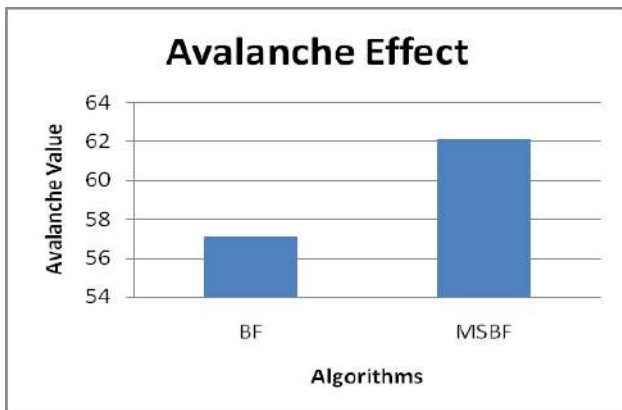


Fig: 2. Comparison of Avalanche Effect

### CONCLUSION

This paper gives a detailed study of the most popular symmetric key encryption algorithm that is Blowfish and discussed about its advantages. Based on the benefits of Blowfish algorithm we have proposed and implemented a new approach to further enhance the existing algorithm to achieve better results in terms of security. The striking feature of Blowfish encryption algorithm is that for the same input plaintext the cipher text generated at each time will be different. This is because every time a new random number gets generated and this as a result gives difference in the application of F function over each round. The advantage of different cipher text generated for the same input is it will greatly enhance the security aspect of blowfish algorithm. The above results clearly indicate that the Avalanche effect of MS-Blowfish is much better than Blowfish algorithm. So it is clear that MS-Blowfish algorithm is very strong, secure and unbreakable than the Blowfish algorithm. The research work can be further extended with other optimization techniques which have potential capacities

### ACKNOWLEDGEMENT

None

## CONFLICT OF INTEREST

No conflict of interest

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] US National Bureau of Standards.[1977]Data encryption standard,"U.S. Fed. Inform. Processing Standards Pub., FIPS PUB 46,January.
- [2] Bruce Schneier.[1996] Applied Cryptography: Protocols, Algorithms, and Source Code in C. 2nd Edition, New York, John Wiley and Sons Inc, 21-27.
- [3] Gene M. Adahl.[1967] Validity of the single processor approach to largescale computing capabilities, in proceedings of the April 18-20, 1967, spring joint computer conferenceAFIPS'67(spring) ACM,Newyork,NY, USA, pp.483-485.
- [4] William Stallings. [2011] Cryptography and Network Security, *Fifth Edition, Pearson Education*, 119-120.
- [5] YevgeniyDodis, ShaiHalevi, Tal Rabin. [2000]A Cryptographic Solution to a Game Theoretic Problem, *Advances in Cryptology Crypto 2000*, Springer-Verlag Berlin,Heidelberg-, 113-130.
- [6] International Encyclopedia Of The Social Sciences, 2nd Edition, pp.290-292.
- [7] Bruce Schneier[1993] Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish), in Cambridge Security Workshop on Fast Software Encryption, Cambridge, UK, December 9-11:191-204.
- [8] Bruce.Schneier[1994] The Blowfish Encryption Algorithm, *Dr. Dobb's Journal*, 19(4): 38-40.
- [9] Bruce Schneier. [1994]Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish) Fast Software Encryption, Cambridge Security Workshop Proceedings (December 1993), Springer-Verlag, 191-204.
- [10] on a Pentium.Retrieved 08:04:58, August 27, 2010 from <http://www.schneier.com/blowfish-speed.html>.
- [11] T Srikanthan et al. Drill – A Flexible Architecture for BlowfishEncryption Using Dynamic Reconfiguration, Replication, Inner-Loop, Pipelining, Loop Folding Techniques, Springer- Verlag Berlin Heidelberg.6256-639.
- [12] AamerNadeem, M YounusJaved.[ 2005] A Performance Comparison of Data Encryption Algorithms,*IEEE, Information and Communication Technologies*, 2005, ICICT 2005.First International Conference, -02-27:84-89.
- [13] <https://www.schneier.com/cryptography/blowfish/speed.html>
- [14] Mingyan Wang,YanwenQue, The Design and Implementation of Password Management System Based on Blowfish Cryptographic Algorithm,IEEE Xplore, International Forum on Computer Science- Technology and Applications,2009, IEEE Computer Society, 978-0-7695-3930-0/09.
- [15] R. Mohan, C.Rajan, Dr. N.Shanthi, "A Stable Mobility Model Evaluation Strategy for MANET Routing Protocols." *International Journal of Advanced Research in Computer Science and Software Engineering*. vol. 2, p.p.58-65, December 2012.
- [16] Krishnamurthy GN, V.Ramaswamy, Leela GH, Ashalatha ME.[2008] Performance enhancement of Blowfish and CAST-128 algorithms and Security Analysis of Improved Blowfish Algorithm Using Avalanche Effect, *IJCSNS*,8 (3): 244-250.

## ABOUT AUTHORS



**Dr.V. Joseph Raj** received the Ph.D. degree from ManonmaniamSundaranar University, Tirunelveli, India and P.G. degree from Anna University, Chennai, India. He worked as an Associate Professor of Department of Computer Engineering in European University of Lefke, North Cyprus for two years. He has been working as a Professor and HOD of Computer Science in Kamaraj College, Thoothukudi, affiliated to ManonmaniamSundaranar University, Tamilnadu, India. He is serving as Chairman of Computer Applications (P.G.) Board of Studies and Chairman of Computer Science Board of Examinations of ManonmaniamSundaranar University. He has been guiding Ph.D. scholars in various Indian Universities and many of his Ph.D. scholars were awarded Ph.D. degree. He has a vast teaching experience of about 29 years and research experience of 21 years. His research interests include Artificial Neural Network, Digital Image Processing, Wireless Networks, Operations Research and Biometrics. He has published several articles in International Journals and Conferences and National Journals and Conferences.



**Mrs. Shamina Ross B.** received the M.Phil. degree from Vinayaka Mission University, Salem, India, P.G. degree from Annamalai University, Chidambaram, India and U.G. degree from ManonmaniamSundaranar University, Tirunelveli, India. She is presently working as Assistant Professor in the Department of Computer Applications in Scott Christian College, Nagercoil, affiliated to ManonmaniamSundaranar University, Tamilnadu, India. She has more than 7 years of teaching experience and research experience of 5 years. She has published papers in International Journals. Her area of interest in research includes Network Security and Artificial Intelligence.

# IMPROVING EFFICIENCY OF SELFISH NODE DETECTION IN AD-HOC NETWORK USING COLLABORATIVE CONTACT BASED WATCHDOG

Meenakshi P\*, V. Jayashreeja, D. Gayathri, C. Sowmya

Vel Tech, Dr. Rangarajan Dr. Sakunthala Engineering College, INDIA

## ABSTRACT

In ad-hoc networks, there are some nodes which refuse to co-operate with network in transferring the data. Such nodes behave in a selfish manner and hence called selfish nodes. Due to the presence of selfish nodes, the problem of false positive and false negative occurs. Therefore detecting those nodes is essential for the overall performance of the network. For detecting such nodes and reducing the positive and negative detections, we introduce a COCOWA model that would launch the collaborative contact-based watchdog. It detects by collecting the forwarding history evidence from its upstream and downstream nodes. To further improve the performance of the proposed collaborative inspection scheme, a reputation system is introduced, in which the inspection probability could vary along with the target node's reputation. Under this system, a node with a good reputation will be checked with a lower probability while a bad reputation node could be checked with a higher probability. Downstream nodes. To further improve the performance of the proposed collaborative inspection scheme, a reputation system is introduced, in which the inspection probability could vary along with the target node's reputation. Under this system, a node with a good reputation will be checked with a lower probability while a bad reputation node could be checked with a higher probability.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

AD-HOC network,  
selfish node detection

\*Corresponding author: Email: [minuraj11@gmail.com](mailto:minuraj11@gmail.com); Tel: +91 9710432387

## INTRODUCTION

The fundamental task of the nodes in the network is to transfer the data from source to destination successfully. Mobile ad-hoc networks constitute the mobile nodes which move with a certain mobility. In mobile ad-hoc networks, the mobile nodes voluntarily cooperate in order to work properly. This is a cost-intensive activity and some nodes can refuse to cooperate leading to selfish node behavior. The above situation can lead to the decrease in network performance. The watchdog is a well-known mechanism to detect selfish nodes, but they can fail, generating false positive and false negative that can induce to wrong operations. Moreover relying on local watchdogs alone can leave to poor performance when detecting selfish nodes in term of precision and speed. This is especially important on networks with sporadic contacts, such as Delay Tolerant Networks (DTNs) where sometimes watchdogs lack of enough time or information to detect the selfish nodes. Thus we propose Collaborative Contact-based watchdog (COCOWA) as a collaborative approach based on the diffusion of local selfish nodes awareness when a contact occurs, so that information about selfish nodes is quickly propagated. Selfishness means that some nodes refuse to forward other nodes' packets to save their own resources. In DTNs, selfish nodes can seriously degrade the performance of packet transmission. In a survey, the number of packet losses is increased by 500 percent when the selfish node ratio increases from 0 to 40%. Another problem is the presence of colluding or malicious nodes. Malicious node intentionally disturbs the correct behavior of the network, so the detection process is necessary for the proper performance of the network.

## EXISTING SYSTEM

The local watchdog does not evaluate the effect of false positives, false negatives and malicious nodes. For example, the approach only transmits positive detections. The problem is that if a false positive is generated it can spread this wrong information very quickly on the network, isolating nodes that are not selfish. Therefore, an approach that includes the diffusion of negative detections as well becomes necessary. Another problem is the



impact of colluding or malicious nodes. Although a reputation system can be useful to mitigate the effect of malicious nodes, it clearly depends on how are combined local and global ratings, as shown in this paper. Another implementation issue is the high imposed overhead due to the flooding process in order to achieve a fast diffusion of the information

### Problem Definition

Attack detection process is not satisfied.

High communication cost overhead due to the data transmission from source to destination.

The data transmission process takes much time.

### PROPOSED SYSTEM

This paper proposes CoCoWa as a collaborative contact-based watchdog to reduce the time and improve the effectiveness of detecting selfish nodes, reducing the harmful effect of false positives, false negatives and malicious nodes. CoCoWa technique is used to detect Sybil attack, black whole attack and redirect attack. Nodes is attacked by (I)Sybil attack, it will forward the data but it won't acknowledge to the source, it won't forward the data but it will acknowledge to the source.(II)Black whole attack, it will interprets the data.(III) Redirect attack is to forward the data to source. CoCoWa can reduce the overall detection time with respect to the original detection time when collaboration scheme is not issued, with a reduced overhead (message cost).

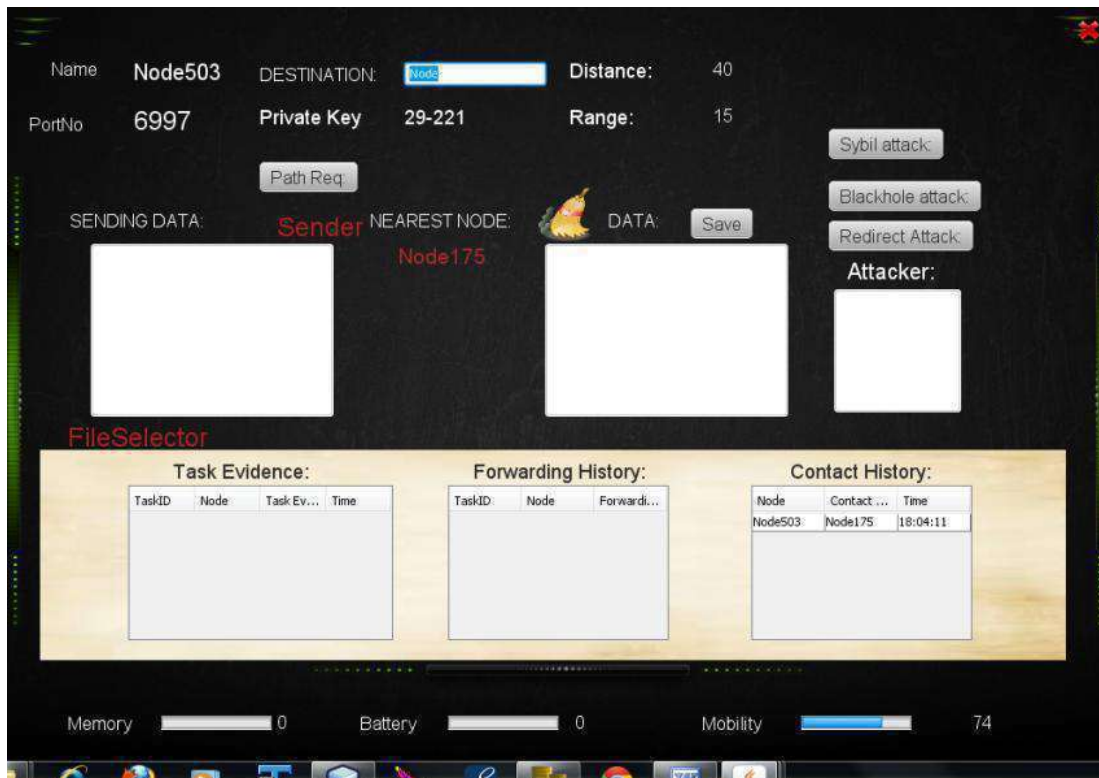
### SNAPSHOTS

#### THE HOME SCREEN





CREATION OF NODES



### PATH REQUEST

Name: Node430 DESTINATION: Node423 Distance: 60  
PortNo: 1478 Private Key: 223-679 Range: 25

Path Req: Node430 -> Node082 -> Node423

SENDING DATA: Sender NEAREST NODE DATA: Save

Attacker:

FileSelector

TaskID	Node	Task Ev...	Time
--------	------	------------	------

TaskID	Node	Forward...
--------	------	------------

Node	Contact ...	Time
Node430	Node503	18-04-35
Node430	Node175	18-04-36
Node430	Node240	18-04-36
Node430	Node082	18-04-38

Memory: 22 Battery: 44 Mobility: 62

### BLACK HOLE ATTACK

Name: Node082 DESTINATION: Node Distance: 20  
PortNo: 9786 Private Key: 2177-2923 Range: 10

Path Req: Node423

SENDING DATA: Sender NEAREST NODE DATA: Save

Attacker:

FileSelector

TaskID	Node	Task Ev...	Time
--------	------	------------	------

TaskID	Node	Forward...
--------	------	------------

Node	Contact ...	Time
Node082	Node423	18-04-14

Memory: 49 Battery: 77 Mobility: 79

### AFTER INSPECTION



## CONCLUSION

Thus we design, to reduce transmission overhead incurred by misbehavior detection and detect the malicious nodes effectively for secure MANET routing.

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

None.

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] S Abbas, M Merabti, D Llewellyn-Jones, and K.Kifayat.[ 2013]Lightweight sybil attack detection in manets, *IEEE Syst J*, 7( 2): 236–248
- [2] S Bansal and M. Baker.[ 2003.] Observation-based cooperation enforcement in ad hoc networks, arXiv:cs.NI/0307012,
- [3] S Buchegger and J.-Y.Le Boudec, Self-policing mobile ad hoc networks by reputation systems, *IEEE Commun.Mag*, 43.

# DESIGN OF SUB SAMPLING FILTER ARCHITECTURE FOR DISCRETE WAVELET TRANSFORM

Nirmala R<sup>1</sup> and Sathiya Sekar K<sup>2</sup>

<sup>1</sup>Dept. of ECE, Vivekanandha college of Engineering for Women, Ellayampalayam, Tiruchengode, INDIA

<sup>2</sup> Dept. of EEE, S A Engineering College, Chennai, INDIA

## ABSTRACT

Power dissipation and area reduction is the main constraint in the present scenario of VLSI circuits. The power dissipation of the circuit is mainly due to static or leakage power. The leakage power contributes to about 50% of the power dissipation in all the devices that are used in our day to day life. Discrete Wavelet Transform (DWT) has more advantages compared with FFT and DCT and it has many applications. Design of VLSI architecture of DWT is very important in the present scenario. To get better efficiency and throughput, DWT architecture is proposed with sub sampling filter with array multiplier and ripple carry adder. In the existing system, Carry save adder and Bough Wooley multiplier were used to design a subsampling filters, resulting that the power dissipation and number of transistors required are more which leads to complexity. In the proposed method the number of transistors required to design the sub sampling filter and power dissipation are measured using LTSpice IV tool which is less compared with existing filter.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

VLSI circuits; Discrete Wavelet Transform; Bough Wooley Multiplier; Transistors; LTSpice IV tool.

\*Corresponding author: Email: [nirdha06@gmail.com](mailto:nirdha06@gmail.com), [ksathiyasekar@gmail.com](mailto:ksathiyasekar@gmail.com); Tel.: +91 9688864777

## INTRODUCTION

In the field of signal/image processing currently we are using wavelet transform instead of Fourier Transform (FT), the Discrete Cosine Transform (DCT) and Discrete Sine Transform (DST). The Discrete Wavelet Transform (DWT) is an efficient platform for multi resolution analysis, with this signals can be decomposed into different sub-bands with excellent characteristics in the time and frequency domain. Comparing with previous DCT, DWT has better coding efficiency and excellent quality of restoration of image with high compression ratio. Multiplier is one of the basic functional unit in digital signal processor. Most of the high speed DSP systems has minimized multiplication units with high data throughput. Most hardware implementations address one or two essential design optimizations to improve their performance in terms of area, throughput or power dissipation. High-throughput and lower-power VLSI implementations are considered two of the essential optimization axes, especially when considering portable and real-time DSP applications. The hardware implementation for the DWT design can be significantly improved by designing application-specific FIR filters. The effort required to design the different filter blocks prolongs the design time and reduces the overall design productivity.

In the paper [1] DWT architecture combines several optimizations that improve the performance of the hardware design in terms of throughput and power dissipation. We designed and analyzed the performance of numerous DWT architectures using pertinent metrics and cost functions that assess the impact of the design optimizations. In the paper [2] the conventional array multiplier is implemented using 16T full adder cell. In conventional array multiplier has a tradeoff between power and area. But array multiplier is synthesized using 10T full adder cell uses 96 less transistor count which reduces 2.82% of total power and increases the speed by 13.24% also 15.69% less power delay product when being compared with the conventional array multiplier. Paper [3] implements the 1D discrete Wavelet Transform architecture using poly phase structure. This structure has high scalability and hardware requirement is very less. In the paper [4] convolution based 1D discrete Wavelet Transform combines polyphase decimated FIR filter with pipelined computation structure to get higher through put and minimized chip area. Paper

[5] implements the 1-bit full adder with metal-oxide-semiconductor (CMOS) logic and transmission gate logic. It explores the power and area improvement.

In our work we have designed the sub sampling filter using multiplier and adder. The multipliers are implemented using Array, Baugh-wooley multiplier concepts and the adder structures used in the MAC are Ripple Carry, Carry Save, The paper is organized as follows, in 2<sup>nd</sup> subdivision DWT and its blocks are explained. In 3<sup>rd</sup> subdivision different types of multipliers and adders are explained. In 4<sup>th</sup> subdivision sub sampling filter design is designed with proposed methods. In 5<sup>th</sup> Results and Discussion and Finally conclusion.

## DISCRETE WAVELET TRANSFORM

In digital image processing field, compression is one of the most effective techniques. A signal can be decomposed into set of basic functions using wavelet transform, those set of basic functions are known as wavelets. Dilations and shifting are used to obtain wavelets from a mother wavelet. It is also known as single prototype wavelet. For the hardware implementation of DWT plenty of VLSI architectures are available. The architectures of 1-D DWT is classified into two types convolution-based and lifting-based. Plenty of VLSI architectures available and proposed for DWT hardware implementation. Convolution based and lifting based architectures are mainly used for 1-D DWT.. Critical path has been reduced using pipelining, will require large number of registers for 1-D structure. Polyphase decomposition is used for proper utilization, there are classified as two types:

Type-I decomposition:

$$H(z) = H_e(z^2) + z^{-1}H_o(z^2)$$

$$G(z) = G_e(z^2) + z^{-1}G_o(z^2)$$

$$\tilde{H}(z) = \tilde{H}_e(z^2) + z^{-1}\tilde{H}_o(z^2)$$

$$\tilde{G}(z) = \tilde{G}_e(z^2) + z^{-1}\tilde{G}_o(z^2)$$

Type-II decomposition

$$H(z) = H_e(z^2) + zH_o(z^2), \quad G(z) = G_e(z^2) + zG_o(z^2)$$

$$\tilde{H}(z) = \tilde{H}_e(z^2) + z\tilde{H}_o(z^2), \quad \tilde{G}(z) = \tilde{G}_e(z^2) + z\tilde{G}_o(z^2)$$

For DWT and IDWT, convolution based architectures can be constructed using the sub sampling and down sampling filters.

## ADDER ARCHITECTURES

Adders are used in digital signal processing applications or in processing elements. Adder unit is not only for addition and subtraction it's also used to perform multiplication and division. Two basic adders are half adder and full adder. The carry propagate adders can be classified as follow:

### Carry Save Adder (CSA)

To perform addition of multiple operands carry save adder is used. The generation of sum and carry in CSA is similar to basic adder structure; the difference is that at next stage the carry is added to the sum. RCA is used to produce the final stage results. **Figure- 1** shows transistor level diagram of carry save adder .

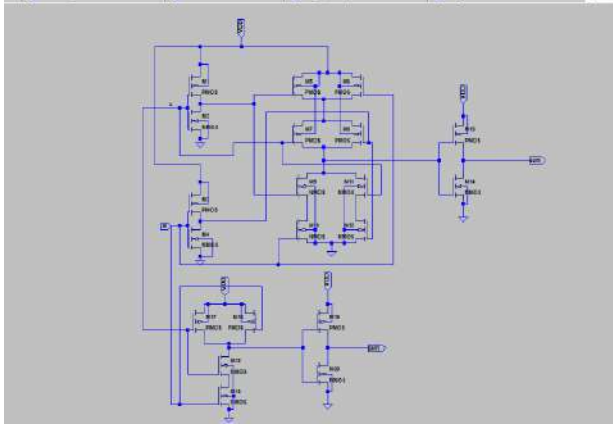


Fig:1. Transistor level diagram of carry save adder

**Ripple Carry Adder (RCA):**

N-bit RCA consist of N full adders where N is the total number of bits, here the carry signal propagate from LSB to MSB, it traverses longest path known as worst case delay path via N-stages. Mathematical equations are used to calculate propagate time **figure- 2,3** shows the implementation.

$$t_{adder} = N - 1 t_{carry} + t_{sum}$$

Where,  $t_{carry}$  – time taken to propagate carry,  $t_{sum}$  – time taken to produce sum.

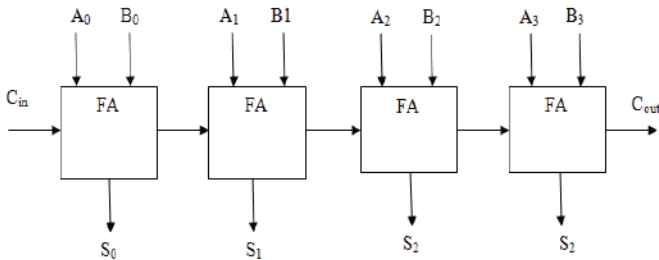


Fig:2. Block diagram of ripple carry adder

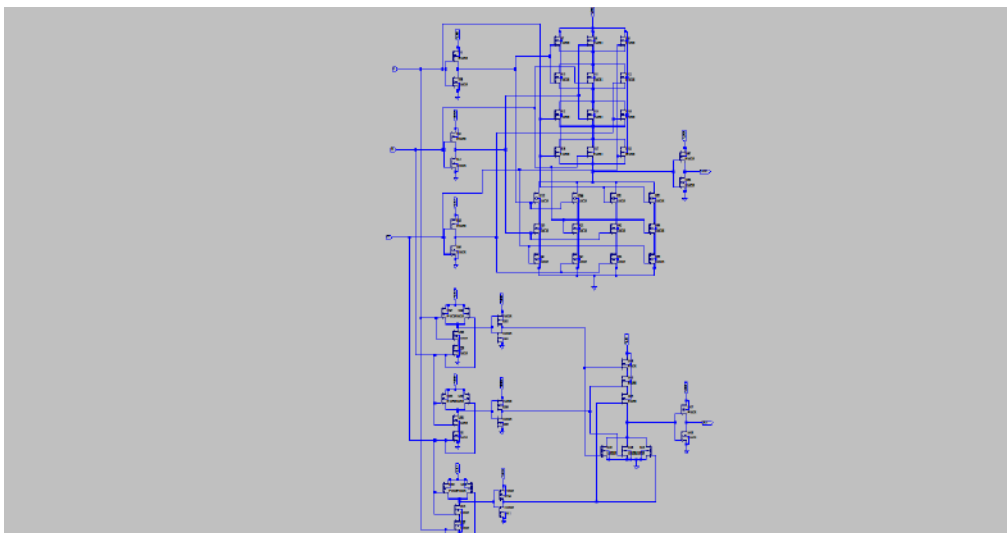


Fig: 3. Transistor level diagram for full adder

Conventional CSA using Ripple Carry Adder. In final stage of conventional carry save adder, RCA is used

### BAUGH WOOLEY MULTIPLIER

To handle the sign bits, Baugh wooley multiplication is the best effective method. To design a regular multiplier this approach has been developed and it's suited for 2's complement number in **Figure- 4,5**.

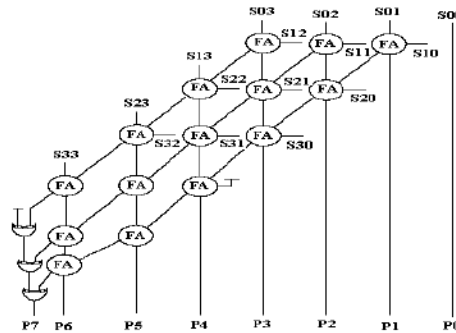


Fig: 4. Block diagram of Bough Wooley Multiplier

The **Figure- 4** describes the unsigned multiplication based on Baugh wooley algorithm. In that algorithm initially AND terms are created, those created AND terms are sent through an half-adders and full-adders with carry outs chained to the next MSB at each level of addition. It's also used to multiply the negative operands. Although the algorithm is not able to work with unsigned inputs, the goal of the project was a signed multiplier.

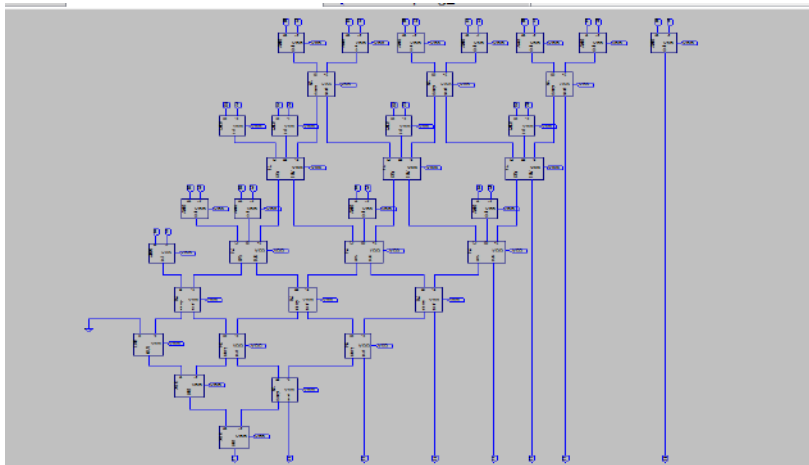


Fig: 5. Transistor level diagram of bough-wooley multiplier

### ARRAY MULTIPLIER

Array multiplier is very familiar because of its regular structure. Array multiplier structure is working based on repeated addition and shifting principle. The multiplication of each multiplier digit with multiplicand generate a partial product. Those partial products are added, before it is shifted into their bit sequence. A normal carry propagation adder is used to perform the summation. In array multiplier, N is the no. of multiplier bits; N-1 adders are required for implementation shows in **Figure -6,7**.



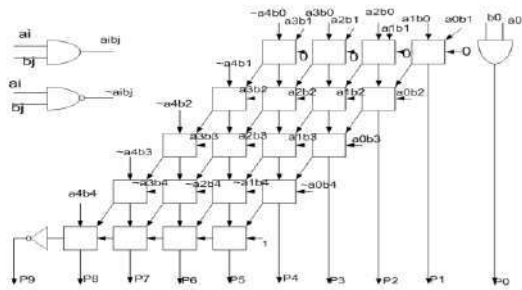


Fig: 6. Block diagram of array multiplier

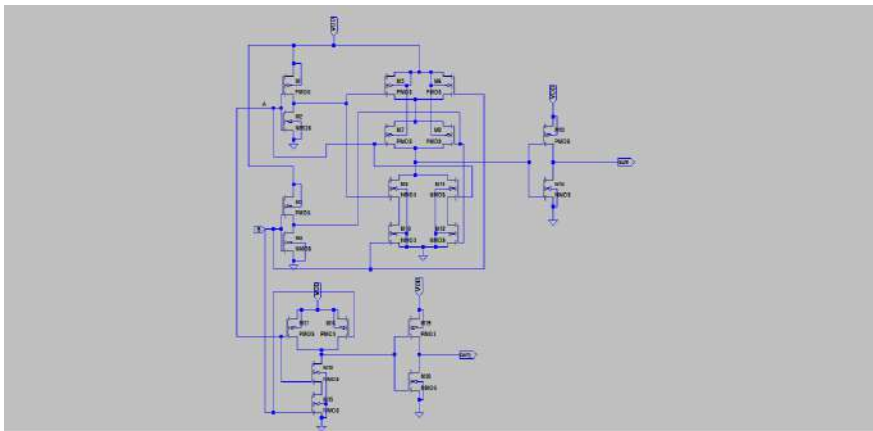


Fig: 7. Transistor Level Diagram of Array Multiplier

The advantage of array multiplier is its regular structure; therefore layout becomes simple and it occupies less area since it has small size. In VLSI, the regular structures can be cemented one over another; this reduces the amount of mistakes and also reduces layout design.

### PROPOSED SUBSAMPLING FILTER

In the DWT structure, the filters called sub sampling filters were used. The sub sampling filter was designed by using different module which consists of multiplexer, adder; multiplier and D- flip flop. For these sub sampling filters inputs are designed and implemented with 4 bits shown in figure 8,9. We implement the design with PTM 90 nm technology. The proposed sub sampling filter consist of 2:1 multiplexers, D Flip flop, Ripple carry adder, Array multiplier. The 4 bit binary data is given as a input for sub sampling filter. The 2:1 Multiplexer consist of two inputs A (4 bits) & B (4 bits) and select line 0 and 1, the output is produced by choosing a select line which is multiplied with input data using array multiplier. This output is added with another multiplexer output by ripple carry adder. This data bits are given to D Flip flop, which produce some delay. In that delays time period other module will perform the above function and produce the output which is given to another multiplexer. This process will be repeated for three module. Here, Ripple carry adder and array multiplier are used to reduce the power dissipation and area.

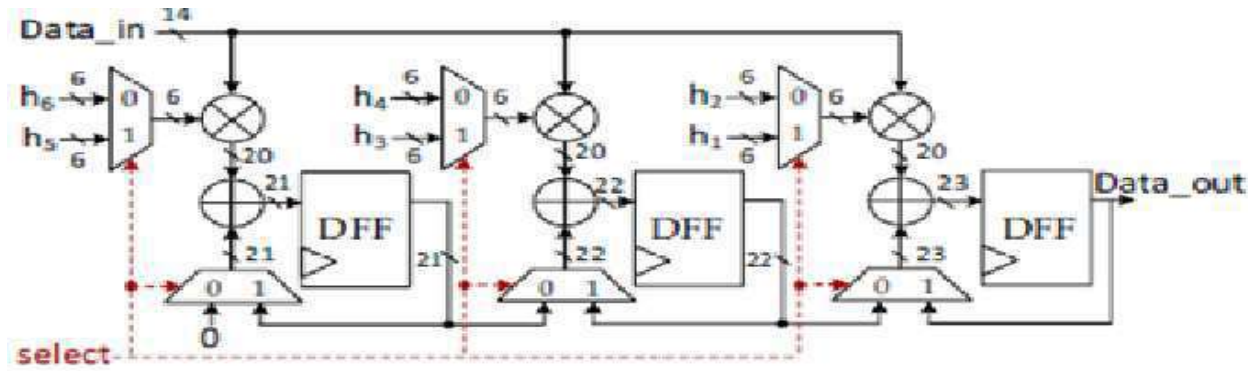


Fig: 8. Proposed Sub sampling filter

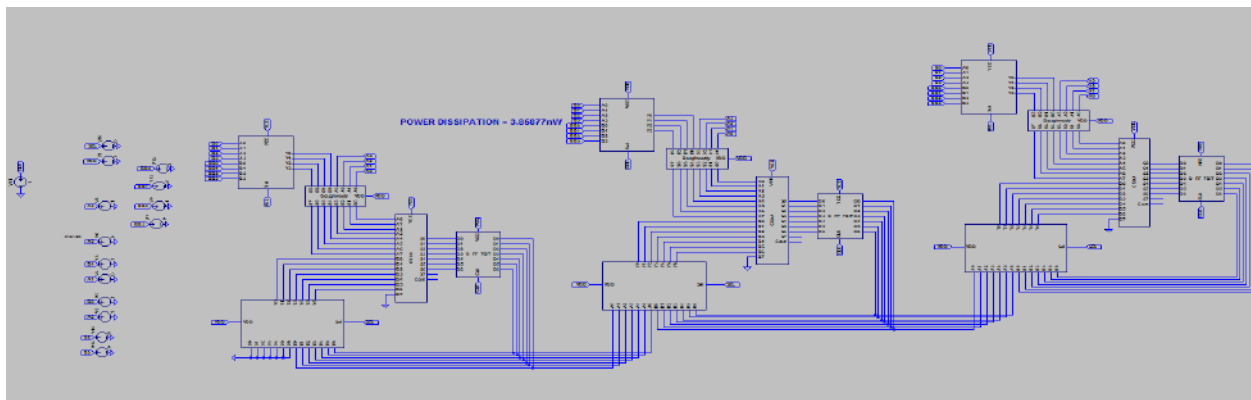


Fig: 9. Schematic View of Sub sampling Filter

## RESULTS AND DISCUSSION

In the 90nm technology, design of sub sampling filter with bough-wooley multiplier and carry save adder requires 3816 transistors and power dissipation is 3.85nW. But in the proposed system with ripple carry adder and array multiplier requires 3486 transistors and power dissipation is 3.48nW Shown in **Table- 1**. **Table- 2** explores the different types of adder and multiplier design with their power dissipation. From the results, the proposed design has significant improvement in reduction both area and power dissipation.

Table: 1. Comparison between existing and proposed system

Description	Subsampling filter with bough-wooley multiplier and carry save adder	Subsampling filter ripple carry adder and array multiplier
Technology	90nm	90nm
No. of. Transistor Used	3816	3486
Power dissipation	3.858 nW	3.4787 nW

Table: 2. Power dissipation for Adders and Multipliers

DESCRIPTION	POWER DISSIPATI
Bough-wooley multiplier	574.34pW
Array multiplier	476.45pW
Carry save adder	470.23pW
Ripple carry adder	151.77pW

### CONCLUSION

In this project, subsampling filter in discrete wavelet transform are designed using LT spice IV software with PTM(Predictive Technology Model) 90nm technology. In the existing system of subsampling filter with bough-wooley multiplier and carry save adder has 3816 transistors and power dissipation is 3.85nW. But in the proposed system with ripple carry adder and array multiplier has 3486 transistors and power dissipation is 3.48nW. From the results the proposed design of subsampling filter for DWT has better efficiency and less power conception.

### ACKNOWLEDGEMENT

None

### CONFLICT OF INTEREST

No conflict of interest

### FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

### REFERENCES

- [1] R Hourani, I Dalal, W Davis, C Doss, and W Alexander.[2008] An efficient VLSI implementation for the 1D convolutional discrete wavelet transform,
- [2] KripaMathew, S AshaLatha,T.Ravi and E Logashanmugam . [2013] Design and Analysis of an Array Multiplier Using an Area Efficient Full Adder Cell in 32nm CMOS Technology .high speed.
- [3] J Chilo and T Lindblad.[2008] Hardware implementation of 1D wavelet transform on an FPGA for infrasound signal classification,” *IEEE Transactions on Nuclear Science*, 55( 1): 9–13.
- [4] R Hourani, I Dalal, W Davis, C Doss, and W Alexander.[2008]An efficient VLSI implementation for the 1D convolutional discrete wavelet transform,” in IEEE International Conference on Midwest Symposium on Circuits and Systems (MWSCAS), Knoxville, Tennessee, U.S.A., August 2008, pp. 870–873.
- [5] Partha Bhattacharyya, Dept. of Electron. & Telecommun. Eng., Indian Inst. of Eng. Sci. & Technol., Howrah, India ; Bijoy Kundu ; Sovan Ghosh ; Vinay Kumar “ Performance Analysis of a Low-Power High-Speed Hybrid 1-bit Full Adder Circuit” in IEEE Transactions on Very Large Scale Integration (VLSI) Systems 23 ( 10 )

# DISTRIBUTED POLICY TRACKING IN WIRELESS SENSOR NETWORKS USING PARALLEL EXECUTION OF TINYPOLICY OVER P2P NETWORKS' LIVE STREAMING VIDEO ON-DEMAND SERVICE

Madeshan Narayanan<sup>1,2\*</sup> and Chokkalingam Arun<sup>3</sup>

<sup>1</sup>Department of Computer Science and Engineering, Sathyabama University, Chennai, INDIA

<sup>2</sup>Department of CSE, Saveetha University, Chennai, Tamil Nadu, INDIA

<sup>3</sup>Dept of Electronics and Communication Engineering, RMK College of Engineering and Technology, Chennai, Tamil Nadu, INDIA

## ABSTRACT

Peer-to-Peer Networks (P2P) play a vital role in all major domains today. This paper discusses how P2P network services are extended to wireless sensor networks (WSNs). In a sensor node's local memory, Policy-based management applications can store only a restricted number of policies in the local memory of a sensor node on WSNs, subject to hardware resource constraints, and are required to be recycled whenever extra policies are needed. To handle these issues, an operation called Parallel Execution of TinyPolicy is developed for storing, locating, accessing and executing policies in a WSN. It is devised to make full use of the memory available in a P2P network and duplication. The result is a more robust policy system against any failure of nodes and single points. To govern and control the embedded devices, the parallel execution of TinyPolicy will facilitate WSNs to solve these difficult issues. Utilizing a P2P architecture, distributed policy-based management and replication of policies, the new framework offers several novel features like dynamic distribution of policies between the sensor nodes. Additionally, the parallel execution of TinyPolicy manages the location of these policies dynamically by thrusting the widely-used policies against the target node, instead of leaving them solitarily in the WSN. This framework is simulated by an NS-2 simulator. In the near future, a WSN-and-P2P system combination may be a guide to developing robust applications. A P2P platform provided by an abstract program for communicating and allowing developers to take charge of functionality is a sign of simplifying the development process of distributed applications.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

Distributed Systems; P2P Networks; Parallel Execution; TinyPolicy; Wireless Sensor Networks

\*Corresponding author: Email: [narayanan\\_baba@yahoo.com](mailto:narayanan_baba@yahoo.com); Tel.: +91 9381485813

## INTRODUCTION

Wireless sensor networks (WSNs) are built for a variety of purposes. Their chief drawback, however, is that their parts have limitations in terms of processing and power. As a result, managing difficult algorithms with information collected by sensors ought to be made in modules outside the WSN. Application design could be a lot more flexible with the introduction of P2P networks. This can surmount the challenges usually related with WSNs as mentioned earlier. However, the design of distributed systems design is a challenging job, with a host of issues relating to communication between components.

In developing applications, P2P networks are a match for WSNs. While P2P networks are better suited for high-end nodes with substantial power, WSNs are well suited for capturing the surrounding information in intense conditions. The challenge is to integrate these architectures to cooperate for functionality's sake.

WSNs play an important role in many systems, assisting people in their usual day-to-day routines and getting them acclimatized to their present circumstances. But they are required to manage all by themselves to discover and configure tools for services, detect and respond to attacks, resolve faults and reconfigure the system to diminish all of these.

Based on the assessments of object attributes, a policy proclaims to choose target objects inside a domain. When all the objects in a domain can be applied by a policy, the simplest case arises. Domain membership can vary invariably and, consequently, a group of objects where the policy is applicable has to be examined during policy interpretation.

An illustration of a management authorization policy is the access rule, which indicates a connection between managers and managed objects in relation to managing operations allowed on objects of a certain kind. The rule also utilizes the scope to choose subsets of objects and, further, describe restrictions on these actions.

By stating why an object has been selected, a manager can indicate the primary membership for exploring a database of a different domain, except this is not offered as component of essential domain service. Policy membership is then required to restrict objects that can be consequently built into the domain or added from another. Additional membership policies are linked to the number of objects allowed in a domain.

A robust mobility-management structure for Internet Protocol version 6 (IPv6) and heterogeneous wireless networks, enabled by policy enforcement, is proposed. Policies are defined, based on infrastructure facility, service agreement and conciliation results. The results of the system's performance confirm that user experience has improved, largely in relation to connectivity. Service providers will therefore be able to hire expensive UMTS economically, and users will always be connected anytime, so long as the distribution of the flow and the selection of the network is visible and flexible [1,14].

This paper is organized in the following manner: In Section II, a step-by-step literature survey related to P2P network services on wireless sensor networks is tabled. In Section III, the proposed model for the parallel execution of TinyPolicy is presented. Details of the experimental results and implementation are presented in Section IV. The final section contains the conclusion and future extensions of the paper.

## LITERATURE REVIEW

WSNs (Wireless sensor networks) have become all-encompassing in day to day life, penetrating into fields like environment, medicine and defence studies. Every WSN comprises of a number of sensors which are accountable for monitoring single or multiple events. A WSN generally operates in various environments where sensors are obtained from several manufacturers which lead to incompatible issues with respect to standards in hardware and software. Even though specific types of sensors may overcome a few of these issues, it comes at a cost with complexity issues. Researchers have, consequently, recommended policy-based management (PBM) platforms as a suitable solution to trounce these challenges and effectually camouflage the complicated workings behind basic network devices.

A starfish framework in sensor nodes targets self-healing policy deployment. The framework comprises a Finger2 policy system for dynamically adapting a library module to make programming the essential functions of nodes less complex, as well as a client-side editor to manage policies. The policies described are for a health-care body network, with self-healing features for sensor networks and re-configuring policies to handle faults. There is also an intention to broaden the concept of self-healing services and incorporate them in self-managed cell architecture [1]. A model based on a policy understanding the concerns of the several actors involved in practical WSN applications is recommended, achieved by enhancing and optimizing the runtime environment. A sample implementation of the model is implemented and examined, using the SunSPOT platform. The results show that the model is adequately lightweight and can be applied with great advantage in WSN environments. The focus will also be on interaction with several parallel programming models [2-3].

Managing several sensor nodes is a daunting task where energy concerns are an associated factor. Though numerous network structuring methods have been suggested, a system to cover the entire structure has yet to be recommended. A management setup for WSNs, known as the SNOWMAN framework, is built to address these shortcomings. It uses an approach based on a policy that lets sensor nodes organize and administer themselves independently. The effectiveness of this model is scrutinized using an NS-2 simulator. The results show that the suggested model permits lesser energy consumption and a longer lifetime than currently existing methods like the LEACH and LEACH-C [4].

Business-level operators can use methods like a policy-based network management (PBNM) to inscribe SLAs in a comprehensible interface without making changes to the codes executed in the controllers. A system for policy authoring to ease the process of configuring SDN architecture is set up. The framework aims at helping business-level operators easily indicate service needs, offer flexibility and permit the said operator to accept or reject suggestions. Even if the QoS classes increase, the toolkit functions well and within the intricacies of SLAs [5].

This paper compares, in different M2M environments, the performance of every PHY mode for IEEE 802.15.4g (SUN) and IEEE 802.11 (Wi-Fi P2P) WSNs. Performances, in terms of various configurations, were examined. The outcome shows that IEEE 802.11 is more susceptible than IEEE 802.15.4g in shadowing channels, relative to the AWGN channel. Hence, the FSK in IEEE 802.15.4g is promising. Whereas, in multipath fading channels, the performance of IEEE 802.11 was superior to that of IEEE 802.15.4g. On the basis of a suitable performance,  $E_b/N_0$ , coverage of service and channel surroundings, an appropriate communication channel can be chosen [6].

CaPI (component and policy infrastructure), a dynamic component that can be reconfigured and acts as a piece of middleware for wireless sensor networks, is presented. CaPI offers two ideas for modern sensor network development and administration. One model encourages embedded developers, while the other supports managing and specifying behavioral apprehensions by managers or domain specialists. CaPI also allows successful customization and active re-configuration of the function and performance of applications. Runtime editions can be ratified effectively because of the survival of fine-grained and coarse-grained methods [7].

There are only a few protocols built for WSNs that are examined on real test beds. WSNs encounter challenges in the form of restricted power supply, little storage capacity, and connectivity problems. The virtual cord protocol (VCP) that utilizes P2P techniques is used for managing information in a WSN and offers effective resource usage. The execution and operation of the VCP as it reveals its performance in a real-time situation using Mica2 motes is discussed here, displaying how a P2P approach can be applied on WSNs for efficient data transport and management [8].

Based on the CIM policy model, a policy management for autonomic computing (PMAC) platform is developed. We provide here the PMAC platform's outline and how it can be utilized in managing network systems. The policy information model accepted by PMAC, and the model for communication between the resource and the policy manager, is presented. The key mechanisms of PMAC - for creating a policy, storing, evaluating and enforcing - are also presented, along with realistic applications of PMAC in managing networks [9].

Managing various nodes in huge numbers is always a tough task. A framework for a safe and policy-based administration of assorted resource-hampered networks of embedded systems is considered. Priority is given to the security of corresponding requests and responses. Messages are transmitted on a web-based approach and, additionally, provided a range of options for secure transmission. The methods used for this purpose depend on the application's requirements [10].

The design and execution of a management system known as SRM (sensor reliability management) for managing the reliability of data in WSNs are explained here. Though designed largely for managing the reliability of the data, it can be simply incorporated in various management services. SRM is made up of modules like user policy specification, evaluation, decision-making and action.

SRM further permits network administrators to communicate with the network by offering management policies. Results prove that SRM not only satisfies reliability requirements, but also decreases energy consumption by half [11].

It is hard to identify, access and administer a sensor node since WSNs normally work in dissimilar environments. There is, consequently, a need to surmount these challenges. The purpose here is to develop a new framework for managing policies in WSNs in a distributed fashion. The proposed work focuses on extending the functionality of WSN management by increasing policy numbers in WSN storage. It also masks the intricacies of policy management processes from users by streamlining those procedures [12]. Wireless sensor networks (WSNs) have inadequate hardware resources, thereby restricting management capabilities and causing volatility and irregularity in the system. The purpose of this work is to build a novel framework for policy-based management for wireless sensor networks (WSNs) to surmount the shortcomings of the current policy-based WSN platforms. The framework comprises major software elements like the monitor, LPDP (local policy decision point), PEP (policy enforcement point), and a set of integrated applications. The framework also consists of a 6-data warehouse [13].

WSN components have restrictions relating to power and processing. As a result, running complex algorithms has to be done using external components. Therefore, we present an amalgamation of WSNs and P2P networks to build systems relying on WSN functions. A programming concept is proposed that permits developers to focus on the operation of the developing method. Using feedback loops as a design tool, and the development of the concept's components, are also suggested. Further, they are required to be compatible, extensible and lowly-fixed [15].

One of the key components of WSNs is PBMS (policy-based management systems). Owing to hardware resource limitations, only a partial number of policies can be stored in a sensor node's local memory by policy-based management applications on WSNs, needing to be recycled if extra policies are needed. To handle this particular issue, a framework known as TinyPolicy is built using a P2P policy storage and operating system called PolicyP2P. It is developed to make use of the memory available on the network as a result of which the policy mechanism is tougher against failure of nodes and solo failure points [16].

## PROPOSED MODEL ON PARALLEL EXECUTION OF TINYPOLICY

The novel policy creation method is started via a policy-user interface on a system connected to the WSN's source peer, which is a highest-level node in the P2P hierarchy address structure. Normally, this peer is the WSN's gateway node and possesses ample power and memory.

Figure-1 represents the Proposed Model for Parallel Execution of TinyPolicy Diagram and the directions for making such a policy are also described. Once the policy is generated by utilizing the Policy GUI, the source peer employs the P2P software part to establish a host node address for the policy. It executes this by building a policy key (policy ID = event ID + sequence ID + session ID) and then messing the policy key to fit inside the sensor network's address space. The source peer's monitor will push the hashed policy key against the node with the best matching address.

After the triggering process, policies are executed in a parallel mode, having already been simplified into minuscules. Policy execution starts once the policy request is made from the required node. The request generally happens from the root node or a remote adjacent node. If it does not exist, the request is then forwarded to the next node and likewise to the root node. Two types of policy repositories are available in this system: a dependent policy repository and an independent policy repository. A dependent policy repository contains dependent policies. For example, a policy 'A' writes two lines in a particular node and policy 'B' writes an additional two lines in that node (with the first two lines of 'A' and the next two new lines of 'B,' we consider that A and B are dependable). If the requested policy is dependable, it searches the dependable repository and if it is undependable, it searches the undependable repository. The requesting time and search time, consequently, are minimized. The question then arises as to how to ascertain whether or not the requested policy is dependable or undependable.

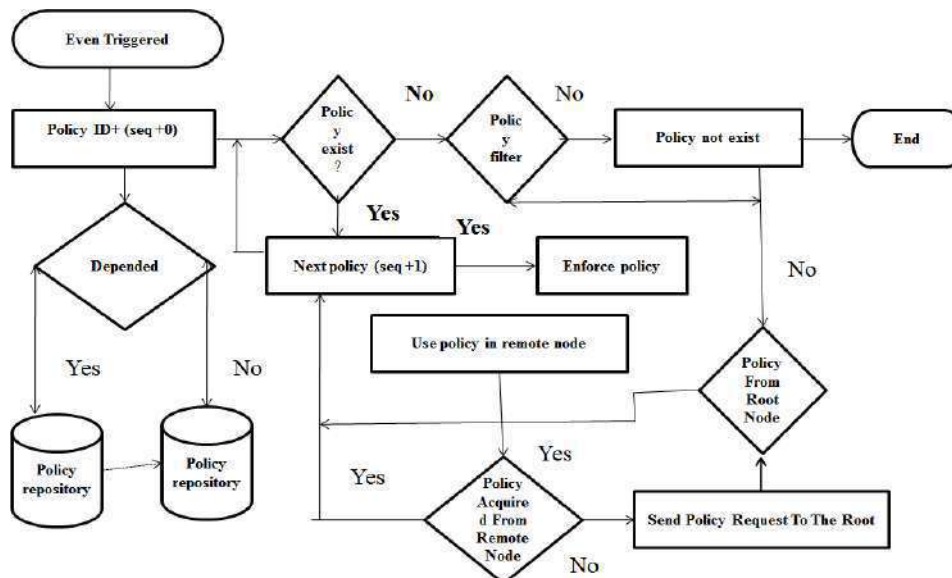


Fig:1. Proposed Model for Parallel Execution of TinyPolicy Diagram

### Policy Creation

Policy ID = Event ID + Sequence ID + Session ID

$$\begin{aligned}
 &= (28+1) \times (28+1) + 28+1 + 2 \text{ byte} \\
 &= 66049 + 257 + 2 \text{ byte} \\
 &= 2 \text{ byte} + 1 \text{ byte} + 2 \text{ byte} = 5 \text{ byte}
 \end{aligned}$$

Event Id – sub task  
 Sequence Id – 10 kb ----each 10kb send to  
 Session id: up completing one work

## RESULT AND IMPLEMENTATION

### Tiny Policy Implementation

We consider 25 nodes in NS-2 simulation. Nodes that try actions such as receiving, sending or storing data need policies. Policies have already been stored in two types of storage areas, dependable and undependable data sources. If a node needs a policy to perform an action, it requests a neighbor node. This request is forwarded to the root node via intermediate nodes. The policy filter module also works to detect whether or not the policy already exists in the intermediate node. If it exists, then it simply forwards the request to the policy hold node; otherwise, it forwards it to the root node. The policy filter also checks whether the required policy is in dependable storage or in undependable storage. The policy is now executed parallelly, with the time taken for overall execution of the policy reduced. The result obtained in this simulation is tabulated below.

### Policy Replacement

Consider a case where the root node sends a video file to the request node at time t1. At that point in time, the root node creates a policy with a session id and destination id. A node requesting the same video file after a while is considered to be at time t2. Now, since the policy is the same, simply update it with the name of the session id and destination id, the policy being stored in two repositories, dependable and undependable. Dependable means that the policy regards current action while undependable means that it does not. Old policies, consequently, are stored in undependable repositories from which a policy can be taken and simply updated with the session id and destination id. This architecture helps in reusing, with certain modifications, policies.

According to the policy structure, we can easily update a current policy from an old one. To update an old policy, we use a pattern-matching algorithm. The [Table- 1] represents parameter and size, the [Table- 2] shows policy table with action and data size.

Table: 1.Parameter and Size

Policy Structure	Data size
[ID] Policy id	5 byte
[IF] Policy condition	3 byte
[THEN]  Do Policy Action	3 byte
[END] End Policy Execution	1 byte
[NEXT] Execute Next Policy	3 byte

### Frequent and Closed Patterns

Table: 2. Policy table with action and data size

Policy	Action	Data size
P1	A1	D1
P2	A1	D2
P3	A1	D1
P4	A2	D2
Pn	An	Dn



Where  $A_1, A_2, A_3 \dots A_n$  is action,  $D_1, D_2, D_3 \dots D_n$   
 Is data size, and  $P_1, P_2, P_3 \dots P_n$  is policy.

Both  $P_1$  and  $p_3$  are equal. Notwithstanding the fact that they are equal, they do not create  $p_3$ , since  $P_3$  is an existing policy

### Frequent pattern covering set

Table: 3. Mapping, destination and frequent action

Frequent Pattern	Destination
{ A1, A2,A3}	{ d1, d2, d3.....dn }
{ A2}	{ d2, d3.....dn }
{ A2,A3}	{ d1, d3.....dn }

Wherein  $A_1, A_2, A_3 \dots A_n$  is action and  $d_1, d_2, d_3 \dots dn$   
 Is destination. The [Table- 3] shows mapping of destination and frequent pattern.

### Determining the Requested Policy Existence

FREQUENT and CLOSED PATTERNS

$TERMSET(Y) = \{t \forall dn \zeta Y \quad t \zeta dn\}$

CLOSURE of X is defined

$CLOSURE(X) = TERMSET(X_1)$

A PATTERN X; if, and only if,  $X = CLOSURE(X)$

Let X be a CLOSED PATTERN.

We can prove that

$supa(X_1) < supa(X)$

FOR all PATTERNS  $X_1 \supset X$ ; OTHERWISE, if  $supa(X_1) = supa(X)$

We have  $X_{11} = X_1$

Wherein  $supa(X_1)$  and  $supa(X)$  are the total support of PATTERN  $X_1$  and X.

Also we have

$CLOSURE(X) = TERMSET(X_1) = TERMSET(X_{11}) \supset X_1 \supset X$

That is,  $Closure(X) \neq X$ .

### Evaluation of policy in terms of number and time

Table: 4. Evaluation of the Policy in Terms of Number and Time

Number of policy	Evaluation time (traditional method)	Evaluation time (proposed method)
0	0.6	0.2
1	1.2	1.1
2	1.4	1.6
3	2.2	1.8
4	3.2	2.0
5	3.8	2.6
6	4.2	3.1
7	4.8	3.5
8	5.2	4.1
9	5.7	4.6
10	6.2	5.1

[Table- 4] represents the evaluation of the policy in terms of number and time. This graph mentions the time of execution in traditional methods and the proposed method between TinyPolicies. Figure-2 shows 5 traditional policies completed within 3.8 seconds but 5 proposed policies completed within 2.6 seconds; and 10 traditional policies completed within 6.2 seconds, with 10 proposed tiny policies completed within 5.1 seconds. An overall indication from the graph is that parallel execution increases the number of policies within a given time, also helping increase the overall performance of the system.

### Evaluation of the policy in terms of size and implementation time

[Table- 5] represents the evaluation of the policy in terms of its size and implementation time. Figure-3 mentions the time of execution between TinyPolicies in traditional methods and the proposed method. It shows that 200-byte sized traditional policies completed within 0.2 seconds, but 300 bytes of policies completed within 2.6 seconds, in both cases. 1000 traditional policies completed within 2.0 seconds and, similarly, 10 proposed TinyPolicies completed within 5.1 seconds. In 5.2 seconds, we discover that both cases have nearly 1000 different policies executed between them. An overall indication from the graph is that parallel execution increases the number of policies within a given time, also helping increase the overall performance of the system.

Table: 5.Evaluation of the Policy In Terms Of Size and Execution Time

Policy size (byte)	Evaluation time (traditional method)	Evaluation time (proposed method)
200	0.6	0.2
300	1.2	1.1
500	1.4	1.6
700	2.2	1.8
1000	3.2	2.0
1200	3.8	2.6
3000	4.2	3.1
3200	4.8	3.5
3600	5.2	4.1
4000	5.7	4.6
5000	6.2	5.1

### CONCLUSION

An extension of the services of P2P networks for wireless sensor networks (WSNs) is evaluated in this paper. The downside of WSNs is that their components are restricted in terms of power and processing, due to which policy-based WSN administration applications can only store a particular amount of policies in the limited memory of sensor nodes. Challenging concerns in relation to govern and control the embedded devices can be resolved by WSNs with the help of the parallel execution of TinyPolicy. P2P networks are fairly flexible in designing applications and can address most of the problematic issues that confront WSNs. Integration could be the key challenge, between these architectural structures, for cooperating in a specific functionality. The new structure uses P2P architecture, distributed policy-based management and a replica of policies for dynamically distributing policies between sensor nodes, a feature supported by this framework. The location of policies can also be managed dynamically. The framework is simulated using the NS-2 simulator. For simulation, 25 nodes are considered. A node planning to perform actions like receiving, sending or storing data requires policies, usually stored in two types of storage: dependable and undependable. Nodes request a neighbor node for a policy to execute an action. This request is then forwarded to the root node through intermediate nodes. The policy filter module detects whether or not the policy already exists in the intermediate node. It also examines if the policy is in dependable or undependable storage. In future, a combination of WSNs and P2P computing may result in the development of robust applications. From a future perspective, we intend to move towards parallel steps to find the existing policy and a new design, with whichever being completed first taking effect. The algorithm will be designed with the same perspective as well.

## ACKNOWLEDGEMENT

We would like to express my deepest gratitude to Saveetha School of Engineering; the Management of Saveetha University; the Principal and Head of the Department of Computer Science & Engineering.

## CONFLICT OF INTEREST

No conflict of interest

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Shen, Chong, Wencai Du, Robert Atkinson, and Kae Hsiang Kwong. [2012] Policy based mobility & flow management for IPv6 heterogeneous wireless networks. *Wireless Personal Communications* 62( 2): 329-361. [9]
- [2] Matthys, Nelson, Christophe Huygens, Danny Hughes, J  Ueyama, Sam Michiels, and Wouter Joosen.[ 2011] Policy-driven tailoring of sensor networks. In *Sensor Systems and Software*, pp. 20-35. Springer Berlin Heidelberg, [10]
- [3] Bourdenas, Themistoklis, and Morris Sloman.[2010] Starfish: policy driven self-management in wireless sensor networks. In Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems, pp. 75-83. ACM, [11]
- [4] Si-Ho, CHA, LEE Jong-Eon, JO Minh, Hee Yong Youn, KANG Seokjoong, and CHO Kuk-Hyun. [2007]Policy-based management for self-managing wireless sensor networks. *IEICE transactions on communications* 90(11) :3024-3033. [12]
- [5] Machado, Cristian Cleder, Juliano Araujo Wickboldt, Lisandro Zambenedetti Granville, and Alberto Schaeffer-Filho.[ 2015] Policy authoring for software-defined networking management. In Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on, *IEEE* pp. 216-224. [13]
- [6] Oh, Eui-Suk, Seong-Hee Lee, Seong-Hyeong Lee, and Seung-Hoon Hwang.[2015] Comparison of SUN and Wi-Fi P2P WSN in M2M Environments. *International Journal of Distributed Sensor Networks* 501 (2015): 791849. [14]
- [7] Matthys, Nelson, Christophe Huygens, Danny Hughes, Sam Michiels, and Wouter Joosen.A component and policy-based approach for efficient sensor network reconfiguration. In Policies for Distributed Systems and Networks (POLICY), 2012 IEEE International Symposium on, pp. 53-60. *IEEE*, 2012. [15]
- [8] Sahar, Syeda Nida, Faisal Karim Shaikh, and Sana Hoor Jokhio.[2011] P2P based Data Management WSNs: Experiences and Lessons Learnt from a Real-world Deployment. *Mehran University Research Journal of Engineering & Technology* 30( 4) [16]
- Agrawal, Dakshi, Kang-Won Lee, and Jorge Lobo. [2005]Policy-based management of networked computing systems. *Communications Magazine, IEEE* 43(10) :69-75.
- Rantos, Konstantinos, Alexandros Papanikolaou, Konstantinos Fysarakis, and Charalampos Manifavas.[2012] Secure policy-based management solutions in heterogeneous embedded systems networks. In Telecommunications and Multimedia (TEMU), 2012 International Conference on, pp. 227-232. *IEEE*.
- Le, Tuan D, Wen Hu, Sanjay Jha, and Peter Corke.[2008] Design and implementation of a policy-based management system for data reliability in wireless sensor networks. In Local Computer Networks, 2008. LCN 2008. 33rd IEEE Conference on, pp. 762-769. *IEEE*.
- Qwasmı, Nidal, and Ramiro Liscano. [2012]Framework for Distributed Policy-Based Management in Wireless Sensor Networks to Support Autonomic Behavior. *Procedia Computer Science* 10 : 232-239.
- Qwasmı, Nidal, and Ramiro Liscano. "Distributed Policy-Based Management for Wireless Sensor Networks. *Procedia Computer Science* 10 (2012): 1208-1212.
- C. Rajan, N. Shanthi, "Swarm optimized multicasting for wireless network", *Life Sci. J*, Vol.4, No. 10, 2013.
- Gutierrez, German, Boris Mejias, Peter Van Roy, Diana Velasco, and Juan Torres. [2008] WSN and P2P: a self-managing marriage.In Self-Adaptive and Self-Organizing Systems Workshops, 2008. SASOW 2008. Second IEEE International Conference on, 198-201. *IEEE*
- Qwasmı, Nidal, and Ramiro Liscano.[ 2015]] TinyPolicy: A distributed policy based management framework for Wireless Sensor Networks. In Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on, pp. 918-921. *IEEE*,

## ABOUT AUTHORS



**M. NARAYANAN** received a B.E. in Computer Science & Engineering from Anna University, Chennai, in 2006, and an M.E. (with distinction) in Computer Science & Engineering from Sathyabama University, Chennai, in April 2010. He has also been pursuing a Ph.D. in Computer Science & Engineering at Sathyabama University since July 2010. Presently he is Assistant Professor, Saveetha School of Engineering, Saveetha University, Chennai, and Tamil Nadu. He has 9 years' teaching experience, with research interests in distributed systems, peer-to-peer networks, and video-on-demand service.



**Dr. C. ARUN** received a B.E in Electronics & Communication Engineering from Anna University, Chennai, as well as an M.E and Ph.D, in Image Processing from Anna University, Chennai, Tamil Nadu, and India. He is, currently, Professor in the Department of Electronics & Communication Engineering, RMK College of Engineering & Technology, Chennai, Tamil Nadu, and India. With 15 years' experience in teaching engineering in several colleges, his research interests lie in image processing, cloud computing, distributed systems, and peer-to-peer networks.

# IMPLEMENTATION OF IMPROVED APRIORI ALGORITHM IN INTERNAL

Vijay Kumar and Veni Devi Gopal\*

Department of Information Technology, KCG College of Technology, Chennai, INDIA

## ABSTRACT

The organizations nowadays mostly use user IDs and passwords as the entry gateway to authenticate the users. But somehow they share or give their credentials to their coworkers for their work and request them to aid co-tasks, thereby losing their security of their credentials and making it as the one of the entry points for an attack. Intruders are basically of two categories. First thing is the external intruder. They are those who are the unauthorized users of the machines they attack and the next are the internal intruders who are those, who have the permission to access and work on the system, but will not have access to some portions of the system and they are hard to find and detect because most of the intrusion detection systems and firewalls can identify and isolate malicious behaviors from the outside the system only. Therefore a security system called the Internal Intrusion Detection and the Protection System referred to as (IIDPS) is made to find and detect the insider who attacked the system and also to keep track of user's habit and finally to determine whether a valid user or not by comparing the current behaviors with the patterns collected and stored in the server before.

Published on: 08<sup>th</sup>- August-2016

### KEY WORDS

Data mining, malicious behavior, user habits, intruder, security.

\*Corresponding author: Email: [kvijay88859@gmail.com](mailto:kvijay88859@gmail.com), [venidevig@gmail.com](mailto:venidevig@gmail.com); Tel.: +91 9500088859

## INTRODUCTION

Data Mining [1-2] is basically a concept of extracting or mining away the knowledge or the information needed from the large block-sets of data. Data mining is one of the way for identifying or to find the the intensive knowledge and information needed for use from those large amounts of data that is stored either in the databases, data warehouses, or the other repositories. Its trust and capacity is to discover valuable data efficiently, non-frequent information from those large databases. But it is certainly sometimes vulnerable to the wrong use of it. So, there might be some confusion among data mining and privacy. Basically, the mining process [3] will be done on this for effectively analyzing and obtaining the results. Privacy [4] basically deals with the extraction of sensitive information with the help of data mining technique. There is an alarming concern about the privacy of the individual users. Each individual has to control their information on their own. The general issues are the usage of other person's credentials or their information. Intrusion detection [5] is a research area where mining algorithms are used and analyzed for the effective detection of and protection for the individual's privacy concern and a brief can be found about it on the guide to IIDPS [6]. The Apriori Algorithm is an effective and suitable algorithm for mining the frequent item-sets for boolean association rules. There are two key concepts which are as follows. 1) The items must have minimum support. 2) The subset contained in the frequent item-set should be frequent.

Some of the possible threats to security are listed below.

*Risk* - Due to the malfunction of the hardware or due to the incomplete design or due to the incorrect software design because of which there may be an exposure of the information or the data or there may be some violation.

*Vulnerability* - Some known or suspected flaw because of which the hardware or software that is being used or the operation of a system that exposes the system to an accidental disclosure.

*Attack* - After the execution of a plan., a threat is carried out as a result

*Penetration* - A successful attack or a successful intruder one who can obtain the unauthorized access as a result to those files and programs to control the state of a computer system.

Intrusion detection systems (IDS) are basically focused on -

- Identifying possible activities of the user and the system.
- Gaining information about them.
- Analyzing the vulnerability.
- Assessing the file contents and the integrity of the system.
- Recognizing malicious activities and patterns that are deviating.
- Giving an alert to the administrator.

In addition, organizations use it generally for things like -

- Identifying issues and the problems associated with those identified issues.
- Analyzing the threats.
- Following the security policies.

## RELATED WORK

Computer forensics [7] science views computer systems in order to identify the pattern, preserve, recover, analyze the results and present important facts and opinions on information collected. It analyzes about the attackers and their behaviors like spreading some computer viruses, some malwares and some malicious codes. The intrusion detection techniques helps on how to detect the malicious network behaviours [8], [9] It refers clearly that, from long searched and generated log files or patterns or the data-sets, these traces or patterns of misuse can be more accurately identified and reproduced when an unknown or a malicious user logs in to a system with the valid user credentials.

Mining of frequent item-sets is an important part in the association mining to search the frequent items from the list of item-set in the database. It is important to find the interesting patterns from the large datasets, such as the association, from the episodes, from the classifier, from the clustering and from the correlation etc. [10]. Apriori [11],[12] is like a subset of the candidate generation approach. It basically generates the candidate item-sets of length (k+1). It is based on the frequent item-sets of length (k). The item-set frequency can also be done by counting their occurrences during the transactions. Then at last in 2000, the Pattern growth was proposed by Han where he did some useful work regarding the FP tree.

## RELATED DETAILS

This IIDPS basically uses two techniques to find the malicious behavior of the user who uses or accesses the account. The first one is basically like the physical entities. So many physical entities will be evaluated to find out about the malicious behavior. The physical entities such as finding where the user is accessing it and what system has been used frequently that would have been recorded before. So with that recorded information the behavior will be evaluated to find the data of who the intruder is.

The few characteristics of the physical entities with which the intruder can be traced out are –

- By finding the physical location of the intruder.
- By identifying the ip address of the system.

Next method is to find out the malicious behavior with the help of user's habits. The user who is accessing the account will have certain characteristics and these things are compared with those that are recorded in the server. From the result after comparing with the characteristics that are in the server, the malicious behavior can be obtained and the same can be reported to the admin as an alert or by mail.

The few characteristics of the user habitual entities with which the intruder can be traced out are

- With the users preferences
- A simple calculation.
- OTP.

So by using few combinations of the above, Security points are made, and the originality of the user can be found. If there is some unusual behavior found after the comparison, then the admin will be intimated from this. So, the admin has the right to stop the access after identifying the malicious behavior. In case if the same is done across in an organization then the admin can ask the user whether to grant or stop the access. In the other case if it's a college the admin can directly stop the access after knowing that malicious behavior of the user who tries to access the system.

## THE IMPROVED APRIORI ALGORITHM

The following is an algorithm for finding the frequent item set using the improved apriori algorithm.[13]

**Input:** transaction database D; min-sup.

**Output:** the set of Frequent L in the database D.

- (1) Find min-sup-count from D.
- (2) Generate L1-candidates.
- (3) Identify L1.
- (4) for (k=2; L<sub>k-1</sub>, k--)
- (5) {for each k-itemset (p, (x<sub>p</sub>) L<sub>k-1</sub> do
- (6) for each k-itemset (q, (x<sub>q</sub>) L<sub>k-1</sub> do
- (7) if (x<sub>p</sub>[1]=x<sub>q</sub>[1])!(x<sub>p</sub>[2]=x<sub>q</sub>[2])!...!(x<sub>p</sub>[k-2]=x<sub>q</sub>[k-2]) then
- (8) {L<sub>k</sub>-candidates.X<sub>k</sub>= X<sub>p</sub>\* X<sub>q</sub>;
- (9) L<sub>k</sub>-candidates.TID-set(X<sub>k</sub>)
- (10) for each k-itemset<X<sub>k</sub>,TID(X<sub>k</sub>)> L<sub>k</sub>-candidates do
- (11) Find sup-count
- (12) Identify L<sub>k</sub>
- (13) set-count=L<sub>k</sub>.item-count
- (14) return L="kL<sub>k</sub>; [14]

Transactions in a database D=10

Table: 1. D of 10 Transactions

TID	Items
T1	It1,It2,tl4
T2	It2,It5
T3	It2,It3
T4	It1,It2,It5
T5	It1,It2
T6	It2,It3
T7	It1,It3
T8	It1,It2,It3,It4
T9	It1,It2,It3
T10	It1,It4

- 1) Scan D for count of each candidate.

Table: 2. Generation of C1

Item Set	Support Count
It1	7
It2	8
It3	4
It4	3
It5	2

- 2) Support count of the candidate-set is then compared with that of the minimum support. Suppose that the minimum transaction support count required is 2.

Table: 3. Generation of L1

Item Set	Support Count
It1	7
It2	8
It3	4
It4	3
It5	2

- 3)
- 4) Generate C2 candidates from L1 and scan D for count of each candidate.

Table :4. Generation of C2

Item Set	Support Count
It1,It2	5
It1,It3	3
It1,It4	3
It1,It5	1
It2,It3	4
It2,It4	2
It2,It5	1
It3,It4	1
It3,It5	0
It4,It5	0

5) Compare candidate support count with minimum support. L2 is determined. Then D2 was determined from L2.

**Table: 5. Generation of L2 Support Count**

Item Set	Support Count
It1,It2	5
It1,It3	3
It1,It4	3
It2,It3	4
It2,It4	2

**Table: 6. D2(Updated Table)**

TID	Items
T1	It1,It2,It4
T4	It1,It2,It5
T8	It1,It2,It3,It4
T9	It1,It2,It3

6) Generate C3 candidates from L2 and scan D2 for count of each candidate..

**Table: 7. Generation of C3**

Items	Support Count
It1,It2,It4	2
It1,It2,It5	1
It1,It2,It3,It4	1
It1,It2,It3	2

**Table: 8. C3 (Updated Table)**

Items	Support Count
It1,It2,It4	2
It1,It2,It3	2

6) Compare candidate support count with that of the minimum-sup. The D2 is scanned in order to determine L3.

**Table: 9. Generation of L3**

Items	Support Count
It1,It2,It4	2
It1,It2,It3	2

7) The algorithm uses L3 to generate a candidate set of 4-itemsets, C4.

**Table: 10. Generation of D3**

TID	Items
T8	It1,It2,It3,It4



## PROPOSED WORK

In this section, we have introduced the IIDPS system and the components of the IIDPS which are described in detail. An algorithm was also presented for generating a user habit file and detecting an internal intruder. The IIDPS system as in **Figure-1**, consists of a mining server, a detection server and three repositories such as user log file repository, user profile repository and an attacker profile repository. So when a user/attacker tries to access the system, the IIDPS will check if it is a valid user by checking it with the admin.

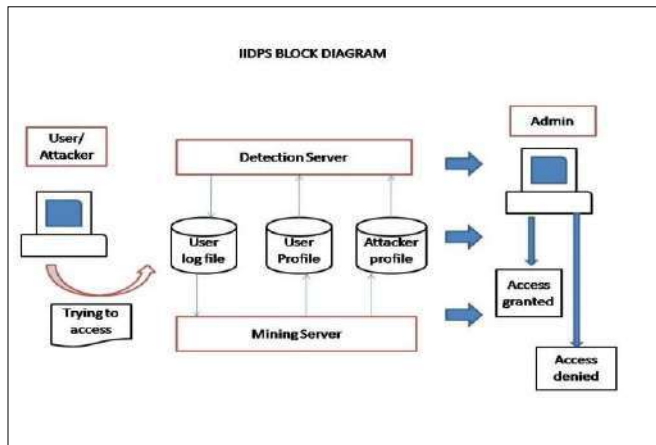


Fig. 1. IIDPS System

The admin will validate and has the power to either grant the access to the user or can revoke the access from the user. By this way, the malicious behaviour if present can be found out with the help of the system.

**Support:** It is defined as rate of occurrence of an item-set in a transaction database.

$$S(I1 \rightarrow I2) = \frac{Tn(I1,I2)}{Tt}$$

Where S is Support, Tn denotes number of transactions containing both item1 (I1) and item2(I2) . Tt denotes total number of transactions.

**Confidence:** It defines the ratio of data item-sets that contains Y in the items-sets of X in all the transactions.

$$C(I1 \rightarrow I2) = \frac{Tn(I1,I2)}{T(I1)}$$

Where C is Confidence , Tn denotes number of transactions containing both item1(I1) and item2(I2). T(I1) denotes number of transactions containing item1 (I1).

## IMPLEMENTATION

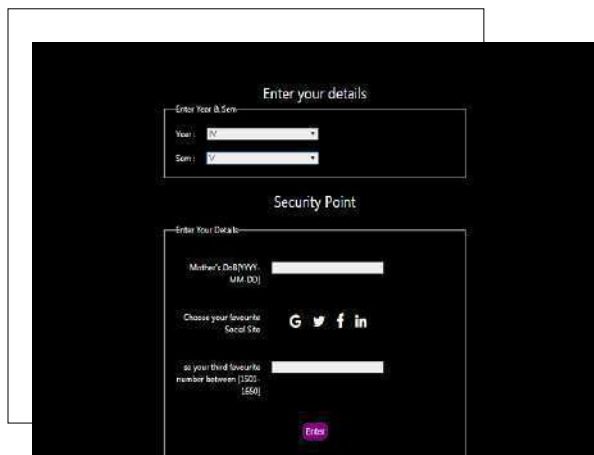


Fig. 2. Security point check



Fig. 3. Frequent item-set

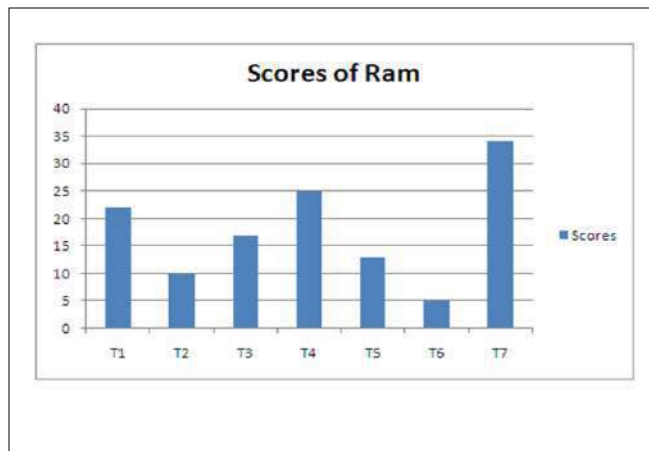


Fig.4. Scores of Ram

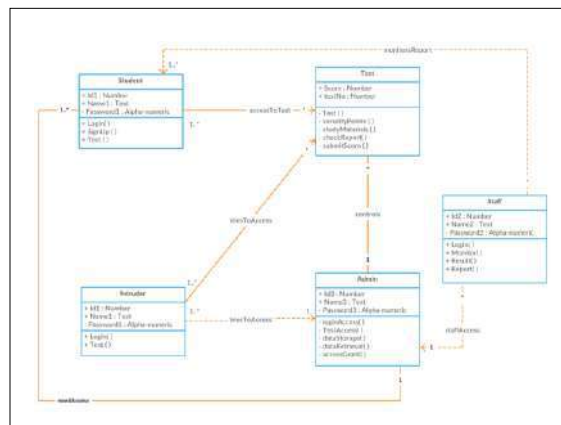
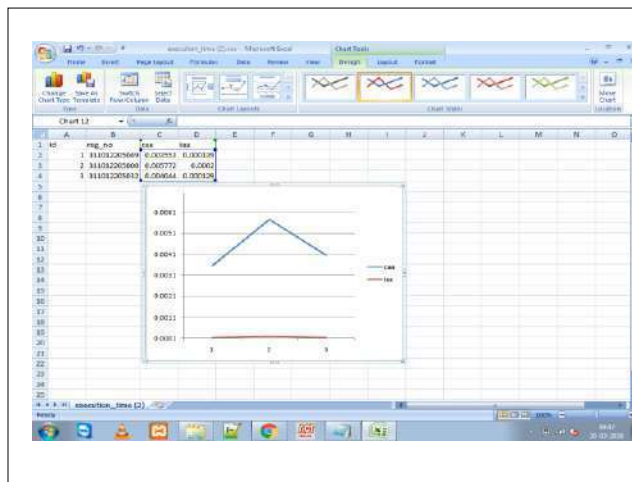


Fig.6. Representation of the overall functions of the system using IIDPS and IAA with the help of the class diagram

The IIDPS system is implemented for an aptitude test system for finding whether the real user of the system is taking up the test. After that, their performance is calculated. The **Figure- 2**, shows the security point which is used to check whether the user is real or not. The frequent item-set is generated after this which is shown in the **Figure-3**, i.e. frequent item-set generation, which is used to monitor the students overall materials that was used by the students.

After that, the staff can login and monitor individual students performance as in **Figure -4** or can monitor the overall performance. A Report can be generated based on the staffs need either semester wise or year wise, to monitor the students performance.

The results of the Classical Apriori Algorithm and the Improved Apriori Algorithm were noted based on its execution time. A comparison was made between those two algorithms and the results were analyzed and a graph was generated as in **Figure- 5**, which clearly indicates that the IAA is more efficient than the CAA.



**Fig:5. Comparative Results between Classical Apriori Algorithm and Improved Apriori Algorithm in terms of Execution time**

The overall functions of the system with the help of IIDPS and IAA are represented with the help of the class diagram as in the **Figure -6**.

## CONCLUSION AND FUTURE WORK

In this paper, we have proposed a system which is called the Internal Intrusion Detection and Protection System (IIDPS). The user's preferences are recorded and compared every time when he logs in to the account. So based on the usage profile and the patterns the IIDPS restricts the intruder. In today's world, internal intrusion detection is one of the major topics in which the research is still going on for its effective development. It can be extended to protect the system even at a higher level by using one of the following two ways. First one is that the IIDPS system can be still improvised by introducing new concepts and thereby protecting it from intrusion efficiently. Second one is by extending the system by using recent technologies such as finger print technology or a face reader technology to easily identify the user, thereby avoiding the malicious activity and to improve the IIDPS's reliability and performance. There are so many researches taking place in this field and a suitable one can be included to effectively develop the IDP system to protect the system.

## ACKNOWLEDGEMENT

None

## CONFLICT OF INTEREST

No conflict of interest

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Fang-Yie Leu, Kun-Lin Tsai, Member, IEEE, Yi-Ting Hsiao, and Chao-Tung Yang "An Internal Intrusion Detection and Protection System by Using Data Mining and Forensic Techniques, *IEEE* 2015.
- [2] J Han and Kamber.[2006] Data Mining: Concepts and Techniques, 2nd ed., The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.
- [3] Pingshui WANG.[2010] Survey on Privacy Preserving Data Mining, *International Journal of Digital Content Technology and Its Applications* 4( 9)
- [4] MB Malik, MA Ghazi and R Ali.[ 2012] Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects, in Proceedings of Third International Conference on Computer and Communication Technology ,*IEEE*.
- [5] Sheetal Thakare<sup>1</sup> ,Pankaj Ingle<sup>2</sup>, BB Meshram.[2012] Intrusion Detection System the Survey of Information Security, *International Journal of Emerging Technology and Advanced Engineering* , 4 (8)
- [6] Guide to Intrusion Detection and Prevention System<sup>1</sup>, NIST, Technology Administration US Department of Commerce.
- [7] ZB Hu, J Su, and VP Shirochin.[ 2007] An Intelligent Lightweight Intrusion Detection System With Forensics Technique, in Proc. IEEE Workshop Intell. *Data Acquisition Adv Comput Syst Technol Appl Dortmund, Germany*, 647–651.
- [8] Q Wang, L Vu, K Nahrstedt, and H Khurana.[ 2010] MIS: Malicious Nodes Identification Scheme in Network-Coding-Based Peer-to-Peer Streaming, in Proc. IEEE INFOCOM, San Diego, CA, USA, , pp. 1–5.
- [9] ZA Baig, "Pattern Recognition for Detecting Distributed Node Exhaustion Attacks in Wireless Sensor Networks, *Compute. Commun.*, Vol. 34, No. 3,pp. 468–484, Mar. 2011.
- [10] Rao S, Gupta R, Implementing Improved Algorithm Over Apriori Data Mining Association Rule Algorithm, *International Journal of Computer Science And Technology*, pp. 489-493, Mar. 2012.
- [11] Xiang Fang.[2013] An Improved Apriori Algorithm on the Frequent Item set, International Conference on Education Technology and Information System (ICETIS 2013)
- [12] SurajP. Patil<sup>1</sup>, U. M. Patil<sup>2</sup> and Sonali Borse.[2012]The Novel Approach for Improving Apriori Algorithm for Mining Association Rule, Proceedings of "ational Conference on Emerging Trends in Computer Technology (NCETCT-2012)"Held at R.C.Patel Institute of Technology, Shirpur, Dist. Dhule, Maharashtra, India. April 21, 2012.
- [13] Akshita Bhandari Ashutosh Gupta Debasis Das.[2014] Improvised Apriori Algorithm using Frequent Pattern Tree for Real Time Applications, *ICICT* 2014.
- [14] Sakshi Aggarwal<sup>1</sup>, Ritu Sindhu.[2015] An Approach of Improvisation in Efficiency of Apriori Algorithm, *Peer Jpreprints*

# AN INTEGRATED APPROACH FOR SUPERVISED LEARNING OF ONLINE USER REVIEWS USING OPINION MINING

Shobana<sup>1</sup> and Anny Leema<sup>2\*</sup>

<sup>1</sup>Dept. of Computer Sc. and Engineering, B.S.Abdur Rahman University, vandalur, Chennai-600100, Tamilnadu, INDIA

<sup>2</sup>Department of MCA, B.S.Abdur Rahman University, vandalur, Chennai-600100, Tamilnadu, INDIA

## ABSTRACT

*Abstract - Internet is a place where people stores and shares information among the other entities reside in the network. In recent years, the amazing development of web technologies, lead to huge quantity of system and user generated information in online systems. This huge amount of information on web platforms enables them to use as data sources, in review making applications based on opinion mining and sentiment analysis. The paper put forward an algorithm for detecting sentiments on user reviews on movie, based on Naive Bayes classifier. We considered the opinion mining domain for the analysis of user reviews and the techniques used in sentimental analysis. We implemented the proposed algorithm to the user reviews and tested its performance, and suggested directions of development.*

Published on: 08<sup>th</sup>- August-2016

### KEY WORDS

*Opinion Mining; Web Content; Mining; Sentiment Analysis; Naive Bayes*

\*Corresponding author: Email: [annyleema@gmail.com](mailto:annyleema@gmail.com); Tel: +91 99 44238480

## INTRODUCTION

The development of web technology and its information sharing among the web evolved as exponential increase in amount of information in online system. The data stored in online consists of various levels and types. Sometimes data analysts and researchers also feel tricky to vary the data that are residing on web. Because of huge volume of information it becomes difficult to process the information by individuals which leads to information overload and affects decision making processes in organizations.

Hence it is important to incorporate a new data analyzing and processing techniques for creation of knowledge. Knowledge discovery and feature extraction from data warehouse is a primary task of organizations in order to develop their products and improve their business strategies. Most of the data stored in online is in the form of text. Storing and transferring of text based data is more convenient and also easier to read by the people. In this context, applying data mining techniques is more reliable to handle data.

Difference between web mining and data mining are important in terms of data collection. In data mining, it is assumed that the data is already collected and stored in databases. In case of web mining, it uses special mechanisms, such as information retrieval - *IR (Information Retrieval)* and information extraction - *IE (Information Extraction)*, to obtain data and to pre-process them to apply data mining techniques. The web mining mechanism is divided into three categories based on its applications

*Web structure mining* - Discovering knowledge from hyperlinks to maximize relative information about the relations between web pages.

*Web usage mining* - Extracting patterns of users and models, from web logs, which is the repository of data access and activities of each visitor to a website.

*Web content mining* - Extracting knowledge and information from web page content.

Text mining technique is mainly applied to discover knowledge from the unstructured text. The outcomes of the text mining on unstructured text are used in research areas like Artificial Intelligence, Natural Language Processing (NLP) and Machine learning. In addition this technique is also applied for content spam detection, document classification and trend analysis.

Users often using web sites will always wishes to post their opinion, ratings and reviews of products. Most of this information is in unstructured format. Hence it is important to impart the knowledge discovering techniques for detection and extraction of opinions or sentiments from textual information.

Analyzing of customer sentiment and their opinion on a newly launched product, based on feedback from web pages is vital for evaluation of impact and making decision on business development. *Opinion mining* is a kind of knowledge discovery deals with habitual methods of detection ,extraction of opinions and classifying the sentiments presented in a text which are given by the users. Application of opinion mining to the raw text data result in creation of effective referral systems, trend analysis, financial analysis, strategic management, market research and product development.

## TECHNIQUES OF OPINIONS MINING

Opinion mining and its uses created a dramatic change in e-commerce. The opinion and review made by the visitors and guests on products enables the organizations to improve their marketing strategy. It has increased big e-commerce sites and recommendations of products and services sites. The large number of reviews on a product promotes easy access to useful and reasonable information to visitors. It can be used to compare offers from different competitors on the market of similar product and make an informed decision about buying a certain offer. It is very difficult for a visitor to read all of lengthy reviews and to form an opinion on a product because:

In some cases the reviews already posted by other existing customers can be very long and only a few sentences may express opinions. Read through of only part of the review may create a false impression about the topic or may give inaccurate result;

The user is not aware about the various metrics used in comparing offers in a certain specialized field. Also, the large number of lengthy reviews makes it difficult for producers and analysis to follow the real expectations of customers. In addition with the lengthy reviews, they face difficulties in follow wide range of products, traded on a variety of web sites. So, it is useful to make a system to detect indicators of performance of a product, and domain specific metrics, to recapitulate the opinions obtained from the large amount of reviews, in several positive and negative aspects.

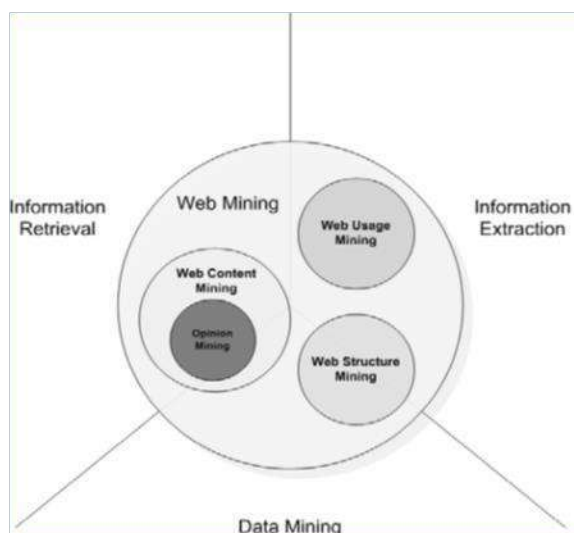


Fig: 1. Data mining

Classifying entire reviews posted by the visitors according to the opinions towards certain product or on object is a sentiment classification. To produce a feature summary, product features are first identified, and positive and negative opinions on them are aggregated. Features are such as components, product attributes like price, availability, warranty, manufacturer, etc.. The effective summarizing of opinions and, grouping of feature expressions is critical. It is very time consuming and tedious for business analysts and organizations to group typically hundreds of feature expressions and reviews that can be identified from the text for an opinion mining application into feature categories. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as the classic text summarization. [2]

For some analyzed entity, we determine the words or phrase that state user opinion. The process is comprised of three stages [10]:

*Entity determination* – identification of texts i.e., user reviews that contain information about the entity/attributes;

*Determination of sentiments* – for the text of the previous stage is considered the content of opinions and sentiments, by searching a set of words carrying sentiments, or by prior training a classifier;

*Determination of entity - sentiment relationship* - at this stage opinions are analyzed and addressed to the entity under review. Usually this is done through a predefined list of patterns.

## SENTIMENT CLASSIFICATION

Polarity classification or Sentiment classification is the binary classification task of expressing either an overall positive or an overall negative opinion by labeling an opinionated document. A typical approach for sentiment classification is to use machine learning algorithms.

### Machine Learning

A system which is capable of automatically acquiring and integrating the knowledge is referred as machine learning. The systems that learn from analytical observation, experience, training, and other means, results in a system with self-improvement, effectiveness and efficiency. Machine learning systems uses knowledge and a corresponding knowledge organization to test the knowledge acquired, interpret and analyze. The machine learning system can be classified into three methods based on the labeling the data. They are supervised learning and unsupervised learning.

**Supervised learning:** Supervised learning produces a function which maps inputs values to desired outputs is called as labels. These labels are assigned by the human experts. Since it is a text classification problem, any supervised learning method can be applied, e.g., Naive Bayes classification, and support vector machines (SVM).

**Unsupervised learning:** Unsupervised learning models a set of inputs, like clustering, and labels are not known during training. Classification of data is performed using some fixed syntactic patterns which are used to express opinions. These mechanisms also have to be done by the human experts. The part-of-speech (POS) tags are used to compose syntactic patterns.

**Semi-supervised learning:** Semi-supervised learning generates an appropriate function or classifier in which both labeled and unlabelled examples are combined. [2, 5]

### Sentiment Analysis Tasks

Sentiment analysis includes a task of classifying the polarity of a given text at the document, sentence or feature level expressing of opinion as positive, negative or neutral. The sentiment analysis can be performed at one of the three levels: the document level, sentence level, feature level.

*Document Level Sentiment Classification:* In document level sentiment analysis, the primary task is to extract informative text for deducing sentiment of the whole document. Because the objective statements are rendered by

subjective statements and complicate further for document categorization task with conflicting sentiment, the learning method creates uncertainty. [6]

**Sentence Level Sentiment Classification:** The sentence level sentiment classification is a fine-grained level than document level sentiment classification. In sentence level sentiment classification polarity of the sentence can be given by three categories as positive, negative and neutral. The challenge faced by sentence level sentiment classification is the identification features indicating whether sentences are on-topic which is kind of co-reference problem [6]

**Feature Level Sentiment Classification:** Product features are defined as product attributes or components. Analysis of such features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach positive or negative opinion is identified from the already extracted features. It is a fine grained analysis model among all other models [2]

## PROPOSED OPINION MINING METHOD

In this case for training, the user comments are collected and extracted from at <http://www.cs.cornell.edu/people/pabo/movie-review-data/> [12]. This collection contains 5331 sentences already classified as positive and 5331 negative opinions from 2000 comments processed and classified in two categories. Comments usually contain several sentences, but opinion will be determined at sentence level, then later determining overall comment opinion. Obtained collection consists of two files, one for each set of positive and negative opinions, containing one sentence per line, making it easy to process. To extract opinions we will use a Naive Bayesian classifier. This type of classifier has the advantage that it is easy to implement, quickly and generate good results.

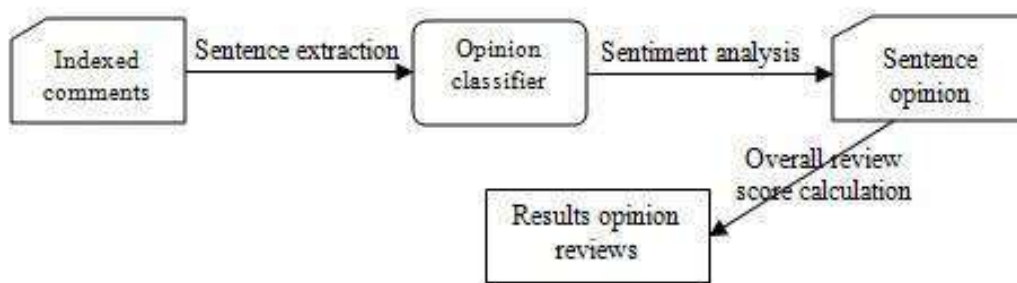


Fig 2. The overall Opinion mining process

### Using Naive Bayes algorithm

The Naive Bayes classifier is a probability classifier which is based on Bayes' theorem. It shows the relation between probability of two events A and B, P (A) and P (B) and conditional probability of event A conditioned by B and event B conditioned by A, P (A | B) and P (B | A). It is given as [13]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes' theorem, we can estimate the probability of an event based on the examples of its occurrence. In this case, we estimate probability that a document is positive or negative, in a certain context, or the likelihood that an event to take place if it was predetermined to be positive or negative. This is facilitated by the collection of positive and negative examples chosen. The process is naive Bayesian because of how we calculate the probability of occurrence of an event - is the product of probability of occurrence of each word in the document. This presumes that there is no connection between the words. This assumption of independence is introduced to facilitate the construction of classifier, it is not entirely true, and there are words that appear together more frequently than individual.



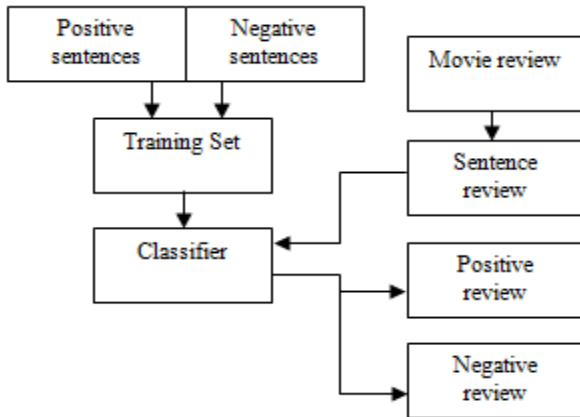
We estimate the probability of a word with positive or negative meaning by analyzing a series of positive and negative examples and calculating the frequency of each of the classes. This learning process is supervised, requiring the existence of pre-classification examples for training. Starting from:

$$P(\text{sentiment}|\text{sentence}) = \frac{P(\text{sentiment})P(\text{sentence}|\text{sentiment})}{P(\text{sentence})}$$

we assume that  $P(\text{sentence}|\text{sentiment})$  is the product of  $P(\text{word}|\text{sentiment})$  for all words in a sentence. We estimate  $P(\text{word}|\text{sentiment})$  as:

$$P(\text{word}|\text{sentiment}) = \frac{\text{no. of word occurrence in class} + 1}{\text{no. of words belonging to a class} + \text{total no. of words}}$$

The steps in the classification method proposed in the paper are presented in **Figure- 3**, below



**Fig: 3. Stages of classification process**

The proposed algorithm has following steps:

```

Initialize P (pos) <- nr_P (pos) / nr_total_probability
Initialize P (neg) <- nr_P (neg) / nr_total_probability
Tokenize sentence in words
For each class of {pos, neg}:
For each word in {phrase}
    P (word | class) <- nr_apartii (word | class) / nr_cuv (class) + nr_total_cuvinte
P (class) <- P (class) * P (word | class) Returns max {P(pos), P(neg)}
  
```

**EVALUATE THE PERFORMANCE OF ALGORITHM**

We use two specific measures for information retrieval systems to evaluate the results of algorithm called precision and the recall. The relation between the precision and recall with respect to positive and negative is given as:

**Table 1. Contingency table of correctly classified reviews**

	Relevant	Irrelevant
Detected opinions	True Positive(TP)	False Positive(FP)
Undetected opinions	False Negative(FN)	True Negative(TN)

*Precision:* Precision is the ratio of the correctly classified extracted opinions and all extracted opinions, the percentage of correctly classified opinions from classified ones:

$$\text{Precision} = \frac{TP}{TP+FP}$$

*Recall:* Recall expresses the ratio of correctly classified extracted opinions and classified opinions in data source, the percent of correctly classified opinions from all opinions in a class:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Another evaluation measure for algorithm may be accuracy, expressing the percentage of correct made classifications, and F-measure, a weighted harmonic mean of precision and recall:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad F = \frac{2 * \text{precision} * \text{recall}}{\text{Precision} + \text{Recall}}$$

We calculate accuracy of classifier, the recall and precision for the two classes, training the algorithm on 5000 sentences for each class of pre-classification test examples and applying it on the rest of the remaining examples.

Analyzing the algorithm efficiency for the above parameters we achieved a 0.814332247557 value of correct classification of opinions. From the result we conclude that a solution to improve the quality of the algorithm is to eliminate insignificant words for classification. Algorithm originally classified words without lexical content, so that besides nouns, verbs, adverbs and adjectives, are considered articles, prepositions and pronouns without semantic value.

We will eliminate these words called stop words in English that can induce noise in the classification. For this we have built an array of four vectors corresponding to those prepositions, conjunctions, articles and pronouns (e.g. articles - the, the, year, conjunctions - and, now, so, still, only, pronouns - who, whom, which, that, this, me, you, ours, prepositions - about, above, across, after, at, around, with, up). After this step we observe that the algorithm efficiency has improved a little, giving a value of: 0.81699346405229.

Algorithm considers that there is no relationship between words in a sentence, but in reality they are interrelated. So there are certain words that occur frequently together in one of two classes. Usually, we observe that the algorithm efficiency increases for values up to a maximum of three groups of words. In this application we tested the introduction in classification of groups of words for the  $n = 2$  and  $n = 3$ . Results are presented in the table below:

**Table 2. Algorithm efficiency**

	Initial Algorithm, groups of n=1 words	Initial groups of n=1 words, stop words	Algorithm, of n=1 eliminate stop words	Algorithm for groups of n=2 words, eliminate stop words	Algorithm for groups of n=3 words, eliminate stop words
Precision	0.8143	0.8169	0.8208	0.7972	
Recall	0.7598	0.7598	0.7659	0.5258	
Accuracy	0.7933	0.7948	0.7993	0.6960	
F-measure	0.7861	0.7874	0.7924	0.6336	
Execution time (s)	1.1535	3.2490	8.0684	14.7787	

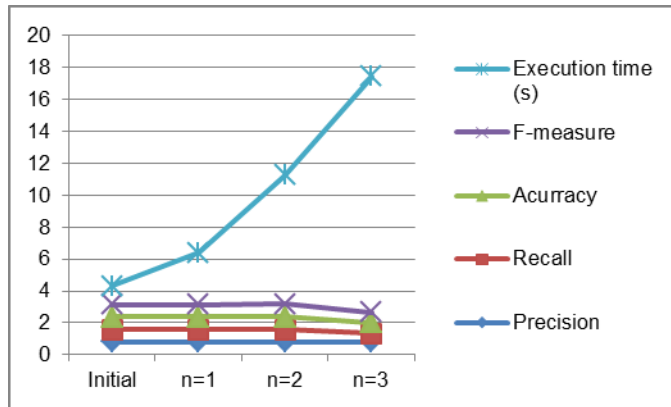


Fig. 4. Algorithm efficiency

## CONCLUSION

The expression of opinions of users in specialized sites for evaluation of products and services, and also on social networking platforms, has become one of the main ways of communication, due to spectacular development of web environment in recent years. The large amount of information on these platforms make them viable for use as data sources, in applications based on opinion mining and sentiment analysis. This paper presents a method of sentiment analysis, on the review made by users to movies. Classification of reviews in both positive and negative classes is done based on a naive Bayes algorithm. As training data we used a collection (pre-classified in positive and negative) of sentences taken from the movie reviews. To improve classification we removed insignificant words and introduced in classification groups of words (n-grams). For  $n = 2$  groups we achieved a substantial improvement in classification. As an extension of the research presented in this paper we want to improve the algorithm, enriching the training set of examples, on the way, with examples classified as strong positive or negative, by an established score of classification. We try to determine, in a review, those sentences which do not express opinions, or determine opinions about the film or the film acto addressed strictly on these items. We try to highlight the main aspects on which opinions are expressed and to extract opinions based on aspects identification. rs and identify opinions

The following chart presents the influence of data training volume on the accuracy of classifications in the method used. We detect a critical number of training data, from which point the increase number of the initial training set, will produce very little influence on precision of the algorithm.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] M Caraciolo (2012, Mar.) Working on sentiment analysis on Twitter with Portuguese language. [Online].<http://aimotion.blogspot.com/2010/07/working-on-sentiment-analysis-on.html>
- [2] Smeureanu A, Diosteanu C, Delcea LA Coffas.[2011] Business Ontology for Evaluating Corporate Social Responsibility, *Amfiteatru Economic*, 29: 28-42
- [3] L Minqing Hu.[ 2004] Mining and Summarizing Customer Reviews," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, USA,.
- [4] Turney P.[ 2002] Thumbs Up or Thumbs Down? semantic orientation applied to unsupervised classification of reviews. ACL'.
- [5] Liu, Web Data Mining - Exploring Hyperlinks, Contents and Usage Data, Second ed.: Springer, 2011.
- [6] Freitag D and McCallum A.[2000] Information extraction with HMM structures learned by stochastic optimization. AAAI-00,
- [7] Hatzivassiloglou V and Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity. COLING'00, 2000.

- [8] Hearst M.[ Direction-based Text Interpretation as an Information Access Refinement. In P Jacobs, editor, Text-Based Intelligent Systems. Lawrence Erlbaum Associates, 1992.
- [9] PD Turney and M.L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus.
- [10] Technical Report ERB-1094, National Research Council Canada, Institute for Information Technology, 2002.
- [11] Janyce Wiebe.[2000] Learning subjective adjectives from corpora. In AAAI/IAAI, pages 735–740.
- [12] M Caraciolo. (2012, Mar.) Working on sentiment analysis on Twitter with Portuguese language. [Online].<http://aimotion.blogspot.com/2010/07/working-on-sentiment-analysis-on.html>
- [13] MF Porter.[1980] An algorithm for suffix stripping. In Program, 14,,: 130–137,
- [14] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In ACL Conference, 2002.
- [15] R.Jaya and Rajan.C , [2016]A Study On Data Mining Techniques, Methods, Tools And Applications In Various Industries, International Journal on Concurrent Applied Research in Engineering and Management,vol.1, No.4, pp. 14-24.
- [16] Ellen Riloff. Automatically generating extraction patterns from untagged text. In Proceedings of AAAI/IAAI, Vol. 2, pages 1044–1049, 1996.
- [17] Nigam, K. and Hurst, M. 2004. Towards a robust metric of opinion. AAAI Spring Symp.on Exploring Attitude and Affect in Text.NLProcessor,2000.
- [18] Pang B, Lee L, and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. EMNLP-02.

# PRIVACY AUGMENTATION BY ACCOMPLISHING TRACEABILITY OVER ORUTA

R. Rajasaranya Kumari<sup>1\*</sup>, R. Vinod<sup>2</sup>, U. Velmurugan<sup>3</sup>, N. Rupavathy<sup>4</sup>

<sup>1,2</sup>Department of Computer Science and Engineering, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, Chennai, INDIA

<sup>3</sup>Department of Computer Science and Engineering, Vel Tech ( Owned by RS Trust), Avadi, Chennai, INDIA

<sup>4</sup>Department of Computer Science and Engineering, Vel Tech Dr.RR & Dr SR Technical university, Avadi, Chennai, INDIA

## ABSTRACT

**Aims:** Cloud Computing is a set of IT Services that are provided to a customer over a network and these services are delivered by third party provider who owns the infrastructure. It is used not only for storing data, but also the outsourced data can be shared by multiple users. Due to this there exists a problem in integrity. **Materials and methods:** The objective is to provide security to the data stored in the public cloud. Several mechanisms have been designed to support public auditing on shared data stored in the cloud. **Results:** With the privacy preserving mechanism, the public auditor audits and verify the integrity of the shared data in the cloud without seeing the data called as public auditing. Along with public auditing security is improved to the data in the public cloud. **Conclusion:** Normally the data is stored only in a single server. But, Here multiple servers are used for storing data for security purpose.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

traceability, data security, public auditing, shared data

\*Corresponding author: Email: [rajasaranya@velhightech.com](mailto:rajasaranya@velhightech.com)

## INTRODUCTION

Mobile app markets are creating a fundamental model shift in the way software is delivered to the end users. The Cloud computing is a new concept of computing technique, by which computer resources are provided dynamically via Internet. It attracts considerable attention and interest from both academia and industry. However, it also has at least three challenges that must be handled before applied to our real life. First of all, data confidentiality should be guaranteed. When sensitive information is stored in cloud servers, which is out of users' control in most cases, risks would rise dramatically. The servers might illegally inspect users' data and access sensitive information. On the other hand, unauthorized users may also be able to intercept someone's data. Secondly, personal information is at risk because one's identity is authenticated according to his information. As people are becoming more concerned about their privacy these days, the privacy-preserving is very important. Preferably, any authority or server alone should not know any client's personal information.

Recently, many mechanisms [2], [3] have been proposed to allow not only a data owner itself but also a public verifier to efficiently perform integrity checking without downloading the entire data from the cloud, which is referred to as public auditing. In these mechanisms, data is divided into many small blocks, where each block is independently signed by the owner; and a random combination of all the blocks instead of the whole data is retrieved during integrity checking [3]. A public verifier could be a data user who would like to utilize the owner's data via the cloud or a third-party auditor (TPA). Moving a step forward, Wang et al. designed an advanced auditing mechanism [2] (named as WWRL in this paper), so that during public auditing on cloud data, the content of private data belonging to a personal user is not disclosed to any public verifiers. That is, there is a leakage of identity privacy.

Recently proposed access control models, such as attribute-based access control, define access control policies based on different attributes of the requester, environment, or the data object. ABE features a mechanism that

enables an access control over encrypted data using access policies and ascribed attributes among private keys and ciphertexts. Especially, ciphertext-policy provides a scalable way of encrypting data such that the encryptor defines the attribute set that the decryptor needs to possess in order to decrypt the ciphertext. Thus, different users are allowed to decrypt different pieces of data per the security policy. This effectively eliminates the need to rely on the storage server for preventing unauthorized data access. Various types of clouds are present such as private cloud, public cloud, hybrid cloud and community cloud. Our aim is to improve security and audit the data in the public cloud.

Oruta is a novel privacy preserving public auditing mechanism. Here Oruta uses ring-signature for public auditing. By using ring signature, public verifier is able to verify the integrity or correctness of shared data without retrieving the original data fully. With the privacy preserving mechanism, the public auditor audits the shared data in the cloud without seeing the data called as public auditing. In this paper, along with public auditing we have improved data security in public cloud. Here we have achieved traceability (tracking the fake users). Data privacy is also improved by blocking the fake user from accessing the data from the public cloud.

This paper proposes a erasure correction code algorithm to protect the users' data stored in the cloud storage from the unauthorized access. The paper is organized as follows. Section II gives the detail on the various issues in cloud data storage ie problem statement. Section III describes about the architecture diagram. Section IV talks about the related work. Section V, gives the conclusion for the paper.

	PD P[9]	WWRL[5]	Proposed system
Public Auditing	✓	✓	✓
Data Privacy	x	✓	✓
Identity Privacy	x	x	✓
Traceability	x	x	✓

Fig: 1. Comparison among Different Mechanisms

### PROBLEM STATEMENT

The system model in this paper involves four parties: the cloud server, original user, a group of users and a public verifier. Cloud storage consists of several servers. The original user outsouce the data to the cloud. He will register the authorized user who can share that data from the cloud. Only that authorized user can share the data from the cloud. The problem exists when any other unknown person or user knows the authorized users' password ie the security problem arise. The shared data can be easily downloaded by the fake user. There is no traceability(ie the fake user cannot be tracked). Due to this Data Privacy in cloud is not preserved. Normally the outsourced data is kept only in private cloud.

Our mechanism should be designed to achieve the following properties: (1) **Public Auditing**: A public verifier is able to check the integrity of shared data without viewing the data from the cloud. (2) **Correctness**: A public verifier is able to correctly verify shared data integrity. (3) **Traceability**: Tracking the fake user from accessing the data from the cloud. (4) **Data Privacy**: Data shared in cloud should not be should not be downloaded by the fake user. (5) **Data security**: Only the data Owner and the authorized user are allowed to access or download the data. The user will be given some priviledges.

## ARCHITECTURE DESIGN

In this paper, we are going to achieve traceability (tracking the fake user). The Original user first register his account for login. The original user registers the users accounts whom he wants to share his data. Those users will be given some privileges. Data will be stored in three different servers for security purpose. Now when the original user uploads the data to the public cloud, the data will get split up into three sub blogs. Then the splitted files will get encrypted and stored in three different servers. The data is splitted and then encrypted using erasure correction code technique. If the user enters and try to download the file, OTP will be generated and send to the mail. If the OTP is valid, it allows the user to download the file. When the OTP is valid, the data is decrypted and then merged using erasure correction code technique and the downloaded as a original file. If the OTP is invalid, the user will be considered as fake user and he is not allowed to download the data or accessing the data from the cloud.

### ERASURE CORRECTION CODE:

The technique used here is erasure correction code. Erasure correction code is a method of data protection in which data is broken into fragments, expanded and encrypted into redundant data pieces and stored across a set of different locations or different servers. This algorithm is also used for decrypting the data and merging the data from different locations or different servers. This technique is described clearly in reference[4].

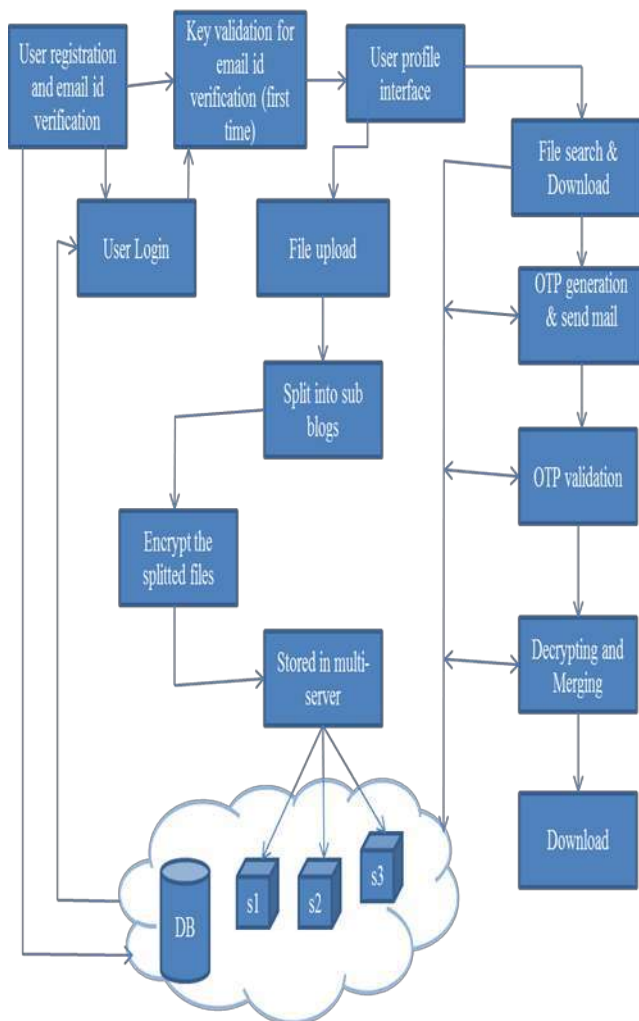


Fig. 3. Tracking the fake user

## RELATED WORK

The paper “Privacy Preserving Public Auditing for Secure Cloud Storage” [5] is used to enable the TPA to perform audits for multiple users simultaneously and efficiently. Combination of HLA and random masking is used, which enables TPA to perform auditing without viewing the content. The paper “PORs: Proofs of Retrievability for Large Files” [10] provides POR’s scheme which is also able to check the correctness of data on an untrusted server. The original file is added with a set of randomly-valued check blocks called sentinels. The Drawback is it focus only on personal data in the cloud. In [4] RDC is a technique by which user can check the integrity of data outsourced in servers, Audit the correctness of data under the multi-server scenario. The methods used here is replication, erasure coding, network coding. Erasure coding is a method of data protection in which data is broken into fragments, encoded and stored in multiple servers. Erasure coding is used in Privacy augmentation. In [3] the verifier is able to publicly audit the integrity of data without retrieving the entire data, which is referred to as public auditing. But, this mechanism is only suitable for auditing the integrity of personal data.

## CONCLUSION

In this paper, we propose a Privacy augmentation by accomplishing traceability over Oruta. A technique is used to improve data security called erasure correction code. Data Privacy in cloud is improved by tracking the fake user and blocking them from downloading the data from cloud. We utilize ring signatures, so that a public verifier is able to audit shared data integrity without viewing the entire data. Along with public auditing, data security and data privacy is improved.

## REFERENCES

- [1] C Wang, Q Wang, K Ren, and W Lou. [2010] Privacy- Preserving Public Auditing for Data Storage Security in Cloud Computing, *Proc. IEEE NFOCOM*, pp. 525-533,.
- [2] G Ateniese, R Burns, R Curtmola, J Herring, L Kissner, Z Peterson, and D Song. [2007] Provable Data Possession at Untrusted Stores, *Proc. 14th ACM Conf. Computer and Comm. Security (CCS ’07)*, 598-610,
- [3] B Chen, R Curtmola, G Ateniese, and R Burns. [ “Remote Data Checking for Network Coding-Based Distributed Storage Systems,” *Proc. ACM Workshop Cloud Computing Security Workshop (CCSW’10)*, pp. 31-42, 2010.
- [4] C Wang, SS Chow, Q Wang, K Ren, and W Lou. [ . 2013] Privacy-Preserving Public Auditing for Secure Cloud Storage,” *IEEE Trans. Computers*, 62( 2): 362-375
- [5] C Wang, Q Wang, K Ren, and W Lou. [2009] Ensuring Data Storage Security in Cloud Computing, *Proc. 17th Int’l Workshop Quality Ensuring Data Storage Security in Cloud Computing*,” *Proc. 17th Int’l Workshop Quality of Service (IWQoS’09)*, pp. 1-9,.
- [6] G Ateniese , RD Pietro, LV Mancini, and G Tsudik. [2008] Scalable and Efficient Provable Data Possession, *Proc. Fourth Int’l Conf. Security and Privacy in Comm. Networks (SecureComm’08)*,.
- [7] B Wang, B Li, and H Li. [2012] Knox: Privacy- Preserving Auditing for Shared Data with Large Groups in the Cloud,” *Proc. 10th Int’l Conf. Applied Cryptography and Network Security (ACNS’12)*, pp. 507-525, June
- [8] RL Rivest, A Shamir, and Y Tauman. [2001] How to Leak a Secret,” *Proc. Seventh Int’l Conf. Theory and Application of Cryptology and Information Security: Advances in Cryptology (ASIACRYPT’01)*, pp. 552-565.
- [9] A Juels and BS Kaliski. [2007] PORs: Proofs of Retrievability for Large Files,” *Proc. 14th ACM Conf. Computer and Comm. Security (CCS’07)*, pp. 584-597,
- [10] B Wang, B Li, and H Li. [2012] Oruta: Privacy- Preserving Public Auditing for Shared Data in *The Proc. IEEE Fifth Int’l Conf. Cloud Computing*, 295- 302.



# ORGANIZING MULTIMEDIA DATA USING SEMANTIC LINK NETWORK

J. Robin Joe, M. Senthil\*, K. Arun Kumar

Department of Computer Science and Engineering, SRM University, Chennai, INDIA

## ABSTRACT

**Aims:** Multimedia resources such as images, audio, video are growing at a high rate due to our daily usage of internet and other digital activities. So organisation of these resources is the biggest challenge in today's world. Whatever application we use in internet, it generates certain amount of data that are needed to be preserved. For example, we use Google or other search engines to search for a thing. In order to give correct output, these search engine providers need a way to organize these multimedia resources. Here organizing refers to efficient way of storing multimedia resources so that while retrieving, it will give us appropriate results. Just adding new images or other resources to a database daily won't provide the best retrieval result. **Materials and Methods:** In this paper, the Semantic Link Network model is used for organizing multimedia resources. A whole model for generating the association relation between multimedia resources using Semantic Link Network model is proposed. Each image in internet has a name and tags associated with it. The tags and the surrounding texts of multimedia resources are used to measure their semantic association. **Results:** Based on this information from the images, this proposed model aims to get a value for each image and classify them according to the range of values. **Conclusions:** This type of organisation of multimedia resources enables one to efficiently store the resources. So while retrieving, it produce appropriate results to all the users who are interested in retrieving the resources.

Published on: 08<sup>th</sup>– August-2016

### KEY WORDS

Multimedia Resources, Semantic Link, Multimedia Resources Organization, Databases, Search Engines

\*Corresponding author: Email: [senthil.m@rmp.srmunv.ac.in](mailto:senthil.m@rmp.srmunv.ac.in); [robin.robin.joe@gmail.com](mailto:robin.robin.joe@gmail.com); Tel.: +91 98 40 638033

## INTRODUCTION

Understanding the semantics of multimedia has been an important component in many multimedia based applications [1]. Manual annotation and tagging has been considered as a reliable source of multimedia semantics. Unfortunately, manual annotation is time-consuming and expensive when dealing with huge scale of multimedia data. The rapid increase number of multimedia resources has brought an urgent need to develop intelligent methods to represent and annotate them [2].

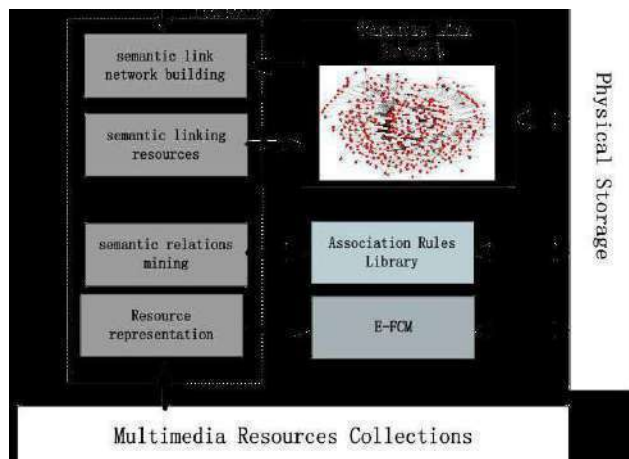
In this paper, the Semantic Link Network (SLN) model is used for organizing multimedia resources with social tags. Semantic Link Network is designed to establish associated relations among various resources (e.g., Webpages or documents in digital library) aiming at extending the loosely connected network of no semantics (e.g., the Web) to an association rich network [3]. The tags and surrounding texts of multimedia resources are used to represent the semantic content. The relatedness between tags and surrounding texts are implemented in the Semantic Link Network model.

## Related works

The Semantic web is an evolving development of the World Wide Web, in which the meanings of information on the web is defined therefore; it is possible for machines to process it. The basic idea of Semantic Web is to use ontological concepts and vocabularies to accurately describe contents in a machine readable way [4]. These concepts and vocabularies can then be shared and retrieved on the web. In the Semantic Web, each fragment of the description is a triple, based on Description Logic. Thus, the implicit connections and semantics within the description fragments can be reasoned using Description Logic theory and ontological definitions. Earlier research work on the Semantic Web focused on defining domain specific ontologies and reasoning technologies [5]. Therefore, data are only meaningful in certain domains and are not connected to each other from the World Wide Web point of view, which certainly limits the contributions of Semantic Web for sharing and retrieving contents within a distributed environment.

The Semantic Link Network (SLN) was proposed as a semantic data model for organizing various Webs resources by extending the Web's hyperlink to a semantic link. SLN is a directed network consisting of semantic nodes and semantic links [6].

The basic mechanism is shown below.



## SEMANTIC LINK NETWORK

The proposed model consists of the following parts,

### Resources representation

Element Fuzzy Cognitive Map (E-FCM) [19] is used to represent multimedia resources with social tags since it does not only reserve resources' keywords but also the relations among them.

### Resources storage mechanism

Database/XML is used to store E-FCM since it is easy to define the mark-up elements.

### SLN generation mechanism

Based on E-FCM and the association rules, ALN can be generated by machine automatically.

### The Basic Heuristics

Based on common sense and our observations on real data, five heuristics that serve as the base of the proposed computation model are given as follow.

**Heuristic 1.** Usually each tag of a multimedia resource appears only one time. Different from writing sentences, users usually annotate a multimedia resource with different tags. For example, the possibility of using tags "apple" for an image is very low. Therefore, in this paper, we do not employ any weighting scheme for tags such as tied.

**Heuristic 2.** The order of the tags may reflect the correlation against the annotated multimedia resource. Different tag reflects the different aspect of a multimedia resource. According to Heuristic 1, the weight of a tag against the image cannot be obtained. Fortunately, the order of the tags can be get since user may provide tags one by one.

**Heuristic 3.** The number of tags of a multimedia resource may not relevant to the annotation correctness. Different users may give different tags about the same multimedia resource. For example, users may give tags such as "apple iPhone "or "iPhone4 mobile "for a same image about iPhone. It is hardly to say which tag is better for annotation though the latter annotation has three tags.

**Heuristic 4.** Usually some tags may be redundant for annotating a multimedia resource. Of course, users may give similar tags for a multimedia resource. For example, the tags "apple iPhone" may be redundant since iPhone is very semantic similar to apple.

**Heuristic 5.** Usually some tags may be noisy for annotating a multimedia resource. Users may give inappropriate or even false tags for a multimedia resource. For example, the tags "iPhone" are false for an image about the iPod.

## GENERATING THE SEMANTIC LINK

The proposed computation model is divided into three steps:

### Tag relatedness computation

In the proposed computation model, each tag can be seen as a concept with explicit meaning. Thus, we use some equations based on co-occurrence of two concepts to measure their semantic relatedness. The core idea is that "you shall know a word by the company it keeps". In this section, four popular co-occurrence measures (i.e., Jacquard, Overlap, Dice, and PMI) are proposed to measure semantic relatedness between tags [7].

Besides co-occurrence measures, the page counts of each tag from search engine are used. Page counts mean the number of web pages containing the query  $q$ . For example, the page counts of the query "Obama" in Google are 1,210,000,000. Moreover, page counts for the query "q AND p" can be considered as a measure of co-occurrence of queries  $q$  and  $p$ .

The page counts for the query "p AND q" should be considered. For example, when we query "Obama" and "United States" in Google, we can find 485,000,000 Web pages, that is, =485,000,000. According to probability and information theory, the mutual information (MI) of two random variables is a quantity that measures the mutual dependence of the two variables. Point wise mutual information (PMI) is a variant of MI.

$$PMI(p,q)=\log(N*N(p\cap q)/N(p)*N(q))/\log N$$

where  $N$  is the number of Web pages in the search engine, which is set to according to the number of indexed pages reported by Google.

Algorithm 1:MaxRel

**Input:** The tags set of two images  $f_1$  and  $f_2$ , which is  $s(f_1)$  and  $s(f_2)$

**Output:** The semantic relatedness of two images  $f_1$  and  $f_2$  for each  $t_i s(f_1) /*page position initial */$

```

N ( s ( f1 ) ) N ( tj );
Pos ( s ( f1 ) ) Pos ( tj );
for each tj, s ( f2 )
N ( s ( f2 ) ) N ( tj );
Pos ( s ( f2 ) ) Pos ( tj );
for each ti, s ( f1 )
for each tj, s ( f2 )
if ( ti, tj ) sr(ti, tj):/*pruning*/
counts and
else sr(ti, tj) = f ( N(ti), N(tj) )
/*relatedness*/
return maxRel(f1,f2)= f ( Pos(ti), Pos(tj),
sr(ti, tj));

```

### Semantic Relatedness Integration

We change the semantic relatedness integration of all tag pairs to the assignment in bipartite graph problem. We want to assign a best matching of the bipartite graph [8].

Adopting the proposed maxRel function, we are sure to find the global maximum relatedness that can be obtained pairing the elements in the two tags sets. Alternative methods are able to find only the local maximum since they scroll the elements in the first set and, after calculating the relatedness with all the elements in the second set, they select the one with the maximum relatedness. Since every element in one set must be connected, at most, at one element in the other set, such a procedure is able to find only the local maximum since it depends on the order in which the comparisons occur. For example, considering the below diagram.

Image f1 with tags Image f2 with tags t1,t2,t3

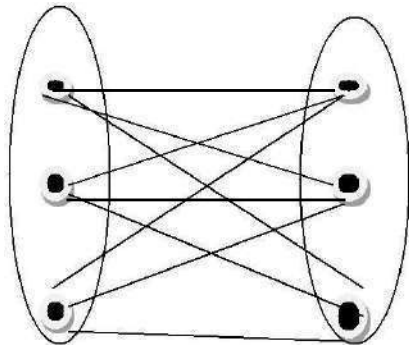


Fig:1 One to one many relationship

t1 will be paired to q1 (weight=1.0). But, when analysing t3 the maximum weight is with q2 (weight=0.9). This means that t2 can no more be paired to q2 even if the weight is maximum, since this is already matched to t3. As a consequence, t2 will be paired to q3 and the average of the selected weights will be  $(1.0+0.3+ 0.9)/3 = 0.73$  which is considerably lower than using maxRel where the sum of the weights was  $(1.0+0.8+0.7)/3=0.83$ .

Overall, the cardinality of two tag sets is used to follow heuristic 3. The one-to-one map of tags pair is used to follow heuristics 4 and 5. The maxRel function is used to match a best semantic relatedness integration of two multimedia resources.

4.3 Tag Order Revision

In this section, the maxRel function proposed in section 4.2 is revised considering the order of tags. For example, the relatedness of tag pair with high position should be enhanced, which is summarized as

**Schema 1:** This schema means that the identical tag pairs of two multimedia resources and should be pruned in maxRel function.

We add a decline factor to the maxRel function, and the detailed steps are:

- (1) According to the maxRel function in section 4.2, the best matching tag pairs are selected, which is denoted as:

$$\text{MaxRel}(f1, f2) = \sum sr(ti, tj)$$

- (2) Computing the position information of each tag, which is denoted as Pos(t1)

$$\text{Pos}(ti) = |s(f) - i| / |s(f)|$$

- (3) Add the position information of each tag to the equation, which can be seen as a decline factor:

$$\text{Sr}(f1, f2) = \sum \text{Pos}(ti) * sr(ti, tj) * \text{Pos}(tj)$$

- (4) Of course, similar to maxRel function, equation should divide the result of the maximization by

$$\text{Sr}(f, f2) = \frac{\sum \text{Pos}(ti) * sr(ti, tj) * \text{Pos}(tj)}{\sum \text{Pos}(ti) * \text{Pos}(tj)}$$

**Schema 2.** Identical tag pruning. This schema means that the identical tag pairs of two multimedia resources and should be pruned in maxRel function. In other words, the semantic relatedness of the same tag of two multimedia resources is set as 0.

The above schema is used to ensure the relatedness measures of two multimedia resources. If we do not prune the identical tag pairs of two multimedia resources, the proposed method will be transformed to the similarity measures. For example, the cosine similarity between two tags is to find the number of identical elements of two vectors. The overall algorithm of the proposed computation mode is presented in algorithm 1.

## IMPLEMENTATION

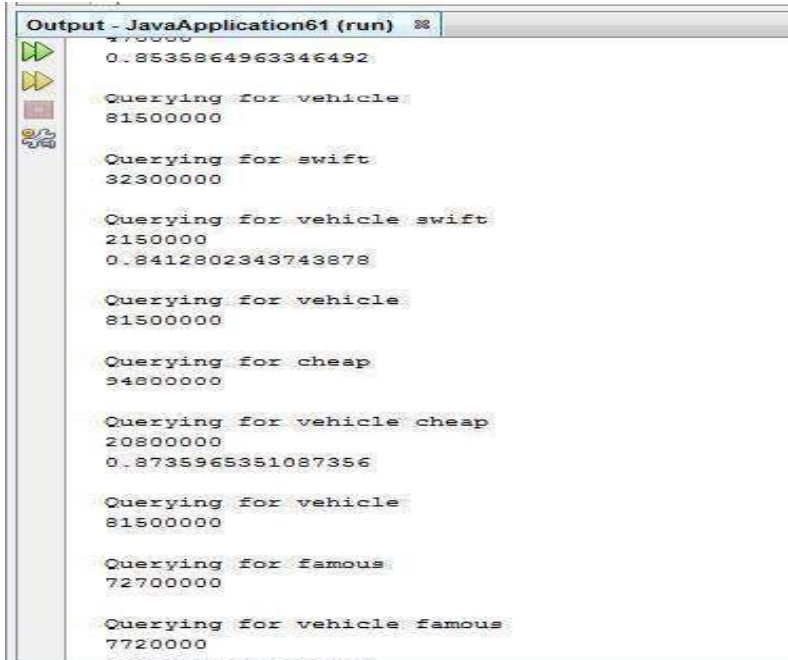
### Google API Code

```
Query =URLEncoder.encode(query, "UTF-8");
```

```
URL url = new URL("http://ajax.googleapis.com/ajax/services/search/web?start=0&rsz=large&v=1.0 &q=" + query);
```

```
URLConnection connection = url.openConnection();
```

```
connection.addRequestProperty("Referer" , HTTP_REFERER);
```



```

Output - JavaApplication61 (run)
0.8535864963346492
Querying for vehicle
81500000
Querying for swift
32300000
Querying for vehicle swift
2150000
0.8412802343743878
Querying for vehicle
81500000
Querying for cheap
34800000
Querying for vehicle cheap
20800000
0.8735965351087356
Querying for vehicle
81500000
Querying for famous
72700000
Querying for vehicle famous
77200000

```

Fig. 2. Implementation java code

## RESULTS FROM PROPOSED MODEL

```

public double compute(int one, int two, int both) {
//To calculate semantic relatedness double n= Math.pow(10, 11);
    double z=Math.log((n*both)/(one*two));
    double x=(double) (z/(Math.log(n)));System.out.println(x);
return x;
}

```

```

----TAG ORDER REVISION----
0.6184452058832405
0.6184452058832405
----TAG ORDER REVISION----
0.8241017833636933
0.5494011889091288
----TAG ORDER REVISION----
0.56453423
0.18817807666666667
----TAG ORDER REVISION----
0.6208857630411686
0.4139238420274457
----TAG ORDER REVISION----
0.8682118356575788
0.3858719269589238
----TAG ORDER REVISION----
0.8732853426766157
0.19406340948369238

```

```

----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.6184452058832405
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.8241017833636933
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.56453423
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.6208857630411686
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.8682118356575788
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.8732853426766157
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.56453423
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.56453423
----SEMANTIC RELATEDNESS BETWEEN TAGS----
0.655553849557736

```

Fig. 3. Semantic values in proposed system

### Evaluation on Image Clustering

In this section, we evaluate the correctness of using tag order. In section 4.3, we add the position information of each tag to the semantic relatedness measures. The tags with high position are treated as the major element for semantic relatedness measures. We evaluate the using of tag order by the clustering task. We employ the proposed semantic relatedness of images into K- means clustering model. Since the K- means model depends on the initial points, we random select core points 100 times.

We evaluate the effectiveness of document clustering with three quality measures: *F-measure*, *Purity*, and *Entropy*. We treat each cluster as if it were the result of the proposed method and each class as if it were the desired set of images. Generally, we would like to maximize the *F-measure* and *Purity*, and minimize the *Entropy* of the clusters to achieve a high- quality document clustering. Moreover, we compare the clustering results between the proposed method using tag order or not.

### Evaluation on image searching

Five queries from group2 are selected as the test set including “Louis Vuitton”, “Gucci”, “Chanel”, “Cartier”, and “Dior”. These queries are searched in Flickr. The top 50 images are obtained as the data set. Moreover, we remove the queries on the tags of each image. For example, the tag “Cartier” of the top 50 images is re-moved of the query “Cartier”. The reason for that operation is that the proposed method is based on the semantic relatedness other than co-occurrence. We choose cut-off point precision to evaluate the proposed method on image searching. The cut-off point precision ( $P_n$ ) means that the percentage of the correct result of the top  $n$  returned results. We compute the  $P1$ ,  $P5$ , and  $P10$  of the group2 test set.

## APPLICATIONS

Content-based image retrieval (CBIR) is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases. "Content-based" means that the search analyzes the contents of the image rather than the metadata such as keywords, tags, or descriptions associated with the image [9]. The term "content" in this context might refer to colors, shapes, textures, or any other information that can be derived from the image itself.

CBIR is desirable because most web-based image search engines rely purely on metadata and this produces a lot of garbage in the results [10]. Also having humans manually enter keywords for images in a large database can be inefficient, expensive and may not capture every keyword that describes the image [11]. Thus a system that can filter images based on their content would provide better indexing and return more accurate results.

The proposed SLN based model can be used for video searching. The ontology based video searching is similar to CBIR, which also focuses on the content of the videos [12 – 14]. Figure. 1 gives the searching interface of the developed tool based on the proposed SLN based model. From Figure.2, 3, the searching procedures for a user are as follow.

- (1) Ontology based queries. Different from web search engines, the proposed SLN based video search constricts the searching method. Users can only select the defined attributes or concepts as the searching queries.
- (2) Associated videos suggestion. Since the video resources are organized by their association relation.

## CONCLUSION

Recent research shows that multimedia resources in the wild are growing at a staggering rate. The rapid increase number of multimedia resources has brought an urgent need to develop intelligent methods to organize and process them. In this paper, the Semantic Link Network model is used for organizing multimedia resources. Semantic Link Network (SLN) is designed to establish associated relations among various resources (e.g., Web pages or documents in digital library) aiming at extending the loosely connected network of no semantics (e.g., the Web) to an association-rich network. Since the theory of cognitive science considers that the associated relations can make one resource more comprehensive to users, the motivation of SLN is to organize the associated resources loosely distributed in the Web for effectively supporting the Web intelligent activities such as browsing, knowledge discovery and publishing, etc. The tags and surrounding texts of multimedia resources are used to represent the semantic content. The relatedness between tags and surrounding texts are implemented in the semantic Link Network model. Two data mining tasks including clustering and searching are performed by the proposed framework, which shows the effectiveness and robustness of the proposed framework.

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the technical support given by M.Senthil, Assistant Professor, Department of Computer Science and Engineering SRM University, Chennai, Tamilnadu, India.

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] L.Adams and Y. Wales. The process of criminal investigation based on grey hazy set. *2010 IEEE International Conference on System Man and Cybernetics*, pp.26-28, 2010.
- [2] E. Michael. Efficient and low-complexity surveillance video compression using backward-channel aware wyner-ziv video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(4):452-465, 2009.
- [3] A. Harris. Random-forests-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(5):649-659, 2008.
- [4] H. Yu, C. Pedrinaci, S. Dietze, and J. Domingue. Using linked data to annotate and search educational video resources for supporting distance learning. *IEEE Transactions on Learning Technologies*, 5(2):130-142, 2012.

- [5] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang. Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, 10(7):1342-1355, 2008.
- [6] Lionel M. Ni, and G. Xue. A Reliability-Oriented Transmission Service in Wireless Sensor Networks. *IEEE Transactions on Parallel and Distributed Systems*, 22(12): 2100-2107- 2011.
- [7] Lionel M. Ni. Opportunity-Based Topology Control in Wireless Sensor Networks. *IEEE Transactions on Parallel and Distributed Systems*, 21(3): 405-416, 2010.
- [8] M. Wigan and R. Clarke. Big data, s big unintended consequences. *Computer*, 46(6):46-53, 2013.
- [9] D. Yuan. A highly practical approach towards achieving minimum datasets storage cost in the cloud. *IEEE Transactions on Parallel and Distributed Systems*, 24(6):1234-1244, 2013.
- [10] Nepal, S. Pandev, and J. Chen. A privacy leakage upper-bound constraint based approach for cost- effective privacy preserving of intermediate datasets in cloud.
- [11] J. Chen. Do we need to handle very temporal violation in scientific workflow system? *ACM Transactions on Software Engineering and Methodology*, early access, 2013.
- [12] J. Chen. A scalable two-phase top- down specialization approach for data anonymiation using map reduce on cloud. *IEEE Transactions on Parallel and Distributed Systems*, early access, 2013.
- [13] "Cisco Visual Networking Index: Forecast and Methodology, 2009-2014," Available: <http://www.cisco.com/en/US/solutions/collateral/>
- [14] "Great Scott! Over 35 Hours of Video Uploaded Every Minute to YouTube," The official YouTube blog, 2013.



## AN EFFECTIVE ONLINE AUCTION SYSTEM

Marella Saibharath, Padmavasani. K, P. Muneeswaran\*

Department of Computer Science and Engineering, SRM University, Ramapuram Campus, Chennai, INDIA

### ABSTRACT

**Aims:** Online auctions are among the most influential e-business applications. Although there have been considerable efforts in setting up market places, online trading still lays in its early stages. Quite a few companies have started projects of their own, trying to improve their purchasing and sales channels. **Materials and Methods:** The most impressing concept of Internet market places is the conduction of online auctions. An online auction system holds online auctions for various products on a website. **Results:** It's place for buyers and sellers to come together and trade almost anything. **Conclusion:** In this system it consists of a web-portal where registered users can propose new auctions, purchase and place bids in order to buy the items on auction.

Published on: 08<sup>th</sup>– August-2016

#### KEY WORDS

Auction, Bid, Buyer and Seller,  
Bid shielding, shill bidding, Web-Portal

\*Corresponding author: Email: [padmavasani@yahoo.co.in](mailto:padmavasani@yahoo.co.in); [saibharath9999@gmail.com](mailto:saibharath9999@gmail.com) Tel.: +91 9176665427

### INTRODUCTION

An *Auction* is Latin work which means augment. Auction is a bid, a method of selling; buying and services offered take place. Online Auctioning System has several other names such as e-Auctions, electronic auction etc. The requirement for online auction or online bidding can be more precisely specified by the client. Online Bidding has become more wide spread in all sorts of industrial usage. It not only includes the product or goods to be sold, it also has services which can be provided. Due to their low cost this spreading out made the system to grow. Bidders can be maintained in a single database according to the preference, and they can be monitored. User's data can be maintained in a confidential way for validity and integrity of contractual documentation. Multiple bidders can be communicating with a great ease. This system allows multiple bids by single users.

The Objective is to develop a user-friendly auctioning site where any kind of product can be auctioned and provide value-added services to the bidders and the sellers.

#### Literature survey

Well-settled principles of law, such as those surrounding fraud in its various forms, have long maintained their vitality, adapting to changes in the legal and business environments through judicial and legislative understanding and intervention. Many of these changes have manifested themselves in the world of commerce [1]. The creation and growth of the Internet has resulted in significant changes in the way people engage in commerce. The increasing popularity of the Internet as a medium of commerce has generated an increase in Internet fraud, raising new and difficult legal issues in areas including online auctions [2].

We explore and analyze the structure of Internet auctions from a logical and an empirical perspective. Such web-based auctions are rapidly rising as a mercantile process of choice in the electronic marketplace. While traditional auction theory focuses on single-item auctions, we observe that a majority of on-line auctions are multi-item auctions. A significant donation of work is the theoretical derivation of the structure of the winning bids in multi-

item progressive online auctions. Additionally, for comparative purposes, we explore the structural characteristics of alternative multi-item auction mechanisms proposed in the auction theory. We derive hypothesis based on our analytical results and compare two different types of auction mechanisms. We test the conventional auction theory assumption regarding the homogeneity of bidders and present the first ever empirically derived classification and performance-comparison of on-line bidders. We test our hypotheses using real-world empirical data obtained by track a premier web-based auction site. Arithmetical analysis of the data indicates that firms may gain by choosing alternative auction mechanisms. We also provide directions for further exploration of emerging but important dimension of electronic commerce [3].

We have recently seen a marvelous number of auctions conducted over the Internet. This form of electronic commerce is rapidly growing, and it is projected to account for 30 % of all E-Commerce by 2002. Using actual bidding transaction data from 324 businesses-to consumer online auctions [4], we analyze the bidder's arrival process during each auction. We find that most bidders like to sign on early in the auction; typically, 70 % of the bidders sign on during the first half. Our statistical analysis reveals that the minimum initial bid is negatively connected with the number of bidders per auction, while the number of units offered and the length of the auction are positively connected with the number of bidders. We also present a model for estimating the expected price as a function of the number of bidders, the mean and variance of the private valuation distribution, and the number of units to be sold in the auction [5]. Our analysis shows that increased dispersion in the bidders values may either increase or decrease the auction price, depending on the bidders overall arrival process, the length of the auction, and the number of units. We calculate the best auction length and show that an auction's profit is a uni-modal function of its duration and the number of units [6].

### Problem Statement

Auctions are used to sell many things in addition to antiques and art. All round the world there are auctions of commodities such as tobacco, cattle, racehorses and just above anything elsewhere there's market of multiple people interested in buying the same thing that's the key to auction-a bunch of people who are interested in buying the same object, and taking turns offering bids on the object. The right to buy that object will go to the highest bidder; It is called as traditional auction [7].

While update of traditional auction is online auction, companies from various industries are moving to online auction say eBay, asteinrete, on sale provided a worldwide platform for bidder by getting it to masses. In these websites they create their own auction by adding their own product to auction. Buyer can bid the product which he/she can win the product if he applies the winning bid [7-9]. The following are some of the disadvantages of existing auction system

- Traditional method is time consuming process
- Date and time plays an important role, as they operate for few hours only.
- There is no separate module for sellers to upload their own Product in online auction system.
- Bid shielding
- Shill bidding

### SYSTEM DESIGN

System architecture is the process of defining the components, modules, interfaces and data for a system to satisfy specified requirements. The following is the architecture for the system

#### Module Description

- Authentication Module
- Buyer Module
- Seller Module
- Administrator Module

### Authentication Module

Authentication module is the module where the user gets authenticated. Authentication is otherwise known as validation. In this module the buyer/seller first gets registered in order to proceed further. Buyer and seller have separate authentication procedure. If the provided information does not meet the criteria, the user is not validated.

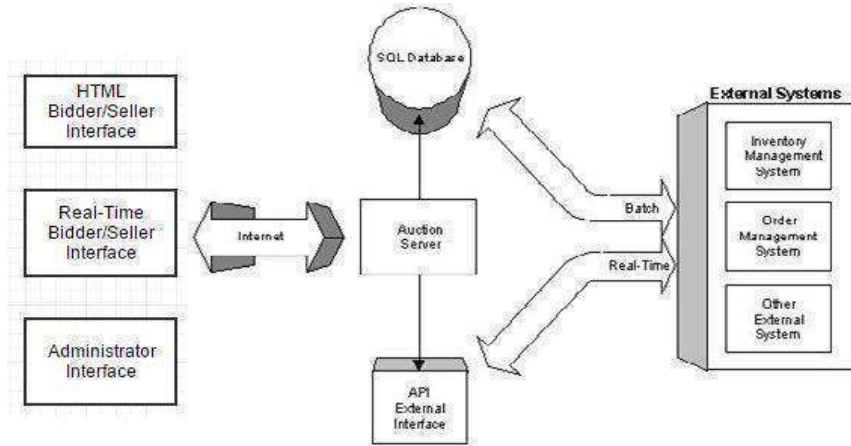


Fig.1.Architectural Diagram

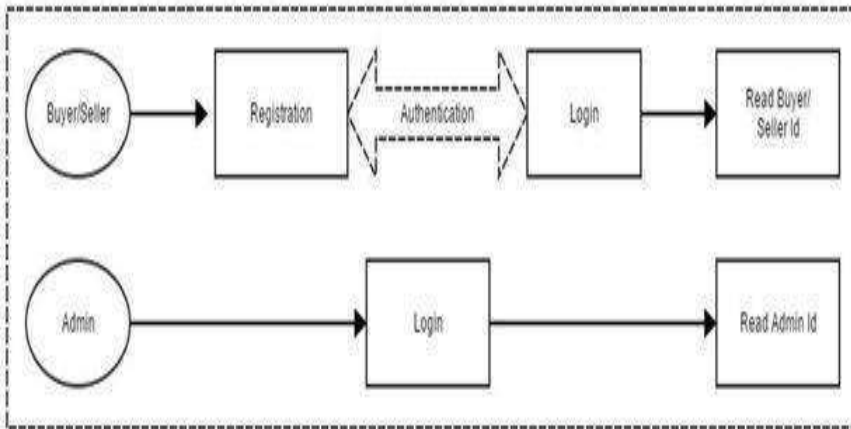


Fig.2.Authentication Module

### Buyer Module

In this module buyer can visit the site. In order to apply the bid the buyer has to login to the system and must buy bid points in order to bid the product. If the buyer wishes to buy the product he can apply the bid. If the bid is unique and large the buyer will get the product. When the buyer won the product he has to make payment to the system.

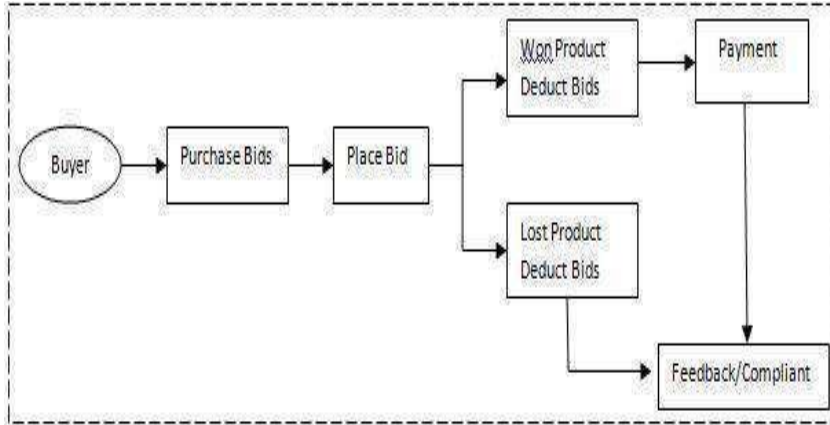


Fig.3.Buyer Module

### Seller Module

In this module seller has to login to the system and register his product to the system for auction. When the auction is completed administrator informs the details of buyer so that seller handover the product to the system. Seller receives the payment from the system.

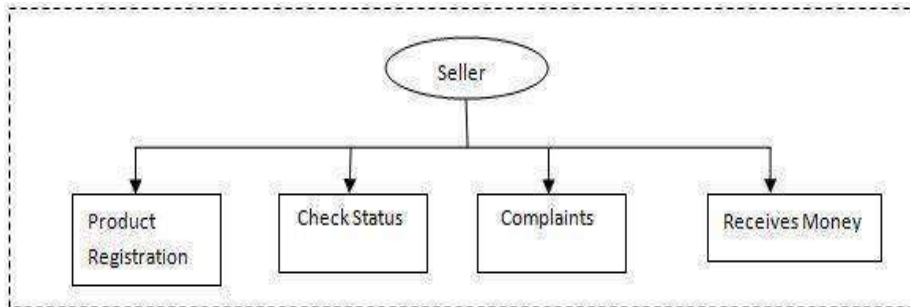
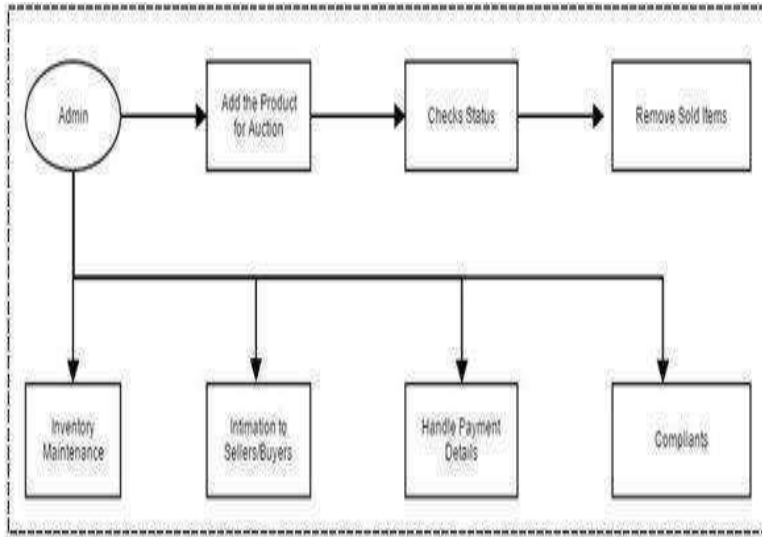


Fig.4.Seller Module

### Administrator Module

Admin module does all the task that enables the user to bid for an item effortlessly. Admin will create and update the categories. Under the categories we can find different items that are up for the auction. Admin will take care of all the information regarding the items under each category. Admin will be responsible for all the actions done by the users. Admin can block the users and can change privileges of the selected user. Admin can delete the categories and can delete the items that are up for the auction. Administrator is responsible for the inventory maintenance.



. Fig.5.Administrator Module

## RESULTS AND DISCUSSION

The existing system has no option for selling their own product in auction and traditional method is time consuming process where date and time plays an important role, as they operate for couple of hours. To overcome this, system creates a way for sellers to upload their product for auction. There are no time constraints in this system like traditional one. Buyers who are applying bid can be monitored by the administrator. Buyer can place more than one bid at a time for multiple products. By this system we can overcome the problems of Bid Shielding and shill bidding. Buyer can easily compare the bid which is applied for a particular product.



Fig.6.Snapshot Representing Restriction of Bid shileding



Fig.7.Snapshot of New Seller Module

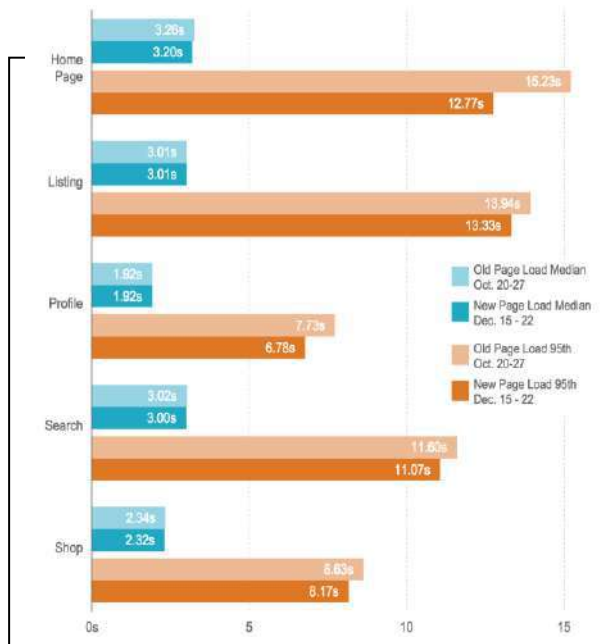


Fig.: 8.Front – End Performance Analysis

## CONCLUSION

Online auction is a system where we participate in a bid for products and service. This auction is made easier by using online software which can regulate processes involved. There are several different auction methods or types and one of the most popular methods is English auction system. This system has been designed to be highly-scalable and capable of supporting large numbers of buyers and sellers in an active auction.

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the technical support given by P.Muneeswaran, Assistant Professor\*, Department of Computer Science and Engineering, SRM University, Ramapuram Campus, Chennai.

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Albert, M. R. (2002). "E-Buyer Beware: Why Online Auction Fraud Should Be Regulated". *American BusinessLawJournal* 39 (4):575. Doi:10.1111/j.1744-1714.2002.tb00306.x
- [2] Ravi Bapna, R.; Goes, P.; Gupta, A. (2001). "Insights and analyses of online auctions". *Communications of the ACM* 44 (11): 42. Doi: 10.1145/384150.384160.
- [3] Vakrat, Y.; Sideman, A. (2000). "Implications of the bidders' arrival process on the design of online auctions". *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*. p. 7. Doi:10.1109/HICSS.2000.926822. ISBN 0-7695-0493-0..
- [4] Milgrom, P.; Weber, R. (1982). "A theory of auctions and competitive bidding". *Econometrica* 50 (5):1089-1122. Doi: 10.2307/1911865. JSTOR 1911865.
- [5] Pinker, E. J.; Sideman, A.; Vakrat, Y. (2003). "Managing Online Auctions: Current Business and Research Issues". *Management Science* 49 (11): 1457. Doi:10.1287/mnsc.49.11.1457.20584
- [6] Simone Pigolotti, Sebastian Bernhardson, Jeppe Juul, Gorm Galster, Pierpaolo Vivo (2012). "Equilibrium strategy and population-size effects in lowest unique bid auctions". Retrieved 2012-10-25.
- [7] Atwood, Jeff (2009-05-25). "Penny Auctions: They're Gambling". *Coding Horror*. Retrieved 2013-01-03.
- [8] "Is Swoop Nothing More Than a Well-Designed Gimmick?". *Technologizer.com*. 2008-09-17. Retrieved 2013-01-03.
- [9] "Auction Scams". *OnlineAuctionReviews.org*.

## VANET BASED TRAFFIC ANALYSIS USING CLOUD SERVER

K. Suresh, R.A.Preethi, T. Sarada Kiranmayee\*

Dept. of Computer science and Engineering, SRM University, Chennai, TN, INDIA

### ABSTRACT

Wireless Sensor Networks (WSNs) can have high demands for real-time data transmission and processing, but this is often constrained by limited resources. Cloud Computing can act as the backend for WSNs to provide processing and storage on demand. Sensor-Cloud infrastructure is becoming popular that can provide an open, flexible, and reconfigurable platform for several monitoring and controlling applications. Most common technique for knowing the location of vehicle is with the help of GPS. In the proposed system introducing the concept of determining the traffic Congestion in way of travelling using pre condition monitoring with sensor nodes.

Published on: 08<sup>th</sup>– August-2016

#### KEY WORDS

WSNs, GPS, VANET, Cloud Computing, Sensor Detector

\*Corresponding author: Email: [Sureshkannan92@hotmail.com](mailto:Sureshkannan92@hotmail.com), [tskiranmayee@gmail.com](mailto:tskiranmayee@gmail.com) Tel.: + 91-9790905348

### INTRODUCTION

Traffic congestion is becoming huge problem. Specially incities. VANETs are wide rising in several developed countries as a less costly, distributive and cooperative hold up system. It essentially carries with it 3 parts: sensing, processing, and communication. WSNs are in the main motive to monitor environmental conditions like temperature, sound, and vibration and pressure .The applications of

WSN is applied in several areas.

WSN components:

- A radio transceiver with associate antenna for transmission and receiving knowledge.
- A microcontroller for interfacing with the sensors.
- Energy sources like batteries or different power provide.

Cloud Computing may be an assortment of virtualized resources that may be allotted on demand. The elastic resources capability of the Cloud is that the main motivation for integration WSNs with the Cloud. Cheap Devices to be incorporated in vehicles itself and communicate with fixed satellite to accumulate the info and fetch into the system. Every node in an exceedingly device network is loaded with a radio transceiver or another wireless communication device, at low microcontroller, associated an energy supply most frequently cells/ battery. The nodes of device network have cooperative capabilities that are sometimes deployed in an exceedingly random manner.

**Figure-1** shows a typical WSN setup for traffic condition monitoring. Sensor Nodes are mounted on cloud computing server .Here Cloud Computing server act as a base station. The sensor nodes communicate with the base station, The base station collates the sensor information from sensor node data and transmits it to the travelling vehicle .So Travelling vehicle communicate directly with the Base station server rather sensor node. So sensor detects the traffic with precondition monitoring to determine congestion.



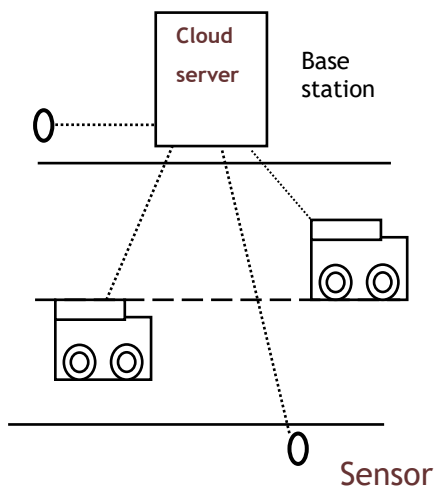


Fig: 1 Sensor node in VANET

## RELATED WORKS

Wireless sensor networks (WSNs) can be used for monitoring the railway infrastructure such as bridges, rail tracks, track beds, and track equipment along with vehicle health monitoring such as chassis, bogies, wheels, and wagons [1]. Condition monitoring reduces human inspection requirements through automated monitoring, reduces maintenance through detecting faults before they escalate, and improves safety and reliability. WSNs enable continuous real-time capture of data. However, WSNs need to be able to handle the harshness of outdoor long-term condition monitoring; often in hostile environments and must minimize energy usage as the nodes are not attached to a wired power supply. They typically use low-power sensors powered by batteries although authors are investigating alternative power supplies such as local energy generation. Hence, the network to enable data capture has to be carefully designed to overcome these factors and prevent transmission errors, latency, network outages, missing data, or corrupted data. Vehicular ad hoc networks (VANETs) vehicle-to-vehicle and vehicle-to-infrastructure communications which can be a reliable and secure system for efficient traffic control [2]. Considering the broadcast nature of the medium, multi-hop routing, multiple communication paradigms and short duration of vehicle to vehicle sessions, the establishment of VANET according to modern day needs can be critical. So while on the road if there is no base station in nearby, there is actually not a problem because due to the ad hoc network structure all the nodes create a network by hopping the signal eventually to the nearest base stations.

Moreover, through VANET, each vehicle can communicate with the other vehicle through V2V network. So, with the ad hoc network created within the traffic can be controlled. Whenever a car will come into a close proximity within a certain region which can make congestion in the road, by V2V the car will send message to the other car and create enough room in the road so that when the green signal turns on every car can move comfortably without making a huge traffic jam due to congestion. Smartphone's serve as a technical interface to the outside world [3]. These devices have embedded on-board sensors (such as accelerometers, WiFi, and GPSs) that can provide valuable information for investigating users' needs and behavioral patterns.

Similarly, computers that are embedded in vehicles are capable of collecting valuable sensor data that can be accessed by smart phones through the use of On-Board Diagnostics (OBD) sensors. This paper describes a prototype of a mobile computing platform that provides access to vehicles' sensors by using smart phones and tablets, without compromising these devices' security. Data such as speed, engine RPM, fuel consumption, GPS locations, etc. are collected from moving vehicles by using a WiFi On-Board Diagnostics (OBD) sensor, and then back hauled to a remote server for both real-time and offline analysis[4]. We describe the design and implementation details of our platform, for which we developed a library for in-vehicle sensor access and created a non-relational database for scalable backend data storage. We propose that our data collection and visualization tools.

Due to the fast development of ICT including smart phone, Internet, computer and wireless communication, the vehicle industry can be revolutionized and shifted to a new era [5]. In this paper, we introduce the concept of cloud computing enabled real time vehicle services with special focus on customized services like healthcare, resource sharing, parking and dining etc. A three-tier V-Cloud architecture is proposed with detailed explanation about each sub-layer. In certain area when parking is a problem, drivers will prefer to pay a little investment on their navigator or smart phone so that they can easily find a suitable parking lot within a short time. More importantly, when the car is stopping with large amount of resources like memory/flash, power and computing capability, it can rent such resources to other vehicle users who is in need and is willing to pay some expense. In summary, the three tier V-Cloud architecture can provide some innovative and real time services based on cloud computing techniques. It is worth noting that BASN with context-aware reasoning and knowledge processing techniques can largely improve drivers' safety, comfort and convenience. Traditionally, the vehicle has been the extension of the man's ambulatory system, docile to the driver's commands. Recent advances in communications, controls and embedded systems have changed this model, paving the way to the Intelligent Vehicle Grid. The car is now a formidable sensor platform, absorbing information from the environment (and from other cars) and feeding it to drivers and infrastructure to assist in safe navigation, pollution control and traffic management [6].

The next step in this evolution is just around the corner: the Internet of Autonomous Vehicles. Pioneered by the Google car, the Internet of Vehicles will be a distributed transport fabric capable to make its own decisions about driving customers to their destinations. Like other important instantiations of the Internet of Things (e.g., the smart building), the Internet of Vehicles will have communications, storage, intelligence, and learning capabilities to anticipate the customers' intentions. The concept that will help transition to the Internet of Vehicles is the Vehicular Cloud, the equivalent of Internet cloud for vehicles, providing all the services required by the autonomous vehicles. In this article, we discuss the evolution from Intelligent Vehicle Grid to Autonomous, Internet-connected Vehicles, and Vehicular Cloud. The urban fleet of vehicles is evolving from a collection of sensor platforms to the Internet of Autonomous Vehicles [7].

Like other instantiations of the Internet of Things, the Internet of Vehicles will have communications, storage, intelligence and learning capabilities to anticipate the customers' intentions. This article claims that the Vehicular Cloud, the equivalent of Internet Cloud for vehicles, will be the core system environment that makes the evolution possible and that the autonomous driving will be the major beneficiary in the cloud architecture. The use of Sensor-Cloud architecture in the context of several applications [8].

The Sensor-Cloud architecture enables the sensor data to be categorized, stored, and processed in such a way that it becomes cost-effective, timely available, and easily accessible. Earlier, most WSN systems which were included to several controlling/monitoring schemes were closed in nature, zero, or less interoperability, specific application oriented and non extensible. However, integrating the existing sensors with cloud will enable an open, extensible, scalable, interoperable, and easy to use, re constructible e network of sensors for numerous applications. However, due to the limitations of WSNs in terms of memory, energy, computation, communication, and scalability, efficient management of the large number of WSNs data in these areas is an important issue to deal with. Sensor-Cloud infrastructure is becoming popular that can provide an open, flexible, and reconfigurable platform for several monitoring and controlling applications [9].

## ALGORITHM

### A. Routing table

VANET play an important role in dissemination of traffic and emergency information of traffic among vehicles moving on roads. However, because of high mobility of vehicular nodes, maintenance of routing table generates high network traffic. Routing table contains information about the topology of the network [10]. The proposed systems use the following table with routing table information. Here routing table contain the vehicle entry, vehicle moving and non-moving vehicle information.

Table-1 : Routing table

Vehicle Entry	Vehicle moving	Non-Moving vehicle
	✓	-
	✓	-
	✓	-
	✓	-
	✓	-

**Congestion control**

The main objective of congestion control is to best exploit the available network resources while preventing sustained overloads of network nodes and links. Congestion control mechanisms are essential to maintain the efficient operation of network. Ensuring congestion control within vehicular ad hoc networks address special challenges, due to the characteristic and specificities of such environment such as high dynamic and mobility of nodes, high rate of topology changes, high variability in nodes density and neighborhood, broadcast/geo-cast communication nature. In this system congestion can be identified by non-moving nodes (10 or 15 vehicles) with same point. The sensor node detect the congestion made in the road network.

**Architecture Diagram**

In [Figure- 2], The sensor detector connected in each sensor node in VANET .The microcontrollers used in sensor nodes are ultralow-power microcontrollers to conserve energy.

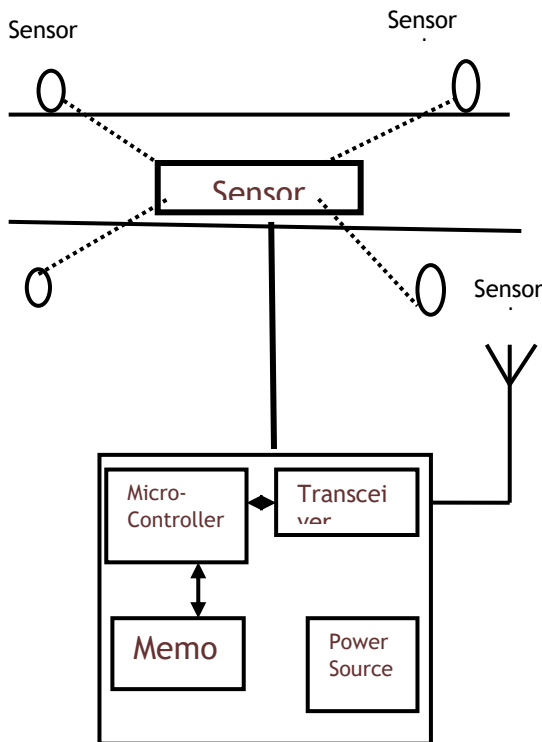


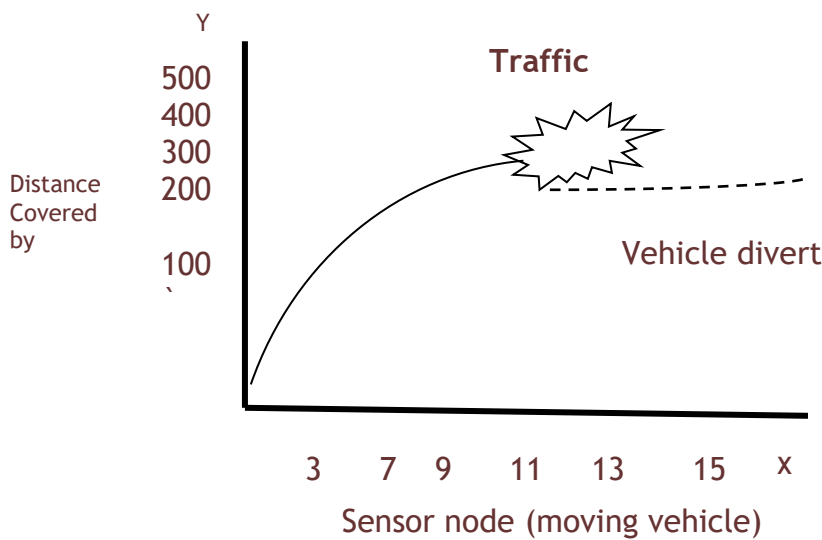
Fig: 2 Architecture Diagram

Figure-3 shows the performance graph between the Distance coverd by with moving vehicle. After some point congestion make occurred in the rode network. When Sensor node Detect the traffic .The vehicle can be diverted in to another way.

Table – 2: Types of Sensor

Sensor node	Description
Accelerometers	To measure vibrations on infrastructure
Fiber-optical sensors	Convert a linear or angular displacement into a signal suitable for recording.
Temperature sensor	To monitor the temperature of the atmosphere
Time domain reflectometer	Converts the travel time of a high frequency

**SIMULATION RESULTS**



**Fig: 3 Performance Graph**

When a vehicle has covered a certain distance if more than ten vehicles are stalled then the cloud server with sensor node notifies the moving vehicle to an alternate path which is as shown in Fig.3

**CONCLUSION**

This paper prevents the traffic Congestion for the use of VANET monitoring system with use of sensor node. Here Cloud computing server act as a base station for to communicate sensor node and travelling vehicle. The sensor detector has the micro controller and transceiver. Finally in this way to determining the traffic Congestion with help of sensor node.

**Future Work**

In the future work based to overcome the difficulty of short life time of battery power , to improve the sensor battery power .

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the technical support given by T.Sarada Kiranmayee, Assistant Professor (OG), Department of Computer Science and Engineering, SRM University, Chennai.

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] J.Victoria Hodge, Simon "Sensor Networks for Condition Monitoring in the Railway Industry:" A Survey", IEEETRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS vol. 16, no. 3, june 2015.
- [2] V. Balas and L. Jain, "World knowledge for sensors and estimators by models and internal models," J. Intell. Fuzzy Syst., vol. 21, no. 1, pp. 79– 88, Apr. 2010.
- [3] E. Berlin and K. van Laerhoven, "Sensor networks for railway monitor-ing: Detecting trains from their distributed vibration footprints," in IEEEInt. Conf. Distrib. Comput. Sens. Syst., Cambridge, MA, USA, 2013,80–87.
- [4] G.sasikala, 2Dr.A.R. Deepti ,"Real Time Services for Cloud Computing Enabled Vehicle Networks",IOSR Journal of Computer Engineering ISSN: 2278-8727Volume 11, Issue 1 (May. - Jun. 2013
- [5] Atif Alamri, Wasai Shadab Ansari, "A Survey on Sensor-Cloud: Architecture, Applications, and Approaches"International Journal of Distributed Sensor Networks Volume 2013
- [6] P. Smith, D. Hughes, K.J. Beven, P. Cross, W. Tych, G. Coulson and G. Blair. "Towards the provision of site specific flood warnings using wireless sensor networks" In Meteorological Applications, Special Issue: Flood Forecasting and Warning, vol.16, no.1, pp.57–64, 2009.
- [7] T. Amirhosein, R. Rouvoy and F. Eliassen, "A Component-based Approach for Service Distribution in Sensor Networks" in Proceedings of the 5th International Workshop on Middleware Tools, Services and Run-Time Support for Sensor Networks, 2010.
- [8] N. B. Priyantha, A. Kansal, M. Goraczko, and F. Zhao, "Tiny WebServices: Design and Implementation of Interoperable and Evolvable Sensor Networks," The 6th ACM Conference on Embedded Network Sensor Systems, Raleigh, NC, USA, pp.253–266, 2008.
- [9] R. Faludi, "Building Wireless Sensor Networks: with ZigBee, XBee, Arduino, and Processing", Sebastopol, CA: O'Reilly Media, 2010.
- [10] W. Dargie, and C. Poellabauer, Fundamentals of wireless sensor networks:theory and practice, pp.168–183, 191–192, 2010.

# DEVELOPMENT OF AN EFFICIENT TECHNIQUE FOR MINING TOP-K CLOSED HIGH UTILITY ITEMSETS

Baby Velayudhan, Sakthivel and Subasree

Dept of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Pampady, Thiruvilwamala, Thrissur, Kerala, INDIA

## ABSTRACT

Discovery of High Utility Itemsets (HUI) or pattern from database is very useful in processing business. Recently proposed two techniques for data mining Closed High Utility Itemsets (CHUIs) and Direct Discovery of High Utility Pattern have been identified and have implemented and analyzed. From the result it is found that CHUI mining is the compact mining technique which reduces the number of itemsets by several orders of magnitude. Still the end user has to select the minimum utility for selecting CHUI which is a tedious process. CHUI refers to discover all itemsets having a utility meeting a user-specified minimum utility threshold  $min\ util$ . User has to set appropriate minimum utility threshold by trial and error, which is a tedious process. If  $min\ util$  is set too low, too many CHUIs will be generated, which may cause the mining process to be very inefficient. On the other hand, if  $min\ util$  is set too high, it is likely that setting  $min\ util$  high, it is likely that no CHUIs will be found. This project addresses the above issues by proposing a new framework for top-k closed high utility itemset mining (TopK-CHUI), where  $k$  is the desired number of CHUIs to be mined. Results show that the user can easily retrieve Closed High Utility Itemset by specifying  $k$ , which is the TopK number of CHUIs. TopK-CHUI is efficient and user friendly. Results on real datasets show that the technique TopK-CHUI is very efficient for the end user. So TopK-CHUI is an user friendly data mining technique.

Published on: 08<sup>th</sup>– August-2016

## KEY WORDS

Closed high utility itemsets, TopK-CHUI mining, High utility itemset, Utility mining, Data mining.

\*Corresponding author: Email: [babyvellayudhan@gmail.com](mailto:babyvellayudhan@gmail.com) ; Tel.: +91 77 36 830775

## INTRODUCTION

Discover the sets of items (itemsets) with high utilities such as high profits is the important requirement of a business. For this utility mining is used. In each transaction, each item has a weight (e.g. unit profit) and appear more than once in each transaction (e.g. purchase quantity), these are utility mining. The important of an itemset is represented by the utility. Utility can be measured in terms of weight, profit, cost, quantity or other information depending on the user preference. Comprehension will be very difficult for the users if the algorithm gives a large number of high utility patterns.

Candidate pattern's Transaction Weighted Utilization (TWU) [2, 5, 7] is the transaction's utility sum. If an itemset's utility is no less than a user-specified minimum utility threshold then it is called a High Utility Itemset (HUI) [2, 4, 5, 10] otherwise, it is called a low utility itemset. Utility mining is an important task and has a wide range of applications such as biomedical applications, website click stream analysis, cross marketing in retail stores and mobile commerce environment.

If a HUI is not a subset of any other HUI then it is said to be maximal [1, 6]. The reason is that without scanning the database the utilities of the subsets of a maximal HUI cannot be known. If a high utility itemset has no proper superset having the same utility then it is said to be closed [1, 6]. For any non-closed high utility itemset  $Y$ ,  $Y$  does not appear in a transaction without its closure  $Z$ . Moreover, the utility (e.g. profit/user preference) of  $Z$  is guaranteed to be higher than the utility of  $Y$ . For these reasons, users are more interested in finding  $Z$  than  $Y$ . Moreover, closed itemsets having high utilities are useful in many applications. For example, in market basket analysis,  $Z$  is the closure of  $Y$  means that no customer purchases  $Y$  without its closure  $Z$ . Thus, when a customer purchases  $Y$ , the retailer can recommend  $Z - Y$  to the customer, to maximize profit.

CHUI integrates high utility itemset mining into the concept of closed itemset. Closed High Utility itemset Discovery (CHUD) [1] algorithm mines CHUIs in a depth-first search by using vertical database. CHUD takes as parameter the

abs min utility threshold and a database  $D$ . In practice it is difficult for users to choose an appropriate minimum utility threshold. The output size can be very small or very large depending on the threshold. To discover the itemsets with the highest utilities and precisely control the output size without setting the thresholds, a promising solution is to redefine the task of mining HUIs as mining Top-k High Utility Itemsets (Top-k HUIs) [3].

Instead of specifying the minimum utility threshold, let the users specify  $k$ , i.e., the number of desired itemsets. Alphabet  $k$  represents the number of itemsets that the users want to find whereas choosing the threshold depends primarily on database characteristics, which are often unknown to users so setting  $k$  is more intuitive than setting the threshold. For many applications it is very desirable that using a parameter  $k$  instead of the  $\text{min\_util}$  threshold. For example, Top-k HUI mining serves as a promising solution for users who desire to know "What are the top-k sets of products (i.e., itemsets) that contribute the highest profits to the company?" and "How to efficiently find these itemsets without setting the  $\text{min\_util}$  threshold?", to analyze customer purchase behavior.

In top-k HUI mining The  $\text{min\_util}$  threshold is not given in advance. The search space can be efficiently pruned by the algorithms by using a given  $\text{min\_util}$  threshold in traditional HUI mining. However, no  $\text{min\_util}$  threshold is provided in advance in the scenario of top-k HUI mining. Therefore, the designed algorithm has to gradually raise the threshold to prune the search space by setting the minimum utility threshold initially to zero [3]. Such a threshold is an internal parameter of the designed algorithm and is called the border minimum utility threshold  $\text{min\_utilKBorder}$  in TopK-CHUI technique.

## BACKGROUND AND PROBLEM DEFINITION

### Problem Definition

CHUI is a compact and lossless representation of HUIs. From the analyses conducted it is found that CHUI is the best techniques which retrieves very less number of HUIs that is why CHUI is compact and lossless representation of HUIs. For mining CHUIs user has to produce the  $\text{min\_util}$  and the dataset, but prediction of  $\text{min\_util}$  is a tedious job for the user. User has to follow trial and error method for selecting the value of  $\text{min\_util}$ . If the user is giving  $\text{min\_util}$  near to zero then mining will produce too many CHUIs and comprehension of these are very difficult. If the user is giving  $\text{min\_util}$  as high value then there may not be any itemset to display. So for using the mining technique efficiently user must know the highest utility value available in the dataset which is not easy. If the user knows the highest value then he can get the CHUIs by giving percentage of highest value as  $\text{min\_util}$ . At the same time if he wants to see the highest  $k$  CHUIs as the output then it is easy for the user to say the value of  $k$ .

The problem statement is: Given a transactional database  $D$  and the desired number of CHUIs  $k$ , the problem of top-k closed high utility itemsets mining is to discover all the itemsets having a utility no less than  $\delta$  in  $D$ .

### Related Works

This subsection introduces related works about top-k closed high utility itemset mining, including closed high utility item-set mining and top-k high utility itemset mining.

1) Closed high utility item discovery: Closed High utility itemset incorporates the concept of closed itemset with high utility itemset mining. Closure on the utility of itemsets can define as a high utility itemset is said to be closed if it has no proper superset having the same utility. In real dataset it is unlikely to achieve a high reduction of the number of extracted itemsets since not many itemsets have exactly the same utility as their supersets. The join order between the closed constraint and the utility constraint is defined as a) Mine all the high utility itemsets first and then apply the closed constraint. b) Mine all the closed itemsets first and then apply the utility constraint. The two constraints can be applied in any order during the mining process.

The closed high utility itemset discovery technique considers vertical database and mines CHUIs in a depth-first search. CHUD takes as parameter a database  $D$  and the abs min utility threshold. CHUD convert  $D$  into a vertical database during the first scan of  $D$ . At the same time, CHUD calculates TWU [2, 5, 7] of items and computes the transaction utility for each transaction  $TR$ . When a transaction is retrieved, its transaction utility and  $Tid$ (transaction identifier) are loaded into a GTU (global TU-Table). If an item's estimated utility (e.g. its TWU) is no less than abs min utility then it is called a promising item. After the database scan, promising items are collected into an ordered list  $O$ , sorted according to a fixed order such as increasing order of support. Since supersets of unpromising items are not CHUIs, only promising items are kept in  $O$ .

CHUD uses Itemset - TidsetpairTree(IT-Tree) [8] to find CHUIs. Each node consists of an itemset  $X$ , Tidset  $g(X)$ , two ordered sets of items  $\text{PREV-SET}(X)$  and  $\text{POST-SET}(X)$  and estimated utility. The TU-Table stores the transaction utility with transaction id. In a recursive manner CHUD generates candidates, starts with candidates containing a single promising item and recursively joins items to them to form larger candidates. This is done by using the total order  $<$ , the complete set of itemsets is divided into  $n$  non-overlapping subspaces, where the  $k^{\text{th}}$  subspace is the set of itemsets containing the item  $a_k$  but no item  $a_i < a_k$ . For each item  $a_k$  belongs to  $O$ , CHUD creates a node  $N(fa_k, g)$  and puts items  $a_i$  to  $a_{k-1}$  into  $\text{PREV-SET}(a_k)$  and items  $a_{k+1}$  to  $a_n$  into  $\text{POST-SET}(a_k)$ .

2) Top-k high utility itemset mining: High utility itemset mining algorithms can be generally categorized into two types: two-phase and one-phase algorithms. The main characteristic of two-phase algorithms is that they consist of two phases. In the first phase, they generate a set of candidates that are potential high utility itemsets. In the second phase, they calculate the exact utility of each candidate found in the first phase to identify high utility itemsets.

Let C be the set of candidates produced in Phase I. Candidates in C are sorted in descending order of their estimated utilities. Thus, candidates with higher estimated utility values will be considered before those having lower estimated utility values. During the phase II, if the utility of a newly considered CHUI X is larger than  $\min_{util_{kBorder}}$ , X and  $EU(X)$  are inserted into a min-heap structure named TopK-CHUI-List. Then,  $\min_{util_{kBorder}}$  is raised to the utility of the k-th HUI in TopK-CHUI-List, and HUIs having a utility lower than  $\min_{util_{kBorder}}$  are removed from TopK-CHUI-List. If the estimated utility of the current candidate Y is less than the raised  $\min_{util_{kBorder}}$ , Y and the remaining candidates do not need to be considered any more because the upper bounds on their utilities are less than  $\min_{util_{kBorder}}$ . When the algorithm completes, TopK-CHUI-List captures all the Top-k CHUIs in the database.

### TOPK-CHUI TECHNIQUE

In this section, an efficient technique named TopK-CHUI (mining Top-K closed utility itemsets) is proposed for discovering Top-k CHUIs without specifying min util. The strategy used in TopK-CHUI technique is that raising the threshold by sorting exact utility of candidates.

```

ALGORITHM: TopK-CHUI
Input: (1) A database D;
       (2) The number of desired CHUIs k; Output: (1) Top K CHUIs;

01. Set  $\min_{util_{kBorder}}=0$ ; TopKCHUI-List=;
02. Calculate the Transaction Weighted Utility
03. Create utility list of items
04. TopK-CHUI-Search( $\min_{util_{kBorder}}$ , closedset, utilitylist, preset, postset);
05. Output TopKCHUI-List
    
```

Fig. 1. TopK-CHUI

Figure-1 shows the pseudo code of TopK-CHUI. Each time a candidate itemset X is found by the CHUI search procedure, the TopK-CHUI algorithm checks whether its utility value is no less than  $\min_{util_{kBorder}}$ . If utility is less than  $\min_{util_{kBorder}}$ , X is not a TopK-CHUI. Otherwise, X is considered as TopK-CHUI. Each time a TopK-CHUI X is found and its utility is higher than  $\min_{util_{kBorder}}$ , X is added into TopKCHUI-List. If there are less than k itemsets in TopKCHUI-List,  $\min_{util_{kBorder}}$  will not change. Once k itemsets are found and the k-th utility value in TopKCHUI-List is higher than  $\min_{util_{kBorder}}$ ,  $\min_{util_{kBorder}}$  is raised to k-th minimum utility value in TopKCHUI-List. The strategy used in TopK-CHUI Technique is raising the threshold by sorting the exact utilities of the candidates. If the utility of a newly considered CHUI X is larger than  $\min_{util_{kBorder}}$ , X is added into TopKCHUI-List.

Experiments were performed on a computer with a 1.7GHz Intel i3 processor and 4GB of memory, running Windows 8.1. Algorithms are implemented in Java. Real datasets were used in the experiments. Datasets were acquired from [archive.ics.uci.edu/ml/datasets.html](http://archive.ics.uci.edu/ml/datasets.html). It includes real datasets with synthetic utility values. Internal utility values generated using a uniform distribution in [9]. External utility values generated using gaussian (normal) distribution.

Table: 1. Dataset Characteristics

Dataset	#Transactions	#Distinct items	Avg. trans. Length
Accident	340183	468	33.8
BMS	59601	497	4.8
Connect	67557	129	43
Foodmart	4141	1559	4.4
Mushroom	8124	119	23

Table- 1 shows characteristics of the datasets. In real life scenarios there are three kinds of datasets that are commonly encountered: (1) dataset containing long transactions, (2) dense dataset, (3) sparse dataset. In the experiments have done by using real-life datasets BMS, Mushroom, Foodmart to respectively represent the above three real cases.



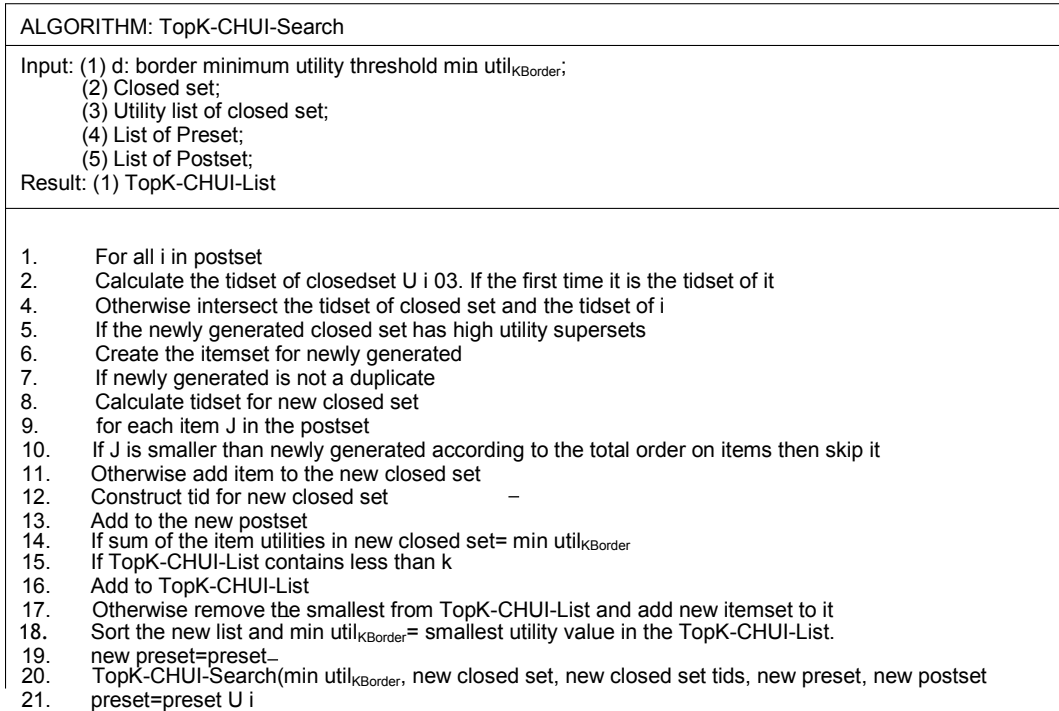


Fig.2. TopK-CHUI-Search

## RESULT AND DISCUSSION

Figure- 3(a) shows the runtime of the algorithm on the datasets with varied k respectively. Figure- 3(b) shows the minimum utility selected for the algorithm on the datasets with varied k respectively. Figure- 3(c) shows the memory used by the algorithm on the datasets with varied k respectively. Execution time varies according to the value of k which is given by the user as input. If user wants only Top10 itemsets then the execution time will be less. Execution time is high when the user uses large datasets, for example when k=10 the execution time is high for Accidents and BMS datasets.

In dense datasets like Mushroom memory usage is high but the execution time is less. For large dataset like Accidents, BMS and connect memory usage is less but the execution time is more. Minimum utility percentage is selected according to the utility in the dataset available. For example BMS have highest utility value. Memory usage is high for Mushroom dataset. Table- 2, Table- 3 and Table- 4 shows the runtime, minimum utility and memory usage on the datasets with varied k respectively. TopK-CHUI technique cannot compare with CHUI because CHUI takes input as the minimum utility but TopK-CHUI as the number of CHUI itemsets needed. So analysis have been done by running the technique with different real datasets for different values for k. It is also found from the analysis that if we are running the CHUI-Miner[1] for the minimum utility selected for k=10 then the execution time is less for TopK-CHUI technique

Table: 2. Execution Time (Seconds)

Database	K=10	K=100	k=1000	k=10000
Accidents	466.17	623.758	694.858	1001.56
BMS	365.54	1008.368	875.521	1290.21
Connect	55.794	60.511	56.882	168.916
Foodmart	3.12	3.727	4.094	6.335
Mushroom	17.97	29.265	35.108	74.034

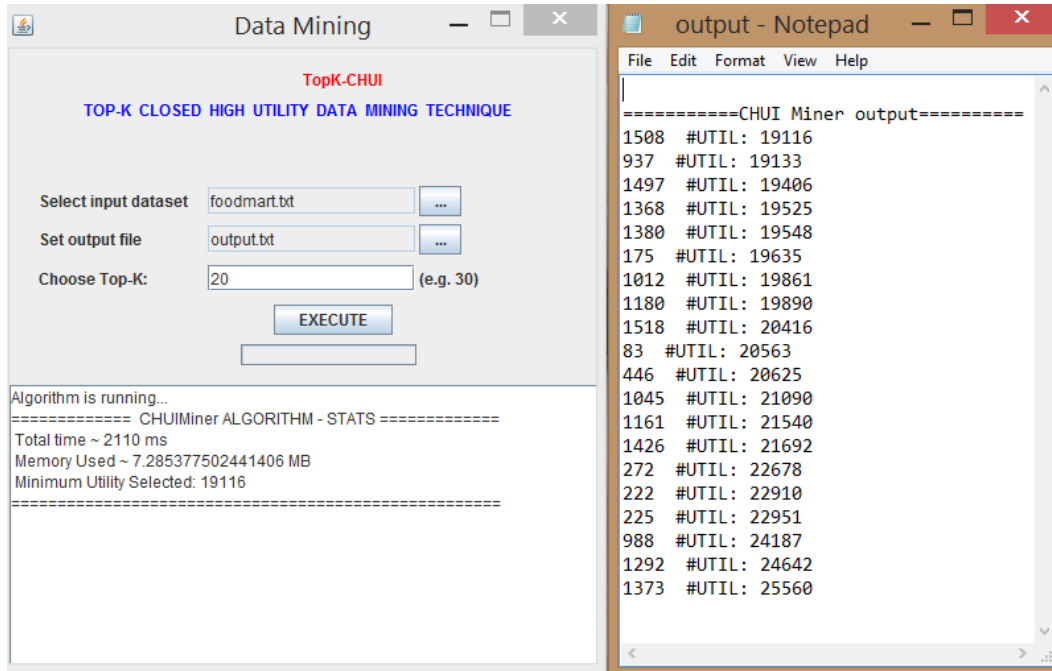


Fig. 3. Sample Output

Table 3. Minimum Utility (%)

Database	K=10	K=100	k=1000	k=10000
Accidents	93.37	82.54	68.34	52.15
BMS	20.53	9.55	5.47	3.47
Connect	98.3	94.94	88.33	71.26
Foodmart	80.69	57.9	24.53	0
Mushroom	73.76	61.08	34.68	12.07

Table 2. Memory Used (MB)

Database	K=10	K=100	k=1000	k=10000
Accidents	7.29	7.29	11.52	11.25
BMS	16.34	14.98	13.29	17.78
Connect	7.1	7.56	17.16	22.52
Foodmart	9.68	7.63	11.66	8.45
Mushroom	39.13	35.42	33.34	22.91

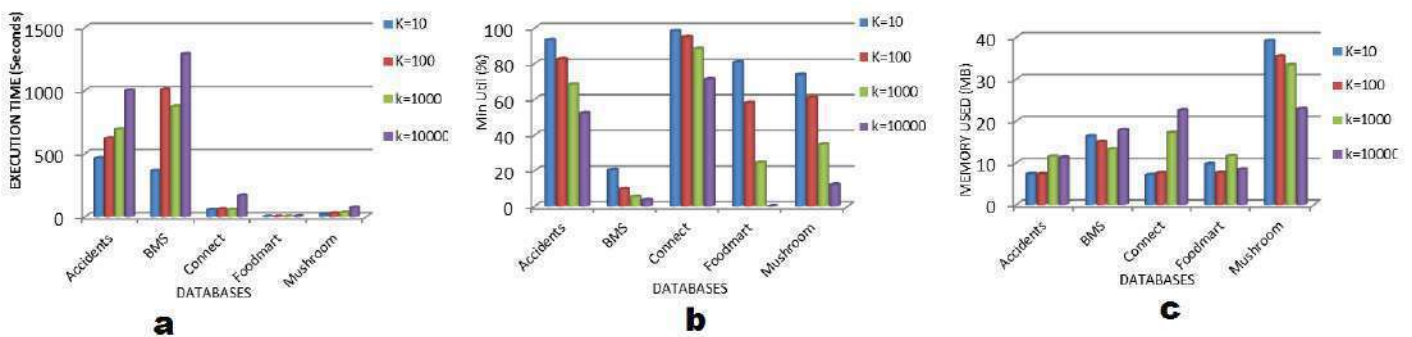


Fig. 4. a) Execution time, b) Minimum Utility and c) Memory used

## CONCLUSION

The problem of top-k closed high utility itemsets mining, where k is the desired number of closed high utility itemsets to be mined is discussed. An efficient technique TopK-CHUI is proposed for mining such itemsets without setting minimum utility threshold. The strategy used with this is raising the threshold by sorting exact utilities of candidates. It is user friendly because user need not to do trial error method for determining the value of minimum utility, end user can directly give the value of k for to k closed high utility itemsets. Evaluations on different types of real datasets show that the proposed algorithm has good scalability on large datasets. It can be incorporated with other mining techniques such as top-k high utility web access patterns and top-k mobile high utility sequential patterns.

## CONFLICT OF INTERESTS

The authors declare that they have no conflicts of interest.

## ACKNOWLEDGEMENT

None.

## FINANCIAL DISCLOSURE

NIL

## REFERENCES

- [1] Vincent S Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, Philip S Yu.[ 2016] Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets, *IEEE Trans. Knowl. Data Eng.*, 27( 3): 726 -739.
- [2] Junqiang Liu, Ke Wang, and Benjamin C.M. Fung. [2016] Mining High Utility Patterns in One Phase without Generating Candidates, *IEEE Trans. Knowl. Data Eng*, 28( 5):1245- 1257
- [3] Vincent S Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S Yu.[ 2016] Efficient Algorithms for Mining Top-K High Utility Itemsets, *IEEE Trans. Knowl. Data Eng*, 28( 1): 54- 67.
- [4] GC Lan, TP Hong, VS Tseng.[ 2014.]An efficient projection based indexing approach for mining high utility itemsets, *KAIS*, 38( 1): 85-107
- [5] VS Tseng, CW Wu, BE Shie, PS Yu.[ 2010] UP- Growth: An efficient algorithm for high utility itemset mining,” in Proc. *ACMSIGKDD Int Conf Knowl Discov Data Mining* , 253-262.
- [6] CW Wu, P Fournier-Viger, PS Yu, VS Tseng.[ 2011]Efficient mining of a concise and lossless representation of high utility itemsets, in *Proc. IEEE Int Conf Data Mining*, 824-833.
- [7] VS Tseng, BESHie, CW Wu, PS Yu.[ 2013] Efficient algorithms for mining high utility itemsets from transactional databases,*IEEE TKDE*, 25( 8):1772-1786,
- [8] CW Lin, TP Hong, WH Lu.[2011] An effective tree structure for mining high utility itemsets, *Expert Syst Appl*, 38( 6): 7419-7424
- [9] R.Jaya and Rajan. C,[2016]A Study On Data Mining Techniques, Methods, Tools And Applications In Various Industries, *International Journal on Concurrent Applied Research in Engineering and Management*, vol.1, No.4, pp. 14-24
- [10] M Liu and J Qu.[ 2012] Mining high utility itemsets without candidate generation, in *CIKM. ACM* , 5564.

## TOUCH DOWN AVIATION ANALYSIS

S. Sageengrana\*, G.Tamilmani, K. Rajathi<sup>3</sup>, P. Karthikeyan<sup>1</sup>Dept of Computer Science and Engineering, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, Chennai, INDIA

## ABSTRACT

**Aims:**Energy harvesting by means of any source of energy become essential for developing countries. Natural and renewable energy could provide clean environment to certain areas like airstrip ( Run way)Materials and methods:Mostly energy harvesting is achieved through wind source become seasonal and will have huge uncertainty in power quality. To overcome and to produce power without uncertainty with enhanced power quality, we would like to present a existing concept for a newer area with minimization. **Results:**Enormous amount of wind is generated during takeoff and landing of an aircraft .In the aim of conserving the wind energy in those areas we have proposed to use low power, lower size wind turbine on both sides of the runway to require power. **Conclusion:**The power generation could provide clean energy and an important parameter called touchdown point of an air craft on the runway (air strip).

Published on: 08<sup>th</sup>– August-2016

## KEY WORDS

Formal verification, static  
analysis, Android, Inter-App  
vulnerabilities

\*Corresponding author: Email: [sageengrana@velhightech.com](mailto:sageengrana@velhightech.com)

## INTRODUCTION

Global Mobile app markets are creating a fundamental model shift in the way software is delivered to the end users.

The increase in demand for power has become more in day to day life. Rising in demand for electricity made the advancement in finding the alternative generation of power resources for the future generations. Energy harvesting system has become more essential for all the developing countries. There are many alternative energy resources like solar, wind and tidal which has become seasonal and during the uncertainties it is much more difficult to generate power with all enhanced power quality.

Electricity use can vary dramatically on short and medium time frames, largely dependent on weather patterns. Energy Demand Management (EDM) which is also known as Demand Side Management (DSM) is an alternative modification made in generating the energy resources during the peak hours. A newer application for DSM can be done in the airport runway system in balancing the intermittent generation from other renewable energy resources when there is a seasonal issue. EDM activities attempt to bring the electricity demand and supply closer to the operational needs for a particular organization and also for the consumers nearby the society.

Reducing energy demand is contrary to energy suppliers, government and private sectors are themselves trending to generate the electricity for the future demands that will increase the efficiency of energy consumption. In this concept the metro airport is making use of an alternative resource in generating electricity for their usage and also to supply power for other residential and commercial sectors which are located nearby. The main aim for this type of alternative resource is to reduce the demand in power and also to use the resources which can be made from easily available physical quantities. The type of generation should also provide a clean environment by not affecting any harm for living creatures.

Utility activities that influence customer use of electricity encompasses the planning, implementation and monitoring if activities designed to encourage consumers to change their electricity usage patterns.

## AEROMACS APPROACH

The existing system uses a normal power supply for the runway power management and more energy is required to make the lighting system effective during night times. The communication with the C- band is at higher rates but the touchdown point is done manually through AeroMACS. There are more difficulties in analyzing the speed of the aircraft between the runways from the point of touchdown. There are chances of angle deviation from touchdown point which leads to overshooting of runways. Improper monitoring of the taxi-ways could lead to accidents and collision of aircraft.

## AEROMACS COMMUNICATION

AeroMACS is based on a specific commercial profile of the Institute of Electrical and Electronics Engineers (IEEE) 802.16 standard known as Wireless Worldwide Interoperability for Microwave Access or WiMAX. To help increase the capacity and efficiency of the nation's airports, a secure wide band wireless communications system is proposed for use on the airport surface.

As the communications, navigation, and surveillance (CNS) facilities for air traffic management (ATM) at an airport grow in number and complexity, the need for communications network connectivity and data capacity increases. Over time, CNS infrastructure ages and requires more extensive and expensive monitoring, maintenance, repair or replacement. Airport construction and unexpected equipment outages also require temporary communications alternatives.

## RUNWAY POWER GENERATION

Runway power generation is done with the normal distribution generation supply from the substation, feeders and transmission lines. Airport runway airstrips need more energy during the night time lighting for the aircraft landing and takeoff. When more power is consumed then there will be energy demand and crisis for the commercial and residential purposes. There are so many alternative methods for generation of power but those have become seasonal with the weather conditions. More advancement must be made to have a alternative sources of energy.

The communication with the C- band is at higher rates but the touchdown point is done manually through AeroMACS. There are more difficulties in analyzing the speed of the aircraft between the runways from the point of touchdown. There are chances of angle deviation from touchdown point which leads to overshooting of runways. Improper monitoring of the taxi-ways could lead to accidents and collision of aircraft. The runway lighting system is given with normal power supply by which more demand for the energy is needed. There must be some improvement made in this existing system so that there will be lesser number of accidents and to find the optimal solutions for the development of alternative resources.

## PROPOSED SYSTEM

The proposed concept Low Power Wind Turbine can installed on both sides of runway to acquire power during aircraft landing and take off.

Generally on run way air crafts movements will be around 400 KMPH. During the movement very great air velocity will occur on runway, the same can be enough to drive small micro wind turbines to produce power.

The proposed concept will have wind turbine, step down transformer, rectifier unit, ultra capacitor, lighting system and a on line wireless transmitter.

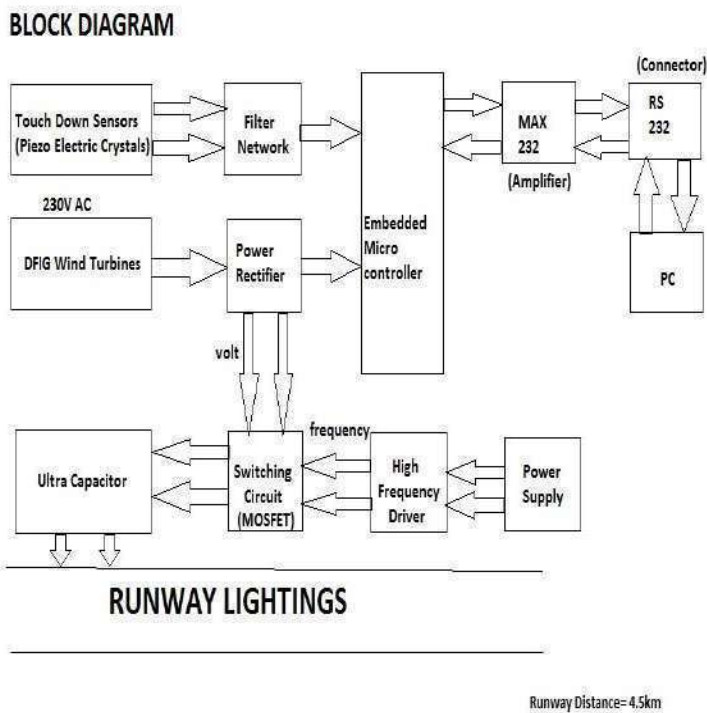


Fig. 1: Block diagram

The LPWT will be a small double fed induction generator (DFIG), which can produce 230v output of AC.

The output of the DFIG can be connected to a step down transformer to reduce voltage and will be fed to a full wave bridge rectifier Ac to Dc conversion.

Output of the rectifier will be applied to ultra capacitor called super capacitor to store and utilize for airport applications like light and signalling.

### PIEZO ELECTRIC SENSOR

A piezoelectric sensor is a device that uses the piezoelectric effect, to measure changes in pressure, acceleration, temperature, strain, or force by converting them to an electrical charge. The prefix piezo- is Greek for 'press' or 'squeeze'.



Fig.2: Piezo Electric Sensors

## SIGNAL CONDITIONER

Signal conditioners are essential to improve field-received signals. Signal conditioner job starts from simple amplification to protection. For our circuit input will be 0v to 1000mv and must be amplified to 5volts. Essentially, we need a signal conditioner to amplify the IR detector output. Remove shell voltage and atmospheric pollution. To remove unwanted frequency.

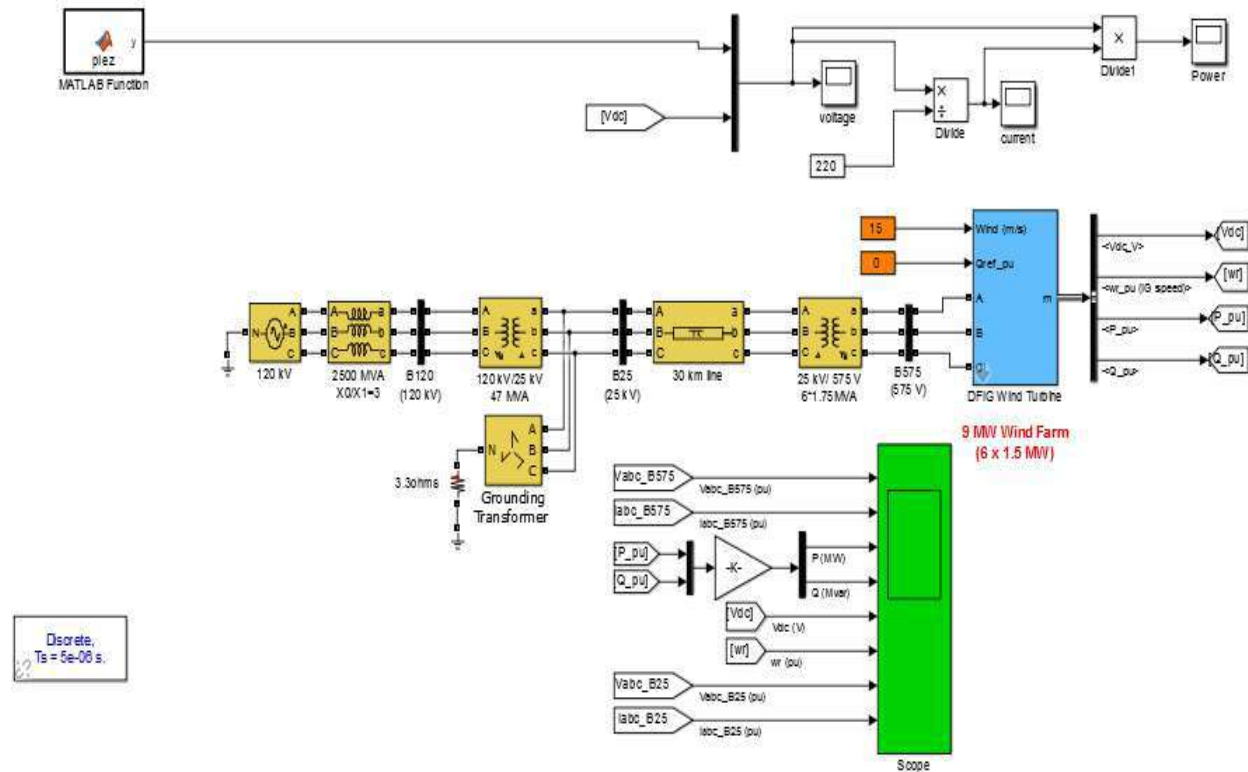


Fig.3: The circuit

voltage and current of interconnect link B575

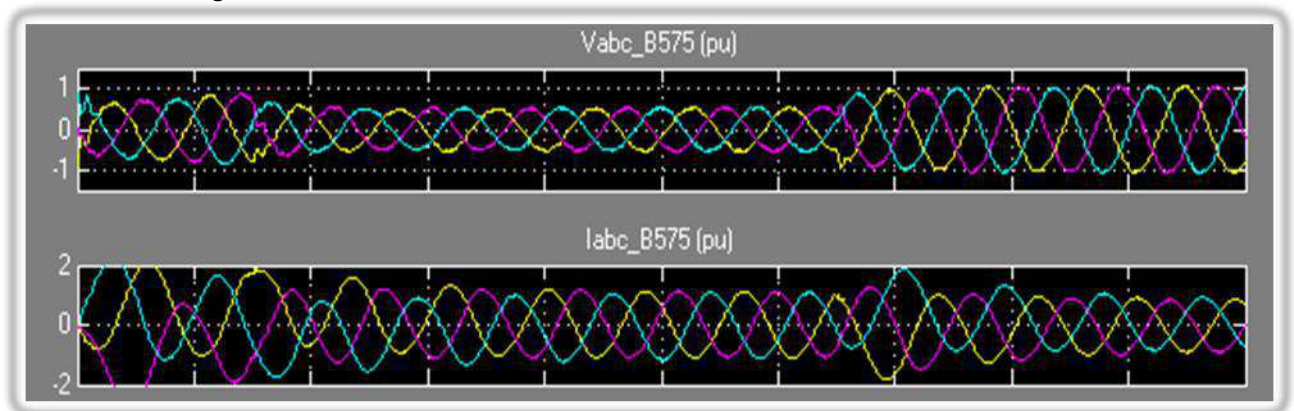


Fig.4: voltage and current of interconnect link B25

## EMBEDDED MICROCONTROLLER

To perform the various operations and conversions required to switch, control and monitor the devices a processor is needed.

Industrial advantages in power electronics like built in ADC, RAM, ROM, ports, USART, DAC. This leads to lesser space occupation by the circuit and also the speed of embedded controllers are more compared to other processors.

The embedded controller selected for this project is PIC16F877A due to its various features.

The PIC 16F877A has five serial ports namely A, B, C, D and E. It has five parallel ports namely:

PSP (Parallel Slave Port 8 bit wide)

SSP (Serial Synchronous Port)

MSP (Master Serial Synchronous Port)

I<sup>2</sup>C (Inter Integrated Circuit)

SPI (Serial Peripheral Interface)

12 bit 10 channel PSP (Parallel Slave Port) -12 bit accuracy

Sleep mode processor

Built in temperature sensor

Built in RAM and EPROM

## RESULTS

DFIG wind turbine is supplied with the input wind (m/s) which makes the turbine rotate. Another parameter is piezo MATLAB function where the piezo crystals are based on the pressure and impact which is coded according to the automation of touchdown point. The turbine voltages generated in the piezo effect are combined in the channel. The total voltage and power from both turbine and piezo effect can be separately viewed in the scope output.

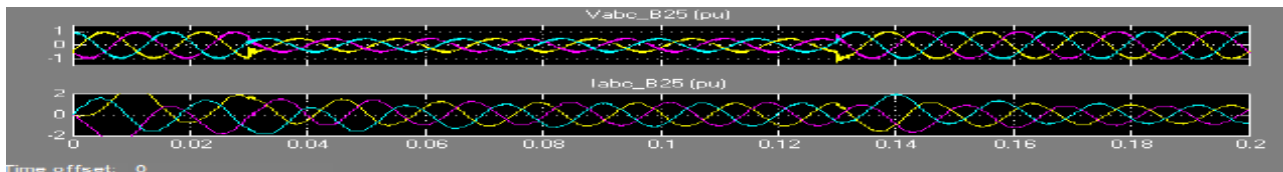


Fig.5: Voltage and speed of DFIG

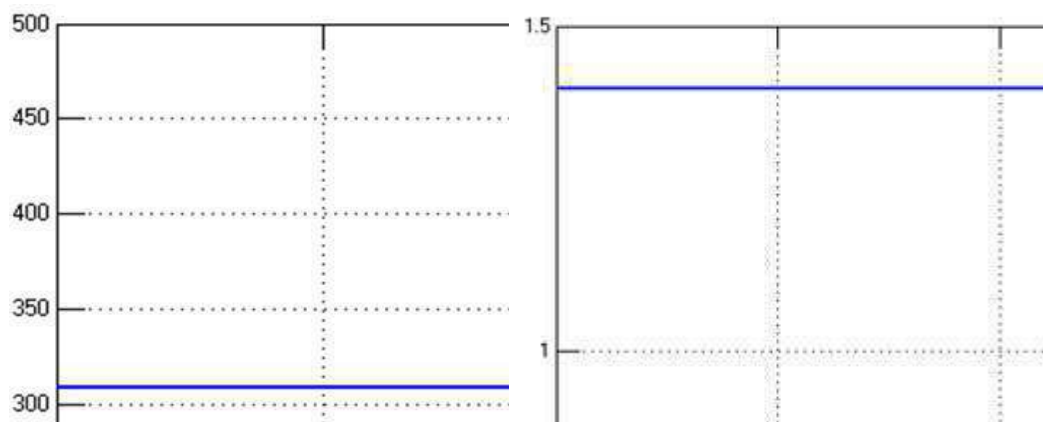


Fig.6: Generated Voltage Generated Current





Fig.7: Generated Power

Hence large amount of power can be generated from the runway which can be utilized for taxi-ways and the excess power generated can be connected to the feeder of transmission lines for onward use in commercial/ residential purpose.

## CONCLUSION

This project could have a complete solution for the existing system with the improved efficiency and automation for the social and safety welfare of the country. It uses touchdown sensors to automate the aircraft landing and power is generated with the pressure and impact from the vibration during landing. The implementation of this system could report analyzing the speed of the aircraft from the point of touchdown which is automated.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Biezd DJ. The propulsive- only flight control problem.[1991] Aerospace and electronics Conference NAECON, *Proceedings of IEEE*, vol.2
- [2] Tom Tucker.[ 1999]Touchdown: The Development of Propulsion Controlled Aircraft At NASA Dryden, Monographs in Aerospace History,
- [3] Mandatory Instructions Signs. [2004]ICAO Recommended Airport Signs, Runway and Taxiway Markings, Case study
- [4] Airport Authority of India.[2006]Manual Of Aeronautical Information Services”, Technical information with case study, Jan 2006.
- [5] Erlich I, Wilch M, and Feltes C.[ 2007]Reactive Power Generation by DFIG based Wind Farms with AC grid Connection”, Power Electronics and application, *European conference IEEE transaction*, Sep
- [6] Provan CA and Atkins SC. [2011]Tactical Airport Configuration Management, Integrated communication, Navigation and surveillance conference, *IEEE conference*
- [7] Eduaro.SAyra. [2011]Risk Analysis of Runway Overrun Excursions at Landing: A Case Study” Department of Statistics and Operations Research, Rey Juan Carlos University,
- [8] De Souza Ribeiro, Gomas De Matos J.[ 2011]] Isolated Micro-grids with renewable Hybrid Generation”, *IEEE Transaction on* 2(1)
- [9] Jin Tian and Tingdi Zhao.[2012]Controllability – involved risk assessment model for carrier landing of aircraft, *IEEE Proceedings*
- [10] Al-AmeenSalih, Zhahir A, Ahmad MT.[ 2013]Modeling and Simulation of a high Accurate Ground based positioning and Landing System”, Space Science and Communication, *IEEE International Conference*
- [11] Aircraft Accident Investigation Report, “NTSC-(National Transport Safety Committee)” case study of transport and safety, 2013
- [12] Abdel- Geliel M, Anany.M.[2014] Modeling and Simulation of Hybrid Power Generation System of Wind Turbine, Micro turbine And Solar Heater Cells”, Control and Automation (ICCA), 11th *IEEE Conference*, June 2014.
- [13] ItikaTandon and Alok Kumar.[ 2014] A unique step Towards Generation of electricity via new Methodology.

- [14] L.GarlinDelphina and VPS Naidu. Detection of Airport Runway Edges using Line Detection Techniques, National Aerospace Laboratories Bangalore.
- [15] Paola Pulini, and Simon Plass.[ 2014] AeroMACS Evolution Analysis during landing, takeoff and approach phases, German Aerospace Center, July
- [16] Report of Federal Aviation Administration, Analysis of aircraft touchdown point and the associated uncertainty. Jan 2008.
- [17] Wind Engineering Retrospect and Prospect papers of ninth International Conference, 1995 vol. 1, Theme- 2 Extreme Winds, Study of extreme wind estimation procedures. ISBN 81-224-0714-5.
- [18] James F. Shackelford and Mandanapalli K. [2005]Murlidhara, Introduction toMaterial Science for Engineers.;Pearson Education Inc- United States, ch 15 ,pp. 507.

## A STUDY OF PROTOCOL SECURITY -CLOUD ARCHITECTURE

Sabapathi\*, Muniyappan, Danu Senthil

Dept. of CSE, Vel Tech High tech DR.RR&amp;DR.SR Engineering College, INDIA

## ABSTRACT

**Aims:** Protocol security which is important concern in the network security, which ensures the security and integrity of data transmission over the internet. Secure network data from any illegitimate attempt to extract content of the data and Cloud Computing is the advance computing technology for the internet users. **Materials and methods:** We can enhance our system without modify or changing our resources by the internet over cloud technology. So, all the resources we can get through cloud system. Our data should be transmit in secure channel. **Results:** This paper focus on Protocol security in private cloud system deployment and security against POODLE (Discovered by GOOGLE TEAM at OCT 2014).

Published on: 10<sup>th</sup>– August-2016

## KEY WORDS

Private Cloud, POODLE, ECC and DH, Open Stack

\*Corresponding author: Email: [sabapathi2000@gmail.com](mailto:sabapathi2000@gmail.com) Tel.: +91-95008-25476; Fax: +91-44-26840 249

## INTRODUCTION

Protocol security, Protocol it's the set of rules which act as channel for connectivity for the data transferring. So, it's a important concern on connection setup over network. For secure communication, in cloud system, it's much more important. SSL (Secure Socket layer), is a protocol used to transfer the private encrypted data and deliver through the secure communication. Two sub protocols were exists in SSL, they are Record and Handshake protocol. When the data transfer, then the format is called Record protocol, and the exchange between client and server are done using the record protocol, which refers as handshake protocol. Cloud computing, Resource centric technology; we can access the resources over the network. Through Cloud computing, Centralization of infrastructure with low cost, we can increase the Peak –load capacity, dynamic allocation of CPU, storage and bandwidth. Virtualization, running multiple Operating System on single physical computer. So all the resources, we can get via Cloud technology. If we get software like Ms –Office from the cloud then it's called Software As A Service (SAAS). If Operating system provide service Platform As A Service (PAAS), If it's resources provide service like storage then it's called Infrastructure As A Service (IAAS). POODLE (Padding Oracle On Downgraded Legacy Encryption), it's Protocol secure break vulnerability on SSL V3. POODLE it's a Man –In – The Middle attack, which is exploit and allows attacker to read cipher text So Against POODLE we use TLS\_FALLBACK\_SCSV, It's a tool using TLS 1.2. OpenStack, it's a Cloud Operating system and its set of software tools for building and managing cloud computing platforms for public and private. Actually OpenStack it's a open source software. HTTPS/TLS 1.2 connection setup, Private cloud deployment using OpenStack, Elliptic Curve Cryptography and Diiffie Hellman using for short key generation and secure connection establishment.

## RELATED WORK-EXISTING PROBLEM

We should analyses the communication channel eavesdrops Https/SSL channel has lot of Man in the Middle Attacks are possible. If I connect to the banking via the Wi Fi, we may think our connection is secure but the Wi-Fi hotspot might connect to bank behalf of ourselves. Wifi hotspot sneakily redirect to an Http page and connect to the bank, eavesdrop may threat our transaction details. other way “Homographic similar https address” and SSL doesn't have perfect forward secrecy SSL connection Andsome establishes two main phases, handshake and secure data transfer phases. This paper should not be against of SSL but we should aware of associated problems with SSL. Private Cloud setup OpenStack can explore with help of the Mass Phishing, brute force and automated exploitation tools there are lot of open source cloud solutions are available to build private cloud with IaaS cloud

service layer. Vorasetal. [1] Devises a set of criteria to evaluate and compare most common open source IaaS cloud solutions. Mahjoubet al. [2] compares the open source technologies to help customers to choose the best cloud offer of open source. Technologies. Most common open source cloud computing platforms are scalable, provide IaaS, and support dynamic plat-Form, Xen virtualization technology, Linux operating system and Java [3],[4]. However, they have different purposes. For example, Eucalyptus [5] fits well to build a public cloud services (IaaS) with homogeneous pool of hypervisors, whileOpenNebula [6] fits well for building private/hybrid cloud with heterogeneous virtualization platforms [7]. Many authors have analyzed the cloud security challenges and propose methodologies for security evaluation of theCloud solutions. Cloud Security Alliance (CSA) announce Cloud Control Matrix Version 1.3 [8] which can assistthe potential cloud customers to assess the overall security risk of a cloud service providers classifying the security Controls according to cloud service layer and architecture. Methodology for security evaluation of on-premise systemsand cloud computing based on ISO 27001:2005 [9] is pro-posed in [10]. The authors in [4] evaluate ISO 27001:2005Control objective importance for on-premise and the three cloud service layers IaaS, PaaS (Platform as a Service) andSaaS (Software as a Service). International Organization for Standardization (ISO) is developing new guidelines ISO/IECWDTS 27017 [11] that will recommend relevant security controls for information security management system(ISMS) implementation in cloud computing. Eucalyptus and CloudStack [12] have integrated the maximum security levelin front of Open Nebula and OpenStack open source cloud solutions [13].

## SYSTEM AND MODELS

Our system model involves cloud service provider which includes cloud system administrators, tenant administrators n (or operators) who manage the tenant virtual machines, and tenant users (or tenant's customers) who use the applications and services running in the tenant virtual machines. Cloud providers are entities such as Amazon EC2 and Microsoft Azure who have a vested interest in protecting their reputations. The cloud system administrators are individuals from these corporations entrusted with system tasks and maintaining cloud infrastructures, who will have access to privileged domains. In our proposal we usingOpenStack as a private cloud system, which controls the large pool of computing, storage, networking resources via the user friendly interface. Openstack has more features such as rolling upgrades, federated Identity service. We assume that as cloud providers have a vested interest in protecting their reputations and resources, the adversaries from following modules.



## DIFFIE HELLMAN SECURE KEY EXCHANGE

**Fig: 1. INTERNAL PROCESSING MODEL– CLOUD SYSTEMS**

In Secure cloud systems should use secure channel for data transfer. In this proposal mainly for against for the POODLE .POODLE it's a kind of Man in the Middle Attack, which can disable Https/SSL base connection. In this connection we have usingencrypted Https/TLS v1.2 [Figure -1] shows channel, when the user login to the browser and sends the request to the server, then the server generated key using elliptic curve cryptography and safely handover the key with response using Diffie Hellman secure handshaking methodology to the user. Hence the channel should be encrypted Https/TLS v1.2 against POODLE concern.

## METHODS

### System Design Model

In our proposed system, we tried full fledge integrity and more secure concern model over the networking system and we designed in private cloud platform of OpenStack .In our secure model provide overall the starting level of communication channel to the end level. Starting from user account creation for the login, account form have more peculiar details query. All the peculiar data stored in the server. The password will be generating by the server, that password should be stronger and individual can remember their password easily and stronger. For example my last name is Birth year is 1989, grandfather name is subramani,my, my favorite symbol is \$,then the password may mani89\$x.So that if I lost my physical id saved thing like my mobile,pendrive,wallet like except my memory, the authentication threat like guessing password vulnerability may prevent. Then the Encrypted Http/TLS v1.2 protocol base connection setup against for POODLE attack, then local server designed as using OpenStack as a cloud base local server. In Openstackinbuild component Keystone also provides the very good authentication additionally deploying ECC base key generating in the OpenStack. Once the Shortest and strong key generated then the key safely transferred to the client via Secure Diffie Hellman Handshaking protocol. In this paper we define five modules, are following

### Entry Module

User Account creation should be more specific, peculiar, ultimate details collection. Based upon of the unique details, the password will be generated. Because of the key should be more unique. Because password guessing hacking or some may lost their ID Proofs, then hacker may try to illegal usage of your account. For example if I lost my wallet within that National ID's. Then hacker tries to guessing password. Once they get, they may the king of your things. So that, strong password should be more required. So that Account creation should have psyche identity, personal interest, National Id proof, Grandfather Name, like. Password should more specific, if hacker try the forget password options, they should get more trouble.

### SEO Analysis For Seeking Secure Protocol

Channel Communication is the more important aspects in the networking system. So, that protocol plays the vital role in the security part, in recently POODLE attack which causing the SSL disability. So Protocol seeking in our system which should more secure and supports more website visibility. In this connection, we Analysis gives the secure channel, reliability, protocol. As per SEO Report encrypted Https will give more optimized for our model. Statistical [Figure -2] Shows, encrypted Https will give standard visibility for the net users. In blackline represents the encrypted Https and and shadowed represent the http users. Significantly increasing the encrypted Https users because of which supporting more number of website.

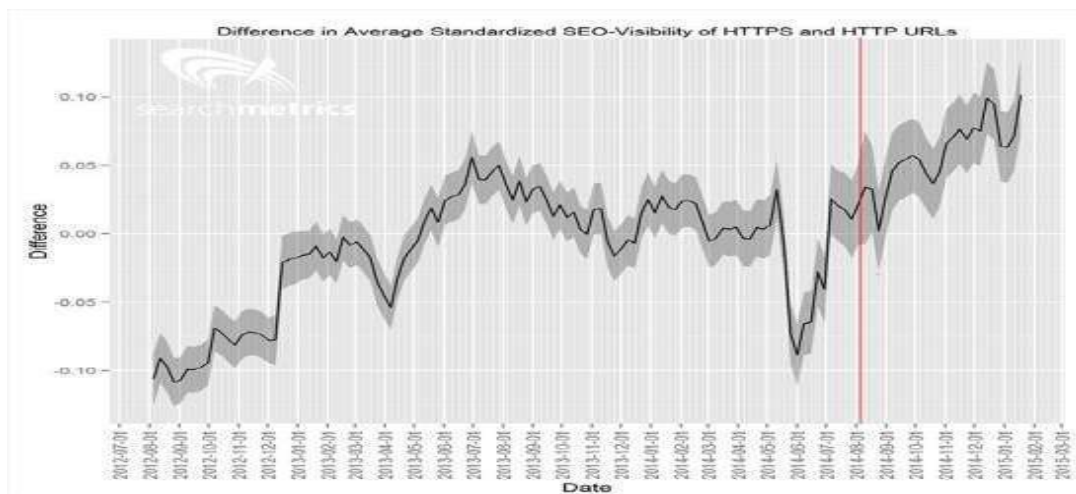


Fig: 2. Https significantly increase their visibility As SEO Report

### *b.iConnection Setup via encrypted HTTP-TLS v1.2*

Communication channel has more important concern securing network. Wherever HTTPS/SSLV3, there is POODLE, Wi-Fi hotspot, likewise more MITM attacks are associated with SSL. In the POODLE (Padding Oracle Downgraded Legacy Encryption), it will disable the SSL base connection and threat the confidential things. Initially used in cloud setup for connection establishment. Hence SSL 3.0 will be disabled by the POODLE Hacker.

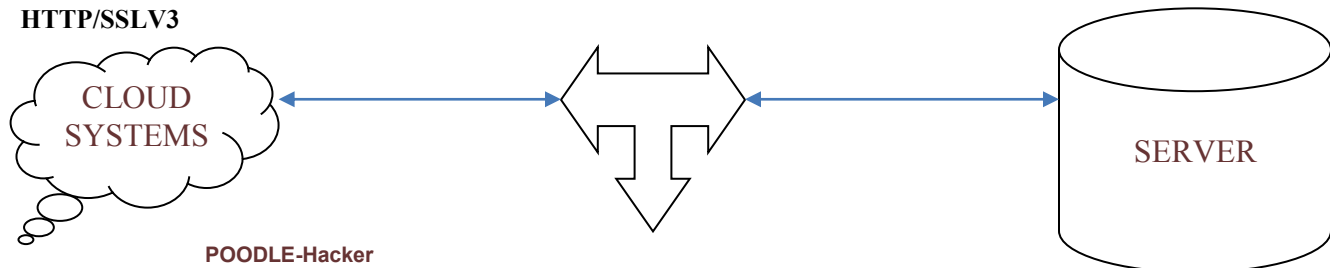


Fig: 3. POODLE –Exploit SSL v3

So that in our proposed system client and server communication channel using encrypted HTTPS/TLS v1.2 were we used for against POODLE. [Figure -3], [Figure -4]



Fig: 4. POODLE –Prevention Using Encrypted HTTPS/TLS 1.2

### Server Authentication

We examined in cloud system. In this project deploy OpenStack- private cloud act as local server. OpenStack is an Open source cloud platform. OpenStack has the set of software that can manage the cloud environment. OpenStack, OpenNebula like lot of open sources are available. So anyone can add components to OpenStack to help it to meet their needs. OpenStack has set of components some of them follow.

OpenStack [Figure -5] provides the very good support for protocol security and efficient way of private cloud platform.

**NOVA**-Primary computing engine behind Open stack. It's a fabric controller which is used for deployment and managing virtual machines and computing tasks.

**SWIFT** – Storage system for object files. This makes scaling easy.

**CINDER** – block storage component

**NEUTRON** – Networking capability for OpenStack

**KEYSTONE** –Provide Identity Service for authentication and authorization

**HORIZON** – It's a Dashboard behind OpenStack. It's a modular web based User Interface (UI)

**CEILOMETER** –Provide single point of contact for billing system.

**HEAT**- Provide Orchestration services for multi composite cloud applications.

**TROVE**- Provide Database as a service

### Unique ID Generating

Client establishes the connection using Http/TLS 1.2 to the OpenStack local server. The server has generate the key Using Elliptic Curve Cryptography, we can generate the around 160 bits providing same security level as 1024 bits. So that Computation speed is high. Less Memory, long term battery life. So ECC will generate key efficiently. ECC will give the good support to the ECC and DH[14]. Based on National Institute and Standard Technology table comparison with their ratio Recommends [Table- 1], we used Elliptic Curve cryptography for short and speedy key generation.

### Key Exchange

Password authenticated key management ((PK) in the form of Secure Diffie Hellman (DH) key exchange algorithm was secure key transfer.PK DH It's a Public key cryptography .It uses two keys, for sending message to server using with private key and server responses via the public key. Receiver side using his private key fordecrypt and response using the public key. DH for secure Handshaking, connection establishes supports.

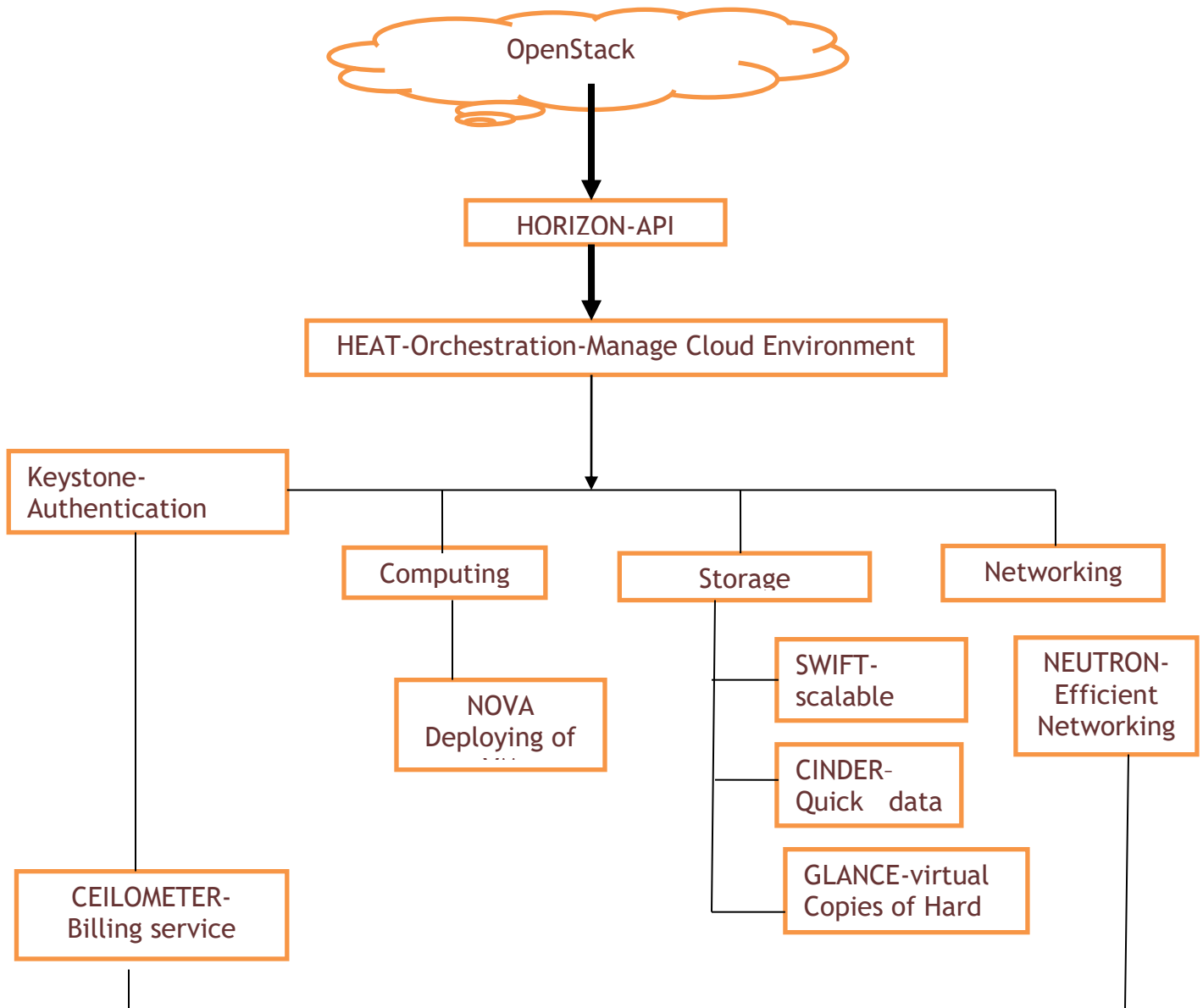


Fig: 5. Open Stack Components

## RESULT AND DISCUSSION

TLS provide very good supporting for secure channel of communication against eavesdropping, and MITM attacks.

### WEBSITE SUPPORT

**Table: 1. PROTOCOL WEB SUPPORT AND SECURITY ANALYSIS**

Protocol Version	Website Support	Security
SSL 2.0	10.4%	Insecure
SSL 3.0	32.6%	Insecure
TLS 1.0	47.2%	GOOD
TLS 1.1	65.7%	Good
TLS 1.2	69%	Very Good

Moreover TLS 1.2 connection setup, website supports are significantly increasing in the Search Engine Optimization world. Protocol support in security [16] are very essential.

We uploading the malware via http to the server, then the malicious Http uploading was detected, but the encrypted HTTPS exploit the malware. Https supporting significantly increase the websites support [15], overall the network and the TLS v1.2 provide more secure and reliable to the protocol security. So in our system we tried to use encrypted https and TLS v1.2 for secure communication channel. As per SEO Report encrypted Https will give more optimized support overall the network and the TLS v1.2 provide more secure and reliable to the protocol security. So in our system we tried to use encrypted https and TLS v1.2 for secure communication channel. In Server part, we concerning short and strong key generation so that ECC algorithm which helps, more memory and time saving. After Key generation which should safely handover to the client using the Diffie Hellman Handshaking protocol

### Output

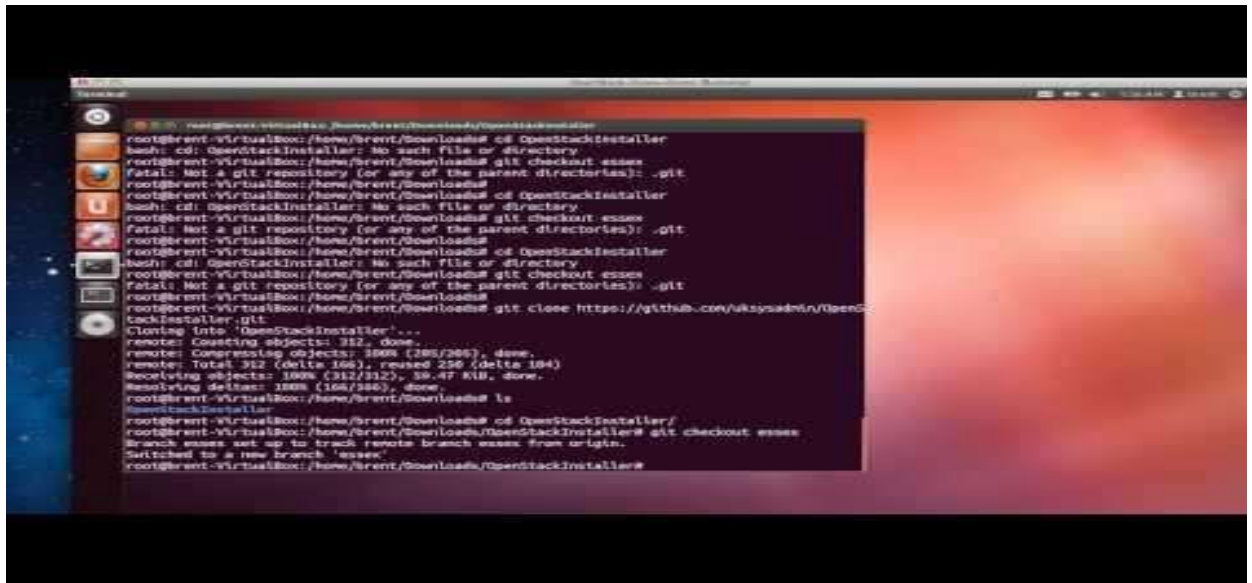
**Table: 2. ECC&RSA Key Comparison**

Symmetric Key Size(bits)	RSA& DH size(bits)	Elliptic Curve Crypto Graph& DH Key & Size(bits)	RSA&ECC Ratio
80	1024	160	7:1
112	2048	224	14:1.5
128	3072	256	21:2

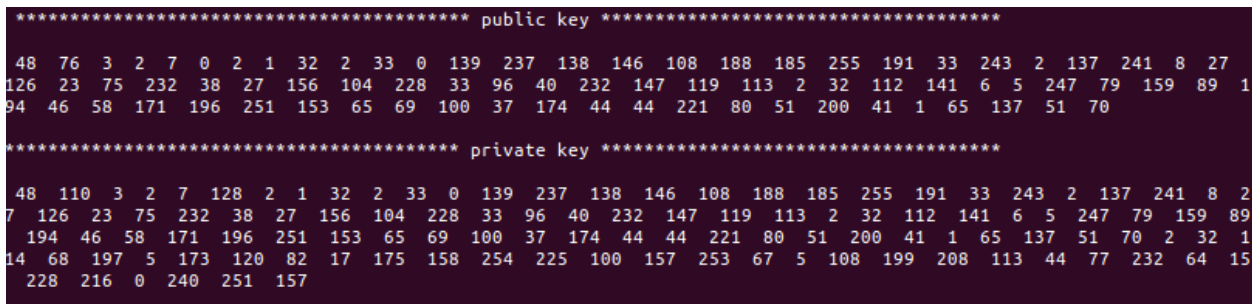
This output shows comparatively RSAVsECC key generation ratios are shown as [Table- 2]. So that ECC&DH combination will gives more secure transmission, short and secure key generation is possible. As per SEO reports Encrypted Https/TLS v1.2 supports, OpenStack Keystone authentication in server side ,then the Elliptic Curve cryptography and Diffie Hellman Secure handshake will provide more strengthen towards the data transmission and secure system



### Screen Shot



Openstack installation



Key Generation

### CONCLUSION

In this paper, we propose Protocol Security using TLS 1.2 in the private cloud system against POODLE attack. In this model concern on design more secure channel communication, Registration Login form with more unique details collecting from user, Unique USER ID generation, Search Engine Optimisation Analysis for the secure protocol for the proper and secure communication channel, OpenStack Private local server, short and strong password generation using Elliptic Curve Cryptography, finally that secure key handover to the client with the help of Diffie Hellman handshaking protocol.

### FUTURE ENHANCEMENTS

We examined in Private cloud system. We try to exploits multi vulnerability attacks like POODLE, Password Guessing vulnerability, And if user forget the password or Loss his ID proof hacker may attack .In OpenStack private cloud deployment then default secure keystone authentication takes place and ECC algorithm for small and speedy unique ID and key is generating. DH for secure handshaking key transfer between user and server. Overall discuss and propose secure model for private Cloud System.In future we can move some public server security system. Security is the first and foremost for all things.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Voras B Mihaljevic, and Orlic M.[2011]Criteria for evaluation of open source cloud computing solutions, in Information Technology Interfaces (ITI), Proceedings of the ITI 33rdInternational Conference on, june 2011, 137 –142.
- [2] Mahjoub M, Mdhaffar A, Halima RB, Jmaiel M.[2011]A Comparative study of the current cloud computing technologies and offers,” in Proceedings of the 2011 First International Symposium on Network Cloud Computing and Applications,ser. NCCA '11. Washington, DC, USA: *IEEE Computer Society*, 131–134.
- [3] Peng J, Zhang X, Lei Z, Zhang B, Zhang W, an Li Q.[ 2009] Comparison of several cloud computing platforms, in Proceedings of the 2009 Second International Symposiumon Information Science and Engineering, ser. ISISE '09.Washington, DC, USA: *IEEE Computer Society*, pp. 23–27.
- [4] Ristov S, Gusev M, and Kostoska M.[2012] Cloud computing security in business information systems,*International Journal of Network Security & Its Applications (IJNSA)*, 4(2): 75–93.
- [5] Eucalyptus. Eucalyptus cloud. [Retrieved: March, 2013].[Online]. Available: <http://www.eucalyptus.com/>
- [6] OpenNebula. Opennebula cloud software. [Retrieved: March,2013]. [Online]. Available: <http://Opennebula.org>
- [7] Cordeiro TD, Damalio DB, NCVN et al.[ 2010]Mangs,Open source cloud computing platforms, in Proceedings of the 2010 Ninth International Conference on Grid and Cloud Computing, ser. GCC '10. Washington, DC, USA: *IEEEComputer Society*, 366–371.
- [8] CSA. Cloud security alliance,” [Retrieved: March, 2013]. [Online]. Available: <http://cloudsecurityalliance.org/>
- [9] ISO/IEC, “ISO/IEC 27001:2005, Information Security Management Systems - Requirements,” [Retrieved: March, 2013]. [Online]. Available: [http://www.iso.org/iso/iso\\_catalogue/catalogue/catalogue\\_detail.htm?csnumber=42103](http://www.iso.org/iso/iso_catalogue/catalogue/catalogue_detail.htm?csnumber=42103)
- [10] Ristov S, Gusev M, and Kostoska M.[2012] A new methodology for security evaluation in cloud computing,” in MIPRO 2012 Proc. of the 35th Int. Convention, *IEEE ConferencePublications*, 2012, pp. 1808–1813.
- [11] ISO/IEC, WDTS 27017, Guidelines on information security controls for the use of cloudcomputing service Retrieved: Mar 2013,Available:<http://www.iso.org/iso/home/store/catalogue/cataloguedetail.htm?csnumber=43757>
- [12] CloudStack. Cloudstack opens source cloud computing. [Retrieved: March, 2013].[Online]. Available: <http://cloudstack.org>
- [13] Ristov S, Gusev M, and Kostoska M.[2012] Security assessment ofopenstack open source cloud solution, in Proceedings of the 7th South East European Doctoral Student Conference (DSC2012), pp. 577–587.
- [14] [Http:ftp://ftp.software.ibm.com/software/iea/content/com.ibm.iea.zos/zos/1.13/Security/zOS\\_V1R13\\_SSL\\_ECC-Support-for-TLS.pdf](http://ftp.software.ibm.com/software/iea/content/com.ibm.iea.zos/zos/1.13/Security/zOS_V1R13_SSL_ECC-Support-for-TLS.pdf)
- [15] <https-vs-http-website-ssl-tls-encryption-ranking-seo-secure-connection/>
- [16] [https:// Transport\\_Layer\\_Security#keyexchange-table](https://Transport_Layer_Security#keyexchange-table)

\*\*DISCLAIMER: This published version is uncorrected proof. plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and checked/reviewed by guest editor.

## IMPROVING PERFORMANCE OF DSR WITH OPTIMIZED FUZZY RULES

Menaka<sup>1\*</sup> and Ranganathan<sup>2</sup>

<sup>1</sup>Dept. of Electronics and Communication Engineering, Dhanalakshmi College of Engineering, Chennai, TN, INDIA

<sup>2</sup>Dean, Dept. of Electrical and Instrumentation, Dr. Mahalingam College of Engineering & Technology, TN, INDIA

### ABSTRACT

In MANET environment the security of every individual network node is crucial due to the pervasive nature of MANETs. Securing the routing algorithms for ad hoc networks is an exceptionally difficult errand due its unmistakable qualities. The DSR Protocol is a source-routed on-demand routing protocol. In this study, Dynamic Source Routing (DSR) protocol is evaluated in untrustworthy environment and to improve performance of multi-objective Particle Swarm Optimization (PSO) and Quantum Annealing (QA) with fuzzy rule selection is proposed. Outcomes reveal that the suggested method performs better than the existing methods.

Published on: 10<sup>th</sup>– August-2016

#### KEY WORDS

MANET, Dynamic Source Routing (DSR) protocol and Multi-Objective Particle Swarm Optimization (MOPSO) with fuzzy rule selection.

\*Corresponding author: Email: [menaka\\_2014@rediffmail.com](mailto:menaka_2014@rediffmail.com)

### INTRODUCTION

MANET is the new developing innovation, which empowers users to communicate with no physical infrastructure regardless of their geographical location that is the reason it is sometimes alluded to as an infrastructure less network. MANETs are self-organizing groups of mobile nodes, which additionally work as router, joined by wireless links. MANETS don't have a centralized infrastructure, namely there is no fixed focal node to organize the task of routing.

Reactive DSR Protocol's operation is split into two stages; path discovery phase as well as path maintenance phase, these phases are activated on demand when a packet needs routing. Route discovery phase floods the network with RREQs if a suitable route is not accessible in the route [1]. DSR utilizes a source routing system to produce a complete route to the destination, this will then be stored briefly in nodes route cache. DSR addresses mobility issues using packet affirmations; failing to obtain an affirmation causes packets to be buffered and path error messages to be transmitted to every upstream node. Path error messages trigger path maintenance phase that expels wrong routes from the route cache and embraces another route discovery phase.

Fuzzy rule-based systems have been successfully applied to solve many classification problems. In many classification problems, fuzzy classification rules are derived from human experts as linguistic knowledge. Because it is not usually easy to derive fuzzy rules from human experts, many approaches have recently been proposed to generate fuzzy rules automatically from the training patterns of the considered classification problem [2]. In order to generate fuzzy rules from training patterns, fuzzy partitions in input spaces are typically regarded for determining premise part of fuzzy classification mode. Grid-type fuzzy partitions of input space and the scatter-type fuzzy partition of the input data have been often used to model fuzzy systems for training patterns. A heuristic method for generating fuzzy rules is applied to the grid-type fuzzy partitions, and a rule selection method, based on PSO algorithms, is then employed to select the relevant fuzzy rules [3] from the generated fuzzy rules. In the PSO-based approach, each individual in the population is considered to represent a fuzzy classification system. Then, a fitness function is implemented for guiding the search process to choose adequate fuzzy classification model as well as number of fuzzy rules.

Optimization issues which have more than one objective function are typical in every field or area of knowledge. In those problems, objectives to be optimized are generally in conflict with one other, which implies that there is

no one solution for these problems. Rather, aimed to find good “trade-off” solutions that represent the best possible compromises among the objectives.

Quantum annealing (QA) is a computational paradigm for searching for the minimal cost function (multi-variable function to be made minimum) through control of quantum fluctuation. Quantum annealing is utilized primarily for combinatorial optimization issues with discrete parameters. In this paper, the traditional DSR protocol is modified in the proposed method to improve Quality of Service (QoS) in untrustworthy environment. PSO and QA with fuzzy rules are proposed for improved performance. Section 2 explains the related works, section 3 explains the methods used for the research, section 4 obtains the results and discussed on it and section 5 concludes the work.

## LITERATURE SURVEY

Chen et al., [4] presented an Entropy-based Fuzzy controllers QoS Routing algorithm in MANET (EFQRM). The key thought of EFQRM algorithm is to build the new metric-entropy and fuzzy controllers with the assistance of entropy metric to decrease the number of route reconstruction in order to give QoS guarantee in the ad hoc network. The simulation results revealed that the suggested methodology and parameters gave an accurate and efficient technique for estimating and assessing the route stability in dynamic MANETs.

Geetha and Thangaraj [5] examined the effect of node mobility on trust establishment is considered and its utilization to propagate trust through a network. This work proposed an enhanced Associativity Based Routing (ABR) with Fuzzy based Trust (Fuzzy-ABR) routing protocol for MANET to enhance QoS and to mitigate network attacks.

Srivastava and Daniel [6] proposed a routing algorithm for the mobile ad hoc networks based on fuzzy logic to find an optimal route for transmitting data packets to the destination. This protocol helped each node in MANET to pick next efficient successor node on the premise of channel variables such as environmental noise as well as signal strength. The protocol enhanced the performance of a route by expanding network life time, diminishing link failure as well as choosing best node for transmitting data packets to next node.

Gupta et al., [7] proposed a routing algorithm based on Fuzzy Logic which is having low communication overhead and storage necessities. The proposed algorithm taken three information variables: signal force, mobility and delay. The supreme value of every parameter could take an extensive range at different points on the network.

Dahiya and Dureja [8] proposed a fuzzy based efficient routing protocol (FBERP) for substantial scale mobile ad-hoc networks that expected to minimize the packet loss rate. Every node in the network is portrayed by its communication parameters. The authors added to a fuzzy logic controller that consolidated these parameters, Packet Loss Rate, Communication Rate, Energy and Delay Parameters. The value acquired, demonstrated the need of a node and it is utilized as a part of route formation. The simulation demonstrated that the proposed protocol outperformed the standard AOMDV routing protocol in minimizing the packet loss.

## METHODOLOGY

### Fuzzy Modeling

Fuzzy Modelling (FM) [12] usually comes with two contradictory requirements to the obtained model: the interpretability, capability to express the behaviour of the real system in an understandable way, and the accuracy, capability to faithfully represent the real system. Since they are contradictory issues, more priority has generally been given to one of them (defined by the problem nature), leaving the other one in the background. Two FM approaches arise depending on the main objective to be considered:

- Linguistic FM, mainly developed by means of linguistic (or Mamdani) FRBSs, which is focused on the interpretability.
- Precise FM, mainly developed by means of Takagi-Sugeno FRBSs, which is focused on the accuracy.

Regardless of the approach, a common scheme has been considered to attain the desired balance between interpretability and accuracy. Firstly, the main objective (interpretability or accuracy) is tackled defining a specific model structure to be used, thus setting the FM approach.

Then, the modelling components (the model structure and/or the modelling process) are improved by means of different mechanisms to compensate for the initial difference between both requirements. Thus, accuracy improvements are proposed in

linguistic FM at the cost of part of the interpretability whilst interpretability improvements are proposed in precise FM at the cost of part of the accuracy.

Actually, the interpretability-accuracy trade-off is a very important branch of research nowadays. Focusing on Linguistic FM with improved accuracy<sup>1</sup> (still nearer of the interpretability) many examples can be find in the existing literature. This approach has been performed by learning/tuning the MFs by defining their parameters or shapes, their types (triangular, trapezoidal, etc.), or their context (defining the whole semantics), learning the granularity (number of linguistic terms) of the fuzzy partitions or extending the model structure by using linguistic modifiers, weights (importance factors for each rule), or hierarchical architectures (mixing rules with different granularities), among others. The main problem of these approaches is that although the system accuracy can be greatly improved (e.g., with a simple tuning of MFs), the original interpretability of the linguistic models is lost to some degree giving way to more complex systems or less interpretable rule structures. Additionally, although rule base reduction and input variable selection [13] processes improve the interpretability, they can also help to improve the accuracy when redundancy and inconsistency criteria are considered (but usually these improvements are not very significant).

Tuning of MFs usually needs an initial model with large number of rules to get an appropriate level of accuracy. Generally, to obtain a good number of initial rules, methods ensuring covering levels higher than needed are used. In this way, rules are obtained that being needed at first could be unnecessary once the tuning is applied or rules that could impede the tuning of the remaining ones in order to obtain the global optimum in terms of the accuracy (better configuration of rules to get the minimum error after tuning of the parameters). Thus, find the following types of rules respect to this global optimum in the complete set of rules: Bad Rules (erroneous or conflicting rules) that degrade the system performance (rules that are not included in the most accurate final solution); Redundant or Irrelevant Rules that do not significantly improve the system performance; Complementary Rules that complement some others slightly improving the system performance; and Important Rules that should not be removed to obtain a reasonable system performance. Obviously, this is a simplification of the problem by only considering in principle the most accurate solution in order to have an idea of the shape of the optimum Pareto. On the other hand, to determine those types of rules in advance is impossible since it directly depends on each concrete configuration of rules and still more on the optimal configuration of the MF parameters for each rule configuration. Therefore, this is impossible to establish any criteria that could be used in the search process. However, by taking into account the possible existence of these kinds of rules, different rule configurations and different tuning parameters, can estimate the following zones in the space of the objectives:

- Zone with Bad Rules, which contains solutions with bad rules. In this zone, the Pareto front does not exist given that removing these kinds of rules would improve the accuracy and these solutions would be dominated by others.
- Zone with Redundant or Irrelevant Rules, which is comprised of solutions without bad rules but still maintaining redundant or irrelevant rules. By deleting these kinds of rules the accuracy would be practically the same.
- Zone with Complementary Rules, comprised of solutions without any bad or redundant rule. By removing these rules the accuracy would be slightly decreased.
- Zone with Important Rules, which contains solutions only comprised of essential rules. By removing these kinds of rules the accuracy is really affected.

### Proposed Particle Swarm Optimization (PSO) -Fuzzy Rules

Particle Swarm Optimization Algorithm (PSO) refers to a population-based optimization technique that finds the optimal solution using a population of particles [9, 10]. All swarms of PSO are solutions in solution space. PSO is fundamentally developed by simulating the flocking of birds. PSO is defined thus: All individual particles  $i$  have the following characteristics: Current positions in search space,  $x_{id}$ , a current velocity,  $p_{id}$ , and a personal best position in search space,  $p_{id}$ .

- The personal best position,  $p_{id}$ , relates to the position in search space wherein particle  $i$  represents smallest error as given by objective function  $f$ , presuming minimization task.
- The global best position denoted by  $p_{gd}$  represents the position yielding the lowest error amongst all the  $p_{gd}$ .

During the iteration all particles in the swarm are updated through equations (1 and 2):

Particles Velocity

$$v_{id} = w * v_{id} + c_1 * rand() * (p_{id} - x_{id}) + c_2 * rand() * (p_{gd} - x_{id}) \quad (1)$$

The current position of particles are updated for obtaining the subsequent position.

Particles Position

$$x(t+1) = x(t) + v(t+1) \quad (2)$$

wherein  $c_1$  and  $c_2$  refer to two positive constants,  $c_1$  and  $c_2$  are two arbitrary numbers within the range [0,1], and  $w$  is the inertia weight.

For applying the PSO scheme to solve multi-objective optimization issues, it is obvious that the original scheme has to be modified. Solution sets of problems with several objectives do not comprise of single solutions (as in global optimization) are

explained. Instead, in multi-objective optimization, aim to find sets of different solutions (known as Pareto optimal set). Generally, when resolving multi-objective issue, there are three primary goals to attain [11]:

1. Make the number of elements found in the Pareto optimal set maximum
2. Make the distance of the Pareto front yielded by the suggested algorithm with respect to true (global) Pareto front minimum
3. Make the spread of solutions discovered maximum, so that it can have a distribution of vectors as smooth and uniform as possible.

First, the swarm is initialized. Next, sets of leaders are also initialized with non-dominated particles from swarms. As previously noted, sets of leaders are typically stored in external archives. Later, certain quality metrics are computed for all leaders for selecting one leader for all particles of swarms. With every generation, leaders are chosen and flight is performed. Almost all existing MOPSOs employ a kind of mutation operator after performing the flight. Later, particles are evaluated and corresponding pbest are updated. A new particle replaces its pbest particle usually when this particle is dominated or if both are incomparable (i.e., they are both non-dominated with respect to one another). After all particles are updated, set of leaders are updated as well. In the end, quality metric of the set of leaders are recomputed. The procedure is iterated for a specified (usually fixed) number of iterations.

### **Pseudo code of a general MOPSO algorithm**

```

Begin
Initialize swarm
Initialize leaders in an external archive
Quality(leaders)
g = 0
While g < gmax
For each particle
Select leader
Update Position (Flight)
Mutation
Evaluation
Update pbest
EndFor
Update leaders in the external archive
Quality(leaders)
g++
EndWhile
Report results in the external archive
End
  
```

All required data regarding rule-base or memberships functions are to be definitely specified. They are used to accurately represent fuzzy logic.

### **PSO based Fuzzy Controller Design Method**

The PSO algorithm [14] is a computation technique proposed by Kennedy and Eberhart. Its development was based on observations of the social behavior of animals such as bird flocking and fish schooling of the swarm theory. If a set  $P^g$  with  $N$  particles is called a population in the  $g$ -th generation and expressed by equation (3):

$$P^g = \{p_1^g, p_2^g, \dots, p_h^g, \dots, p_N^g\} \quad (3)$$

The position vector and velocity vector of the  $h$ -th particle ( $h \in \{1, 2, \dots, N\}$ ) in the  $g$ -th generation ( $g \in \{1, 2, \dots, G\}$ ) are respectively denoted by equation (4 and 5):

$$p_h^g = (p_{(h,1)}^g, p_{(h,2)}^g, \dots, p_{(h,j)}^g, \dots, p_{(h,n)}^g) \quad (4)$$

And

$$v_h^g = (v_{(h,1)}^g, v_{(h,2)}^g, \dots, v_{(h,j)}^g, \dots, v_{(h,n)}^g) \quad (5)$$

where  $n$  is the number of searching parameters and  $p_{(h,j)}^g$  denotes the position of the  $j$ -th parameter ( $j \in \{1, 2, \dots, n\}$ ) of the  $h$ -th particle in the  $g$ -th generation.

The procedure of PSO algorithm can be described as follows:

Step 1: Initialize PSO by setting  $g=1$ ,  $F_1^{pbest} = F_2^{pbest} = \dots = F_N^{pbest} = 0$ , the maximum number of generation (G), the number of particles (N), and four parameter values of  $c_1, c_2, \omega_{max}$  and  $\omega_{min}$ .

Step 2: Generate the initial position vector  $p_h^1 = (p_{(h,1)}^1, p_{(h,2)}^1, \dots, p_{(h,j)}^1, \dots, p_{(h,n)}^1)$  and the initial velocity vector  $v_h^1 = (v_{(h,1)}^1, v_{(h,2)}^1, \dots, v_{(h,j)}^1, \dots, v_{(h,n)}^1)$  of N particles randomly by equation (6):

$$p_{(h,j)}^1 = p_j^{\min} + (p_j^{\max} - p_j^{\min}) \text{rand}()$$

and

$$v_{(h,j)}^1 = \frac{(v_j^{\max} - v_j^{\min})}{20} \text{rand}()$$
(6)

where  $p_j^{\max}$  and  $p_j^{\min}$  are respectively the maximum value and minimum value of the j-th parameter.  $v_j^{\max}$  and  $v_j^{\min}$  are the maximum velocity and minimum velocity the j-th parameter.  $\text{rand}()$  is an uniformly distributed random number in [0,1].

Step 3: Calculate the fitness value of each particle in the g-th generation by equation (7):

$$F(p_h^g) = \text{fit}(p_h^g), h = 1, 2, \dots, N$$
(7)

Where  $\text{fit}(\cdot)$  is the fitness function.

Step 4: Determine  $F_h^{pbest}$  and  $p_h^{pbest}$  for each particle by equation (8):

$$F_h^{pbest} = \begin{cases} F_h^g, & \text{if } F_h^{pbest} \leq F_h^g \\ F_h^{pbest}, & \text{otherwise} \end{cases}, h \in \{1, 2, \dots, N\}$$

and

$$p_h^{pbest} = \begin{cases} p_h^g, & \text{if } F_h^{pbest} \leq F_h^g \\ p_h^{pbest}, & \text{otherwise} \end{cases}, h \in \{1, 2, \dots, N\}$$
(8)

Where  $p_h^{pbest}$  is the position vector of the h-th particle with the personal best fitness value  $F_h^{pbest}$  from the initial to the current generation.

Step 5: Identify an index q of the particle with the highest fitness by equation (9):

$$q = \arg \max_{h \in \{1, 2, \dots, N\}} F_h^{pbest}$$

and determine  $F^{Gbest}$  and  $p^{Gbest}$  by

$$F^{Gbest} = F_q^{pbest} = \max_{h \in \{1, 2, \dots, N\}} F_h^{pbest}$$

and

$$p^{Gbest} = p_q^{pbest}$$
(9)

Wherein  $p^{Gbest}$  refers to position vector of the particle with the global best fitness value  $F^{Gbest}$  from the beginning to the current generation.

Step 6: If  $g=G$ , then go to Step 12, Otherwise, go to Step 7.

Step 7: Update the velocity vector of each particle by equation (10):

$$v_h^{g+1} = \omega v_h^g + c_1 \cdot \text{rand}1() \cdot (p^{Gbest} - p_h^g) + c_2 \cdot \text{rand}2() \cdot (p^{Gbest} - p_h^g)$$
(10)

Where  $v_h^g$  and is the current velocity vector of the h-th particle in the g-th generation.  $v_h^{g+1}$  is the next velocity vector of the h-th particle in the (g+1)-th generation.  $\text{rand}1()$  and  $\text{rand}2()$  refer to two uniformly distributed arbitrary numbers in [0,1].  $c_1$  and  $c_2$  are constant values.  $\omega$  is a weight value and defined by equation (11):

$$\omega = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{G} \cdot g \quad (11)$$

Where  $\omega_{\max}$  and  $\omega_{\min}$  are respectively a maximum value and a minimum value of  $\omega$ . Step 8: Check the velocity constraint by equation (12):

$$v_{(h,j)}^{g+1} = \begin{cases} v_j^{\max}, & \text{if } v_{(h,j)}^{g+1} > v_j^{\max} \\ v_{(h,j)}^{g+1}, & \text{if } v_j^{\min} \leq v_{(h,j)}^{g+1} \leq v_j^{\max} \\ v_j^{\min}, & \text{if } v_{(h,j)}^{g+1} < v_j^{\min} \end{cases} \quad (12)$$

$H=1,2,\dots,N, j=1,2,\dots,n$ .

Step 9: Update position vectors of all particles by equation (13):

$$p_h^{g+1} = p_h^g + v_h^{g+1} \quad (13)$$

Where  $p_h^g$  is the current position vector of the  $h$ -th particle in the  $g$ -th generation.  $v_h^{g+1}$  is the next position vector of the  $h$ -th particle in the  $(g+1)$ -th generation.

Step 10: Bound the updated position vector of each particle in the searching range by equation (14):

$$p_{(h,j)}^{g+1} = \begin{cases} p_j^{\max}, & \text{if } p_{(h,j)}^{g+1} > p_j^{\max} \\ p_{(h,j)}^{g+1}, & \text{if } p_j^{\min} \leq p_{(h,j)}^{g+1} \leq p_j^{\max} \\ p_j^{\min}, & \text{if } p_{(h,j)}^{g+1} < p_j^{\min} \end{cases} \quad (14)$$

$H=1,2,\dots,N, j=1,2,\dots,n$ .

Step 11: Let  $g=g+1$  and go to Step 3.

Step 12: Determine the corresponding fuzzy controller based on the position of the particle  $p^{G_{best}}$  with the best fitness value  $F^{G_{best}}$ .

### Quantum Annealing (QA)

Quantum mechanics works with wave-functions that can equally well sample wide regions of phase-space. Instead of thermal fluctuations, one exploits here the quantum fluctuations provided by a suitably introduced – and equally artificial – kinetic energy. Annealing is then performed by slowly reducing to zero the amount of quantum fluctuations introduced. Quantum fluctuations have, in many respects, an effect similar to that of thermal fluctuations – they cause, for instance, solid helium to melt even at the lowest temperatures – but they differ considerably in other respects. In particular, quantum systems can tunnel through classically impenetrable potential barriers between energy valleys, a process that might prove more effective than waiting for them to be overcome thermally.

Quantum PSO (QPSO) is that which drops velocity vector of the original PSO and subsequently changes updating scheme of particles' positions for making searches more simple as well as effective [15]. Pseudo code for proposed method is:



*Begin*

*step1*: initialization

PSO

Initialize a swarm by the proposed schedule initialization algorithm

give initial value:  $w_{max}$ ,  $w_{min}$ ,  $c_1$ ,  $c_2$ ,  $Gen$  and generation=0;

set indicator  $m=0$

QA

set initial temperature  $T$ , final temperature  $T_0$

*Step*: 2 Iteration process

do {

*Generate* next swarm;

find new gbest and pbest;

update gbest of swarm and pbest of each particle;

generation++;

if(gbest is not improved)

$m++$ ;

end if

if ( $m==Gen$ )

{

while ( $T>T_0$ )do

{

$s' \leftarrow Produce\_Solution(s)$ ;

*Compute*( $s, s'$ )

if ( $\Delta E \leq 0$ ) or ( $\exp(-\Delta E / T) > rand(0,1)$ )

$s \leftarrow s', T \leftarrow BT$

end if

}

end while

$m=0$ ;

} endif

} while (termination condition is not satisfied)

Step 3: Output Optimization solution

End

## RESULTS AND DISCUSSION

Simulations were carried out using OPNET in a rectangular network of size 4 Square Kilometer. The simulation parameters are given in [Table- 1]. Experiments are conducted with 25 nodes with random mobility. Throughput, end to end delay and percentage of malicious node detected are calculated and the simulation results are drawn in graphs and in tables. To imitate real-time situation, untrustworthy nodes are added to the network.

Table: 1. Simulation Setup

Total area of the network		4 sq.km
Number of nodes	25	
Mobility of the nodes	constant speed of 10 kmph, 20 kmph, 30 kmph, 40 kmph, and 50 kmph	
Data rate	2Mbps	
Routing Protocol	DSR & Proposed	
Node mobility	10, 20, 30, 40 and 50 Kmph	
Transmission range of node	200 m	
Data type	Constant Bit Rate	

[Table-2, 3, 4] and [Figure- 1], [Figure- 2] , [Figure- 3] shows the throughput, end to end delay and percentage of malicious node detected respectively that are obtained from different experiments.

Table: 2. Throughput

Throughput x 100 %	FR-DSR	FR-DSR in untrustworthy environment	QAFR-DSR	QAFR-DSR-Untrustworthy	MOPSOFR-DSR	MOPSOFR-DSR in untrustworthy environment
No mobility	0.9715	0.9465	0.9784	0.9563	0.9853	0.9627
15 kmph	0.9372	0.9275	0.9442	0.9408	0.9553	0.9416
30 kmph	0.9246	0.9087	0.9329	0.9152	0.939	0.9216
45 kmph	0.9026	0.8726	0.9126	0.8849	0.915	0.8868
60 kmph	0.8547	0.7847	0.8674	0.7959	0.8713	0.7967
75 kmph	0.8109	0.7672	0.8177	0.7738	0.8229	0.7781

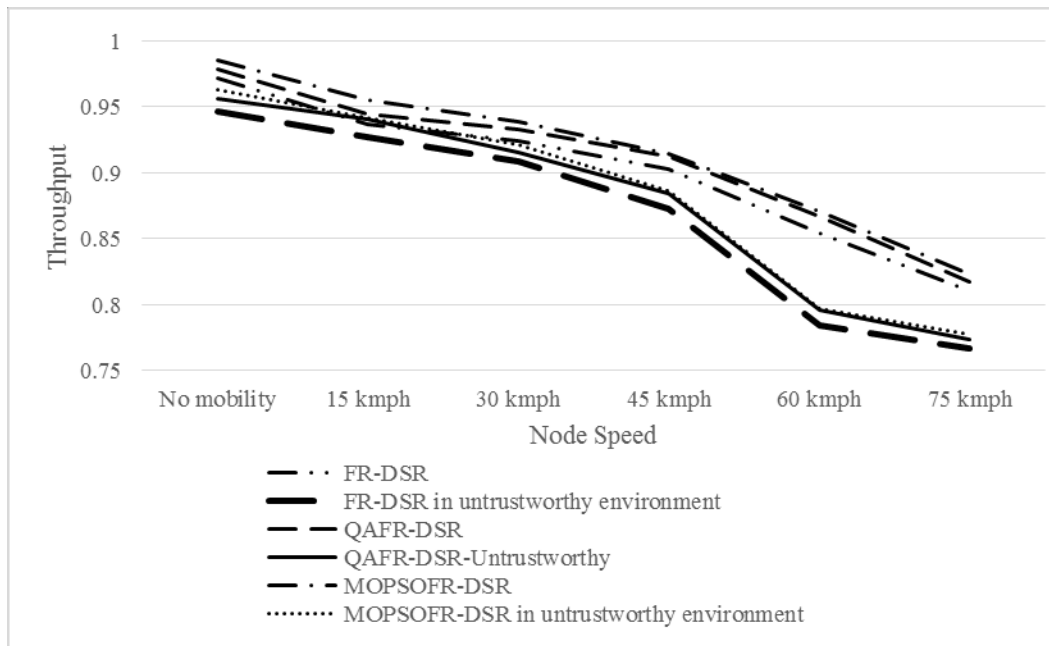
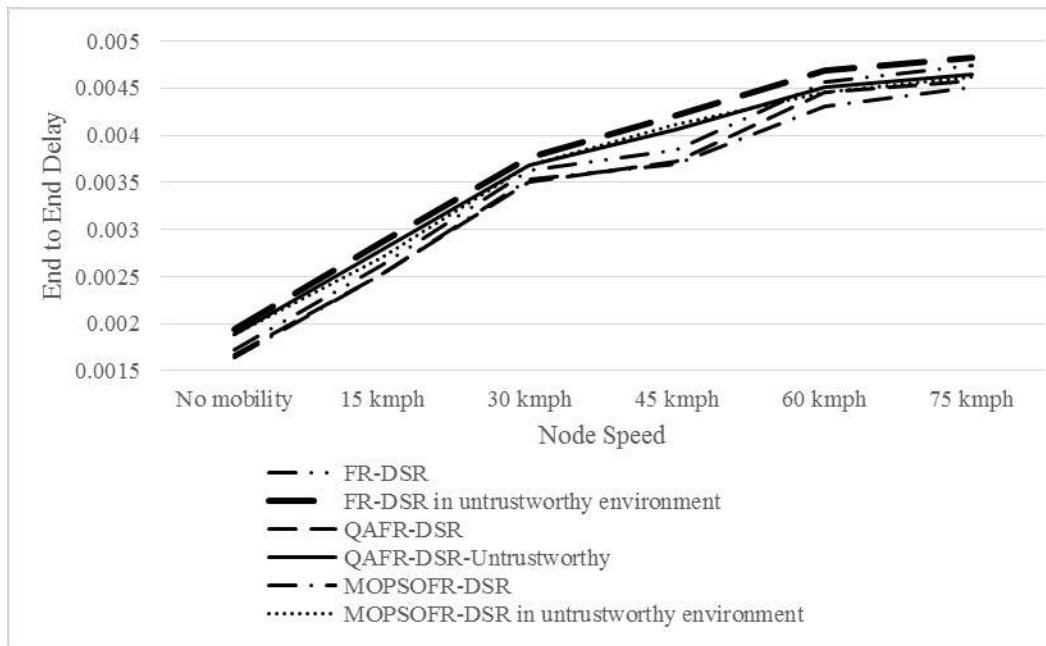


Fig: 1. Throughput

From **Table- 2** and **Figure- 1** it is observed that the throughput of MOPSOFR-DSR performs better by 1.41% than FR-DSR, by 4.02% than FR-DSR in untrustworthy environment, by 0.7% than QAFR-DSR, by 2.99% than QAFR-DSR untrustworthy, by 2.32% than MOPSOFR-DSR in untrustworthy environment for no mobility. The throughput gets decreased when the node speed increases. For node speed 75 kmph, the throughput of MOPSOFR-DSR performs better by 1.47% than FR-DSR, by 7% than FR-DSR in untrustworthy environment, by 0.63% than QAFR-DSR, by 6.15% than QAFR-DSR untrustworthy, by 5.59% than MOPSOFR-DSR in untrustworthy environment.

**Table: 3. End to End Delay**

End to End Delay ms	FR-DSR	FR-DSR in untrustworthy environment	QAFR-DSR	QAFR-DSR- Untrustworthy	MOPSOFR-DSR	MOPSOFR-DSR in untrustworthy environment
No mobility	0.00173	0.00194	0.00167	0.00189	0.001641	0.00189
15 kmph	0.00262	0.00287	0.00253	0.00279	0.002532	0.002702
30 kmph	0.00363	0.00376	0.00351	0.00368	0.003527	0.003681
45 kmph	0.00384	0.00421	0.00372	0.00406	0.003689	0.004114
60 kmph	0.00457	0.00468	0.00445	0.00451	0.004307	0.00445
75 kmph	0.00474	0.00482	0.00458	0.00464	0.004516	0.00462



**Fig: 2. End to End Delay**

From **Table- 3** and **Figure- 2** it is observed that the End to End Delay of MOPSOFR-DSR performs better by 5.28% than FR-DSR, by 16.69% than FR-DSR in untrustworthy environment, by 1.75% than QAFR-DSR, by 14.0% than QAFR-DSR untrustworthy, by 14.1% than MOPSOFR-DSR in untrustworthy environment for no mobility. The end to end delay gets increased when the node speed increases. For node speed 75 kmph, the end to end delay of MOPSOFR-DSR performs better by 4.84% than FR-DSR, by 6.5% than FR-DSR in untrustworthy environment, by 1.41% than QAFR-DSR, by 2.71% than QAFR-DSR untrustworthy, by 2.28% than MOPSOFR-DSR in untrustworthy environment.

Table: 4. Percentage of Malicious Node Detected

Percentage of Malicious node detected	Initial state of DSR	Steady state of FR-DSR	Initial state of QAFR-DSR	Steady state of QAFR-DSR-Untrustworthy	MOPSOFR-DSR	MOPSOFR-DSR in untrustworthy environment
No mobility	0.4	0.9	0.47	0.94	0.5	0.9
15 kmph	0.4	0.8	0.45	0.94	0.4	0.9
30 kmph	0.3	0.8	0.34	0.93	0.4	0.8
45 kmph	0.3	0.7	0.35	0.83	0.3	0.7
60 kmph	0.2	0.6	0.22	0.69	0.3	0.7
75 kmph	0.1	0.4	0.12	0.45	0.2	0.5

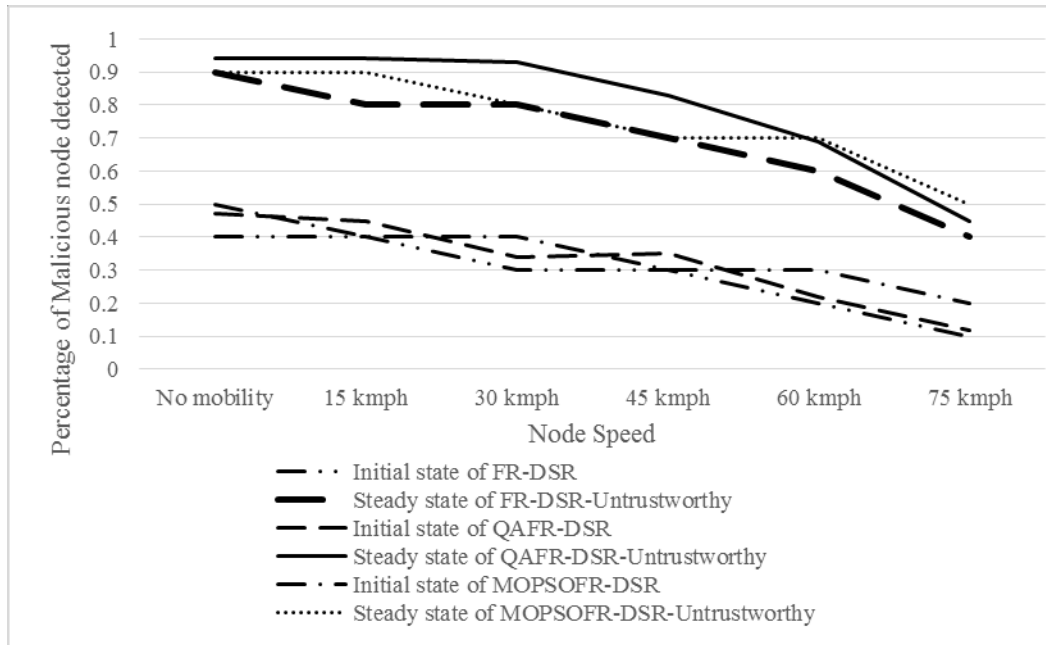


Fig: 3. Percentage of Malicious Node Detected

From Table- 4 and Figure- 3 it is observed that the Percentage of Malicious Node Detected of Initial state of MOPSOFR-DSR performs better by 22.22% than initial state of FR-DSR and by 6.19% than initial state of QAFR-DSR, the steady state of MOPSOFR-DSR performs equally to steady state of FR-DSR and by 4.35% than steady state of QAFR-DSR for no mobility. The Percentage of Malicious Node Detected gets decreased when the node speed increases. For node speed 75 kmph, the Percentage of Malicious Node Detected of MOPSOFR-DSR performs better by 6.66% than initial state of FR-DSR and by 50% than initial state of QAFR-DSR, the steady state of MOPSOFR-DSR performs better by 22.22% than steady state of FR-DSR and by 10.53% than steady state of QAFR-DSR.

### CONCLUSION

All nodes are to maintain excellent reputation values for receiving network services. Only by forwarding other nodes' packets a node can maintain a high reputation value. It also affects the reality of the data transmission. In such untrustworthy environment, it is difficult to achieve high throughput. In this proposed model, the trust level of all the nodes are calculated and analyzed to identify the capability of nodes. This calculated trust values along with node mobility and inference from previous records are also used to calculate the nodes reputation value. Results show that the throughput of MOPSOFR-DSR performs better by 1.41% than FR-DSR, by 4.02% than FR-DSR in untrustworthy environment, by 0.7% than QAFR-DSR, by 2.99% than QAFR-DSR untrustworthy, by 2.32% than MOPSOFR-DSR in untrustworthy environment for no mobility. The throughput gets decreased when the node speed increases. For node speed 75 kmph, the throughput of MOPSOFR-DSR performs better by 1.47%

than FR-DSR, by 7% than FR-DSR in untrustworthy environment, by 0.63% than QAFR-DSR, by 6.15% than QAFR-DSR untrustworthy, by 5.59% than MOPSOFR-DSR in untrustworthy environment.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

### REFERENCES

- [1] Johnson DB, Maltz DA. [1996] Dynamic Source Routing in Ad Hoc Wireless Networks, Mobile Computing, T. Imielinski and H. Korth, Ed. *Kluwer Academic Publishers*, 5: 153-181.
- [2] Ishibuchi H, Nozaki K, Yamamoto N and Tanaka, H. [1995] Selecting Fuzzy If-Then Rules for Classification Problems Using Genetic Algorithms, *IEEE Trans Fuzzy Systems*, 3L: 260270
- [3] Chen C. [2006] Design of PSO-based fuzzy classification systems. *Tamkang Journal of Science and Engineering*, 9(1): 63
- [4] Chen H, Sun B, Zeng Y, He X. [2009] An entropy-based fuzzy controllers QoS routing algorithm in MANET. In Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on , *IEEE* 3: 235-239)..
- [5] Geetha K., Thangaraj P. [2015] An Enhanced Associativity Based Routing with Fuzzy Based Trust to Mitigate Network Attacks.
- [6] Srivastava S, Daniel AK. [2013] An Efficient Routing Protocol under Noisy Environment for Mobile Ad Hoc Networks using Fuzzy Logic. *INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ARTIFICIAL INTELLIGENCE*, 2(6).
- [7] Gupta AK, Kumar R, Gupta NK. [2014] A trust based secure gateway selection and authentication scheme in MANET. In Contemporary Computing and Informatics (IC3I), 2014 International Conference on (pp. 1087-1093). *IEEE*.
- [8] Dahiya R, Dureja A. [2014] Fuzzy Based Efficient Routing Protocol for Route Recovery In MANET. *International Journal Of Engineering And Computer Science* ISSN: 2319-7242 3(6).
- [9] Kennedy J and Eberhart RC. [1995] Particle swarm optimization" , Proceeding of the 1995 IEEE International Conference on Neural Networks (Perth, Australia), *IEEE Service Centre, Piscataway, NJ*, Iv: 1942-1948.
- [10] Permana KE, Hashim SZM. [2010] Fuzzy membership function generation using particle swarm optimization. *Int. J Open Problems Compt Math*, 3(1), 27-41.
- [11] Reyes-Sierra M, Coello CC. [2006] Multi-objective particle swarm optimizers: A survey of the state-of-the-art. *International journal of computational intelligence research*, 2(3), 287-308.
- [12] Alcalá R, Gacto MJ, Herrera F, Alcalá-Fdez J. [2007] A multi-objective genetic algorithm for tuning and rule selection to obtain accurate and compact linguistic fuzzy rule-based systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(05): 539-557.
- [13] HM Lee, CM Chen JM Chen and YL Jou. [2001] An efficient fuzzy classifier with feature selection based on fuzzy entropy, *IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics* 31:3 :426-432.
- [14] Wong CC, Wang HY, Li SA. [2008] PSO-based motion fuzzy controller design for mobile robots. *International Journal of fuzzy systems*, 10(1):24.
- [15] Sun B, Gui C, Zhang Q, Chen H. [2009] Fuzzy controller based QoS routing algorithm with a multiclass scheme for MANET. *International Journal of Computers, Communications & Control*, 4(4): 427-438.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ, the article is published as provided by author and checked/reviewed by guest editor.

# OPTIMIZING SUPPORT VECTOR MACHINE USING MODIFIED CLONAL SELECTION FOR BRAIN COMPUTER INTERFACE

Padmavath<sup>1\*</sup> and Ranganathan<sup>2</sup>

<sup>1</sup>Dept. of Electronics and Communication Engineering, Dhanalakshmi Srinivasan College of Engineering and echnology, Mamallapuram, Chennai., TN, INDIA

<sup>2</sup>Dept. of Electrical and Instrumentation, Dr. Mahalingam College of Engineering & Technology, Pollachi, TN, INDIA

## ABSTRACT

Brain Computer Interface's (BCI) central element, is a translation algorithm converting electrophysiological input from user into output capable of controlling external devices. Many studies over the past two decades have shown that people and animals can use brain signals to convey their intent to a computer using BCIs. This is possible through use of sensors that capture signals in the brain, corresponding to certain thought forms. The kernel parameters setting for SVM in training process impacts on the classification accuracy. The Modified Clonal Selection Algorithm (CLONALG) is one such system inspired by the clonal selection theory of acquired immunity, which has shown success on broad range of engineering problem domains.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Brain Computer Interfaces (BCIs),  
ElectroCorticoGraphy (ECoG),  
Support Vector Machine (SVM),  
Clonal Selection Algorithm  
(CLONALG).

\*Corresponding author: Email: [ssmadhu80@gmail.com](mailto:ssmadhu80@gmail.com), [profsks@rediffmail.com](mailto:profsks@rediffmail.com)

## INTRODUCTION

A BCI is a hardware that allows humans to interact with a computer through brainwaves. Presently, BCI is used for healthcare, and education based on neural-feedback which is a type of brainwave using bio-feedback. BCI controls computers using human brain waves. BCIs convert brain signals into outputs communicating user's intent [1]. As this new communication channel is independent of peripheral nerves and muscles, it is resorted to by those with severe motor disabilities.

To achieve this, a BCI system consists of four sequential components are signal acquisition, feature extraction, feature translation, and device output. These four components are controlled by an operating protocol that defines the onset and timing of operation, the details of signal processing, the nature of the device commands, and the oversight of performance. An effective operating protocol allows a BCI system to be flexible and to serve the specific needs of each user.

BCIs use invasive and non-invasive methods. ElectroEncephaloGraphic activity (EEG) [2] from the scalp is used by non-invasive BCIs. Though convenient, safe and inexpensive, they are susceptible to artifacts like electromyography (EMG) signals, which have low spatial resolution and so need much user training. Single-neuron activity recorded in the brain is used by invasive BCIs. Though having higher spatial resolution and providing control signals with much freedom, BCIs still are dependent on electrodes in the cortex and so have problems ensuring stable long-term recordings.

ECoG is acute recording of electrical activity directly from cortical surface during exposure in surgical treatment of epilepsy [3, 4]. Recent studies emphasized the intraoperative ECoG importance for precise epileptic focus localization and good surgical outcome. ECoG is not invasive, as neuronal recordings as the brain is not entered into. It has a higher Signal-to-Noise Ratio (SNR) than EEG and also higher spectral/spatial resolution [5] which necessitates re-engineering of signal processing and classification techniques found in conventional EEG-based

BCIs. Extreme data scarcity due to limited time available for volunteering patients is an obstacle to characterize information in ECoG signals.

A huge challenge in designing BCI is the selection of relevant features from a huge set of potential features. High dimensional features vectors are not good because of the curse of dimensionality in training classification protocols. Features selection may be carried out studying all potential subsets of features. But the quantity of possibilities increases in an exponential fashion, making extensive searches impractical for even moderate quantities of features. Certain more effective optimization protocols may be employed with the objective of decreasing quantity of features and at the same time improving classification performance [6].

The typical CLONALG model implies the choosing of antibodies (candidate solutions) on the basis of affinity either by matching antigen patterns or through evaluation of patterns by cost functions. Chosen antibodies are vulnerable to cloning proportional to affinity, and hyper-mutation of clones inversely proportional to clone affinity. Resulting clonal-set competes with antibody populations for membership in subsequent generations and finally, low-affinity population members are substituted by arbitrarily created antibodies [7]. CLONALG is the abbreviation of the clonal algorithm and has been inspired by the following elements of the clonal selection theory [8]

Maintenance of a specific memory set

- Selection and cloning of most stimulated antibodies
- Death of non-stimulated antibodies
- Affinity maturation (mutation)
- Re-selection of clones proportional to affinity with antigen
- Generation and maintenance of diversity

A mixture kernel function based on radial based and polynomial kernel was introduced and the parameters of this new kernel function were optimized. Their algorithm gives the better results than normal SVM in fault diagnosis. But it has some disadvantages. Firstly, their immune optimization method refers to crossover parameter. But original immune optimization algorithm (CLONALG) has not crossover operator. A clonal selection algorithm whose name is determined as CLONALG by them. The objective function of the immune optimization method is calculated based on training of SVM [9]. Because of these properties, clonal selection converges faster than genetic algorithm and does not catch local minimum.

Hybrid algorithms [10] are the combination of exact algorithms and Meta-heuristics. In the scientific community, the term “meta-heuristic” refer to general purpose approximated optimization methods, such as Tabu search, evolutionary computation, and simulated annealing, among others. A general classification of meta-heuristic algorithms grouped into two categories: Collaborative combinations: In an environment of collaboration, the algorithms exchange information, but are independent. The exact and meta-heuristic algorithms may be executed sequentially, in parallel or intertwined. Integrated combinations: In integrated methods, an algorithm is a subordinated component of another algorithm. In the integrated combinations category, there are two subcategories: the meta-heuristic algorithm is the master and controls the calls to the exact algorithm and the exact algorithm is the master and calls the meta-heuristic algorithm.

In this paper, proposed the optimized SVM using modified CLONALG. Section 2 deals with literature related to this work, section 3 describes the methods used in the work, section 4 deals with results and discusses obtained results and finally section 5 concludes the work.

## RELATED WORKS

BCI using Electroencephalogram signal was discussed by Shende&Jabade [11]. If BCI aims to control robotics machinery with better accuracy. The use of BCI is to control the wheelchair/robotic limb movement for a disable person and to access security systems. BCI can discover emotions which is can control the surrounding environment like controlling the interior of a car/house.

A low-cost non-invasive BCI hybridized with eye tracking described by Kim et al., [12] also discussed its feasibility through a Fitts' law-based quantitative evaluation method. Non-invasive BCI received a lot of attention recently. BCI applications in real life need to be user-friendly and easily portable. In the new work, an approach to

realize a real-world BCI, EEG-based BCI combined with eye tracking was investigated. The new hybrid BCI system was discussed regarding a practical interface scheme. Though further advancement was required, the new hybrid BCI system had the potential to be useful in a natural/intuitive manner.

A novel driver-vehicle interface for the individuals with severe neuromuscular disabilities to use intelligent vehicles through P300 and steady-state visual evoked potential (SSVEP) BCIs to select destination and test its performance in the lab and in real driving conditions was proposed by Fan et al., [13]. The new interface has 2 components: a selection component based on a P300 BCI and a confirmation component based on an SSVEPBCI. The proposed system improved destination selection accuracy compared to a single P300 BCI-based selection system, specifically for participants with relatively low accuracy in using P300 BCI.

Krusiński et al., [14] presented a preliminary analysis of the relationship between EEG and ECoG event-related potentials (ERPs) recorded from a single patient using a BCI speller. The patient carried out one experimental session through usage of BCI spelling paradigms controlled by scalp-recorded EEG before ECoG grid implantations as well as one identical session controlled by ECoG post grid implantations. The patient was capable of achieving near perfect spelling precision through EEG as well as ECoG. An offline analysis of the average ERPs was performed to assess how accurately the average EEG ERPs could be predicted from the ECoG data. The preliminary results indicated that EEG ERPs can be accurately estimated from proximal asynchronous ECoG data using simple linear spatial models.

The robust nature of Least-Square Support Vector Machines (LS-SVMs) for classifying multi-class self-paced MI temporal features while tuning hyper parameters automatically was investigated by Hamedí et al., [15]. MI EEG signals were pre-processed/segmented into non-overlapped distinctive time slots. The new method was evaluated/compared to three classifiers. Results indicated LS-SVM's high potential to classify different MI's by getting average  $89.88 \pm 8.00$  classification accuracy when using Sign Slope Changes (SSC) features.

A multi-ganglion Artificial Neural Network (ANN) based Feature Learning (ANNFL) method to extract the deep feature structure of a single-trial multi-channel ERP signals and improve classification accuracy proposed by Gao et al., [16] extracted feature vectors and classified them using SVM. The method outperformed a PCA and conventional three-layer auto-encoder leading to higher classification accuracies in five subjects' BCI signals than with using single-channel temporal features. ANNFL is an unsupervised feature learning method, which automatically learns feature vector from EEG data providing more effective feature representation than the PCA and single-channel temporal feature extraction methods.

Parallel multi-objective optimization methods for coping with high-dimensional features selection problems were proposed by Kimovski et al., [17]. Many parallel multi-objective evolutionary alternatives were proposed and evaluated using synthetic/BCI benchmarks. Results showed that cooperation of parallel evolving subpopulations improved solution quality and computing time speedups based on a parallel alternative and data profile.

A neural classifier optimized using the Backtracking Search optimization Algorithm (BSANN) to classify 3 mental tasks consisting of a right or left hand movement imagination and word generation was presented by Agarwal et al., [18]. The new method BSANN was tested on a publicly available BCI Competition 3-5 datasets. Result showed that BSANN exhibited better results than 21 other algorithms for mental tasks classification regarding classification accuracy.

Ding [19] proposed a new strategy combining with the SVM classifier for features selection that retains sufficient information for classification purpose. For improving classification precision, the variables optimization of the penalty constant  $C$  as well as the bandwidth of Radial Basis Function (RBF) kernels is a significant step in the establishment of effective as well as high-performance SVM model. Aiming at optimizing the parameters of SVM, also presented a grid based Ant Colony Optimization (ACO) algorithm to choose parameters  $C$  and  $\hat{\lambda}$  automatically for SVM instead of selecting parameters randomly by human's experience and traditional grid searching algorithm, so that the classification feature numbers can be reduced and the classification performance may be enhanced concurrently. Experiments prove the feasibility as well as efficacy of the method.

Gonzalez et al., [20] proposed a method for classifying single-trial ERPs using a combination of the Lifting Wavelet Transform (LWT), SVM and Particle Swarm Optimization (PSO). In particular, the LWT filters, the set of EEG channels and SVM parameters that maximize the classification accuracy are searched using PSO. The



authors evaluated the method's performance through offline analysed on the datasets from the BCI Competitions II and III. The proposed method achieved in most cases a similar or higher classification accuracy than that achieved by other methods, and adapted wavelet basis functions and channel sets that match the time-frequency and spatial properties of the P300 ERP.

Wang et al., [21] used GA-SVM hybrid algorithm with two purposes: Selecting of the optimal feature subset and deciding the parameters for SVM classifier after the features extracted through the algorithm called Sample Entropy. Compared with GA-based feature selection and GA-based parameters optimization for SVM, the GA-SVM hybrid algorithm has fewer input features and gain much higher classification accuracy.

Rathipriya et al., [22] suggested a hybrid algorithm to advance the classification achievement rate of MI-based ECoG in BCIs. To verify the effectiveness of the suggested classifier, the authors restored the SVM classifier with the identical features extracted from the cross-correlation method for the classification. The performances of those procedures are assessed with classification correctness through a 10-fold cross-validation procedure. The authors furthermore consider the performance of the suggested procedure by comparing it with existing system.

## METHODOLOGY

In this section, GA and Modified CLONALG are described. Genetic algorithm is a heuristic approach for resolving optimization issues. Currently, the approach is utilized in several research areas, but it initiated in the genetic sciences. Therefore, most terms which are utilized for describing the optimization problems are inherited directly from biological terms. To run the optimization process with a GA, first, the problem environment has to be defined, that is, a way of encoding problem solutions to the form of genetic algorithm individuals, fitness functions which are utilized for evaluation of individuals in all generations, genetic operations which are utilized for mixing as well as modifying individuals, an approach for choosing individuals as well as other extra genetic algorithm variables. Once problem environment is defined, the selected GA is applied to process individuals by a given number of generations.

### Genetic Algorithm (GA)

The general scheme of the classic GA, i.e. Holland algorithm from 1975, may be delineated as given below. In the initial stages of the protocol, a set of arbitrarily selected individuals, each coding one solution of the problem at hand, is created. The individuals are ranked as per selected criteria, in the form of fitness functions. Later, solutions of small values of fitness functions are discarded from the set of solutions. They are replaced by new solutions, which are created by combining together parts of solutions of high fitness (crossover stage). From time to time random alterations are made in the existing solutions. The alterations permit exploration of completely new regions of the problem space. The whole procedure is iterated till adequate solutions are discovered.

Different GAs can be used for feature selection. From all the protocols, the typically utilized one is the protocol which necessitates coding of all extricated features. As per this method, all genes of an individual relate to a single feature and contains the data as to whether the feature exists in the specified individual or not. The protocol is compatible with the traditional Holland algorithm which implies that it begins from arbitrary population and it employs one-gene mutation as well as one-point crossover. The quality of individuals generated in consequent iterations is appraised as per the accuracy of classifiers created separately for all individuals [23].

Because of arbitrary selection of genes to individuals of the initial population, all individuals contain around half of all potential features (assuming uniform distribution). In the case of spaces comprised of features extricated from raw EEG signals (comprised generally of a minimum of 102-103 features), initiating a procedure of looking for the optimum set of features from the middle of the set is not a profitable solution as it can disable considerable decrease of features from the set. This is because of direction of optimization procedure to increase classification precision that favors individual of greater accuracy, i.e. individuals that code solutions generating classifiers of higher number of free parameters (and so, higher number of features).

Theoretically, the optimization process has not to be guided purely by the classifier results. It is possible, for instance, to equip genetic algorithm fitness function with penalty term that penalizes individuals coding too huge quantity of features. It is possible to develop certain specialized genetic operators converting these unwelcome individuals to individuals carrying smaller number of features. In practice, however, the scale of the required reduction of the feature set is so large that it is extremely difficult to develop the stable function penalizing individuals which carry several features or functions to convert the individuals. Improved solution is to run genetic algorithm with individuals of restricted quantity of features, coding merely small subsets of the entire features set.

For applying the solution, certain alterations are to be made in the genotype. Firstly, it is not necessary to stick to binary coding; a better solution is to utilize integer genes. Second, all genes ought to encode index of a single feature from an entire set of features. With this method, a single individual comprises indexes of features which are to be delivered to classifier inputs. The adequate quantity of features (that is, the quantity of genes contained in a single individual) is set by user prior to launch of the protocol, with respect to the quantity of recorded observations as well as applied classifiers.

When these coding methods are employed, individuals with two or more equal genes may occur because of genetic operations or because of arbitrary selection of initial population. In some applications, e.g. in the travelling salesman problem, such an individual indicating a double visit in one city would be eliminated as an incorrect one. But in the case of features selection issue, guided by the classification accuracy, such an individual is not regarded as defective – rather, it may even be required. Repair is necessary as several usages of the same feature in the classifier do not make sense, but the repair involves discarding all but one of the genes coding the feature. Such individuals are desirable because if the precision of the classifier utilizing features coded in the individual was considerably high to permit the individual surviving the selection procedure, it would denote that further decrease in the quantity of features is possible.

The genetic algorithm controls a population of potential solutions to problems. The solutions are coded as binary chains. The group of chains represent the genetic material of a set of individuals. Artificial operators of selection, crossovers as well as mutations are employed in a stochastic search procedure for finding best individual through simulation of natural evolutionary procedure. All candidate solutions are linked to fitness values that measure the excellence of solutions. Hence, the fitness simulates environmental pressure of Darwin's natural evolution. A simplified GA pseudo-code structure [24]:

1. *Initialization of population*

2. *Evaluation of population*

3. *While Better fitness < Fitness Required do*

*Selection of parents*

*Crosses and mutations*

*Evaluation of population*

*End While*

GA is adaptive heuristic search protocol on the basis of evolutionary concepts of natural selection as well as genetics. The fundamental notion of GA is that it is formulated to mimic procedures in natural systems necessary for evolution. The main operator of GA to search in pool of possible solutions is Crossover, Mutation and selection.

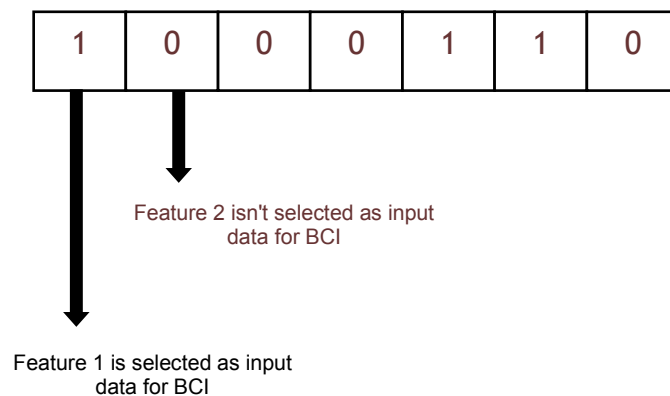
The genetic search process is iterative: evaluating, selection and recombining string in the population during each one of iterations (generation) until reaching some termination condition. Evaluating all strings is based on fitness functions that are problem dependent. It defines which of the potential solutions are better. It corresponds to environmental determination of survivability in natural selection.

Selection of a string, which represents a point in the search space, depends on the string's fitness relative to those of other strings in the population, those points that have relatively low fitness [25].

Mutation, like in natural environments, is a low probability operator and merely flips bit. The objective of mutation is the introduction of new genetic material into an existing individual; that is, to add diversity to the genetic characteristics of the population. Mutation is used in support of crossover to ensure that the full range of allele is accessible for each gene.

Crossover, whereas, is employed with great probability. It is a randomized though structured operator that permits information exchange between points. The goal is the preservation of fittest individual without introducing any new value.

The suggested method to the utilization of genetic algorithms for feature selection involves encoding a set of  $d$ , Feature  $s$  as a binary string of  $d$  elements, in which a 0 in the string indicates that the corresponding Feature has to be omitted, and 1 that it has to be included. The coding strategy denotes presence or absence of a certain Feature from the Feature space [ Figure 1 ]. The length of chromosome equal to Feature space dimensions.



**Fig. 1. Schema of the proposed GA-based feature selection approach**

GAs are computational models inspired by evolution. These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure, and apply recombination operators to these structures in such a way as to preserve critical information. Genetic algorithms are typically seen as function optimizers, though the range of issues to which genetic algorithms are employed is vast [26].

The present study uses a GA for feature selection. Thus, each member of the population was encoded with a binary string of length equal to the feature set size. Each bit of these strings represented one specific feature. If the bit value was '1', this feature was used for classification, if the value was '0' it was not used for classification. Therefore, each member of the population represented a feature subset.

The fitness value of each member of the population was calculated as the kappa coefficient achieved using the corresponding feature subset for classifying. This criterion of accuracy has also into account the distribution of wrong classifications and it was chosen because it was used to evaluate the submissions to the dataset. Population size was equal to the feature set size. The elitist selection was set to 2 and the roulette selection method was used. Single-point crossover with probability of 0.8 and uniform mutation with probability of 0.1 were applied to every generation in order to create the next population. The number of generations was set to 50. The GA searches for all feature subset sizes smaller than 15 features.

A GA with aggressive mutation has been proposed, where a detailed description of it can be found. In short, the most important algorithm features are as follows:

Step 1: An individual is composed of integer genes. The number of genes in an individual ( $N$ ) is fixed for the whole algorithm and is set either by the user or automatically based on the number of features, observations, and classifier parameters. Each gene either contains an index of one feature from the feature space or is equal to zero

Step 2: An initial population of  $M$  individuals is created randomly by choosing values from the interval  $f_0; 1; 2; \dots; P_g$ , where the values from 1 to  $P$  correspond to the feature indexes and the value zero corresponds to "none feature" state.

Step 3: The order of two main GA steps is reversed comparing to the classic scheme first the reproduction takes place and then the selection is performed.

Step 4: The basic genetic operation used in the algorithm is a mutation. This is a very aggressive form of mutation, as not only is each individual from the mother population mutated, but also each gene of that individual. The mutation scheme is as follows [27]:

```

for  $i = 1$  to  $M$ 
  take an individual  $i$ 
  for  $g = 1$  to  $N$ 
    take a gene  $g$ 
    assign a random value from the
    interval  $\{0, 1, 2, \dots, P\}$  to the gene  $g$ 
  save the individual  $i$  as a new individual
  
```

Step 5: The second genetic operation used in the algorithm is the classic Holland crossover performed on the mother population.

Step 6: After reproduction, the population is composed of:  $M$  mother individuals,  $NM$  of new individuals created during mutation and  $M$  new individuals created during crossover. All of these individuals are then evaluated according to their classification capabilities (i.e. a classifier is implemented and validated for each individual).

Step 7: The selection step is based on the discarding strategy in which only  $M$  individuals providing the highest classification accuracy remain in the population. Since all the best individuals from the last algorithm step (individuals from the mother population) take part in the selection process, this strategy guarantees that the best individual from the next population has at least the same fitness value as the best individual from the previous population. The population created in the selection process is a mother population for the next algorithm step.

Step 8: The reproduction and selection steps are repeated by a predetermined number of iterations.

The flowchart of GA as shown in [Figure- 2] [28]:

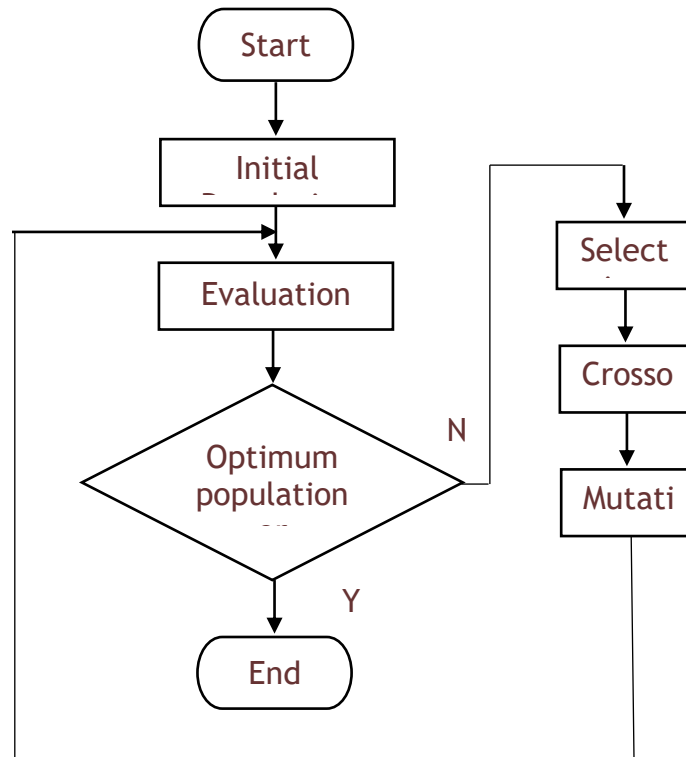


Fig: 2. Flowchart of Genetic Algorithm

## PROPOSED MODIFIED CLONALG

The proposed Modified CLONALG consists in an outer (GA) search loop where the current population is checked for constraint violation and then divided into feasible (antigens) and infeasible individuals (antibodies). If there are no feasible individuals, the best infeasible one (that with the lowest constraint violation) is moved to the antigen population. Here, AIS is given as inner loop wherein antibodies are first cloned and then mutated. Next, the distances (affinities) between antibodies and antigens are computed. Those with higher affinity (smaller sum of distances) are selected thus defining the new antibodies (closer to the feasible region). This (CLONALG) cycle is repeated a number of times. The resulting antibody population is then passed to the GA where constraint violations are computed as well as fitness function values for the feasible individuals. The selection operation is then performed in order to apply recombination and mutation operators to the selected parents producing a new population and finishing the external genetic algorithm loop. The selection process in the genetic algorithm comprises in binary tournaments wherein all individuals are chosen once and the opponent is arbitrarily drawn, with substitutes from the population. The tournament rules are as follows:

- Feasible individuals are preferred to infeasible ones
- Between two feasible individuals, that with the greater fitness value is chosen, and
- Between two infeasible individuals, that with the lesser constraint violation is selected. It is to be observed here the affinity is computed from the sum of phenotypical distances between individuals, employing a standard Euclidean vector norm.

Pseudo-code for the suggested hybrid is as follows:

```

Begin
for i = 0 to number Of Generations GA do
computeViolation();
dividePopulation();
antibodies < - infeasiblePop();
antigens < - TopFeasible();
for j = 0 to number Of Iterations CLONALG do
cloneAntibodies();
mutateAntibodies(); computeDistanceAntibodiesAntigens();
antibodies < - selectBetterAntibodies();
end - do;
computeViolationAntibodies();
computeFitnessFeasiblePop();
tournamentSelection();
crossover();
mutation();
end - do;
End
  
```

The main goal of the research reported in this paper was to explore alternative constraint-handling schemes for GAs used in global optimization. In the past, it have explored the use of penalty functions in engineering optimization. However, this previous work indicated a high correlation between performance of the GA and the fine-tuning of its penalty factors. Additionally, an important issue for us was not to increase in an important way, the number of fitness function evaluations (as when using, for example, the coevolutionary penalties, which do not require a manual fine-tuning of the penalty factors, but whose computational cost is extremely high).

The previous requirements led us to the development of the approach proposed in this paper. In the proposed approach, it use the search engine of the GA to conduct the search towards the global optimum. However, the GA is hybridized with a scheme inspired on an artificial immune model, which acts as a local search mechanism that helps the GA to reach the feasible region in a more efficient way. Since this local search mechanism is based only on similarities between chromosomic strings, no additional evaluations of the fitness function are required. Thus, it keep a low computational cost for the approach, which was one of its main design goals.

The proposed algorithm emulates the invaders recognition process by combining antibodies' libraries in order to attain antigen specificity. Furthermore, the purpose is to learn to identify the proper antibodies. The search process of the approach is led by a GA. Thus, what it propose is a scheme in which a simple emulation of an artificial immune system is embedded into a GA. Note however that the computational complexity of the approach is not really  $O(N^2)$ , because the internal scheme (i.e., the artificial immune system) does not evaluate the original fitness function of the problem as it will see later on [Figure 1]. This internal scheme (which is indeed another GA) guides its search based on string similarities and not on objective function values [29].

**A Serial Version of the Proposed Algorithm:** The serial algorithm version to handle constraints using an immune system is described below [Figure- 1]:

**Step 1:** Generate randomly an initial population for the GA.

BEGIN inner GA

**Step 2:** If the initial population contains a mixture of feasible and infeasible individuals, then it divide the population in two groups. The first group contains the infeasible individuals, which are denominated "antibodies", and the second contains the feasible individuals, which are called "antigens".

**Step 3:** If none of the individuals in the initial population is feasible, then it use the magnitude of constraint violation of each individual as its fitness. Then, it use the best individual in the population as the "antigen", where "best" refers to the individual with the lowest amount of constraint violation.

**Step 4:** Select randomly a sample of antibodies of size  $\sigma$ .

**Step 5:** The fitness of the sample of antibodies is computed according to their similarity with a set of antigens in the following way:

- An antigen is randomly selected from the antigens population.
- Each antibody in the sample is compared against the antigen selected, and it compute the result of the comparison, to which it will call Z (matching magnitude). Z represents a distance (normally but not necessarily Euclidean) measured at the genotype level (i.e., at the level of the chromosomal encoding). Z is computed using:

$$Z = \sum_{i=1}^L t_i$$

Where  $t_i = 1$  if there is a matching at position  $i=1, \dots, L$  ( $L$  is the length of the chromosome), or zero if there is no match. A large  $Z$  value means a high matching between the two strings compared and, therefore, a high fitness value.

**Step 6:** Based on the fitness computed in the previous step, the population of antibodies is reproduced in a traditional GA (using crossover and mutation).

**Step 7:** The process is repeated from the fourth step until convergence (e.g., when the mean and the maximum fitness in the population are practically the same) or until it reach a maximum number of iterations.

**Step 8:** Individuals are returned to the external GA and it proceed in the conventional way.

END inner GA

**Step 9:** Apply binary tournament selection (with the objective function of the problem) using special rules (as described below).

**Step 10:** Apply crossover and mutation in a conventional way.

**Step 11:** The process is repeated from step 2 until reaching stopping condition.

The binary tournament used in step 9 is defined in the following way (two individuals are compared each time):

- If one individual is infeasible and the other one is feasible, then the feasible individual wins.
- If both individuals are feasible, then the one with the highest fitness value is the winner.
- If both individuals are infeasible, then the winner is the one with the lowest constraint violation value.

Few problems are required to be mentioned. First, as it indicated before, the approach is really using a GA embedded inside another GA used to optimize a certain function. However, the GA that is run with the emulation of the immune system does not use the fitness function directly; it only computes Hamming distances, which are very inexpensive with respect to evaluating the objective function of the problem. Also, the implicit premise of the technique is that, under certain conditions, the reduction of genotypic differences between two individuals will produce, as a consequence, a phenotypic similarity, which, in the case, will make that an infeasible individual approaches the feasible region. The algorithm is an extension of the proposal of Hajela & Lee.

To clarify the way in which the approach works, it provide next both, the internal and the external GA's adopted:

#### Internal GA

**Step 1:** Initialize the fitness of all antibodies to zero.

**Step 2:** Compute the fitness of the antibody pool based on similarity to the antigens (or based on complementarity); this requires the following specific steps:

- An antigen is selected at random.
- A sample of antibodies of size  $\mu$  is selected from the antibody pool without replacement.
- The match score of each antibody is computed by comparing against the selected antigen, and the antibody with the highest score has the match score added to its fitness value; the fitness of the other antibodies is unchanged.
- The antibodies are then returned to the antibody population, and the process is repeated a number of times (typically two or three times the antibody population size).

**Step 3:** Based on the fitness computed in Step 2, a GA simulation is conducted with prescribed probabilities of crossover and mutation to evolve the antibody population through one generation of evolution.

**Step 4:** The process is then repeated from Step 1 until convergence in the antibody population is attained.

#### External GA

**Step 1:** A population of designs is randomly generated.

**Step 2:** The fitness function, a composite of the objective function and a penalty associated with constraint violation, is obtained for the entire population.

**Step 3:** Members within the top 3% of the population obtained at the end of Step 2 are designated as antigens, and the entire population (including the antigens) is defined as the starting population of antibodies.

**Step 4:** Using an antibody sample size  $\mu$  smaller than the number of antigens, the degree of match  $Z$  is obtained for each member of the population according to the steps described before.

**Step 5:** The match score of each design is used as a fitness measure in a traditional selection or reproduction operation. During this reproduction operation, the size of the population is unchanged.

**Step 6:** The crossover and mutation operations are performed on the new population of antibodies formed in Step 5.

**Step 7:** The process is then repeated from Step 2 with an intent of evolving the population to maximize the  $Z$  function and cycled to convergence.

Note that the proposed approach is designed to operate only on binary strings. Although it know that the binary alphabet can be used to encode any type of decision variables, it may be useful in some cases to use alternative encodings. Should that be the case, the proposed approach is not directly applicable, and its generalization to alphabets of higher cardinality (e.g., real-numbers encoding) remains as an open research area.

At this point, it is important to clarify the main weaknesses that it identified in Hajela's algorithm and that led us to develop the algorithm proposed herein:

Hajela's approach requires a penalty function in order to sort the population and assign the antigens. This makes necessary to evaluate twice the objective function of the problem (per individual) at each generation of the external GA. This may become considerably expensive (computationally speaking) when dealing with real-world applications. In contrast, the approach only evaluates once the objective function (for each individual) per generation, since it do not use a penalty function. It relate the values of the antigens to the constraint violation of each solution. This keeps us from evaluating the fitness function more than once and makes unnecessary to sort the population.

In Hajela's approach, the computation of the fitness Z in the internal GA is performed through a cycle in which the population of antibodies must be traversed several times. In the case, it compute Z in the internal GA by performing a single traversal of the antibodies.

The approach of Hajela& Lee is only validated with a few engineering optimization problems, and no information about its computational cost is provided. In the case, it have used some benchmarks reported in the evolutionary computation literature and it have compared the results against a highly competitive constraint-handling technique which is representative of the state-of-the-art in the area (the homomorphous maps).

A schematic of the serial version of the algorithm based on the artificial immune system as shown in [Figure- 3]

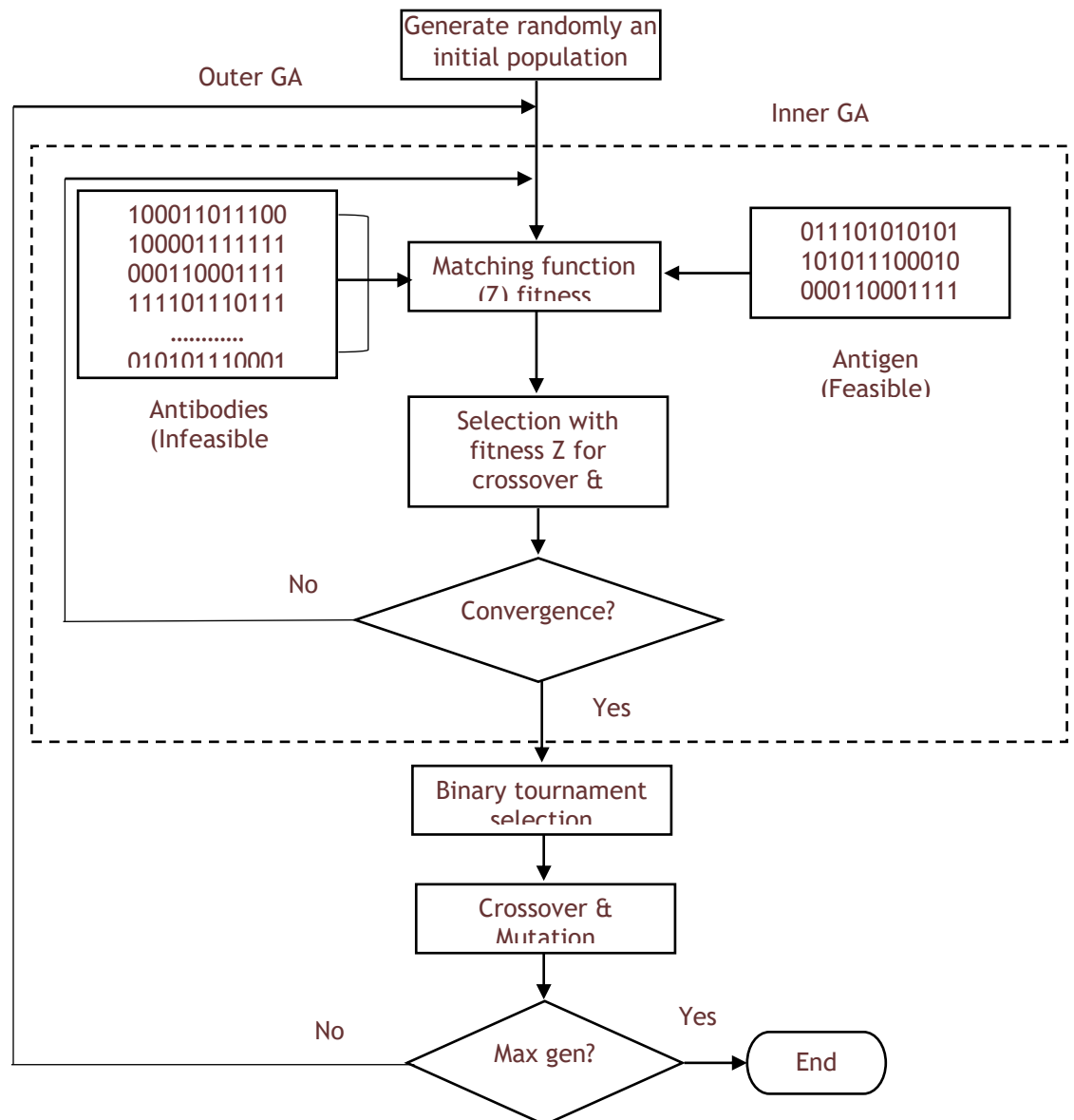


Fig: 3 (A). schematic of the serial version of the algorithm based on the artificial immune system

## RESULTS

In this section, the Autoregressive - Wavelet fused features, Modified CLONALG Feature Selection, SVM classifier (ARWAMCSVM), Autoregressive - Wavelet fused features, Hybrid CLONALG Feature Selection, SVM classifier (ARWAHCSVM), Autoregressive - Wavelet fused features, Modified CLONALG Feature Selection, Optimized SVM classifier (ARWAMCSVM-Opt) and Autoregressive - Wavelet fused features, Hybrid CLONALG Feature Selection, Hybrid Optimized SVM classifier (ARWAHCSVM-HOpt) are evaluated. [Table-1] shows the summary of results obtained. [Figure- 4], [Figure- 5], [Figure- 6] shows the classification accuracy, precision and recall respectively.

Table: 1. Summary of Results

	Classification Accuracy	Precision	Recall
ARWAMCSVM	84.17	0.84205	0.8417
ARWAHCSVM	92.81	0.92815	0.9281
ARWAMCSVM-Opt	95.68	0.95695	0.9568
ARWAHCSVM-HOpt	96.76	0.9678	0.9676

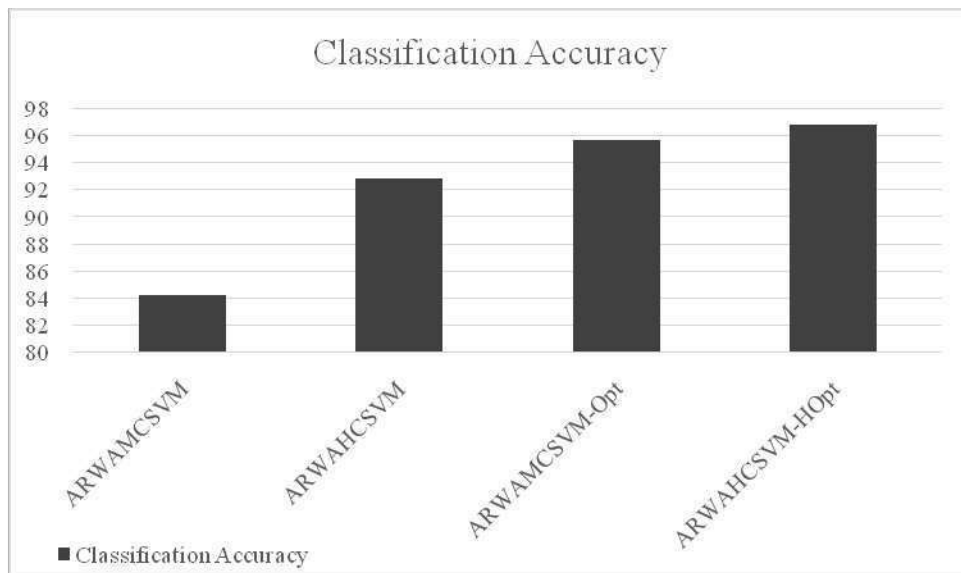


Fig: 4 .Classification Accuracy

From Table- 1 and [Figure -4] it is observed that the classification accuracy of ARWAHCSVM-HOpt performs better by 13.9% than ARWAMCSVM, by 4.17% than ARWAHCSVM and by 1.12% than ARWAMCSVM-Opt.



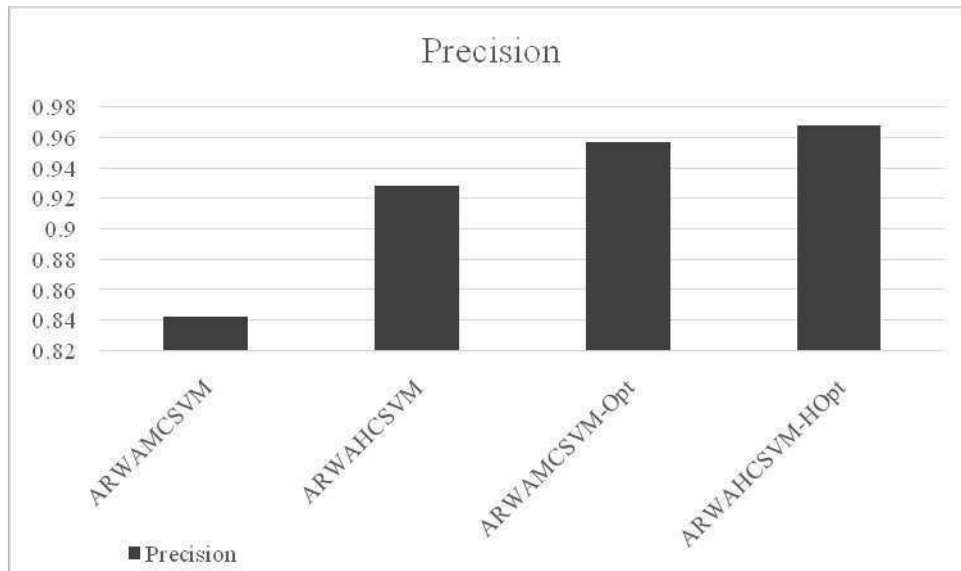


Fig: 5. Precision

From Table- 1 and [Figure- -5] it is observed that the precision of ARWAHCSVM-HOpt performs better by 13.9% than ARWAMCSVM, by 4.18% than ARWAHCSVM and by 1.13% than ARWAMCSVM-Opt.

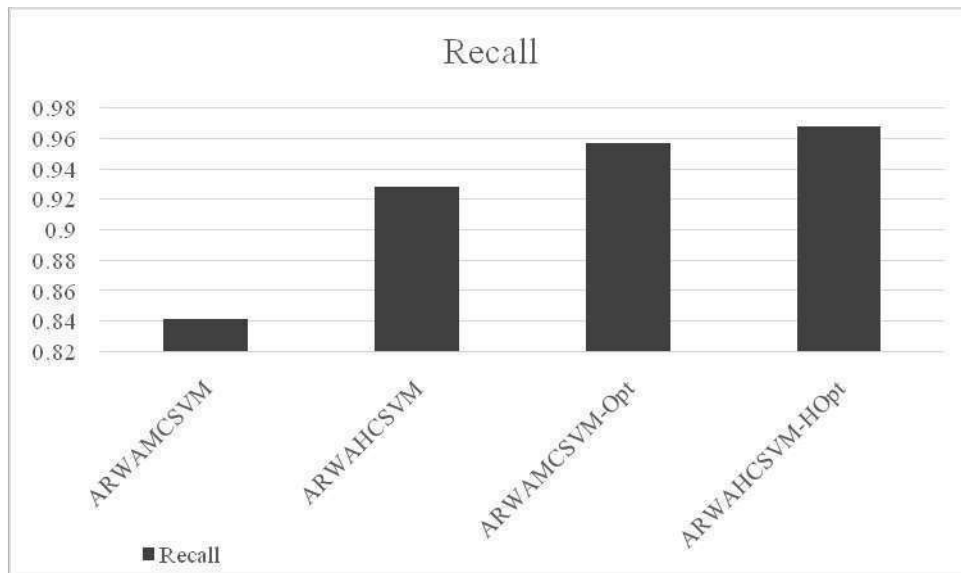


Fig: 6.Recall

From Table- 1 and [Figure- 6] it is observed that the recall of ARWAHCSVM-HOpt performs better by 13.9% than ARWAMCSVM, by 4.17% than ARWAHCSVM and by 1.12% than ARWAMCSVM-Opt.

## CONCLUSION

ECG includes a spatial scale between EEG and intra-cortical microelectrode recording, and ECoG offers a balance between invasiveness, spatiotemporal resolution, and signal stability for BCI applications. BCI has progressed, but it is slowed by many factors including noise in brain signals, muscular artefacts and inconsistency and variability of user attention/intentions. In this paper proposed modified CLONALG optimizes SVM. For

intensification, the strategy works with many clones to improve. Experiments were undertaken through tenfold cross validation and accuracy achieved is satisfactory but further work is needed for classification accuracy improvement.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Ko M, Bae K, Oh G, Ryu T. [2009] A study on new gameplay based on brain-computer interface. In Breaking New Ground: Innovation in Games, Play, Practice and Theory: Proceedings of the 2009 Digital Games Research Association Conference, Brunel University.
- [2] Leuthardt EC, Schalk G, Wolpaw JR, Ojemann J G, Moran DW. [2004] A brain-computer interface using electrocorticographic signals in humans, *Journal of neural engineering*, 1(2):63.
- [3] Miller KJ, Abel TJ, Hebb AO, Ojemann JG. [2011] Rapid online language mapping with electrocorticography. *Journal of neurosurgery Pediatrics*, 7(5):482.
- [4] Al-Qudah AA, Tamimi AF, Ghanem S. [2000] Electrocorticography in the management of surgically treated epileptic patients. *Neurosciences*, 5(1):22-25.
- [5] Shenoy P, Miller KJ, Ojemann JG, Rao RP. [2008] Generalized features for electrocorticographic BCIs. *Biomedical Engineering, IEEE Transactions on*, 55(1):273-280.
- [6] Nicolas-Alonso LF, Gomez-Gil J. [2012] Brain computer interfaces, a review, *Sensors*, 12(2):1211-1279.
- [7] Brownlee J. [2007] Clonal selection algorithms. Complex Intelligent Systems Laboratory, Swinburne University of Technology, Australia.
- [8] Koklu M, Kahramanli H, Allahverdi N. [2012] A New Approach To Classification Rule Extraction Problem by the Real Value Coding. *International Journal of Innovative Computing, Information and Control*, 8(9).
- [9] Aydin I, Karakose M, Akin E. [2011] A multi-objective artificial immune algorithm for parameter optimization in support vector machine. *Applied Soft Computing*, 11(1): 120-129.
- [10] Buzón-Cantera IE, Mora-Vargas J, Ruiz, A, Soriano P. [2015] A Hybrid Optimization Model: An Approach for the Humanitarian Aid Distribution Problem, *Applied Mathematical Sciences*, 9(127): 6329-6346.
- [11] Shende PM, Jabade VS. [2015, January] Literature review of brain computer interface (BCI) using Electroencephalogram signal. In Pervasive Computing (ICPC), 2015 International Conference on (pp. 1-5). IEEE.
- [12] Kim M, Kim BH, Jo S. [2015] Quantitative Evaluation of a Low-Cost Noninvasive Hybrid Interface Based on EEG and Eye Movement. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 23(2): 159-168.
- [13] Fan XA, Bi L, Teng T, Ding H, Liu Y. [2015] A brain-computer interface-based vehicle destination selection system using P300 and SSVEP signals. *Intelligent Transportation Systems, IEEE Transactions on*, 16(1): 274-283.
- [14] Krusienski DJ, Shih JJ. [2010, August] A case study on the relation between electroencephalographic and electrocorticographic event-related potentials. In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, 6019-6022.
- [15] Hamedi M, Salleh S, Ting CM. [2014, September] Multiclass self-paced motor imagery temporal features classification using least-square support vector machine. In Functional Electrical Stimulation Society Annual Conference (IFESS), 2014 IEEE 19th International (pp. 1-5). IEEE.
- [16] Gao W, Guan JA, Gao J, Zhou D. [2015] Multi-ganglion ANN based feature learning with application to P300-BCI signal classification. *Biomedical Signal Processing and Control*, 18: 127-137.
- [17] Kimovski D, Ortega J, Ortiz A, Baños, R. [2015] Parallel alternatives for evolutionary multi-objective optimization in unsupervised feature selection. *Expert Systems with Applications*, 42(9): 4239-4252.
- [18] Agarwal SK, Shah S, Kumar R. [2015] Classification of mental tasks from EEG data using backtracking search optimization based neural classifier. *Neurocomputing*.
- [19] Ding S. [2009, November] Feature selection based F-score and ACO algorithm in support vector machine. In Knowledge Acquisition and Modeling, 2009. KAM'09. Second International Symposium on IEEE, 1: 19-23).
- [20] Gonzalez A, Nambu I, Hokari H, Iwahashi M, Wada Y. [2013, October] Towards the Classification of Single-Trial Event-Related Potentials Using Adapted Wavelets and Particle Swarm Optimization. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 3089-3094). IEEE.
- [21] Wang L, Xu G, Wang J, Yang S, Guo L, Yan W. [2011, July] GA-SVM based feature selection and parameters optimization for BCI research. In Natural Computation (ICNC), 2011 Seventh International Conference on, 1: 580-583). IEEE.
- [22] Rathipriya N, Deepajothi S, Rajendran T. [2013] Classification of motor imagery ecog signals using support vector machine for brain computer interface. In Advanced Computing (ICoAC), 2013 Fifth International Conference on (pp. 63-66). IEEE.
- [23] Rejer I, Lorenz K. [2013] Genetic algorithm and forward method for feature selection in EEG feature space.

*Journal of Theoretical and Applied Computer Science*,  
7(2): 72-82.

- [24] Atum Y, Gareis I, Gentiletti G, Acevedo R, Rufiner L. [2010] Genetic feature selection to optimally detect P300 in brain computer interfaces. In Engineering in Medicine and Biology Society (EMBC), 2010 *Annual International Conference of the IEEE ,IEEE*, 3289-3292.
- [25] Ghanbari AA, Broum A, Navidi H, Ahmadi A. [2012] Brain computer interface with genetic algorithm. *International Journal of Information and Communication Technology Research*, 2.
- [26] Corralejo R, Hornero R, Alvarez D. [2011, August] Feature selection using a genetic algorithm in a motor imagery-based Brain Computer Interface. In Engineering in Medicine and Biology Society, EMBC, 2011 *Annual International Conference of the IEEE* (pp. 7703-7706). *IEEE*.
- [27] Rejer I. [2015] Genetic Algorithms for Feature Selection for Brain-Computer Interface. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(05):1559008.
- [28] Moon SN, Bawane N. [2015] OPTIMAL FEATURE SELECTION BY GENETIC ALGORITHM FOR CLASSIFICATION USING NEURAL NETWORK.
- [29] Coello CAC, Cortés NC. [2004] Hybridizing a genetic algorithm with an artificial immune system for global optimization. *Engineering Optimization*. Doi: 10.1080/03052150410001704845

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and checked/reviewed by guest editor.

# AN ENHANCED POWER BASED MAC PROTOCOL FOR WIRELESS NETWORKS

Prathapa Reddy<sup>1\*</sup>, Subramanyam<sup>2</sup>, Satyaprasad<sup>3</sup>

<sup>1</sup>JNTUK, Kakinada, INDIA

<sup>2</sup>Santhiram Engineering College, Nandyal, A.P, INDIA

<sup>3</sup>JNT University Kakinada, Kakinada, A.P, INDIA

## ABSTRACT

Power based connectivity is an ad hoc research topic. Implementing power control mechanisms enhance network life and to improve throughput, locate cost effective routes and spatial reuse. This paper uses MAC layer performs flow control, error detection, framing, correction, and physical addressing. This layer includes functions/procedures required for transmitting data between two or more network nodes. MAC protocol design should tackle issues generated by node roaming and unreliable time varying channels. MAC protocol is used to implement coordination functions and power control mechanisms through the use of fuzzy rules generated by upper network layers using two input variables like link quality and node neighborhood count with optimal power consumption level as the output variable. Experiments using fuzzy rules are compared to DSR routing and two hop power control methods. Result showed that two hop power control with fuzzy logic method yields lower route discovery time, increased cache replies, minimum simulation time and end to end delay when compared to DSR routing and two hop routing protocol.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Power Control, Medium Access Control (MAC), Energy Saving, Fuzzy logic

## INTRODUCTION

An ad hoc network consists of small, mobile devices using wireless medium to communicate amongst them. Most ad hoc networks have no fixed infrastructure or centralized administration. As nodes join/leave or roam the network, network topology is dynamic. The network self organizes and self-configures and so differs from wired networks. Ad hoc network applications range from disaster recovery, distributed computing, and battlefield surveillance. Most nodes have limited battery power and bandwidth. Routing protocols are not required [1] when 2 nodes are within transmission range of each other.

Some ad hoc network advantages are the following:

- No central network administration,
- Self-configuring nature (nodes act as routers),
- Self-healing through periodic re-configuration,
- Scalable when nodes are added,
- Flexibility (being accessible using Internet from varied, multiple locations).
- Its limitations include:
  - Every node should cooperate to ensure a highperformance,
  - Throughput depends on system load,
  - Reliability maintenance requires enough available nodes,
  - Huge latency/time delay is experienced in large networks [2]

As wireless medium is used by nodes, it results in collision without successful transmission when multiple hosts attempt to forward data simultaneously. Hence, recourse is taken to medium access control (MAC) protocol [3].

Carrier Sense Multiple Access (CSMA) avoids collisions. When a node wants to send a packet it listens to medium, for a fixed time. When no traffic is sensed, it then starts to transmit. When 2 two nodes transmit simultaneously without knowledge about other's transmission, it leads to a collision. But CSMA-style MAC protocols ensure channel utilization of around 50% - 80% based on access policy used [5].

As ad hoc network nodes are battery operated devices, energy conservation, and power control are the main aims. Ad hoc networks power consumption is reduced by either controlling transmission power or choosing optimal transmission [6] routes. Power based connectivity is a recent concept in wireless ad-hoc networks. Reducing the power consumption at each node will increase the end-to-end throughput. A node's power availability controls transmission power which in turn affects signal quality and determines neighboring nodes status. Hence, this affects network layer, as interference leads to congestion affecting transport layer.

MAC protocols are designed to decide maximum transmission power to forward request-to-send (RTS), and clear-to-send (CTS) packets to save energy thereby determining minimum power needed for data transmission and packet acknowledgement (ACK). Power required for communication is specified through three components: They are  $P_{Rxelec}$ ,  $P_{Txelec}$ , and  $P_{TxRad}(p)$ .  $P_{Rxelec}$  is power consumed by node receiver,  $P_{Txelec}$  is power consumed by transmitter electronics, and  $P_{TxRad}(p)$  is power consumed by power amplifier for packet transmission at power level, where  $p$  is the actual power radiated [7].

Energy consumption when packet is transmitted through the use of hops from sender node to receiver node with distance  $d$ , is given by

$$(d/r)(P_{Rxelec} + P_{Txelec} + cr^\alpha) \quad (1)$$

which is minimized at,

$$r_{crit} = \alpha \sqrt[\alpha]{\frac{P_{Rxelec} + P_{Txelec}}{c(\alpha - 1)}} \quad (2)$$

To satisfy network connectivity, available power range should be greater than  $r_{crit}$  [8,9].

Power Saving processes proposes the use of 2 varying power states at a node [10]:

**Awake:** wireless interface of a node is powered to transmit/receive. In this state, a node transmits/receives or is idle.

**Doze:** The node's wireless interface is powered down when it can neither transmit/receive.

## RELATED WORK

Energy Efficient Routing Protocol with Adaptive Fuzzy Threshold Energy for MANETs was presented by Hiremath and Joshi [11]. MANET's life was affected by node life. A new on-demand routing based protocol was proposed to conserve energy in mobile nodes, to increase MANET life, the suggested methodology being based on adaptive fuzzy threshold holding of residual nodes energy, participating in route discovery from sender to receiver. Experiments were undertaken, and results compared to Load-Aware Energy Efficient Protocol (LAEE) protocol which proved that AFTE was better than LAEE. Improvement in average network life was 13% at the first node failure, 15% when 50% node failure was considered and 23% when 100% node failure was considered in comparison to LAEE.

Fuzzy-controlled Power-aware Routing Protocol (FPRP) for Mobile Ad Hoc Networks was presented by Banerjee and Dutta [12]. Routing decisions dynamically at nodes to form a closer-optimal power-efficient end-to end route to forward data packets in Fuzzy-controlled power aware routing protocol (FPRP). The protocol was distributed so that only neighboring nodes location information was exploited in every routing node. Routers life status was measured through a fuzzy controller called route decider. Fuzzy controller used rate of depletion, residual charge, communication load, and node proximity. Simulation showed FPRP produced major improvements when compared with other power aware ad hoc network routing protocols, even when node numbers exceeded 2000.

A reliable Energy-Efficient Multi-Level Routing Algorithm for Sensor Networks was presented by Yu [13], Fuzzy Petri nets selected cluster heads in this knowledge-based inference approach. Fuzzy logic's reasoning method calculated reliability degree in route budding tree from cluster heads to the base station. Hence, the best and reliable route among cluster heads was constructed. The algorithm provided an idea to balance each node's energy load, providing global reliability for the network. Every iteration has three phases: clustering phase, multi-hop routing phase and data transmission phase. In clustering phase, a cluster heads set was elected, and remaining nodes were cluster members. In the second phase, multi-hop route was generated while in the data transmission

phase, every cluster member node forwarded a specific data amount to the cluster head. All cluster heads aggregated received data and forwarded them to the base station through multi-hop routing. Simulation results demonstrated that network life was prolonged, and energy consumption reduced.

An Adaptive Power Control Based Spectrum Handover for Cognitive Radio Networks was presented by Lu, et al., [14] which proposed a spectrum handover scheme combining dynamic spectrum allocation and power control to reduce unnecessary handovers. This procedure enhanced overall network performance, improving spectral efficiency. Users were split into primary users (PU) and secondary users (SUs). SUs used licensed channels till primary users (PUs) aggregate interference, did not exceed predetermined thresholds. When a PU arrived, an SU calculated the maximum transmission power that the SU did not interfere with the PU, if the SU could reach its receiver and continue transmission with reduced power. Otherwise, it switched over to an idle band. Experiments showed this scheme reduced spectrum handover ratio and improved effective data rate by 30%.

A Fuzzy Logic Approach to Beaconing for Vehicular Ad hoc Networks was presented by Ghafoor, et al. [15]. Vehicular Ad Hoc Network (VANET) is a new technology used for intra-vehicular communication when fixed infrastructures were absent. An Adaptive Beaconing Rate (ABR) approach was for VANETs based on fuzzy logic to control beaconing frequency by considering current traffic characteristics. ABR took one direction vehicles percentage and their status as fuzzy decision making system inputs. Beaconing rate tuning is based on vehicular traffic characteristics.

## MATERIALS AND METHOD

Fuzzy logic (FL) is a reasoning approach, specifying degrees of truthiness instead of Boolean value (true or false) used by computers. Fuzzy logic has 0 and 1 as extreme cases of true and false respectively and also includes truth's various states in between. FL nearly resembles human thinking.

FL is implemented as follows [16,17]:

- Fuzzification – This transforms crisp data into fuzzy data/Membership Functions.
- Fuzzy Inference Process – This combines membership functions with if-then rules to reach a fuzzy output.
- Defuzzification – This uses various methods to calculate different outputs storing them in a lookup table. When an application is executed, output is taken from the lookup table based on the current input variables

A fuzzy [18] IF-THEN rule includes an IF part (antecedent) and THEN part (consequent) where antecedent combines two or more terms, and consequent one term.

Fuzzy logic/fuzzy sets are techniques to control uncertainty in many applications. Fuzzy logic is used by MAC layer to control nodes power consumption in this work. A network layers upper layer generates fuzzy rules. Neighborhood node count and link quality parameters are inputs in fuzzy rule formation. The neighborhood node count's numeric values are represented in 3 terms, low, medium and high. The same 3 terms also represent link quality. Power level to be used by a node is specified by states of very low, low, medium, high and very high link quality. A node's power usage is based on link quality and neighborhood nodes number.

Block diagram for the work is given:

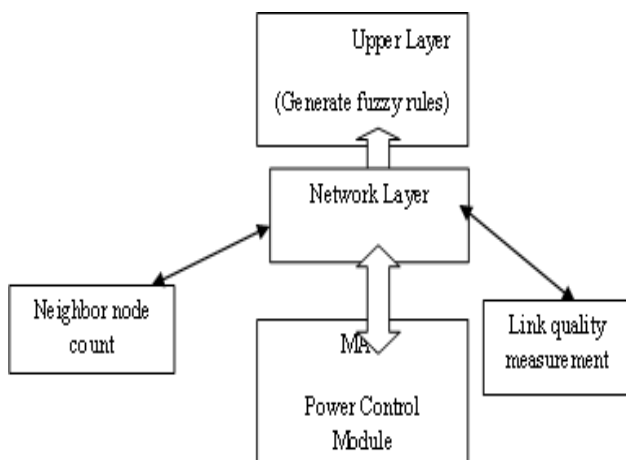


Fig: 1. Block diagram of proposed approach

## RESULTS

For the experiments link quality and neighbor node count are taken from the definition points, and shape parameters given below and termed as three states low, medium and high. Output variable power is taken from the definition points and termed as states very low, low, medium, high and very high.

- Input Variable "LQ"

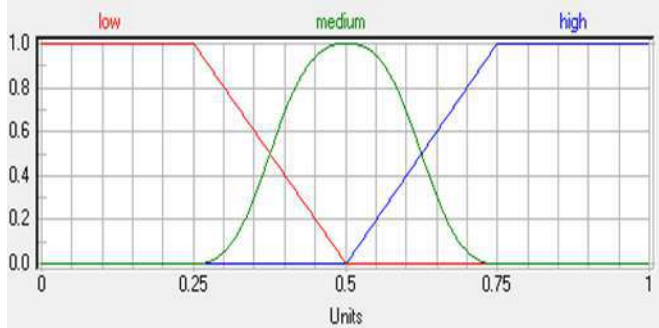


Fig: 2. MBF of "LQ"

Table: 1. Definition Points of MBF "LQ"

Term Name	Shape/Par.	Definition Points (x, y)		
Low	Linear	(0, 1)	(0.25, 1)	(0.5, 0)
		(1, 0)		
Medium	S-Shape/0.50	(0, 0)	(0.25, 0)	(0.5, 1)
		(0.75, 0)	(1, 0)	
High	Linear	(0, 0)	(0.5, 0)	(0.75, 1)
		(1, 1)		

Input Variable "NNC"

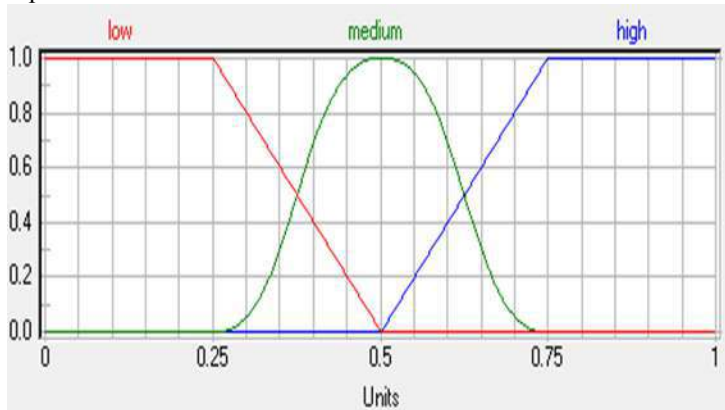


Fig: 3. MBF of "NNC"

Table: 2. Definition Points of MBF "NNC"

Term Name	Shape/Par.	Definition Points (x, y)		
Low	Linear	(0, 1)	(0.25, 1)	(0.5, 0)
		(1, 0)		
Medium	S-Shape/0.50	(0, 0)	(0.25, 0)	(0.5, 1)
		(0.75, 0)	(1, 0)	
High	Linear	(0, 0)	(0.5, 0)	(0.75, 1)
		(1, 1)		

- Output Variable "Power"

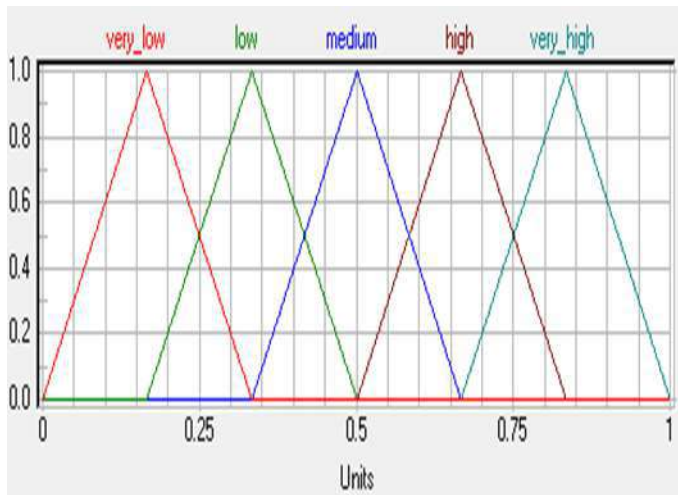


Fig:4. MBF of "Power"

Table: 3. Definition Points of MBF "Power"

Term Name	Shape/ Par.	Definition Points (x, y)		
very_low	Linear	(0, 0)	(0.16666, 1)	(0.33334, 0)
		(1, 0)		
Low	Linear	(0, 0)	(0.16666, 0)	(0.33334, 1)
		(0.5, 0)	(1, 0)	
medium	Linear	(0, 0)	(0.33334, 0)	(0.5, 1)
		(0.66666, 0)	(1, 0)	
High	Linear	(0, 0)	(0.5, 0)	(0.66666, 1)
		(0.83334, 0)	(1, 0)	
very_high	Linear	(0, 0)	(0.66666, 0)	(0.83334, 1)
		(1, 0)		

- Rule Block "RB1"

Using various combinations of link quality and neighborhood node count levels, minmax aggregation is used to form a set of rules.

Parameter Table: 4. Aggregation

Aggregation:	MINMAX
Parameter:	0.00
Result Aggregation:	MAX
Number of Inputs:	2
Number of Outputs:	1
Number of Rules:	45



Table: 5. Rules of the Rule Block "RB1"

IF		THEN	
LQ	NNC	DoS	Power
Low	low	0.20	very_low
Low	low	0.63	Low
Low	low	0.14	Medium
Low	low	0.74	High
Low	low	0.27	very_high
Low	medium	0.13	very_low
Low	medium	0.51	Low
Low	medium	0.16	Medium
Low	medium	0.23	High
Low	medium	0.12	very_high
Low	high	0.09	very_low
Low	high	0.96	Low
Low	high	0.48	Medium
Low	high	0.38	High
Low	high	0.29	very_high
Medium	low	0.27	very_low
Medium	low	0.30	Low
Medium	low	0.60	Medium
Medium	low	0.91	High
Medium	low	0.55	very_high
Medium	medium	0.84	very_low
Medium	medium	0.13	Low
Medium	medium	0.97	Medium
Medium	medium	0.17	High
Medium	medium	0.60	very_high
Medium	high	0.20	very_low
Medium	high	0.18	Low
Medium	high	0.14	Medium
Medium	high	0.34	High
Medium	high	0.68	very_high
High	low	0.80	very_low
High	low	0.52	Low
High	low	0.92	Medium
High	low	0.34	High
High	low	0.09	very_high
High	medium	0.62	very_low
High	medium	0.86	Low
High	medium	0.71	Medium
High	medium	0.61	High
High	medium	0.47	very_high
High	high	0.09	very_low
High	high	0.51	Low
High	high	0.30	Medium
High	high	0.59	High
High	high	0.73	very_high

For comparing, the performance of the proposed fuzzy logic two hop power conserving method, average route discovery time, end to end delay and simulation time are used as parameters. Results are shown from [Figure -5], [Figure- 6], [Figure- 7], [Figure- 8]. Result reveals that proposed fuzzy logic method yields less route discovery time, increased cache replies, minimum simulation time and end to end delay when comparing to DSR routing and two hop routing protocol.

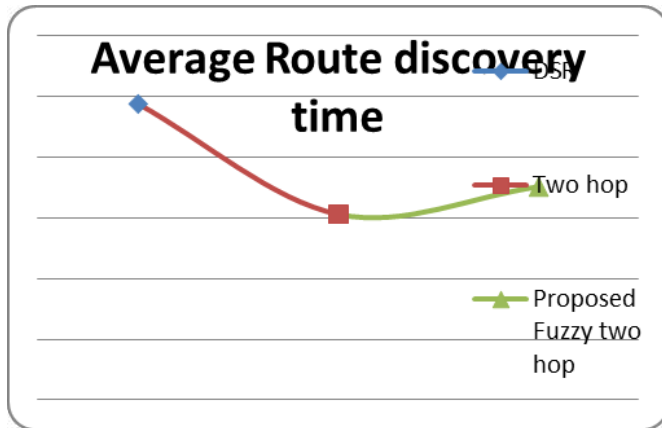


Fig: 5. Average Route discovery time

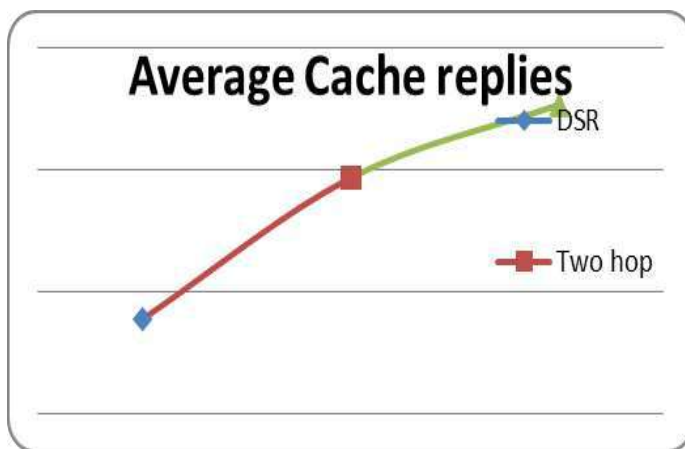


Fig: 6. Average cache replies

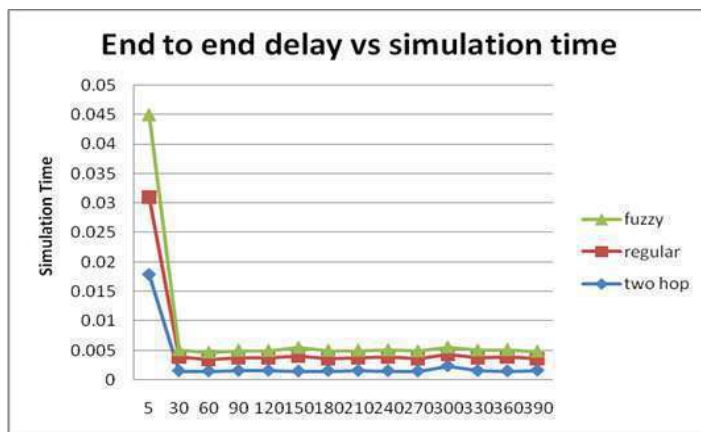


Fig: 7. end to end delay Vs Simulation time

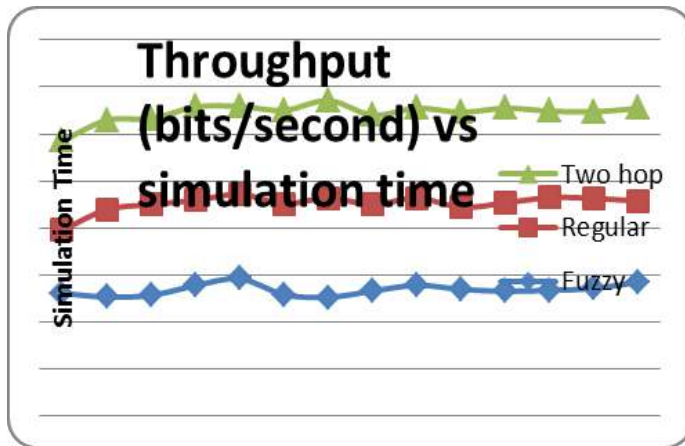


Fig. 8. Throughput Vs simulation time

## CONCLUSION

Every ad hoc network node is a battery powered device and so energy conservation and power reduction are major goals. Ad hoc networks power consumption is controlled by controlling transmission power or choosing optimal transmission routes. MAC protocol to control access to shared wireless medium, avoiding collisions, and maintaining coordinated medium access. Fuzzy logic is similar to human reasoning when some things are uncertain. Fuzzy Logic derives power control at each network node through fuzzy rules generated by input variables link quality and neighborhood count in this work. Experiments using fuzzy rules are compared to DSR routing and simple 2 hop power control procedures. The results revealed that 2 hop power controls with fuzzy logic procedures used reduced route discovery time, increased cache replies with limited simulation time and end to end delay when compared to DSR routing and 2 hop routing protocol.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Wu J, Li H. [2000] Domination and its applications in ad hoc wireless networks with unidirectional links. In Parallel Processing, 2000. Proceedings. 2000 International Conference on (pp. 189-197). *IEEE*.
- [2] Breed G. [2007] Wireless ad hoc networks: Basic concepts. *High frequency electronics*, 44-46.
- [3] Gummalla ACV, Limb JO. [2000] Wireless medium access control protocols. *Communications Surveys & Tutorials, IEEE*, 3(2): 2-15.
- [4] Kumar S, Raghavan VS, Deng J. [2006] Medium Access Control protocols for ad hoc wireless networks: A survey. *Ad Hoc Networks*, 4(3): 326-358.
- [5] Langendoen K, Halkes G. [2005] Energy-efficient medium access control. *Embedded systems handbook*, 34-41.
- [6] Abbas MM, Mahmood H. Power Control in Ad Hoc Networks.
- [7] ElBatt TA, Krishnamurthy SV, Connors D, Dao S. [2000] Power management for throughput enhancement in wireless ad-hoc networks. In *Communications, 2000. ICC 2000. 2000 IEEE International Conference, IEEE*, 3: 1506-1513.
- [8] Almotairi KH, Shen XS. [2012] Distributed power control over multiple channels for ad hoc wireless networks. *Wireless Communications and Mobile Computing*.
- [9] Kawadia V, Kumar PR. [2005] Principles and protocols for power control in wireless ad hoc networks. *Selected Areas in Communications, IEEE Journal on*, 23(1): 76-88.
- [10] Belghith A, Akkari W. [2008, July] Power saving mechanisms for ad hoc networks based on handshaking information tapping. In *Proceedings*

- of the Second international conference on Verification and Evaluation of Computer and Communication Systems (pp. 50-60). British Computer Society.
- [11] Hiremath PS, Joshi SM. [2012]Energy Efficient Routing Protocol with Adaptive Fuzzy Threshold Energy for MANETs. *International Journal of Computer Networks and Wireless Communications (IJCNWC)*, ISSN: 2250-3501 Vol, 2.
  - [12] Banerjee A, Dutta P. [2010] Fuzzy-controlled Power-aware Routing Protocol (FPRP) for Mobile Ad Hoc Networks. *International Journal of Computer Applications*, 11(7):39-43.
  - [13] Yu Z, Fu X, Cai Y, Vuran MC. [2011] A reliable energy-efficient multi-level routing algorithm for wireless sensor networks using fuzzy Petri nets, *Sensors*, 11(3): 3381-3400.
  - [14] Lu D, Huang X, Liu C, Fan J. [2010] Adaptive power control based spectrum handover for cognitive radio networks. In *Wireless Communications and Networking Conference (WCNC), 2010 IEEE* (pp. 1-5). *IEEE*.
  - [15] Zuo J, Ng SX, Hanzo L. [2010, September]. Fuzzy logic aided dynamic source routing in cross-layer operation assisted ad hoc networks. In *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd* (pp. 1-5). *IEEE*.
  - [16] Casillas J, Cordon O, Herrera F. [2000] Learning fuzzy rules using ant colony optimization algorithms. In *Proc. 2nd International Workshop on Ant Algorithms* (pp. 13-21).
  - [17] ThangarajP, K.Geetha , An Enhanced Associativity Based Routing with Fuzzy Based Trust to Mitigate Network Attacks, *World Academy of Science, Engineering and Technology*, Volume 9, Issue 8, Pages 1614 – 1622, 2015.
  - [18] Agarwal S, Hitzler P.[2005]Modeling fuzzy rules with description logics. In *Proceedings of Workshop on OWL Experiences and Directions*. Galway, Ireland.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# AN HYBRID ROUTING PROTOCOL FOR EFFICIENT POWER MANAGEMENT AND CONTROL IN MANET

V. Ramakrishnan\* and SP. Manigandan

Dept of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, INDIA

## ABSTRACT

*As nodes in MANET are predominantly battery driven depletion of energy is a critical concern and to address the issues related to increasing the throughput, efficiency, energy consumption and to reduce the network load we propose a concept to combining power aware routing and channel allocation strategy that solemnly address these concerns. The utilization of the resources such as bandwidth and energy depends on a number of conditions such as network size, node density, and load distribution. These conditions are uncontrollable and often vary throughout the operation of the network. The scope of the work is to bring out the commonalities present separately in channel allocation and power aware routing by joint optimization dynamic channel allocation mechanisms and routing possibilities.*

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

*Manet, Routing, Channel allocation, Residual Power, Cross Layer Optimization*

## INTRODUCTION

AN AD-HOC network is an on demand network which is deployed based on resources available at the particular time. This network has various concerns regarding to the deployment, routing, power management, channel allocation, channel control etc. These are issues which are either addressed partially or relatively in logical manner to increase the overall efficiency and throughput of the deployed network. To manage these resources efficiently there are various methods used in order to negate the issues to a particular extent when these methods are coordinated we can annihilate problems which affects the network to a greater extent. In this paper we propose a concept of combining the routing and channel allocation approach to create an energy efficient load balancing network. The major contention related to channel allocation is decisive

Handling of bandwidth in the event of non-uniform loads. To support uneven or non-uniform loads maximization of spatial reuse and multicasting is enabled at the link layer, that facilitate the use of resources in tangible manner. This annihilates the need for several transmissions of similar loads. In MANET the aware of power heterogeneity is an important technical challenging problem to increase the energy efficiency of each node. The mobile nodes in MANET have different transmission power and power heterogeneity.

Efficient Energy Routing Protocols such as EPAR (Efficient Power Aware Routing protocol) mainly considers the node capacity by its remaining battery power and the expected energy spent for forwarding data packets reliably. EPAR uses mini-max formulation method for the selection of the route that has maximum packet delivery ratio at the smallest Residual Battery Power.

The transmission of information across the network occurs as relaying of data packets from one node to other node within the network. Due to the mobility of node, topology in the network can change dynamically and nodes can be added and removed at any time in the network. To perform proficient handling of resources in network we use dynamic channel allocation algorithm coalesced with routing algorithm to bring out a hybrid scenario called optimized DCA for efficient power management and channel control in the network.

## RELATED WORK

The existing work done in the in addressing issues related to our work is as follows According to Bora Karaoglu and Wendi Heinzelman (2015) a lightweight dynamic channel allocation mechanism and a cooperative load balancing strategy that are applicable to cluster based MANETs can be used to address the issues in channel allocation. They present protocols that utilize these mechanisms to improve performance in terms throughput, energy consumption and inter-packet delay variation (IPDV). Through extensive they showed that both dynamic channel allocation and cooperative load balancing improve the bandwidth efficiency under non-uniform load distributions compared to protocols that do not use these mechanisms as well as compared to the IEEE 802.15.4 protocol with GTS mechanism and the IEEE 802.11 uncoordinated protocol.

According Angel Lozano (2012) DCA algorithms, some important issues have been neglected because of the complexity involved in their study. In particular, the impact of user motion on the performance of DCA systems has not received enough attention. He quantify the impact of motion on the capacity and cost in terms of average number of reassignments per call of a variety of representative distributed fixed-power DCA algorithms. A novel adaptive algorithm especially suited for mobility environments is proposed, which achieves high capacity while controlling the reassignment rate. He also proved that most of this capacity can be effectively realized with a reduced number of radio transceivers per base station.

According to Kunal Gaurav and Mani Upadhay (2014). In MANET each and every process consumes power. Power consumption is one of the most crucial design concerns in Mobile Ad-hoc networks as the nodes in MANET are battery limited. To increase the life time of nodes as well as network energy management is necessary. The life time of a node can be increased when less power is consumed by the nodes during active as well as during inactive communication. In order to increase the life time of a node, traffic should be routed in a way that, power consumption is minimized. DSR is a routing protocol used in MANET. They used two mechanisms in routing Route Discovery and Route Maintenance. During route discovery it selects the path to establish communication between source and destination. The flooding of routes creates overhead in routing traffic which cause extra consumption of energy. They developed new route discovery mechanism is proposed to conserve the energy of nodes during route discovery and data transmission phase compare to original DSR protocol.

## DYNAMIC CHANNEL ALLOCATION ALGORITHM

The first mechanism that we propose is a dynamic channel allocation (DCA) algorithm analogous to the ones that exist in cellular systems. Under uneven and non-uniform loads, it is essential for the MAC protocol to be flexible enough to let the vacant and unused bandwidth to be allocated to the controllers in the heavily loaded regions.

The working of algorithm does channel allocation on behalf of mobile hosts in the cell. Each and every request is marked and time stamped and sent to nearest service stations to be assigned for communication session. The nature of the algorithm is finite and due to non-deterministic propagation, channel interference occurs to avoid this interference multiple service station approval is deployed in this algorithm. As the channel allocation varies with time, the algorithm uses temporal and spatial changes to control the varying load distribution the channel due to the highly dynamic and changing behavior of the network we adopt a dynamic channel borrowing scheme that employ spectrum sensing.

In this algorithm, the channel controllers invariably look out the power level in all the available channels in the network and assess the availability of the channels by correlate the measured power levels with a threshold. If local load surge beyond certain local capacity if the measured power level is minimum, the channel coordinator commences using the channel with the lowest power remains. Once the channel coordinator starts using the channel, its transmission increases the power level measurement of that channel for close by controllers, which in turn blocks them from securing the same channel. If the local network load decreases, controllers that do not need some channels and block the transmissions in that channel, making it reachable for other controllers. Channel coordinators react to the increasing local network load by increasing their share of bandwidth. Although being proficient in lending support for non-uniform network loads, the proactive response taken by the channel coordinators increases the interference in the entire system.

The DCA algorithm approaches the problem of non-uniform load distribution from the perspective of the channel coordinators. The same problem can also be approached from the perspective of ordinary nodes in the network. This cooperative behavior smooths out mild non-uniformities in the load distribution without the need for the adjustments at the channel coordinator side.

The load on the channel coordinators arise from the insistence of the ordinary nodes. Many nodes in a network have access to more than one channel coordinator. The underlying idea of the dynamic channel allocation algorithm is that the active nodes can continuously monitor the channel usage and switch from heavily loaded coordinators to the ones with available resources. These nodes can reveal that the channels available at the channel coordinator are drained and shift their load to the channel coordinators with more available resources. The resources departed by the nodes that switch can be used for other nodes that do not have admittance to any other channel coordinators. This increases the total number of nodes that access the channel and hence increases the throughput.

## POWER AWARE ALGORITHM

The algorithm is deployed on networks in which power is a restricted resource. Only a bounded number of messages can be disseminated between any two hosts. This issue is solved by routing messages so as to augment the battery lives of the hosts in the system. The course of a network with respect to a sequence of messages is the earliest time when a message cannot be sent because of saturated nodes. This metric under the assumption considers that all messages are important. However it can be relaxed to accommodate up to message delivery failures.

Several metrics can be used to enhance power routing for a sequence of messages. Minimizing the energy consumed for each message is an evident solution that revises locally consumed power. Other useful metrics include lessen the variance in each node power level, lessen the ratio of cost/packet, and minimizing the maximum node cost. A drawback of these metrics is that they focus on individual nodes in the system instead of the system as a whole. Therefore, routing messages in accord to these metrics might promptly lead to a system in which nodes have high unconsumed power but the system is not connected because some critical nodes have been depleted of power. The prime focus is on global metric by maximizing the lifetime of the network. This metric is very useful for on demand networks in which each message is decisive and the networks are sparsely deployed.

This problem does not have a constant competitive ratio to the offline optimal algorithm that knows the message sequence. It is an approximation algorithm for power aware message routing that enhances the lifetime of the network and examines its limits. This algorithm combines the benefits of enumerating the path with the minimum power consumption and the path that maximizes the minimal enduring power in the nodes of the network. The power aware algorithm has a good competitive ratio in practice, approaching the performance of the optimal off-line routing algorithm under realistic conditions.

## HYBRID ROUTING CHANNEL ALLOCATION ALGORITHM

### OPTIMIZED DCA

- Begin
- Initial Power = 0
- Compute  $Power/\Delta P_t$  for every host
- Calculate the minimal  $Power/\Delta P_t$  among all nodes
- if some host is sopped then
- exitRequest the measurement of TX power  $Power_{min}$ , of user I in the time slot k of cell  $C_j$
- If  $Power_{ciu} > Power_{cmaxu}$
- for  $l = 1 : \max(\text{Neighboring\_cells})$
- for  $n = 1 : \max(\text{TS\_in\_use})$  where n belongs set of used TS's
- if  $TS_{cIn} == \text{RX time slot}$
- determine interference from nodes:

- $I_{bln} = (\text{Power}_{clu})_n (M_{Bcl,cj})_n$
- else
- determine interference from BS:
- $I_{bln} = (\text{Power}_{clu})_n (B_{Bcl,cj})_n$
- end if
- if  $(((I_{bln} < I_{b_l^k}) \& (\text{direction}(TS_l^n) \neq \text{direction}(TS_l^k))))$
- exchange  $TS_n$  for  $TS_k$  in cell  $l$
- end if
- end for
- end for
- else
- Assign channel
- end if
- END

An important factor in the optimized DCA algorithm is the parameter power that evaluates the initial power level in transmitting nodes it works based on the consumed power used in transmitting the messages measures the tradeoff between the max– min path and the minimal power path. It calculates power for collection of messages that has to transmit without being any interference appearance such that will lead to a longer lifetime for the network than each of the max – min and minimal power algorithms.

The algorithm starts when the mobile is requested to transmit with higher power than the maximum power permitted, i.e. the state at which outage or service degradation would occur. The algorithm steps in assuming that a MS uses at least two TS's for the communication to the BS. It monitors the interference in all  $n$  TS's of all neighboring cells. Two cases can then be distinguished:

If TS in cell  $l$  is used for RX (from the BS point of view) the interference from this particular neighboring TS is caused entirely from its MS's since  $\alpha = 0$  and ideal integration is assumed. Furthermore, it is assumed that the MS's in the neighboring cell are able to determine the path loss to their neighboring BS's. This may be adopted by a fixed transmission power on the pilot channel. The MS's report their transmission power and path loss measurements to the BS which makes it available to the other channel. Hence, the information about the path gain matrix of the mobiles in cell to the BS in cell  $l$ ,  $P_{cl}$ , is assumed to be available to the optimized DCA algorithm.

If TS  $n$  at the BS is used for transmission the interference contribution from cell deals only from the BS (same entity) interference as  $\alpha = 1$ . The transmission powers at the BS's are known and can easily be expressed to the channel and so can the path loss to the neighboring BS's,  $B_{Bcl,cj}$ .

A check is made to examine if there is one TS in the neighboring cell which would cause less interference than the current TS  $k$ . If this is true and TS  $n$  is used for RX while TS  $k$  was used for TX, or vice versa, then the neighboring cell,  $cl$ , interchanges TS  $n$  with TS  $k$ . These results in TS opposing time slots with respect to the  $C_j$ . The algorithm is used by inferring the power level from the nodes by enumerating residual power level from the nodes. Thus using this inferred value to efficiently allocate channel resources.

## RESULTS

This section will use the NS2 network simulator to evaluate the optimized DCA algorithm. Simulation of all nodes in the stationary distribution in the  $1000 * 1000m^2$  two-dimensional plane, grid is divided into regular network topology and random distribution. Data transmission channel distance of each node is 250m, the distance of 500m interference. Control channel transmission distance is two times the data channel. Each of the channel capacity of 2Mbps, each node has 5 interface, at network initialization, an interface distribution control channel, the other interface randomly assigned data channel. There are 11 channel systems, of which one is a common channel. With the RTS/CTS IEEE in the MAC layer of 802.11 DCF collision avoidance mechanism. In a grid like uniform arrangement of the  $5*5$  node in the regular network, in the scene 30 nodes randomly distributed random distribution, node arrangement will no longer move. In event of measure the performance of routing scheme proposed in this paper, firstly selected by multi-channel multi interface wireless networks are commonly used



inrouting strategy as a comparison object, only considered without considering the inter stream interference flow; choose another interference aware in experiment was performed multiple analog averaging as criterion, in the network initialization data channel randomly distributed over the nodes on the interface when the interface assignment, channel no neighbors same distribution channel is to re select the channel assignment. Throughput in different network environment .The system throughput increases when flow number of access network increases. It is also easy to see that our proposed has higher throughput in both two environments. This main reason is optimized DCA adopt the adaptive switch method to achieve the success transferring, and considers disturbing of in-flow and out-flow, so the performance is lower. In [Figure -1], the results of these algorithms are presented, It is visible that optimized DCA algorithm has better channel efficiency and throughput and has a better power utilization factor.

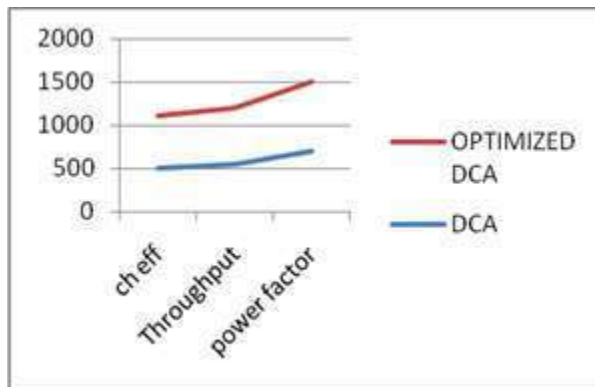


Fig: 1. Comparison of optimized DCA and DCA

## CONCLUSION

We proposed an optimized DCA routing algorithm protocol that merges the functionalities of dynamic channel allocation and power aware routing thus enables the new frame work more efficient in terms of throughput, bandwidth utilization consumption factor. It also enumerates the use of more efficient means of routing and channel allocation mechanisms that will enhance the performance of the ad-hoc networks. The future enhancements that can be included are multi-hop extensions, handover issues, overhead issues and multicasting probabilities.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Numanoglu T, Tavli B, and Heinzelman W.[2005] An analysis of coordinated and non-coordinated medium access control protocols under channel noise. Military Communications Conference, 2005. MILCOM 2005. *IEEE*, 4: 42–48.
- [2] A Chandra, V Gummalla, and JO Limb.[2000] Wireless medium access control protocols. *IEEE Communications Surveys and Tutorials*, 3:2–15.
- [3] P Mohapatra, J Li, C Gui.Qos.[2003] in mobile adhoc networks.*IEEE Wireless Communications Magazine*, 10:44–52.
- [4] B Tavli and WB Heinzelman.[2004.]MH-TRACE:Multihop timereservation using adaptive control for energy efficiency. *IEEE Journal on Selected Areas of Communications*, 22(5):942–953.
- [5] J Karaoguz.High-ratewirelesspersonalarea networks. *Communications Magazine, IEEE*, 39(12):96–102, Dec 2001.
- [6] T Numanoglu, B Tavli, W.B.Heinzelman.Theeffectsofchannel errors on coordinated and non-coordinated medium access control protocols.In Proceedings of IEEE International Conference on Wireless and Mobile Computing, 1: 58–65, Aug 2005.
- [7] Bora Karaoglu, Tolga Numanoglu, and Wendi Heinzelman. Analytical per- formance of soft clustering protocols. *Ad Hoc Networks*, 9(4):635 – 651, 2011.

- [8] Lifei Huang and Ten-Hwang Lai. On the scalability of ieee 802.11 ad hoc networks. In Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing, MobiHoc '02, pages 173–182, New york, NY, USA, 2002.ACM.
- [9] C. Rajan, N. Shanthi, Genetic based Optimization for multicast Routing algorithm for Manet' Sadhana - Academy Proceedings in Engineering Science, Volume 40 Issue 7, 2341-2352, 2015.
- [10] IEEE 802.15.3 Working Group. Part 15.3: Wireless medium access control (MAC) and physical layer (PHY) specifications for high rate wireless personal area networks (WPAN). IEEE Draft Standard, Draft P802.15.3/D16, Feb 2003.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# BINARY PARTICLE SWARM OPTIMIZED CLUSTERS FOR WIRELESS SENSOR NETWORK

Ramana Rao<sup>1\*</sup> and Adilakshmi<sup>2</sup>

<sup>1</sup>Dept. of CSE, University College of Engineering, Osmania University, Hyderabad, INDIA

<sup>2</sup>Dept. of CSE, Vasavi College of Engineering, Osmania University, Hyderabad, INDIA

## ABSTRACT

Wireless Sensor Networks (WSN) is widely applied to many domains, including smart spaces, environmental monitoring, medical systems and robotic explorations. The networks constitutes of scattered sensor nodes that arrange automatically into multi-hop wireless networks. A node possesses one or more sensors, embedded processors, low-power radios and a battery. Data is transmitted through the network by routing. Clustering of nodes in the network helps to conserve energy of the nodes. In this paper, it is proposed to cluster B-MAC protocol with Binary Particle Swarm Optimization (BPSO). The suggested protocol's performance is tested for packet delivery ratio, end to end delay, hops and jitter. The outcome reveals that the new BPSO with cluster BMAC performs better than BMAC in either static or dynamic scenarios. Results show that the proposed method improved the PDR by 3.22% when compared to the BMAC (flooding) protocol at 40 kmph. The proposed method improved the average PDR by 3.30% when compared to the BMAC (flooding) protocol in various node mobility scenarios.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Wireless Sensor Networks (WSN), Cluster Head (CH), Medium Access Control (MAC), Binary Particle Swarm Optimization (BPSO).

## INTRODUCTION

Wireless Sensor Networks (WSN) possesses innumerable sensors distributed in a remote location. Their primary role is that collecting data from the surroundings and pass it on to a base stations [1]. WSN's disadvantage lies in the low energy of the sensors, and this remains a huge restriction for the employment of WSN. Network efficacy relies on the life span of sensor nodes and network coverage.

Once a network is positioned, it is impossible to change batteries. Hence, when WSN is designed, this has to be taken into account. Secondly, the speed at which data is moved from a node to the base station is important. The energy consumed may be decreased by utilizing a couple of nodes to connect with the base station. These nodes are cluster heads and information is collected in them prior to being transferred to a base station [2].

Network efficacy relies on cluster head selection and parameters for the selections. Clustering assures lesser communication overheads and effective resource allotments and so, decreasing the total amount of energy consumed. Disturbances amongst sensor nodes are decreased as well [3]. The obstacle in clustering is in the selection of appropriate nodes to be made cluster heads or gateways.

Cluster heads are hierarchically ordered. The main benefit of such hierarchical protocols is that cluster heads execute in-network data aggregation. Routing moves forward by passing on packets up a hierarchy until a sink node is reached. Flat routing protocols attempt to discover good quality routes from the source to the sink through flooding. Because flooding is an expensive operation in networks that are typically starved for resources, smart algorithms constrain flooding on specific areas [4]. Certain algorithms employ heuristics based probabilistic methods for establishing routing paths.

Flooding policies primary roles are to determine meaning as well as calculate node ranks apart from employing a state machine governing singular packets life cycle on nodes [5]. Flooding policies are generally sorted according to these two characteristics.

WSNs typically have scarce resources. Conserving energy is crucial. A sensor nodes' radio utilized immense energy. Creating low power electronic gadgets to decrease energy consumed by sensor nodes is an area widely researched. Because of limits in hardware, additional energy efficacy can only be accomplished through the planning of energy efficient communication protocols.

Medium Access Control (MAC) is a method that ensured success in network executions. The primary role of MAC is to avoid conflicts from intervening nodes. Creating energy efficient MAC protocol is the method to extend network life [6].

WSN's MAC protocols have two goals. Firstly, for the purpose of creating sensor network infrastructure, many sensor nodes are employed and the MAC system constructs a contact link among the sensor nodes. Second is the sharing of the communication medium in a fair and effective manner.

However, there are certain features apart from energy efficacy that need to be deliberated upon. Effective utilization of energy improves network life. Some more features are how sensor networks respond to alterations in the size of the network and how it adjusts to changing environment dynamics. Alterations to topology, quantity of nodes or total network size ought not to have an effect on typical sensor network functions. MAC protocol ought to be capable of adapting to such features [7]. But, there are few attributes other than energy efficiency which should also be considered. Efficient energy use increases network life. Other attributes are how sensor networks react to changes in network size and how it adapts to changing environment dynamics. Changes to topology, nodes quantity or overall network dimensions should not affect normal sensor network functioning. MAC protocol should adapt to such characteristics.

B-MAC is a method grounded in CSMA, utilizing minimal power listening as well as a drawn out preamble for the purpose of minimal power communications. Nodes possess both 'awake' and 'sleep' sessions and each node have an autonomous timetable. If a node wishes to transfer data, it includes an information packet with a preamble that is longer than the recipient's sleep cycle. In the span of the awake period, a node skims a medium and once the preamble is spotted, it stays awake in order to get the information [8]. Because of the long preamble, the transmitter is sure that the recipient will definitely wake up during the preamble, spot it and stay awake to get information. With B-MAC, an interface where a program may modify sleep schedules to suit the varying traffic is ensured. Despite the excellent performance of B-MAC, there is the challenge of overhearing as well as the energy wasted in long preambles.

This paper suggests cluster based B-MAC with BPSO. Section 2 reviews related literature. Section 3 describes the methods employed in this work. Section 4 reveals experiment results and Section 5 concludes the work.

## RELATED WORK

Utilizing receiver-based MAC protocols (RB-MAC) and adaptable RB-MAC in routing protocol for less energy and lossy WSN protocol in low-power/ lossy wireless channel studied by Akhavan et al., [9] led to a cross-layer method in which routing decisions depend on MAC-layer functions. Outcomes proved that RB-MAC and adaptive RB-MAC performed better than current topmost sender-based MAC protocols with regard to end-to-end energy-efficiency, delay and reliability. Simulations revealed that utilizing receiver-based MAC protocols in RPL-based lossy channel meant decreased retransmissions as opposed to sender-based MAC.

A model for Power-Control and Delay-aware routing and MAC protocol (PCDARM) for Energy constraint and delay sensitive WSN applications created by Rachamalla and Kancharla [10] plans to achieve routing of packets through multiple paths with the delay previously known and with energy/link reliability constraints, in the routing phase and TDMA based slot assignment in an energy efficient manner during MAC phase. The routing phase plan employs traffic splitting on multiple paths. WSN's differential delay, link stability and power restrictions are considered. Slot allotment in TDMA frames with sleep and awake cycles for network nodes lead to effective

MAC phase energy control. Experimental outcomes showed the advantages of the suggested protocol over DEAR. Simulations revealed that packet delay decreased in PCDARM. It performed better in power consumption as well.

A duty-cycled, directional adaptive MAC protocol suggested by Bin et al., [11] called DA-MAC considered that negative impacts of movement whereas particular selected nodes with appropriate features as forwarders, attained excellent end-to-end latency. Experimental outcomes using NS-2 revealed that the DA-MAC showed amazing performance in end-to-end latency as well as power conservation in mobile WSN.

A new cross-layer duty-cycled MAC protocol having data forwarding to pipeline features called P-MAC suggested by Tong et al., [12] for WSNs, separates network sensor nodes into various grades around a sink. PMAC/RMAC performance is assessed with regard to packet delivery latency and energy efficacy using OPNET simulation.

A structure where MAC and routing protocols join hands in order to identify as well as reserve routes, order nodes in clusters and schedule availability to a transmission medium in a coordinated time-shared manner was suggested by Ruiz et al., [13]. The result was named QUALity-of-service-capable clusTer-based Timeshared ROUTing-assisted (QUATTRO) MAC protocol. It was assessed through simulation investigating multiple settings where varying numbers and node densities were taken into account. Outcomes revealed that the protocol had considerable overhead.

Taranovs and Zagursky [14] suggested a completely collision-free TDMA-based MAC protocol for a WSN cluster in which each node is one-hop away from neighbours in all clusters making sure that no routing protocols are required. It allows utilization of restricted sources sensor nodes extending WSN life. Higher level content denotes automatic runtime choosing of services executions and network sources to implement application specifications in resources-effective and context-aware scenarios.

Liang et al., [15] proposed a cross-layer protocol named CL-LEACH for effective power usage of clustered WSNs, grounded in Low Energy Adaptive Clustering Hierarchy algorithm (LEACH). Through the introduction of Frequency Shift Keying (FSK) as well as the consideration of constellation size as a global cross-layer variance, CL-LEACH regards physical, link, MAC and routing layers, integrating a battery non-linearity discharge model called FSK Orthogonal modulation scheme, TDMA and CSMA MAC protocols, and cluster routing algorithm. Experiments reveal that CL-LEACH reduces power usage by a great deal and extends clustered WSNs life as opposed to LEACH protocol.

MAC and routing protocols are a major component in WSNs for end-to-end data transmission. Additionally these protocols utilize a hierarchical architecture for end-to-end, dependable data delivery. Too much time and power is utilized in order to create a hierarchical architecture sequentially. In order to offset these obstacles, Dash et al., [16] proposed a distributed algorithm to create an architecture, whose hierarchy is such that the nodes possess two parent nodes apart from the root node. The novel approach decreased the average power usage of a node by setting other nodes in sleep mode in a dispersed manner as opposed to contemporary methods. The Castalia simulator was utilized to simulate the method and the results were contrasted with the contemporary sequential method.

An optimal, maximized configurable power saving protocol, named Green-MAC for corona based WSNs suggested by Lin et al., [17] possessed many advantageous characteristics such that Green-MAC utilizing a generalized Chinese remainder theorem ensures that two sensors in adjacent coronas wake up at the same bounded time, with no regard to their schedule or cycle lengths; in a cycle length, ATF-ratio (fraction of a cycle's awake time frames) of sensors attains a theoretical minimum; minimum ATF-ratio restrictions, amount of configurable ATF-ratios for sensors attains a theoretical maximum; ATF-ratio configuration scheme is proposed for Green-MAC so that WSN energy usage is decreased while worst event-to-sink delay need is accomplished with high probability. Theoretical study as well as simulation revealed that Green-MAC performed better than contemporary energy conserving protocols for corona-based WSNs, as well as Q-MAC/Queen-MAC with regard to configurability, ATF-ratio, delay violation ratio, network life and event-to-sink throughput.

A dependable cross layer routing scheme (CL - RS) to balance power to extend life by governed restricted power usage was suggested by Kusumamba and Kumar [18]. A combined optimization design was designed as a linear

programming problem. Simulations revealed that joining CL-RS and CL-MAC algorithms at every layer improved network life considerably and that a relationship is present between network life maximization and dependability restrictions. The suggested method's performance was assessed under varying scenarios through ns-2. Out comes established that the new method performed better than layered AODV with regard to end-to-end delay, packet loss ratio, control overhead and power consumption.

A MAC-aware Routing protocol for WSNs (MAR-WSN) in which next hop decisions were dependent on TDMA scheduling and two-hop neighbourhood knowledge was suggested by Louail et al., [19]. Coherent routing protocol decisions in space, with those by MAC protocol, in time, showed itself to be effective against metrics such as power usage, delay and hop number. Simulations revealed the new method's excellent performance in medium/high density networks as opposed to current approaches.

A novel WSN routing method where routing is dependent on radio-aware metric possessing radio information at the MAC layer suggested by Tariq et al., [20] had data being forwarded to a determined neighbour, rather than flooding exploratory data into the entire network with directed diffusion. Simulations revealed the suggested routing method was around 4.3 times more effective with regard to power usage and 2.6 times more dependable with regard to data delivery as opposed to directed diffusion. The new scheme depended on interactions among routing/MAC layers to attain dependability and power efficacy using cross layer optimization.

## METHODOLOGY

Binary PSO with cluster BMAC protocol is proposed.

### Berkeley MAC (Flooding)

Berkeley Media Access Control (B-MAC) utilized in WSNs is a contention based MAC protocol. B-MAC is like Aloha with Preamble Sampling where BMAC duty cycles radio transceiver i.e. Sensor nodes turn ON/OFF continuously without missing the data packets. Preamble length is a parameter to upper layer guaranteeing optimal trade-off between power savings and latency or throughput. BMAC is almost identical to CSMA protocol having Low Power Consumption. BMAC utilizes unsynchronized duty cycling and long preambles to wake up receivers. BMAC improves dependability and channel assessment through a filter.

Sensor node change protocol operating variables such as back off values allowing a flexibility interface. BMAC utilizes an adaptive preamble sampling scheme to decrease idle listening and cut down duty cycle. B-MAC duty cycles a radio through periodic channel sampling called Low Power Listening (LPL). BMAC uses clear channel assessment (CCA) to determine if a packet arrives when node wakes up. If no packet has arrived a timeout puts the node to sleep again. BMAC utilizes CCA and packet back-offs for channel arbitration and link layer acknowledgments for dependability [21].

The first active node sends out control messages when its reconfiguration ends and other nodes flood once to coordinate with their neighbours in this approach. A node spends energy to send one up message and on receipt of multiple up messages from other nodes, polls the channel and sleeps for the remaining time.

It assumes polling interval for LPL during reconfiguration is  $T_p$ . Remember that  $T_p$  can be different than  $T_{lpl}$ . To wake up neighbours, nodes flood up messages with preamble  $T_p$ .

During flooding, a node needs to forward up message once. Let's assume average carrier sense is  $t_{cs}$ , and transmission time for up packet is  $t_{up}$ . A node's energy spent on transmission is

$$P_l t_{cs} + P_s (T_p + t_{up})$$

A node receives  $n$  packets from  $n$  neighbours. And on average it overhears  $T_p/2$  preamble for a packet. Energy it spends in receiving is

$$nP_l (T_p/2 + t_{up})$$

As nodes reboot in uniform distribution, average waiting period before flooding for a node is  $T_d$ . Thus low-power listening cost on each node is

$$P_{poll} t_p T_d / T_p$$

The final component of energy is sleep cost:

$$P_{slp} (T_p - t_p) T_d / T_p$$

Substituting Equations it obtains mean energy cost during reconfiguration as

$$\begin{aligned}
 E_{flood} = & P_l t_{cs} + P_s (T_p + t_{up}) \\
 & + nP_l (T_p / 2 + t_{up}) \\
 & + P_{poll} t_p T_d / T_p \\
 & + P_{slp} (T_p - t_p) T_d / T_p
 \end{aligned}$$

Above equation shows a trade off with  $T_p$ . Increasing  $T_p$  reduces channel sampling frequency, and saves nodes from spending energy on polling. But it increases preamble length, thereby increasing transmission/overhearing cost. To reduce

$E_{flood}$ , it needs to obtain an optimal  $T_p$  from the following equation

$$\frac{dE_{flood}}{dT_p} = 0$$

Depending on data rate, B-MAC proposes a like approach to optimize polling period. But the analysis is grounded in periodic data traffic and does not guarantee a closed form formula. Rather during LPL with flooding network does not create periodic data and flooding of up messages [22] is the only traffic.

## Proposed Binary Particle Swarm Optimization Cluster B-MAC

To use a mobility metric for clustering - a distributed, lowest mobility clustering algorithm, MOBIC, similar in execution to Lowest-ID algorithm - except that mobility metric is basis for cluster formation instead of ID information, is used. The algorithm is described in these steps:

- All nodes send/receive "Hello" messages to/from neighbours. A node measures received power levels of two successive "Hello" message transmissions from a neighbour, and calculates pair wise relative mobility metrics using (1). Before sending next broadcast packet to neighbours, a node computes aggregate relative mobility metric M using (2). M is represented through a double precision floating point number.
- All nodes start in Cluster Undecided state. A node broadcasts its own mobility metric, M (initialized to 0 at the start of computations) through "Hello" or "I'm Alive" messages to its 1-hop neighbours in every Broadcast Interval (BI). It is stored in a neighbour table of every neighbour with a timeout period (TP) arranged. The algorithm is implemented in a distributed manner. So, a node receives aggregate mobility values from neighbouring nodes, and compares its own mobility value with them.
- When a node has lowest value of M (aggregate relative mobility) among neighbours, it assumes Cluster Head status; otherwise it is a Cluster Member. The algorithm forms clusters which are two hops in diameter at the most. When a node is neighbour to two cluster heads, it becomes a "gateway" node. When two neighbouring nodes in a Cluster Undecided state have same M value it compares IDs and follows Lowest-ID algorithm.
- If mobility metrics of two cluster head nodes is same, and both contend to retain Cluster Head status, then cluster head selection is based on Lowest-ID algorithm wherein a node with lowest ID gets status of Cluster Head. When a node with Cluster Member status and low mobility moves into the range of another Cluster Head node with higher mobility, re-clustering does not happen (similar to LCC).
- In a mobile scenario, if two nodes with status Cluster Head move into each other's ranges, re-clustering is postponed for Cluster Contention Interval (CCI) to admit incidental contacts between passing nodes. If nodes are in transmission range of each other even after CCI timer expires, re-clustering is triggered, and a node with lower mobility metric assumes Cluster Head status.

A discrete binary version of PSO for binary problem was proposed by Kennedy and Eberhart where a particle's personal best and global best is modernized. The difference is the particles velocities. Such velocity has to be controlled in a range of [0, 1]. Each particle's velocity got using the equation:

$$\left\{ \begin{aligned}
 v_{i,j}(t+1) &= \eta \left( v_{i,j}(t) + c_1 r_1 (p_{ibest,j}(t) - x_{i,j}(t)) \right) \\
 &\quad + c_2 r_2 (g_{ibest,j}(t) - x_{i,j}(t)) \\
 x_{i,j}(t+1) &= x_{i,j}(t) + v_{i,j}(t+1)
 \end{aligned} \right.$$

where  $v_{i,j}$  is particle velocity,  $x_{i,j}$  is position of particle,  $t$  the number of iterations,  $c_1$ , and  $c_2$  are two positive constants, called respectively as cognitive and social acceleration factors,  $r_1$  and  $r_2$  are random numbers in the range  $[0,1]$ , and  $\eta$  is inertia weight. A particle's best position (pbest) is denoted as  $p_{ibest,j}$  best position of all particles in a swarm is denoted as  $g_{ibest,j}$ .

$$\left\{ h(v_i(t+1)) = \eta \begin{cases} v_{\max}, & \text{if } v_i(t+1) > v_{\max} \\ v_i(t+1), & \text{if } |v_i(t+1)| \leq v_{\max} \\ v_{\min}, & \text{if } v_i(t+1) < v_{\min} \end{cases} \right\}$$

Here  $v_i$  depends on the sigmoid function

$$\text{sig}(v_i) = \frac{1}{1 + e^{-v_i}}$$

Then the particle's position is updated as [23]

$$v_{i,j} = \begin{cases} v_{i,j}^1, & \text{if } \dots X_{i,j} = 0 \\ v_{i,j}^0, & \text{if } \dots X_{i,j} = 1 \end{cases}$$

## RESULTS

In this study, BMAC (Flooding) and proposed BPSO cluster BMAC protocol are evaluated for WSN at static and various mobility levels. The results are analyzed from the following simulation values. The simulations are conducted to evaluate the performance of BMAC (Flooding), proposed BPSO cluster BMAC protocols under static and dynamic conditions. The Random Way point (RWP) mobility model is used, and the mobility is varied from 10 Km/h to 40 Km/h. Performance of BMAC (Flooding) and proposed BPSO cluster BMAC protocols were evaluated based on the Packet Delivery Ratio (PDR), End to End Delay, Number of hops, and Jitter for various mobility level in WSN.

Table: 1. Packet Delivery Ratio

Node mobility	BMAC (Flooding)	BPSO Cluster BMAC
Static	0.94	0.97
10 KMPH	0.89	0.92
20 KMPH	0.87	0.91
30 KMPH	0.83	0.86
40 KMPH	0.77	0.79



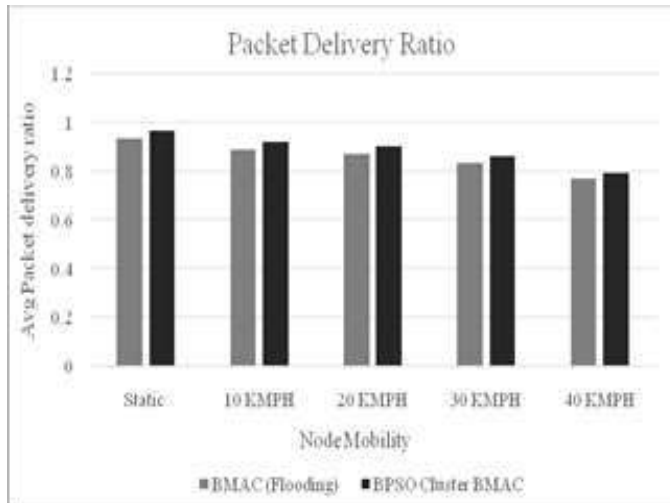


Fig: 1. Packet Delivery Ratio

The proposed method improved the PDR by 3.22% when compared to the BMAC (flooding) protocol at 40 kmph. The proposed method improved the average PDR by 3.30% when compared to the BMAC (flooding) protocol in various node mobility scenarios.

Table: 2. End to End Delay

Node mobility	BMAC (Flooding)	BPSO Cluster BMAC
Static	0.0012	0.0011
10 KMPH	0.0011	0.0011
20 KMPH	0.0011	0.0010
30 KMPH	0.0014	0.0014
40 KMPH	0.0072	0.0067

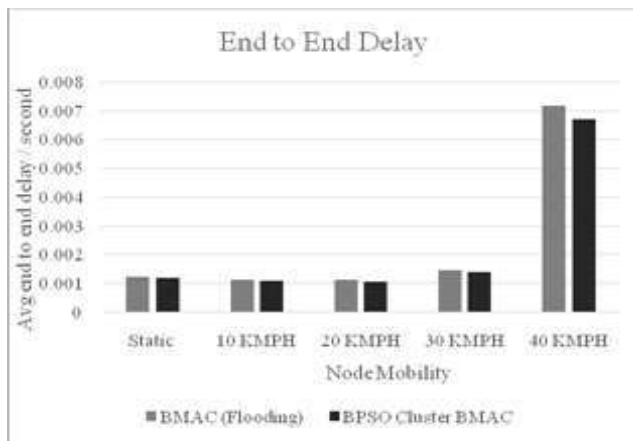


Fig: 2. End to End Delay

The proposed method reduced the end to end delay by 6.78 % when compared to the BMAC (flooding) protocol in 40 kmph. The proposed method reduced average end to end delay by 4.26 % when compared to the BMAC (flooding) protocol in various node mobility scenarios.

Table: 3. Number of Hops

Node mobility	BMAC (Flooding)	BPSO Cluster BMAC
Static	4.1	4.4
10 KMPH	4.2	4.6
20 KMPH	6.0	6.9
30 KMPH	7.6	8.0
40 KMPH	7.8	9.2

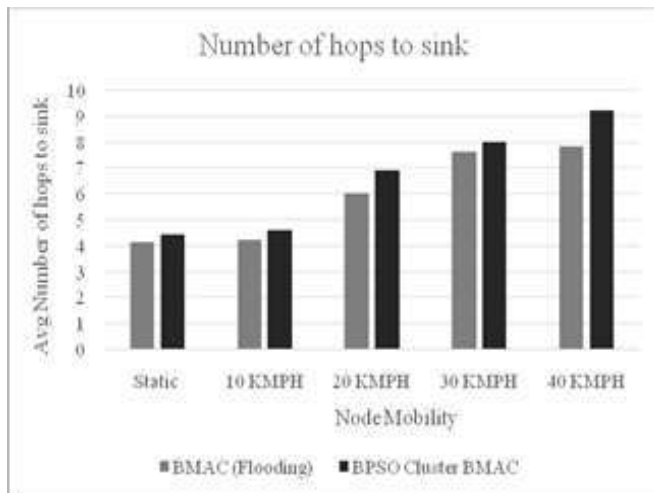


Fig: 3. Number of Hops

The proposed method increased number of hops to sink by 16.47 % when compared to the BMAC (flooding) protocol at 40kmph. The proposed method increased average number of hops to sink by 10.83% when compared to the BMAC (flooding) protocol in various node mobility scenarios.

Table: 4. Jitter

Node mobility	BMAC (Flooding)	BPSO Cluster BMAC
Static	0.0005	0.0005
10 KMPH	0.0012	0.0011
20 KMPH	0.0012	0.0011
30 KMPH	0.0013	0.0012
40 KMPH	0.0019	0.0018

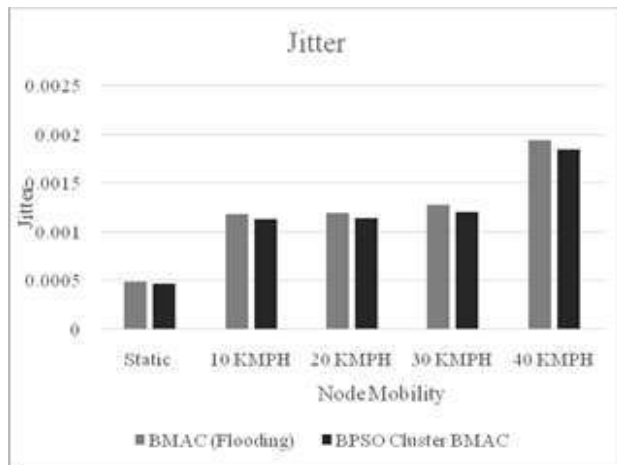


Fig: 4. Jitter

The proposed method reduced jitter by 4.88% when compared to the BMAC (flooding) protocol at 40 kmph. The proposed method reduced jitter by 6.23% when compared to the BMAC (flooding) protocol at 30 kmph.

## CONCLUSION

Clustering of the network relies on the cluster heads to transmit data to the base station. This reduces energy spent by a sensor node to relay data from other nodes to a base station, which, in turn, potentially results in increased network life and larger amount of data delivery during network life. In this work, a BPSO based clustering for BMAC is proposed to improve the lifetime of the WSN and to improve the QoS. BMAC (flooding) and BPSO cluster BMAC protocols performance was evaluated based on Number of hops, End to End Delay, Packet Delivery Ratio (PDR), and Jitter for various mobility WSN levels. Results show that new BPSO cluster BMAC achieved better performance than BMAC (flooding) in static/dynamic scenarios. It was seen that high mobility degraded routing performance.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Uma maheswari AS, Pushpalatha S. [2014] Cluster Head Selection Based On Genetic Algorithm Using AHYMN Approaches in WSN
- [2] Khan FU, Shah IA, Jan S, Khan I, Mehmood MA. [2015] Fuzzy logic based cluster head selection for homogeneous wireless sensor networks. In 2015 International Conference on Open Source Systems & Technologies (ICOSST) (pp. 41-45). *IEEE*.
- [3] Prasant K, Sivakumar P. [2015] Energy Efficient Improvement Geocast Forwarding In Manet Based On A Clustered Structure. *Journal of Engineering Science and Technology*, 10(9):1224-1238.
- [4] Singh SP, Sharma SC. [2015] A survey on cluster based routing protocols in wireless sensor networks. *Procedia Computer Science*, 45:687-695.
- [5] Mahmood MA, Seah WK, Welch I. [2015] Reliability in wireless sensor networks: A survey and challenges ahead. *Computer Networks*, 79:166-187.
- [6] Verma A, Singh MP, Singh JP, Kumar P. [2015] Survey of MAC Protocol for Wireless Sensor Networks. In *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on IEEE*, pp.92-95
- [7] Khatarkar S, Kamble R. [2013] Wireless Sensor Network MAC Protocol: SMAC & TMAC. *Indian Journal of Computer Science and Engineering (IJCSE)*, 4(4): 304-310.

- [8] Lee S. (2015) Stochastic Polling Interval Adaptation in Duty-Cycled Wireless Sensor Networks. *International Journal of Distributed Sensor Networks*, 2015.
- [9] Akhavan MR, Aijaz A, Choobkar S, Aghvami, A. H. (2014) On the multi-hop performance of receiver based MAC protocol in routing protocol for low-power and lossy networks-based low power and lossy wireless sensor networks. *IET Wireless Sensor Systems*, 5(1):42-49.
- [10] Rachamalla S, &Kancharla, A. S. (2015, April). Power-Control Delay-aware routing and MAC protocol for Wireless Sensor Networks. In *Networking, Sensing and Control (ICNSC)*, 2015 IEEE 12th International Conference on (pp. 527-532). IEEE.
- [11] Bin C, Linlin C, Minghua Y, Taolin G, Chengping T. [2013] DA-MAC: A duty-cycled, directional adaptive MAC protocol for airborne mobile sensor network. In *Digital Manufacturing and Automation (ICDMA)*, 2013 Fourth International Conference on (pp. 389-392). IEEE.
- [12] Tong F, Tang W, Xie R, Shu L, Kim YC. [2011] P-mac: a cross-layer duty cycle mac protocol towards pipelining for wireless sensor networks. In *Communications (ICC)*, 2011 IEEE International Conference on (pp. 1-5). IEEE.
- [13] Ruiz J, Gallardo JR, Villasenor-Gonzalez L, Makrakis D, Mouftah HT. [2009] QUATTRO: QoS-capable cross-layer MAC protocol for wireless sensor networks. In *Global Telecommunications Conference, 2009. GLOBECOM 2009*. IEEE (pp. 1-6). IEEE.
- [14] Taranovs R, Zagursky V. [2011] Medium access protocol for efficient communication in clustered wireless sensor networks. In *Telecommunications Forum (TELFOR)*, 2011 19th (pp. 582-585). IEEE.
- [15] Liang Y, Tang Q, Yue X, Li X, Liao Y. [2012] A cross-layer protocol based on leach in wireless sensor networks.
- [16] Dash S, Mallick SS, Hansdah RC, Swain ARA Distributed Approach to Construct Hierarchical Structure for Routing with Balanced Energy Consumption in WSNs.
- [17] Lin YH, Chou ZT, Yu CW, Jan RH. Optimal and Maximized Configurable Power Saving Protocols for Corona-Based Wireless Sensor Networks.
- [18] Kusumamba S, Kumar SM. [2015] A reliable cross layer routing scheme (CL-RS) for wireless sensor networks to prolong network lifetime. In *Advance Computing Conference (IACC)*, 2015 IEEE International (pp. 1050-1055). IEEE.
- [19] Louail L, Felea V, Bernard J, Guyennet H. [2015], MAC-aware routing in wireless sensor networks. In *Communications and Networking (BlackSeaCom)*, 2015 IEEE International Black Sea Conference on (pp. 225-229). IEEE.
- [20] Tariq M, Kim YP, Kim JH, Park YJ, Jung EH. [2009] Energy efficient and reliable routing scheme for wireless sensor networks. In *Communication Software and Networks, 2009. ICCSN'09. International Conference on* (pp. 181-185). IEEE.
- [21] Narain B, Sharma A, Kumar S, and Patle V. [2011]. Energy efficient mac protocols for wireless sensor networks: A survey. *International Journal of Computer Science & Engineering Survey (IJCSES) Vol, 2*.
- [22] Jacobsson M, Orfanidis C. (2015) Using software-defined networking principles for wireless sensor networks. In *11th Swedish National Computer Networking Workshop (SNCNW)*, May 28-29, 2015, Karlstad, Sweden.
- [23] RejinaParvin J, Vasanthanayaki C. [2015], Particle Swarm Optimization-Based Clustering by Preventing Residual Nodes in Wireless Sensor Networks. *Sensors Journal, IEEE*, 15(8):4264-4274.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# A NOVEL ROUTING MECHANISM USING SNR AND CBDM TECHNIQUE TO DEFEND AGAINST DDOS ATTACK

Sathya Priya<sup>1\*</sup> and Bharathi<sup>2</sup>

<sup>1</sup>Velammal Engineering College, INDIA

<sup>2</sup>Panimalar Engineering College, INDIA

## ABSTRACT

The growth of internet technology and increase in network attacks are directly proportional to each other. The internet is always targeted by various types of network threats. Amongst all the other type of attacks, Distributed Denial of Service (DDoS) is considered to be top most network attack. A DDoS attack is an attempt to make a service unavailable or unusable to intend user at intend time and there are no limitations in the number of systems that can launch the attack. Since this type of attack launch attacks by wide range of IP addresses, it is very hard to block and detect at the network firewall level. DDoS attacks remain a serious security problem; the mitigation of threat is very hard to the highly distributed network attacks. In this paper, we proposed DDOS detection using efficient router mechanism with the help of signal to noise ratio deviations and using technique Clustering Based Data Mining (CBDM) to monitor and calculate any deviation from the trained routing data and the same as been alarmed as attack. The evaluation based on anomaly detection using extensive simulations shows effectiveness and low overhead. The proposed work also supports for incremental deployment in real networks.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

DDoS, CBDM, Clustering, Anomaly, Score, Multi-Level

## INTRODUCTION

The increased development of e-commerce leads to the increased development of security on network resource. DDoS attack aims on total packets on the network, which in turn delay the user from accessing their network resource. The DDoS attack runs on a client machine and tries to compromise other systems in the standard network configurations and makes it as weak configurations.. There are two types of DDoS attack such as flood attack and crash attack. The vast development of internet technology resulted in large scale need of IP routers [1].

Initially attacker tries to gain vulnerabilities of weak network. The attacker tries through computer systems network port to gain unauthorized access. The number of ports that are open, means there are more chance of entry into the network. The compromised machines are now instructed to control another set of compromised machines [2]. These are called the agents or daemons. By doing this it is very difficult to track the actual attacker and place of occurrence of attack on the Internet. Therefore, it is most difficult to provide global security for the entire network. Any deviation in security leads to the loss of information. This unprotected environment results in unsatisfied customers and fall in reputation of organization.

In this paper, we propose a novel mechanism for detecting DDoS using clustering based approach. As our network is large, for earlier and accurate detection we need to sub group our network as finite number of clusters [3]. Each cluster is controlled and monitored by cluster lead. The cluster lead monitors the number of users logging in at particular time duration and the total number of packets sent by each user. Before implementing the actual concept, training has been provided to the network to differentiate between normal packets and the abnormal packets [4].

To distinguish normal packet from abnormal packet it considers the score obtained from the parameters such as number of entries of single user at particular interval of time and number of packets sent by him on that particular time period. For the simulation purpose we considered random amount packets can be sent in 3ms. The fore coming analysis of the work structured as follows. Section 2 gives discussion on related work. Section 3 explains

the system model. Section 4 summarizes the computations being performed to detect DDoS at router level. Section 5 provides details about the result discussion. Section 6 concludes with the future work

## RELATED WORK

The increased damage caused by DDoS attacks leads to the increased development of attack detection mechanisms. These approaches vary depending on the techniques being used. Many of the methods were implemented based on anomaly detection mechanism. Reyhaneh and Ahmad [5] proposed an anomaly based DDoS detection based on selected features of attacker packets and classified attacks as normal or abnormal, but failed to differentiate the type of attacks. Basheer Nayef [6], computed the correlation difference based on the outgoing and incoming packets of a network to detect DDoS attack.

Thwe Thwe Oo and Thandar Phyu [7], considered a statistical based approach to observe certain features of a network packet. The proposed system implemented novel routing mechanism to detect DDoS. It first calculates the cost function based upon different parameters of packet transmission. Then using the cost function, it measures the signal to noise ratio. Based on the result obtained scores has been set for each transmission. Then the scores are grouped under threshold values to predict the DDoS attack. Moreover, the proposed work also considers the count of each user entry. If the user entry exceeds with in the limit of predefined time duration, then it is identified as initiative of attack. Then it identifies normal and attack packets from matching the data with the predefined database values.

Thwe Thwe Oo and Thandar Phyu [8], the proposed work shows data mining approach to detect DDoS attack. Here, traffic features are calculated from network and then clustered into normal and abnormal attack traffic using data mining approach. This paper performs extensive computations. Now, in our paper, the proposed system is SNR with CBDM presents novel approach of data mining classification algorithm and score computation using mathematical calculation to detect DDoS. The calculation is based on light weight operations being performed in the network routers. The main aim of dividing network as clusters of different levels [9] helps easier identification of attack and also earlier detection of attack. Diagnosis earlier means prevention in future [10]. The proposed method shows better results with low over lead.

Barati et al., [11] proposed architecture of a detection system for DDoS attack. Genetic Algorithm (GA) and Artificial Neural Network (ANN) are deployed for feature selection and attack detection respectively in the hybrid method. Wrapper method using GA was deployed to be selected the most efficient features and then DDoS attack detection rate was improved by applying Multi-Layer Perceptron (MLP) of ANN. Results demonstrated that the proposed method was able to be detected DDoS attack with high accuracy and deniable False Alarm.

Katkar and Bhatia [12] evaluated the effect of various data preprocessing methods on the detection accuracy of DoS/DDoS attack detection Intrusion Detection System (IDS) and proved that numeric to binary preprocessing method performs better compared to other methods. Experimental results obtained using KDD 99 dataset are provided to support the efficiency of proposed combination.

Bhaya and Manaa [13] presented a hybrid approach called centroid-based rules to be detected and prevented a real-world DDoS attacks collected from "CAIDA UCSD" DDoS Attack 2007 Dataset" and normal traffic traces from "CAIDA Anonymized Internet Traces 2008 Dataset" using unsupervised k-means data mining clustering techniques with proactive rules method. Centroid-based rules are used to effectively detect the DDoS attack in an efficient time. The Result of experiments shows that the centroid-based rules method perform better than the centroid-based method in term of accuracy and detection rate. In term of false alarm rates, the proposed solution obtains very low false positive rate in the training process and testing phases. Results of accuracy were more than 99% in training and testing processes. The proposed centroid-based rules method can be used in a real-time monitoring as DDoS defense system.

## PROPOSED SYSTEM MODEL

The network is composed of collection of systems under an organizational entity. Each system has edge routers to get connected with the network. Every system in the network communicates with the help of Border Gateway Protocol (BGP). BGP passes information about routes to the routers [14]. At each stage of transmission routing

information updating takes place. In the proposed work, we compute path information in router table and pass same to the cluster representatives' who leads the network clusters [15]. For assumption, it is being simulated each cluster can communicate with other clusters using implicit signaling concept. [Figure -1] explains the following.

The Proposed system model works in two phases:

- **Training/ Computation Phase.**  
For training packets sent in TCP connection, offline operations are considered to identify normal user from abnormal user. In order to make implementation easier the network is divided in to sub groups as clusters. The clusters monitors the activity of user with respect to data sent and enter of each user into the network based on the monitoring result the cluster lead checks with predefined value of normal behavior. Deviation in this is noted as attack by cluster and creates score for that particular transmission. Transmission with less congestion will be given high score and vice-versa in order to avoid false alarm [16].
- **Detection Phase.**  
In the Detection phase cluster lead signals the next level cluster as congestion takes place [17]. Hence it disallows the packet transmission to next router. If the obtained level falls in-between normal and abnormal, at that situation, router gets incremented and same procedure has been followed in next cluster. After the confirmation of actual DDoS takes place, it drops packets. This helps to reduce unnecessary packet drops.

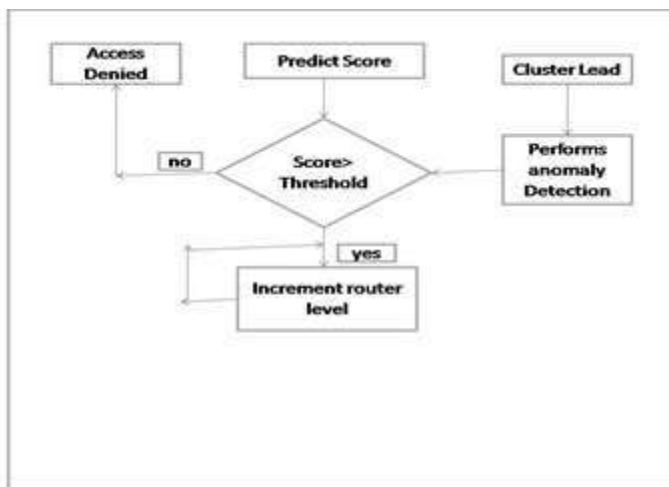


Fig: 1. System Architecture

## ANALYSIS AND COMPUTATIONS

Due to the distributed nature, DDoS attack seems to be more severe than all the other type of attack. . It is also not easy to identify the exact point of attack in the network. So in order to detect exact point of attack and necessarily detection at early stage, we aim to divide the network as clusters. The identification of attack in one cluster, which is nearer to ingress router, prevents the whole network from congestion. Earlier detection helps to prevent the server system being shutdown or overflow. In our previous work, we concentrated on detecting DDoS attack at the client side during authentication process. Now we would like to perform computation inside the router network to detect DDoS attack. For the result accuracy, we trained routers based anomaly based attack detection mechanisms.

Depending on the training provided to the user, score has been given to the each user and the same has been marked in the identification field [18]. Deviation is the predefined computed threshold value based on score, results in the rejection or acceptance of the packet from that particular user IP.

Given set of routers in the network  $R = \{ r_1, r_2, r_3, \dots, r_n \}$  where  $n$  is the total number of routers. Set of packets traverse through the network is termed as packet size  $ps = \{ ps_1, ps_2, ps_3, \dots, ps_n \}$ . The whole network is divided

in to finite group called cluster(k). The ith number of cluster is denoted as  $K_i$ , hence  $KI$  is  $k_i \in n$ . The number of packets from single user is recorded with its time duration and the length of the path it traverse. It is determined by  $R_{jy} - R_{ix} = dxy$ , where  $d$  is the distance travelled in a network.

#### Algorithm Inputs

Number of packets sent as packet size (PS)  
 Time Duration to monitor number of packets received from same IP as TD  
 Predefined Threshold vale for DDoS attack as Total packet Count FC.  
 Threshold value computed based on anomaly detection as score S.  
 Cluster as the group of routers in network as C  
 $p$  is the probability used for calculation purpose  
 User with IP address is represented as U for  $i$  number of times  
 The proposed method works as follows:  
 Step 1. Computation of score Metric.  
 Score has been computed in each router depending on the number of packets received.  
 Read: Packet Size PS, Time Duration TD, Total packet Count FC.  
 Result: Score based on probabilities  $p$   
 $S \rightarrow$  TCP (Anomaly Detection Mechanism)  
 Return  $S \forall PS$

Each packet contains its source and destination address along with relevant topological information. All the nodes participate in forwarding packets and maintain node strength. The analyzing of packet being traversed is monitored by cluster lead.

Cost function (CF) is calculated, based upon the score given by cluster lead, and the number of packets (FC) being sent in particular time interval (TD), packet size (PS) and score (S).

$$CF = \alpha.PS + \beta.FC + \gamma.TD + S. \quad (1)$$

$\alpha, \beta, \gamma$ . Are the expected weights assigned.

Depending on flow of packets under conditions like during attacks and no attacks different values may be assigned to the considered parameter metrics.

The node strength is a measure of sum of weights assigned to and the effects of attacks based on anomaly based detection are to be considered. Each node computes its individual strength by using signal to noise ratio [19]

$$S_{cn} = (1 - \alpha)SNR_{ps} + (1 - \beta)SNR_{fc} + (1 - \gamma)SNR_{td} \quad (2)$$

$S_{cn}$  is the score provided by the cluster and  $SNR_{ps}$ ,  $SNR_{fc}$ ,  $SNR_{td}$  are the cumulative packet flood obtained over a period of time and the score result has been set to 1 to 10 depending on the parameter considered. That is the range from 6 to 10, no Flood, 4 and 5 initiate for the DDoS attack alarm and 3 means the flood, 2 DDoS flood and 1 is DDoS crash.

When the  $S_{cn}$  falls below a critical threshold, it returns score value of 1, 2 and 3 depending on the severity. when the  $S_{cn}$  computed based on SNR is within the range of threshold 9 to 10 it proceeds with the successful packet transmission.  $S_{cn}$  returns the value of 1 to 10 based on the measure of SNR to tell about the node strength:

$$\begin{aligned} S &= 1 \text{ to } 3 \text{ if } CF \geq 100000 \\ S &= 4 \text{ \& } 5 \text{ if } CF > 50000 \text{ \& } \leq 100000 \\ S &= 6 \text{ to } 10 \text{ if } CF \leq 50000 \end{aligned} \quad (3)$$

Network not only affected by bulk bandwidth to induce DDoS attack but also large number of light weight entry can also flood the network. If a user enter more than the  $n$  number of permitted times, then the particular user has been put in blacklist database of cluster lead to keep an eye on particular user in his future transmission.



$$U = \frac{\sigma \sum A(i) * VH(i)}{n}$$

where, n is number of users Scores are updated finally in the score list.

As soon as process gets started, IP address and corresponding physical address of each users are noted [20]. User entry is maintained in A and  $\sigma$  is some constant metric added with user entries. If A, entry is greater than the predefined (n), then access for further transmission is denied. Else, it goes to next vertical cluster [21].

Step 2. Comparison with predefined Threshold.

In anomaly based detection mechanism, it keeps track of all the users entering into the network at the entry point and also counts the total number of packets being sent at particular time duration.

Step 3. Total packet Detection Pseudo code

```

Read: Routers R, Score S
Result: Alarm Signal
1.0 → R.
2. ∀ PS in TD do
    Determine Score → S based on SNR
    if S falls on Predefined Threshold of PS
//Predefined Threshold of PS (PS && U(i) = max & TD <= admitted value)
    compute SNR
    increment R
    else pause R for some (1-p) probability of time
    end if
    do SNR based on CF
    compute threshold
    signal Ci
    Ci alarms and warns Ci+1
    end do
    End.
3.End
    
```

Depending on the following parameters, scores has been fixed for the router loop execution.

Table: 1. Score Metric Calculation

User Entry	Packet Size	Distance Traversed in router (Hop Count)	Score
10.1.125.44	110000	15	1
54.121.54.11	10500	40	5
69.12.56.89	9500	15	8
58.129.102.9	76500	20	4
55.12.13.56	97800	14	2
54.121.55.6	82500	10	4
10.1.75.22	3600	15	3
10.1.25.11	77800	13	3
56.12.33.55	55000	30	5
45.23.66.78	4500	45	7

Whenever same user entry and number of packets sent are high and Distance traversed is less than the score provided will be less. High Score will be given chance of further transmission through the network [22]. [Figure - 2], [Figure - 3] shows network model and cluster formation.



Fig: 2. Model Simulation

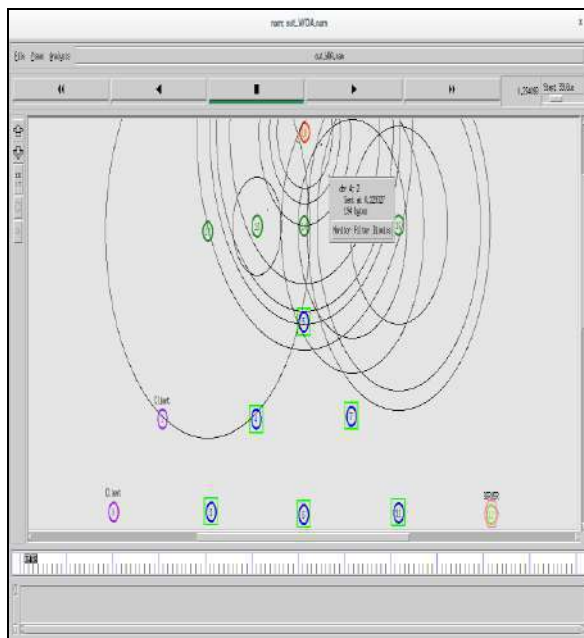


Fig: 3.Cluster Communications

## RESULTS AND DISCUSSIONS

Rank with highest number will be given chance to allow packets to the next level router. Performance measure based on cluster based data mining technique to detect DDoS shows better results with false positive and false negative. The Complexity of mechanism is less as it involves only a set of iterative same simple light weight operations. Until the predefined threshold value exceeds, it continues looks up and computation being performed by cluster lead. Also by considering different parameters at light weight computation rate and monitoring at different levels ensure accuracy in attack detection. The Proposed method can easily detect DDoS attack from flash crowd. The [Figure - 4] simulated the detection procedure with attack and without attack with respect to data loss and data delivery rate.



Fig. 4. Data loss with attack

## CONCLUSION

In this paper, we have proposed a data mining approach for the DDoS attack detection by using clustering based network. Before performing the actual implementation, we proposed to train the network for anomaly based detection as done in the real practical network detection approach. We considered the parameters such as packet count, time duration, hop count. The approach also very effective to be implemented under mobile adhoc network with very less over lead. We simulated the concept under both the normal case and the attack case. We mounted the most powerful DDoS attack changing attack types, so we could get the attack traffic of various types. As the outcome of experiment, we compared the misbehavior user from normal one at early stage by clustering the network as different level of cluster. Score calculation based on different parameters at different level shows the proposed approach is more effective compared with the existing multiple computations method. The future works focus on comparative experiments using different data mining technologies and statistic approach.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Elliott Karpilovsky, Mathew Caesar, Jennifer Rexford, Aman Shaikh, Jacobus van der Merwe.[2012] Practical Network-Wide Compression of IP Routing Tables, *IEEE transactions on Network and Service Management*, 9(4).
- [2] Boswell Steven, Calvert, Ben and Campbell, Paul. [2003] Security+ Guide to Network Security Fundamentals. *Thomas Course Technology: Canada*,
- [3] Keunsoo Lee, Juhyun Kim, Ki Hoon Kwon, Younggoo Han, Sehun Kim.[2007] DDoS attack detection method using cluster analysis, *Expert Systems with Applications*, Elsevier, 34(3):1659-1665.
- [4] Thwe Thwe Oo, and Thandar Phyu, Analysis of DDoS Detection System based on Anomaly Detection System, *International Conference on Advances in Engineering and Technology (ICAET'2014)* March 29-30, 2014 Singapore
- [5] K Reyhaneh, F Ahmad.[2011] An Anomaly-Based Method for DDoS Attacks Detection using RBF Neural Networks, *International Conference on Network and Electronics Engineering IPCST* vol.11, IACSIT Press, 2011, Singapore.
- [6] AD Basheer Nayef.[2005] Mitigation and traceback countermeasures for DDoS attacks” , Iowa State University,
- [7] Thwe Thwe Oo, Thandar Phyu.[2013] A Statistical Approach to Classify and Identify DDoS Attacks using UCLA Dataset” , *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 2( 5).
- [8] Thwe Thwe Oo, Thandar Phyu.[2013] DDoS Detection System based on a Combined Data

- mining Approach”, *4 th International Conference on Science and Engineering*,
- [9] Jérôme François, Issam Aib, and Raouf Boutaba.[2012] FireCol: A Collaborative Protection Network for the Detection of Flooding DDoS Attacks, *IEEE/ACM TRANSACTIONSONNETWORKING* 20(6).
- [10] A El-Atawy, E Al-Shaer, T Tran, and R Boutaba.[2009] Adaptive early packet filtering for defending firewalls against DoS attacks ,” in Proc. *IEEE INFOCOM*, pp. 2437–2445,
- [11] Barati M, Abdullah A, Udzir NI, Mahmud R, Mustapha N. [2014] Distributed Denial of Service detection using hybrid machine learning technique. In *Biometrics and Security Technologies (ISBAST), 2014 International Symposium on* (pp. 268-273). *IEEE*.
- [12] Katkar VD, Bhatia DS. [2014] Lightweight approach for detection of denial of service attacks using numeric to binary preprocessing. In *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on* (pp. 207-212). *IEEE*.
- [13] Bhaya W, Manaa ME. [2014]. A Proactive DDoS Attack Detection Approach Using Data Mining Cluster Analysis. *Journal of Next Generation Information Technology*, 5(4): 36.
- [14] Stefano Vissicchio, Laurent Vanbever, Cristel Pelsser, Luca Cittadini, Pierre Francois, and Olivier Bonaventure.[2013] Improving Network Agility With Seamless BGP Reconfigurations,*IEEE/ACM TRANSACTIONSONNETWORKING*, 21(3).
- [15] Wesam Bhaya, Mehdi Ebady Manaa.[ 2014] Review Clustering Mechanisms of Distributed Denial Of Service Attacks, *Journal of Computer Science, Science Publications*, 10( 10): 2037-2046
- [16] Hari Om, Aritra Kundu.[2012] A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System, in proceeding of the international conference on *Recent Advances in Information Technology (RAIT)*, pp. 15-17.
- [17] Basappa B Kodada, Gaurav Prasad, Alwyn R Pais, Protection Against DDoS and Data Modification Attack in Computational Grid Cluster Environment ,*I.J. Computer Network and Information Security*, 12-18, 2012.
- [18] Wu Xin-Wen, Zi Lifang, Yearwood John.[2010] Adaptive Clustering with Feature Ranking for DDoS Attacks Detection, in proceeding of the international conference on Network and System Security (NSS). 281-286,
- [19] R Dube, CD Rais K.-Y. Wang and SK Tripathi. [1997] Signal stability-based adaptive routing (SSA) for ad hoc mobile net-works, *IEEE Personal Communications*, 4( 1): 36–45,
- [20] A Sardana, R Joshi, and T hoon Kim.[ 2008] Deciding optimal entropic thresholds to calibrate the detection mechanism for variable rate DDoS attacks in ISP domain, in Proc. *ISA*, 270–275
- [21] Koutepas F, Stamatelopoulos and B Maglaris.[2004] Distributed management architecture for cooperative detection and reaction to DDoS attacks, *J Netw Syst Manage*, 12:73–942010

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

## CHASED APPROACH TO DETECT DDOS

Sathya Priya<sup>1\*</sup>, Rajagopalan<sup>2</sup>, Ramakrishnan<sup>3</sup>

<sup>1</sup>Dept. of CSE, Anna University, Chennai, INDIA

<sup>2</sup>Dept. of CSE, GKM Engineering College, Chennai, INDIA

<sup>3</sup>Dept. of IT, Madurai Kamraj University, Chennai, INDIA

### ABSTRACT

*Distributed Denial of Service (DDoS) poses severe threat to the network. DDoS seems to be more severe than all the other attacks, as there is no need to do complex cryptographic techniques to enter the network and corrupt data. This paper focuses on detecting DDoS attack at an early stage using cryptographic hashing technique. The use of SHA-1 implementation in place of normal hashing or non-cryptographic hash function prevents attackers from degrading network performance by finding hash collisions. Moreover for attack free communication between the sender and receiver, a pre-shared key authentication mechanism is followed. This paper proposed Cryptographic Hash based Edge router Deployment (CHASED) approach to detect DDoS attacks and prevent IP spoofing by avoiding hash collisions.*

Published on: 10<sup>th</sup>– August-2016

#### KEY WORDS

DDoS, Pre-shared key, SHA-1, Hash Collision.

### INTRODUCTION

DDoS is one of the challenging network attacks which exploit the network resources [1]. In DDoS attacks, most of the websites were made virtually unreachable to the internet users, hence results in heavy financial loss, fall in reputation of the organization, many unsatisfied customers and so on [2]. Denial of Service (Dos) or DDoS attacks have become a serious threat and great nuisance that destabilizes the internet. DDoS attacks have become one of the major research issues in the field of network security. DDoS attacks are handled by zombies knowingly or unknowingly and attacks the victim [3]. It is best to divide DDoS attacks as local and remote (network based) in order to gain actual knowledge about it. In local attacks, a form of malicious software resides in the computer system affects the other system in the network.

Remote based attacks also called as a network based DDoS attack, which disturbs the client accessing the server from the remote means. Examples of remote based attacks are syn flood attacks, Smurf attacks and so on. DDoS can be implemented by compromising any one of the system in the network, and through that compromise other nodes called zombies and send unwanted or exhausted packets to the server to make it as a victim.

For Example, to introduce attack on the popular website, the attacker can send false HTTP requests over the same network [4]. This type of request is same as that of one made by the intend user. Thus the attacker bypasses the network through any security mechanisms [5]. This paper explores the idea of detecting DDoS using cryptographic hash technique. Cryptographic algorithms are developed using computational hardness, ensuring such algorithms are very hard to break by the adversaries. It can be breakable by theoretical concepts, but it is infeasible for the practical means so these techniques are termed as computationally secured one [6]. This paper focuses on SHA-1 cryptographic hashing technique to detect the DDoS attack in early stage. The below [Figure - 1] shows how a DDoS attack takes place.

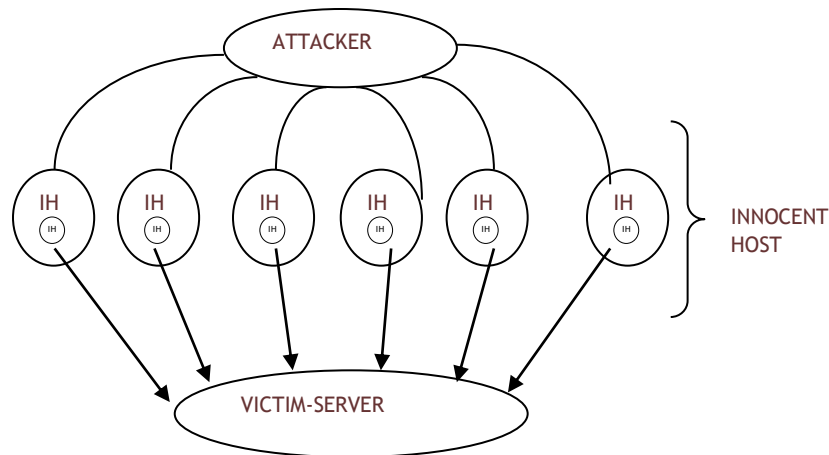


Fig. 1. DDoS Attack

The [Figure - 1] also shows how multiple attackers can attack simultaneously and the existence of multiple attack paths. It is not necessary, that attack can take place only at ingress edge, the attacker can also attack through intermediate routers in the network [7]. Because of the non-stable characteristics of the internet and anonymous behavior, there is no evidence of record of the transmission path of the packet. This paper focuses to detect DDoS attack from the identification of first non-matching packet itself and to break the path in which flooding takes place. For the best results, it is expected each router participating in the network must work in coordinating manner to identify the attacker's path.

## RELATED WORK

Bin Xiao et al, [8] proposed a system to consist of a client detector and a server detector. The client detector is implemented on the side of the innocent clients and utilizes Bloom filter-based detection strategy for generating precise detection outcomes yet utilizes least amount of storage as well as computation resources. Server detectors may actively monitor and control the warning process by sending requests to innocent hosts. A counter is implemented to calculate the SYN and SYN-FIN handshake signals. When only the SYN handshake signal reaches a threshold value, the client detector generates an alarm to inform the server detector. The server detector drops the packets on receiving the alarm. The throughput of this paper in detecting DDoS attack is good but results in hash collision.

Wang H et al., [9], instead of monitoring the ongoing traffic at the front end (like firewall or proxy) or a victim server itself, this detection mechanism detects the SYN flooding attacks at leaf routers that connect end hosts to the Internet. The detection mechanism is based on the protocol behavior of TCP SYN-FIN (RST) pairs. To make the detection mechanism insensitive to site and access pattern, a non-parametric Cumulative Sum (CUSUM) method is applied. From this paper, the idea of implementing the algorithm in the intermediate routers instead of ingress and egress router was proved to defend the DDoS attack and cause less effect on the server in the network. Then a scheme for filtering spoofed packets (DDoS attack) which is a combination of path identification (Pi) and client puzzle (CP) concepts was implemented. In ingress router, client puzzle for the client is placed. In this each IP packet, a unique Pi is added and router marks forwarding packets to generate unique identifiers corresponding to different paths. With this Pi, hop count value is also added. Thus the victim can use the <Pi, HC> tuple to identify and discard malicious packets from the attacker.

Praveena V et al, [10], gave the idea of including hop count in the cryptographic hash table to provide a better way to secure the server from attack. DDoS defense systems are deployed in the network to detect DDoS attacks independently. A communication mechanism is used to exchange information about network assaults among the independent detection nodes for aggregating data regarding overall network assaults noted. We assume that the Internet is composed of a set of Autonomous Systems (AS). The system is implemented with the overlay network to share the attack information using a gossip protocol based on epidemic algorithm over the Internet. Using the aggregated information, the individual defense nodes have approximate information about global network attacks

and can stop them more effectively and accurately. Guangsen Zhang et al, [11] derived the concept of adding more information for the routers to detect the DDoS attack.

Keromytis A et al, [12] gave the concept of intense filtering. An architecture called Secure Overlay Services (SOS) proactively prevents DoS attacks. Here reduced the probability of successful attacks by (i) performing intensive filtering near protected network edges, and (ii) introduction of arbitrariness as well as anonymity in forwarding path to a specific SOS-protected destination.

Almost all recent DDoS assault detection as well as prevention strategies are employed either on victim servers (assault source) or between the two [13]. The spoofed packets could be distinguished from normal ones by the Hop- Count deviation [14].

Source side mechanisms to detect as well as prevent DDoS assaults may be challenging to deploy. Source-end implemented methods works better but are difficult to deploy. After attacks are identified, attack sources can be discovered through traceback [15] as well as pushback technique. Most traceback schemes are implemented by either marking some packets in their routing paths or by sending special packets. By tracking these special marks, it is easy to reconstruct the real routing path reconstructed and locate the true source IP [16].

Once real routes of spoofed packets are detected, pushback method performs advanced filtering and works at the last few routers before the malicious traffic reaches the target victim [17]. Hence, it is not easy to detect abnormal deviations until the DDoS attack is at the final traffic-bursting stage [18].

Existing solutions can fail to raise accurate alarms when DDoS occurs and results in false positives and also using simple hash techniques results in large number of collisions [19]. So, cryptographic hash function is used in place of normal hash function to prevent collision attacks and to reduce false positive at high level.

## METHODOLOGY

The proposed work mainly based on cryptographic hashing technique over the edge routers. Unlike the previous research works on different type of hashing techniques, cryptographic puzzle, special cryptographic masking and so on, the explored work shows better results, as it is collision resistant.

The proposed work is implemented by considering the following assumptions in order to make the approach more effective and practical.

- Edge routers were implemented with cryptographic hashing technique.
- Routers are stable enough to perform hashing computation, which is collision free.
- Threshold value is set to identify normal user from abnormal user.
- Pre shared key authentication mechanism is used to avoid other nodes does not interfere the intend path for packet transmission.
- Multiple attackers can attack at the same or different time, hence multiple attack paths exist.
- It is assumed all routers should work in a coordinated manner, in order to produce good results.
- Each router is implemented with the alarm, in order to raise the alarm whenever it faces DDoS attack. In our simulation, it intimates by changing its color.
- Attackers can generate any volume of attacks and hence tries to flood or crash the server.

### Methodology of Proposed Work

From the previous work [21], it is observed that the hash based technique is comparatively good for identification of DDoS attack. Even though hash based technique has the limitation of collision attacks.

So, in this paper we implemented the DDOS detection approach in Cryptographic Hashing technique, which is highly collision resistant. . As the work simulated in IPV4, the SHA1 is applicable to be used than the SHA3 or other Cryptographic Hashing technique. In this, the IP address of the authorized users in the network is implemented with cryptographic hashing. So for the packets sent/received in a network, the identity should be unique in each router hash table. Hence, it is difficult for the attacker to break or even trying to make collision attacks in cryptographic based hashing technique

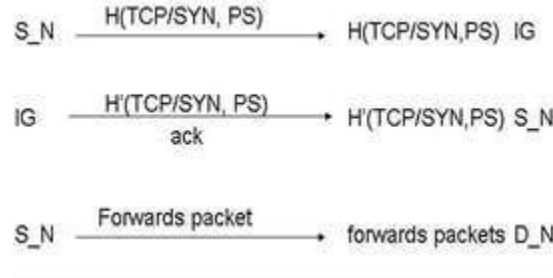
The work considers the following factors:-

- Packets sent / received must meet the threshold value which is already set.
- Pre-shared key authentication mechanism.
- Cryptographic hashed IP address traversed through each router.

Pre Shared Key authentication mechanism:

The main concern of pre shared key authentication mechanism is to overcome IP spoofed packets.

Sender Node(S\_N) sends request using TCP/SYN Hashed (H) with pre shared key(PS) given by routers to Ingress routers (IG). Ingress router checks authentication by Dehash (H') TCP/SYN using pre shared key and sends acknowledgement to sender node. Now, communication starts between the sender node to the destination node (D\_N).



Cryptographic Hash function:-

In Ingress router cryptographic hash function (SHA1) is implemented. Once authentication mechanism starts between sender node to the destination node, then Ingress router cryptographically hash the IP address coming through it and forwards the packet to the next router in the network. SHA 1 cryptographic hash function hashes the IP address of the nodes and generates a unique value. [20]. Cryptographic hash function helps to prevent IP spoofing.

Example for SHA-1 algorithm:

Source IP Address: 198.162.1.4

Hashed Source IP Address:

0xe138a664841494de5a5154981f5a499095e3c18b7

The hashed IP address is 160 bit value. It is in hexadecimal format

An attacker compromises few innocent hosts to perform DDoS attack by making the innocent host create traffic in the network. The innocent host starts creating traffic in the network by passing numerous amounts of packets to the routers that is to be passed to the server. [Figure-3] depicts the following

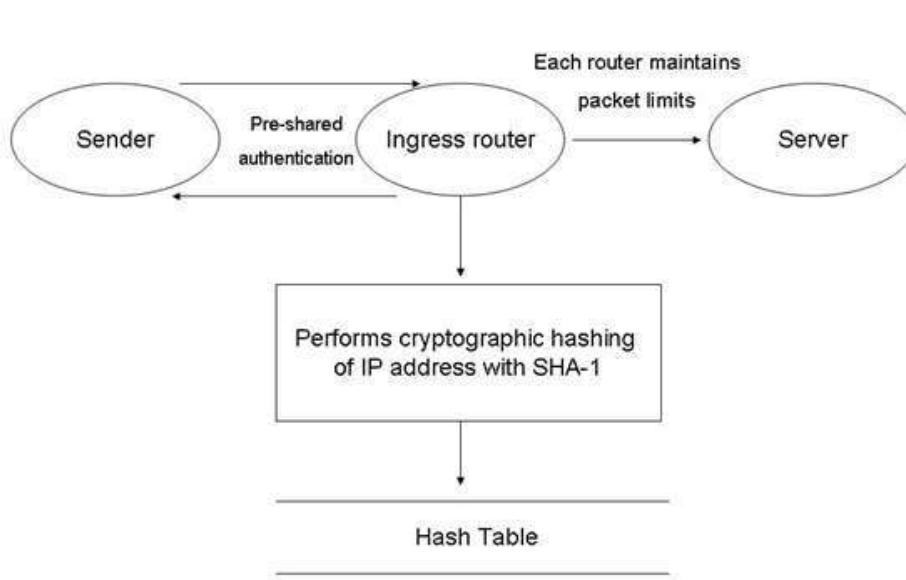


Figure 3: Proposed System Architecture

- Before transmitting the packet it performs pre shared key authentication
- When the packet reaches the ingress router, the router analyses the packets and hashes the source address using the hashing algorithm.
- The router checks the hashed source address with the source address in the hash table. If it matches, the router forwards the packet.
- Each router maintains the limit of packet in can hold. When the threshold value exceeds a certain value, the router starts dropping the packets which prevents exploitation of resources in the victim server.



**Algorithm for DDOS detection:**

**Participants:** S<sub>N</sub>, IG, D<sub>N</sub>, H, H', PS

Let us assume Sender node registered with router using preshared key.

•S<sub>N</sub> sends sendpacket (req) by Hashing (H) TCP/SYN with PS

•IG Dehash(H') and sends back ack ().

•Connection established, session starts, send packet.

•IG performs cryptographic hashing of all packets with SHA-1.

•Each router checks uniqueness of IP address. (Collision resistant)

•Threshold value  $\leq$  Limit, forwards packets else discards and signals as DDOS

### Router-Hash table Implementation

A router performs 'traffic direction' function on the Internet. Data packets are normally transmitted from one router to the next over networks which comprise the internetwork till they reach destination nodes [22].

The route consists of a table in which there are counters to count the number of packets of a particular source IP address. All counters are initialized to 0. When a key a (such as a source IP address) is inserted, the value of the counters are increased by 1 accordingly at the table addresses h1 (a), h2 (a), . . . ,hk (a). If an IP address b is stored in the hash table, the counters at the addresses h1 (b), h2 (b), . . . ,hk (b) in the table are all non-zero. This allows us to monitor the current statistic of control packets flowing through a router towards the server.

When the packet reaches the router, it analyzes the packet and gets the source IP address. When another packet reaches the router, it checks the hashed IP address in the hash table. If the packets are from the same IP address then the packet count is increased else the IP address is hashed and stored in the hash table.

The threshold value that is set can be differed according to various applications. In our proposed system, the threshold value is 200. So if any node sends packets that exceed 200 packets to the server, those packets of the node are dropped. Then the DDOS alarm has raised.

### Securing the network

DDoS attacks impact great challenge for the availability of resources for Internet Service Provider (ISP). A virtual security ring is now implemented around the network. Now, each router is implemented with hash table that tracks the normal user from the abnormal user. Normally with Ingress and egress filtering the router capable of filtering all traffic coming from the normal user and IP Spoofers are identified. As this proposed paper works on predefined threshold value insecure cryptographic hashing environment once the number of packets is higher than that is expected from the normal user, then the server stops sending the CTS and quits the connection with that particular client. By this, attackers from outside the network could be minimized. Setting continuous monitoring with the hash table mapping, the network could reduce threats to the great extent.

## RESULTS AND DISCUSSION

Our proposed system is compared with existing techniques to detect DDoS attack based on data delivery rate and data loss parameters. It shows reduced false positive. Implementing cryptographic hashing all over the network is cost effective but effective for the expected secure network. It is expected the approach extends its support to detect DDoS in effective manner with the collision resistant capability. [Figure 5] [Figure 6] below show the data loss for our proposed system produces better results with cryptographic hashing technique.

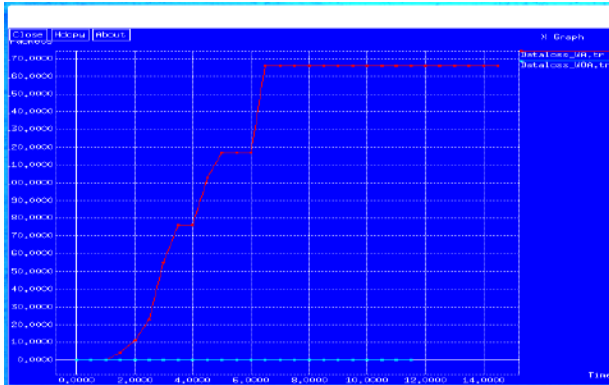


Fig: 5. Data Delivery Rate

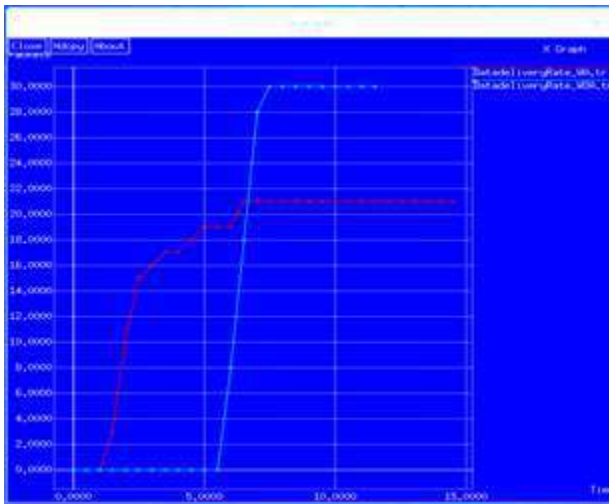


Fig:6. Data loss Rate

## CONCLUSION AND FUTURE WORK

Our system makes use of cryptographic hash techniques that prevents server overloading. Collisions are minimized to a greater extent in the hash table. It also ensures that the information from the client reaches the server without any loss. Based on a particular application, the specified packet count can be deferred. The latency of our proposed system is high since there are various computations involved in the routers. The future work might involve reducing the latency of the system.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Ben-Porat et al. [2013], Vulnerability of Network Mechanisms to Sophisticated DDoS attacks, *IEEE Transactions on Computers*, 62( 5): 1031-1043.
- [2] Bin Xiao et al.[2006] A Novel Approach to Detecting DDoS attacks at an Early Stage, *Super Comp* 36: 325-248.

- [3] Houle K.J and Weaver GM. [2001] Trends in Denial of Service Attack Technology, CERT Coordination Center, pp 1-21. [http://www.cert.org/archieve/pdf/Dos\\_trends.pdf](http://www.cert.org/archieve/pdf/Dos_trends.pdf).
- [4] Mehdi and Angela. [2012] Review of SYN Flooding attack Detection Mechanism, *IJDPS*,3 (1)
- [5] Changhua S et al.[2007] A Novel Router based scheme to Mitigate SYN flooding DDoS attacks, *IEEE Infocom*.
- [6] Dalip Kumar. [2013] Analysis of IP Spoofed DDoS Attack by Cryptography, *IJCEM*,16( 2).
- [7] Belenky A et al.[2003]Tracing multiple attackers with deterministic packetmarking (DPM), *Proc IEEE PACRIM'03*, Victoria, BC, Canada, ] 49–52.
- [8] Bin Xiao et al.[2006] A Novel Approach to Detecting DDoS attacks at an Early Stage, *Super Comp* 36:325-248.
- [9] Wang H et al.[2002] Detecting SYN Flooding Attacks
- [10] Praveena V et al [ 2012], Mitigating Technique to Overcome DDOS Attack.
- [11] Guangsen Zhang et al. [2005] Cooperative Defense against Network Attacks.
- [12] Keromytis A et al, 2004,SOS: An architecture for mitigating DDoS attacks.
- [13] Ferguson P and Senie D,[1998] RFC-2267- Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing.
- [14] Robert Beverly and Steven Bauer.[2005] The spoofer project: Inferring the extent of Source Address Filtering on the Intenet, *USENIX SRUTI: Steps to reducing unwanted traffic on the internet workshop*, (2):53-59.
- [15] Song D.X and Perrig A.[2001] Advanced and Authenticated Marking schemes for IP Traceback, proceedings of the IEEE Infocom *IEEE Computer Society*, Los Alamitos, Calif
- [16] uvari B and Govindarasu M. [ 2006] Novel hybrid schemes employing packet marking and logging for IP traceback, *IEEE Trans. Parallel Distributed Syst*, 17( 5): 403–418.
- [17] John Ioannidis et al. [2002] Implementing Pushback: Router based defense against DDoS Attack, Internet Society.
- [18] Wei-Shen Lai et al. [2008] Using Adaptive bandwidth a location approach to defend DDoS attacks, *International Journal of Software Enginewering and Its Applications*,2(4): 61-72.
- [19] Pyungkoo et al. [2012] A Pseudo State based Distributed DoS Detection Mechanism using Dynamic Hashing, *Springer* , 22-29.
- [20] Zaihongzhou et al,[009] A Novel Distributed Scheme against DDoS attack, *Journal of Networks*.4(9).
- [21] Santhanam et al.[2006] Taxonomy of IP Traceback, *Journal of Information Assurance and Security* 1:. 79-94
- [22] Nashet D et al.[ 2008] Router based Detection for low rate DDoS attacks, *IntConf of HPS*, 177-182.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# IMPROVING GABOR FILTER BANK DESIGN AND SVM OPTIMIZATION THROUGH STOCHASTIC DIFFUSION SEARCH FOR MILD COGNITIVE IMPAIRMENT CLASSIFICATION

Shanthy<sup>1\*</sup> and Karthikeyan<sup>2</sup>

<sup>1</sup>Dept. of Computer Science Engineering, Tamilnadu College of Engineering, Coimbatore, TN, INDIA

<sup>2</sup>Dept. of Electronics and Communication Engineering, Tamilnadu College of Engineering, Coimbatore, TN, INDIA

## ABSTRACT

Mild Cognitive Impairment (MCI) is considered a precursor to Dementia in most cases, manifesting as one of the first symptoms to appear in an individual. Alzheimer's disease (AD) is the topmost reason behind dementia among elderly patients. Advanced neuroimaging methods are exhaustively utilized for diagnosing primary Alzheimer's disease as well as Mild Cognitive Impairment affected individuals with amnesia. In the current study, automatic labelling method which effectively classifies Magnetic Resonance Images (MRIs) as normal or anomalous is suggested through the use of machine learning methods. Textural attributes of MRIs are extricated through Gabor filters because of their excellent performance and textural analyses capacity for analysing spatial frequencies. Choosing best filters is crucial to enhancing performances of machine learning methods and it is NP-hard. The current study suggests a systematic Gabor filter optimization method which produced better as well as problem-specific filter sets utilizing Stochastic Diffusion Searches (SDSs), which are able to discover the location of predetermined patterns or in event that they do not exist, their most optimal instantiation within the search space. This is attained through parallel explorations of entire search spaces by groups of agents exploring in competitive yet cooperative fashion. The filters are valued through Support Vector Machine (SVM)-based application-oriented fitness criteria. Outcomes revealed that better performance of the suggested Gabor filter as well as SDS optimized SVM.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Magnetic Resonance Imaging (MRI), Mild Cognitive Impairment (MCI), Alzheimer's Disease (AD), Gabor Filters, Support Vector Machine (SVM), Stochastic Diffusion Search (SDS)

\*Corresponding author: Email: [shanthy.as.bir@gmail.com](mailto:shanthy.as.bir@gmail.com)

## INTRODUCTION

Alzheimer's disease is a neurodegenerative disorder that progresses for a long time before the manifestation of clinical symptoms. Even though there has been extensive research on AD, there is still some degree of uncertainty with respect to its prodromal stages. The symptoms of Mild Cognitive Impairment, on the other hand, may be identified several years prior to the actual diagnosis. This implies that there is a prolonged preclinical phase which precedes the clinical manifestation of AD. Timely treatment and precocious diagnosis is crucial here, since the progression of the disease may be slowed down and the development of new symptoms delayed [1].

For the early detection of Alzheimer's as well as the prodromal state of dementia, MCI holds great clinical importance. MCI is a heterogeneous syndrome which is often undiagnosed since it is challenging for clinicians to detect cognitive impairment be it at any stage. In later stage dementia up to 50% may escape recognition. The screening tests prevalently used including the Mini-Mental State Examination (MMSE) fail to reliably recognize subtle cognitive impairments in patients in the early stages. Word list, narrative recall and other linguistic memory tests exhibit greater efficacy in the detecting MCI, but they run the risk of yielding false positive diagnosis, which is undesirable [2].

The three phases of Alzheimer's disease are preclinical, MCI and dementia. The starting stage is the preclinical stage, MCI is characterized by mild changes in memory and dementia indicates severe affectation of the disease. AD patients may exhibit symptoms that vary from person to person. Following are few of the common symptoms:

- Loss of memory that inhibits the performance of day-to-day chores.
- Difficulty in problem-solving or planning.

- Time and place disorientation.
- Challenges in comprehending visual content and spatial relationships.
- Impaired judgment.
- Withdrawal from and loss of interest in work and other social interactions.

Initially, cross-sectional studies employed structural MRI to distinguish individuals with MCI from healthy subjects. A majority of the previous MR neuroimaging research concentrated on investigating the grey matter utilizing voxel-based morphometry in identifying MCI. There were significant volume differences distributed primarily within the precuneus and cingulate gyrus between patients afflicted with MCI and those in the control groups, as revealed by a series of such studies. New contributions towards research on neurodegenerative diseases suggest that using DTI to assess the changes in WM microstructure might be a more reliable parameter as compared to grey matter data. This approach is more sensitive to mild structural changes that may take place during the initial stages of degenerative process.

Greater accuracy in structural neuroimaging analysis can be gained by using high tissue contrast yielded by T1-weighted (T1w) MRI as a potential surrogate biomarker that can be used in the diagnosis and prediction of AD. Image processing techniques so far have failed to predict with accuracy the probability of contracting AD in the future for patients who have MCI. In the investigation of various diseases and disorders including dementia, OCD and schizophrenia, by examining cortical structural changes and differences, measurements of cortical thickness gained from MRI is used which exhibits high sensitivity to even minor structure modifications over the cortex.

Results derived from previous studies, however, indicate that the performance of using cortical thickness measurements is poorer than other methods in predicting Alzheimer's disease in patients with MCI, getting accuracy rates of 56% to 70% based on the technique. Since the measurement of cortical thickness is at a great resolution (ranging up to tens of thousands of points on the cerebral cortex), prediction in a discriminatory model utilizing such a vast number of measurements may run the risk of over-fitting.

The aim of feature extraction stage is the extraction of important image-based attributes from MRIs for all subjects in references as well as study group. Firstly, raw scans are adjusted for intensity non-homogeneities apart from noise being removed through non-local means methods. The scans are scaled in a linear fashion in grey-level intensity over all subjects for matching mean levels of reference images, set in standardized target template spaces optimizing global as well as local alignments between reference as well as subjects through consequent modifications [3].

Selecting features is automated features selections mostly, with relevance to predictive modelling problems. It incorporates the selecting of features subsets which have relevance, and are utilized for constructing models. Selecting features is different from reducing dimensionalities. Both decrease features in datasets, but reducing dimensionalities is done through the creation of novel features combination while selecting features involves the inclusion or exclusion of data features with no changes. Selecting features is the identification of non-relevant or repetitive features from data which offer nothing to the predictive method's precision or actually reduces the method's precision and discard them. Three classes of features selection methods are present which are filters, wrappers as well as embedded techniques. Filter approaches utilize statistical metrics to rank attributes and on the basis of the ranks, attributes are retained or discarded from datasets. With wrappers, selecting features sets are regarded as search issues where predictive models evaluate attribute subset. The embedded approach learns which attributes offer most to the method's precision during the construction of the models.

Latest classification techniques were built so that they permit individual classes estimations. Amongst them, machine learning methods are suggested for distinguishing MRIs from two sets of subjects that is healthy versus sick individuals. All the methods need training set that is already classified subjects such as healthy individuals as well as individuals with confirmed diagnoses for the categorizing of fresh subjects who are part of the test populations, into any of the classes which subjects of training sets are part of. One or more attribute variables are needed for the differentiating of the two sets which are being studied.

Particularly, SVMs are being used in recent times for assisting in the distinguishing of Alzheimer's disease afflicted individuals from control subjects through the use of anatomical MRIs. Classification techniques are employed for classifying MCI afflicted individuals in contrast to control individuals or in assisting in the differentiation of Alzheimer's disease from fronto-temporal lobar degradation. Although attribute variables may

be delineated from the entire brain, the variables might not possess related physio-pathological interpretations or merely partial sets of most discriminatory voxels or areas are gradually utilized for the classification of subjects [4].

In the current work, a MCI classification using Gabor filters, SDS and SVM methods. The paper's structure is thus: Section 2 reviews relevant literature. Section 3 elaborates on methods employed; Section 4 exposes outcomes from experiments and Section 5 provides the conclusion.

## RELATED WORKS

Roman and Pascual [5] surveyed the latest discoveries within the field of neuroimaging related to diagnosing Alzheimer's disease as well as Vascular Dementia (VaD). MRIs as well as Computerized Tomographies (CT) have been offered precise demonstrations of locations as well as degree of advancement of atrophic alterations impacting brains due to Alzheimer's disease as well as the several kinds of vascular lesions noted in mixed dementias as well as in pure vascular dementias. Quantifying cortical thicknesses permits earlier diagnoses as well as rates of advancement from MCIs to dementias. Quantification of white matter may be carried out by Diffusion Tensor Imagings (DTI) and functional MRIs (fMRI), functional connectivities, and MR Spectroscopies (MRS).

Zhou et al., [6] suggested CAD based technique on the basis of wavelet-entropies of attribute space method as well as a Naïve Bayes classification technique for the enhancement of brain diagnoses' precision through NMR scans. The attribute that was relevant the most was taken as wavelet entropies that was utilized to for training Naïve Bayes classifiers. Outcomes revealed that the suggested classifier identified anomalous from normal control brains excellently and was on par with recent techniques.

Zhuang et al., [7] utilized DTIs for detecting white matter structure modifications in MCIs as well as its sub-kinds and focused on the examination of whether DTIs may be utilized as possible imaging markers of MCIs. Ability of DTI in discerning MCIs from CNs was tested through binary logistic regression models.

Liu et al., [8] suggested a new multi-task features selection technique for preserving complementary inter-modality data. Particularly, it considered features selection from all modalities as distinct tasks and moreover imposed constraints for the preservation of inter-modality relations, apart from ensuring sparsity of chosen attributes from all modalities. Once features are selected, multi-kernel SVMs were utilized for the integration of chosen attributes from all modalities for classifications. The technique was tested through baseline PET scans as well as MRIs of subjects got from the Alzheimer's disease Neuroimaging Initiative (ADNI) database.

Zhang et al., [9] suggested a new hybrid method for the classification of provided MRIs as normal or anomalous. The suggested technique initially utilized DWT for extracting features and later PCA for reducing features space. Later, Kernel Support Vector Machines (KSVM) with RBF kernels, utilizing Particle Swarm Optimization (PSO) for optimizations was built. Five-fold cross-validations were used for obviating over-fitting.

## METHODOLOGY

Textural attributes of MRIs brain images are extricated through Gabor filters. In this section, the Gabor filter, SDS proposed optimization of the Gabor filter and SVM methods are described.

### Gabor filters

Gabor filters are band-pass filters that possess both orientation-selective as well as frequency-selective characteristics as well as best joint resolutions in spatial as well as frequency fields. Through the application of adequately tuned Gabor filters to signature images, textural data may be created. The accentuated textural data may be utilized for the generation of features vectors. Gabor filters are utilized with great success in segmenting fingerprints as well as palm prints, apart from their detection [10].

1D Gabor filters are given as the product of cosine/sine (even/odd) waves with Gaussian windows thus,

$$g_e(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \text{Cos}(2\pi w_o x) \quad (1)$$

$$g_o(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \sin(2\pi w_o x) \quad (2)$$

Wherein  $w_o$  is center frequency (frequency wherein filters yield best response) and  $\sigma$  refers to the spread of Gaussian windows [11].

Gabor filters are obtained through the modulation of sinusoids with Gaussians. For 1D signals, 1D sinusoids are modulated with Gaussians. Filters respond to certain frequencies, though merely in signals' localized parts. Let  $g(x, y, \theta, \phi)$  be function delineating Gabor filter centred at origin with  $\theta$  as spatial frequency as well as  $\phi$  as orientation. Gabor filter is given by

$$g(x, y, \theta, \phi) = \exp\left(-\frac{x^2 + y^2}{\sigma^2}\right) \exp(2\pi\theta i(x \cos \phi + y \sin \phi)) \quad (3)$$

It is revealed that standard deviation of Gaussian kernels relies on spatial frequencies assessed that is  $\theta$ . 2d Gabor functions comprise of sinusoidal plane waves of certain frequencies as well as orientations, with modulation by 2D Gaussian envelopes. 'Canonical' Gabor filters in spatial domains are as follows:

$$h(x, y) = \exp\left\{-\frac{1}{2}\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right]\right\} \cos(2\pi\mu_0 x + \phi) \quad (4)$$

Wherein  $\mu_0$  and  $\phi$  are frequencies as well as phases of sinusoidal plane waves along z-axes while  $\sigma_x$  and  $\sigma_y$  are space constants of Gaussian envelopes along x-axes and y-axes, correspondingly. Gabor filters with random orientations  $\theta_0$  may be got through rigid rotations of x-y coordinate systems. The 2D functions are revealed to be excellent fits with corresponding field profiles of generic cells in striate cortices.

Frequency-selective as well as orientation-selective characteristics of Gabor filters are more direct in their frequency domains' representations. When  $\phi = 0$ , Fourier transforms of Gabor functions in (5) are real-valued and become

$$H(u, v) = A \left( \exp\left\{-\frac{1}{2}\left[\frac{u - u_0}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right]\right\} + \exp\left\{-\frac{1}{2}\left[\frac{u + u_0}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right]\right\} \right) \quad (5)$$

Wherein

$$\sigma_u = \frac{1}{\sqrt{2\pi}\sigma_x}, \quad \sigma_v = \frac{1}{\sqrt{2\pi}\sigma_y}, \quad \text{and} \quad A = 2\pi\sigma_x\sigma_y \quad (6)$$

Fourier domains' representations in (6) specify the quantity by which filters modify or modulate all frequency components of inputted images. These representations are known as Modulation Transfer Functions (MTF). Choosing best filters is crucial to enhancing performances of machine learning methods and it is NP-hard.

### Stochastic Diffusion Search (SDS)

The current study suggests a systematic Gabor filter optimization method which produced better as well as problem-specific filter sets utilizing Stochastic Diffusion Searches (SDSs), which are able to discover the location of predetermined patterns or in event that they do not exist, their most optimal instantiation within the search space. SDS may be utilized for pattern searches as well as matchings. These issues may be considered with regard to optimizations through the definition of objective functions  $F(x)$ , for hypotheses  $x$  regarding locations of solutions, because similarities between target patterns as well as respective regions at  $x$  in search spaces as well as discovering  $x$  so that  $F(x)$  is maximum. Generally, SDS may be employed with ease to several optimization issues wherein objective functions are capable of being divided into units which may be valued in an independent fashion:

$$F(x) = \sum_{i=1}^n F_i(x) \quad (7)$$

Wherein  $F_i(x)$  is given as the  $i^{\text{th}}$  partial evaluation of  $F(x)$ .

For locating optimum of specified objective functions, SDS uses populations of  $k$  agents, all of which maintain hypotheses regarding optimum. During operations, the model entails iterations of Test as well as Diffusion stages till agents perform convergence on best hypothesis.

SDS Algorithm comprises [12]:

Initialising agents()  
 while(terminating criterion is not fulfilled)  
 Testing hypothesis()  
 Diffusion hypothesis()  
 Stop

**Initialisation:** Generally, first hypotheses of all agents are chosen evenly arbitrarily across search spaces. If data regarding potential solutions are accessible in an a priori fashion, it may be utilized for biasing original choosing of hypotheses.

**Test Function:** Boolean test functions reveal if arbitrarily chosen partial valuation of objective functions denote 'good' hypothesis or not.

**Test Phase:** All agents employ test functions to their current hypotheses. When test functions return true, agents are active, else they are inactive. **Diffusion Phase:** All inactive agents (A) choose one more agent (B) arbitrarily. If they are active, then their hypotheses are duplicated by A. If they are inactive, then A chooses one more arbitrarily across search space. **Convergence:** With the advancement of iterations, clusters or agents with identical hypotheses are generated. When converged, biggest cluster of units gives the best solution.

SDS efficiently carried out most optimal match among already present objects in search spaces as well as descriptions of targets. It is given that SDS will be capable of discovering targets if they exist in the search space else they identify objects with most identical descriptions of targets. Spaces as well as objects are delineated with regard to Atomic Data Units (ADUs) that comprise set of fundamental attributes. All objects in search spaces as well as targets are defined with regard to ADU and are not capable of possessing attributes ADUs may be regarded as single pixel intensities when search spaces are bitmap images or may comprise few higher level characteristics such as vertical or horizontal lines, angles or even semicircles. When search spaces as well as targets are delineated with regard to these characteristics or they may be letters or nodes in graphs.

All agents act in an autonomous manner as well as in parallel with others attempt to identify the location of targets in search spaces. The location of targets are denotes as coordinates of predetermined reference points in targets' descriptions. Transmissions or dispersal of data ensures that units are able to interact with one another and allot operational resources in a dynamic fashion on the basis of results of searches. On the basis of the performance in searches, agents may become active if they reveal possibly accurate locations in search spaces else, they are inactive. Every agent has access to search spaces as well as descriptions of targets [13].

Originally every agent is arbitrarily initialized to a reference point in search space. Additionally, they are first set as inactive. All agents, independent of one another perform probabilistic checks of data at reference points through comparison of arbitrary ADU from targets with respective ADU in search spaces. If tests are successful, agents become active else, they are inactive.

In conclusion, activities of agents indicate the probability that they point to accurate location. But because of partial testing, probability of false positives is not discarded and neither is the likelihood of false negatives. In this manner, SDS may obviate local minimum through correspondence to objects which have partial matches to descriptions of targets. Consequently during diffusion, all inactive agents arbitrarily choose one more agent to interact with. On the basis of whether the selected unit is active or not, the selecting agent points to the same point as the one that is active else arbitrarily resets its position, if it is also inactive. Active units do not perform sampling of other agents for transmissions but they go through fresh testing stage and on the basis of it, maintain active status or become inactive.

The procedure continues till statistical equilibrium is attained. Terminating criteria utilized and supervised the most quantity of agents showing same location in search space. When quantity of agents in the cluster is greater than a specified threshold and it within particular boundaries for a set of iterations, it is described as SDS reaching equilibrium while process is stopped. Although agents perform in an autonomous manner and merely weak forms of probabilistic couplings exist, it ensures that agents build cooperative nature.

### Support Vector Machine (SVM)

SVM is a group of monitored learning mechanisms utilized in classifications as well as regressions. It belongs to a set of generic linear classifications. Specific characteristic of SVMs are that they concurrently reduce empirical classification errors to a minimum while bringing to a maximum the geometric margin. Hence, SVMs are known as maximum margin classifiers. SVMs are grounded in Structural Risk Minimization (SRM). SVM maps input vectors to high dimensional spaces wherein maximal separating hyperplanes are created. Two parallel hyperplanes are created on both sides of hyperplanes which keep information separate. Separating hyperplanes are those which make the distance between two parallel hyperplanes maximum. A presumption that is made is that the greater the margin between parallel hyperplanes, the more improved the generalization error of classifiers [14].

It regards data points in the format

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots, (x_n, y_n)\} \quad (8)$$



Wherein  $y_n = 1/-1$ , a constant representing class to which point  $x_n$  is a part of.  $n$  = number of sample. Every  $x_n$  is a  $p$ -dimensional real vector. Scaling is significant in guarding against variables with greater variance. The training data may be viewed through separating hyperplanes that take

$$w \cdot x + b = 0 \tag{9}$$

Wherein  $b$  is scalar and  $w$  is  $p$ -dimensional Vector. Vector  $w$  is perpendicular to the separating hyperplane. Appending offset variable  $b$  permits the expansion of margins. Absent of  $b$ , hyperplane is made to pass through origin, limiting solution. Parallel hyperplanes may be given as

$$\begin{aligned} w \cdot x + b &= 1 \\ w \cdot x + b &= -1 \end{aligned} \tag{10}$$

If training data is linearly separable, it is capable of selecting hyperplanes such that there are no points between them and later attempts the maximization of distances. Geometrically, it discovers the distance between hyperplanes as  $2/|w|$ . Hence,  $|w|$  is to be minimized. For excitation of data points, it is required to be ensured that for every  $i$  either

$$w \cdot x_i \cdot b \geq 1 \text{ or } w \cdot x_i \cdot b \leq -1 \tag{11}$$

This may be given as

$$y_i (w \cdot x_i + b) \geq 1, \quad 1 \leq i \leq n \tag{12}$$

Samples along hyperplanes are known as Support Vectors (SVs). Separating hyperplanes with biggest margin delineated by  $M = 2/|w|$  which defines supports support vectors implying training data points nearest to it. This has to fulfil:

$$y_j [w^T \cdot x_j + b] = 1, \quad i = 1 \tag{13}$$

Optimal Canonical Hyperplanes (OCH) are canonical hyperplanes possessing most margin. OCHs ought to fulfil the restrictions given below:

$$y_i [w^T \cdot x_i + b] \geq 1 \quad ; i = 1, 2, \dots, 1 \tag{14}$$

Wherein  $l$  is Number of Training data point. For discovering best separating hyperplanes possessing most margins, learning machines ought to make minimum the  $\|w\|^2$

The issue was resolved by saddle points of Lagrange's Function:

$$\begin{aligned} L_p = L_{(w,b,\alpha)} &= \frac{1}{2} \|w\|^2 - \sum_{i=1}^1 \alpha_i (y_i (w^T x_i + b) - 1) \\ &= \frac{1}{2} w^T w - \sum_{i=1}^1 \alpha_i (y_i (w^T x_i + b) - 1) \end{aligned} \tag{16}$$

Wherein  $\alpha_i$  is a Lagranges multiplier. Searching for best saddle points ( $w_0, b_0, \alpha_0$ ) is required as Lagranges should be made minimum in terms of  $w$  and  $b$  and should be made maximum in terms of nonnegative  $\alpha_i$  ( $\alpha_i \geq 0$ ). The issue may be resolved in primal or dual forms. The two formulae are convex and KKT conditions that are required as well as adequate criteria for maximums of equations. Partially differentiated equations in terms of saddle points ( $w_0, b_0, \alpha_0$ ) [15]:

$$\partial L / \partial w_0 = 0 \tag{17}$$

$$w_0 = \sum_{i=1}^1 \alpha_i y_i x_i \tag{18}$$

$$\partial L / \partial b_0 = 0 \tag{19}$$

$$\sum_{i=1}^1 \alpha_i y_i = 0 \tag{20}$$

Replacing above equation, it changes primal to dual form.

$$L_d(\alpha) = \sum \alpha_i - 1/2 \sum_{i=1}^1 \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{21}$$

To discover optimal hyperplanes, dual Lagrangians ( $L_d$ ) have to be made maximum in terms of nonnegative  $\alpha_i$  ( $i$  .e.  $\alpha_i$  should be in nonnegative quadrant) as well as in terms of equality restrictions as given:

$$\alpha_i \geq 0, i=1,2,\dots,1$$

$$\sum_{i=1}^1 \alpha_i y_i = 0 \tag{22}$$

It is to be noted that dual Lagrangians  $L_d(\alpha)$  are given with regard to training data as well as depending solely on scalar products of input pattern  $(x_i^T x_j)$ .

Several kernel functions assist SVM in obtaining best solutions. RBF is employed often because it is capable of classifying multi-dimensional data. With RBF kernels, two variables C that represent costs of penalties as well as  $\gamma$  impact splitting result in features space ought to be set adequately.

### RESULTS AND DISCUSSION

Performance efficacy of suggested methods for the classification of MRIs as MCI is tested through 135 images taken from individuals between 20 and 65 years of age, with around 84 anomalous scans revealing MCI. Feature extraction is carried out on the MRIs through Gabor filters as well as the suggested optimization technique. The outcomes got through Gabor filters with no optimization but filter banks with orientations of 0, 45, 90 and 135 degrees on 13 by 13 windows. [Table 1 - 5] and [Figure 1 - 5] as shown below:

Table 1. Classification Accuracy

Techniques used	Classification Accuracy (%)
CSGabor - SVM(Poly)	89.95
CSGabor - SVM(RBF)	93.15
SDSGabor- SVM(Poly)	88.13
SDSGabor- SVM(RBF)	92.69
CSSDS- SVM(Poly)	94.93
SCSDS- SVM(RBF)	95.85

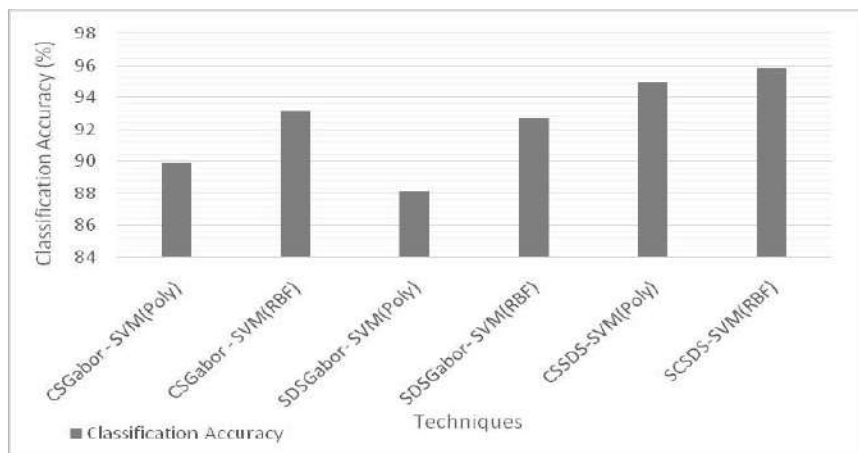


Fig. 1. Classification Accuracies

From the [Figure- 1], it is seen that the SCSDS-SVM (RBF) technique increased classification accuracies by 195.78%, 195.92%, 8.39%, 3.35% & 0.96% when compared with various number of CSGabor – SVM (Poly), CSGabor – SVM (RBF), SDSGabor- SVM (Poly), SDSGabor- SVM (RBF) and CSSDS-SVM (Poly) methods.

Table: 2. Sensitivity for normal

Techniques used	Sensitivity for normal
CSGabor - SVM(Poly)	0.9111
CSGabor - SVM(RBF)	0.9556
SDSGabor-SVM(Poly)	0.8741
SDSGabor-SVM(RBF)	0.9333
CSSDS-SVM(Poly)	0.9699
SCSDS-SVM(RBF)	0.9699

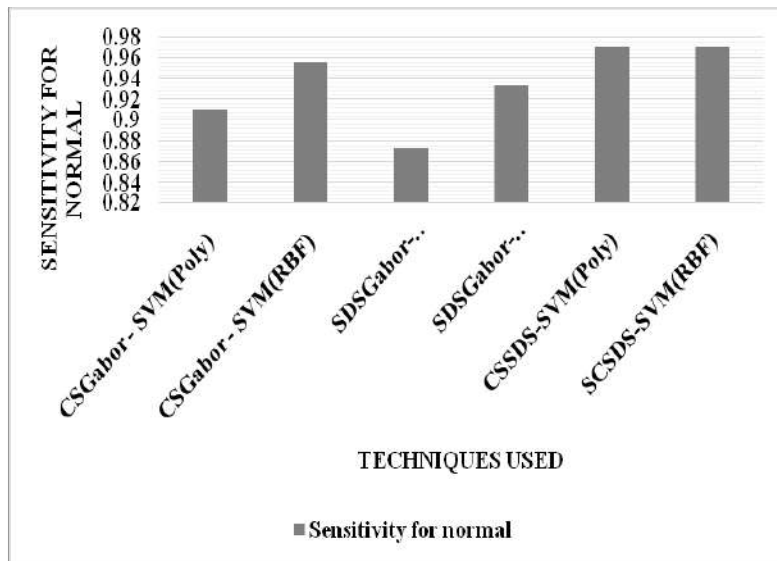


Fig: 2. Sensitivity for normal

From the [Figure- 2], it is seen that the SCSDS-SVM (RBF) technique improved sensitivities for normal by 6.25%, 1.48%, 10.39%, 3.84% & 0% when compared with various number of CSGabor – SVM (Poly), CSGabor – SVM (RBF), SDSGabor- SVM (Poly), SDSGabor- SVM (RBF) and CSSDS-SVM (Poly) methods.

Table: 3. Sensitivity for abnormal

Techniques used	Sensitivity for abnormal
CSGabor - SVM(Poly)	0.881
CSGabor - SVM(RBF)	0.8929
SDSGabor-SVM(Poly)	0.8929
SDSGabor-SVM(RBF)	0.9167
CSSDS-SVM(Poly)	0.9167
SCSDS-SVM(RBF)	0.9405

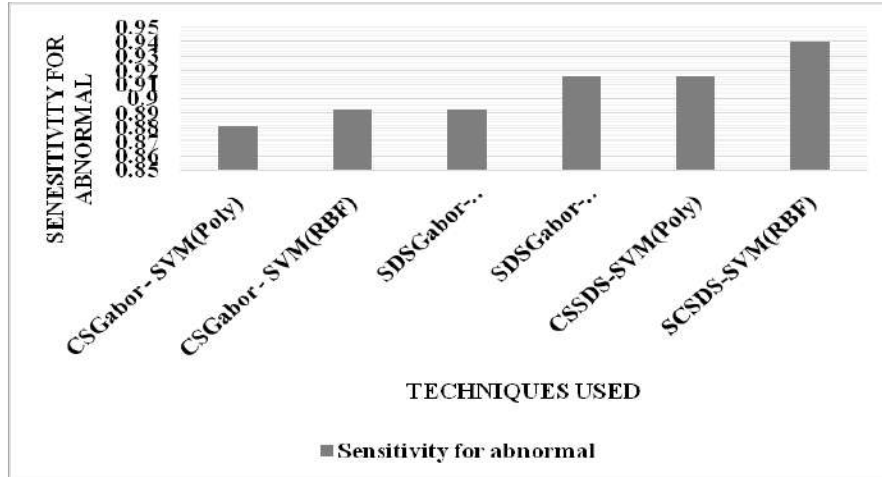


Fig: 3. Sensitivity for abnormal

From the [Figure- 3], it is seen that SCSDS-SVM (RBF) technique improved sensitivities for abnormal by 6.53%, 5.19%, 5.19%, 2.56% & 2.56% when compared with various number of CSGabor – SVM (Poly), CSGabor – SVM (RBF), SDSGabor- SVM (Poly), SDSGabor- SVM (RBF) and CSSDS-SVM (Poly) methods.

Table: 1. Specificity for normal

Techniques used	Specificity for normal
CSGabor - SVM(Poly)	0.881
CSGabor - SVM(RBF)	0.8929
SDSGabor- SVM(Poly)	0.8929
SDSGabor- SVM(RBF)	0.9167
CSSDS- SVM(Poly)	0.9167
SCSDS- SVM(RBF)	0.9405

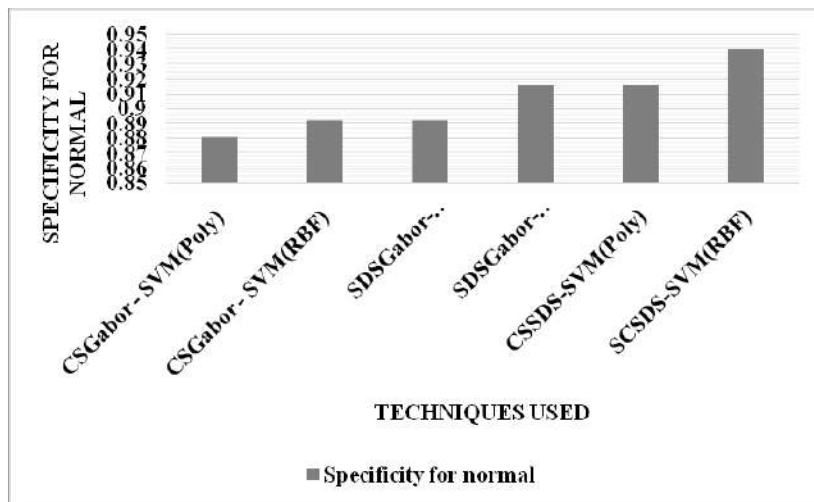


Fig:1. Specificity for normal

From the [Figure 4], it can be observed that the SCSDS-SVM (RBF) method increased Specificity for normal by 6.53%, 5.19%, 5.19%, 2.56% & 2.56% when compared with various number of CSGabor – SVM (Poly), CSGabor – SVM (RBF), SDSGabor- SVM (Poly), SDSGabor- SVM (RBF) and CSSDS-SVM (Poly) methods.

Table: 2. Specificity for abnormal

Techniques used	Specificity for abnormal
CSGabor - SVM(Poly)	0.9111
CSGabor - SVM(RBF)	0.9556
SDSGabor- SVM(Poly)	0.8741
SDSGabor- SVM(RBF)	0.9333
CSSDS- SVM(Poly)	0.9699
SCSDS- SVM(RBF)	0.9699

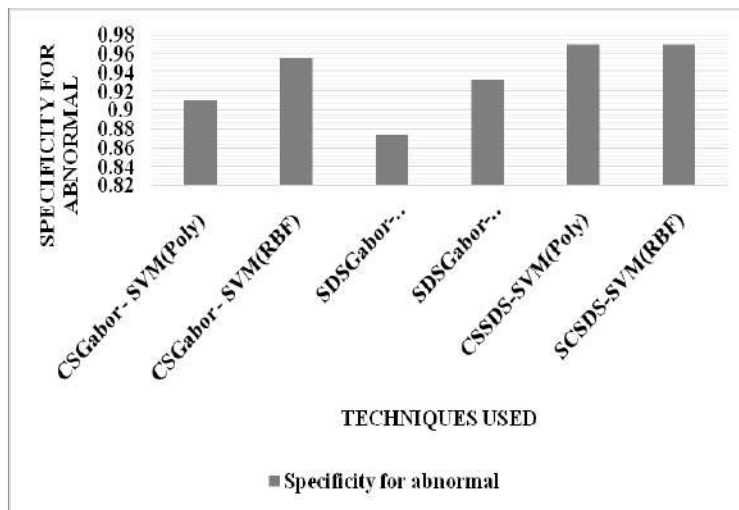


Fig:2. Specificity for abnormal

From the [Figure- 5], it can be observed that the SCSDS-SVM (RBF) method increased Specificity for abnormal by 6.25%, 1.48%, 10.39%, 3.84% & 0% when compared with various number of CSGabor – SVM (Poly), CSGabor – SVM (RBF), SDSGabor- SVM (Poly), SDSGabor- SVM (RBF) and CSSDS-SVM (Poly) methods.

### CONCLUSION

To optimize variables, several search as well as optimizing methods are utilized because it is NP-hard. The current work utilized SDS for selecting variables in SVM and for optimizing variable selections for Gabor filters. Gabor filter banks are also built through SDS with the goal of most textural attributes discriminations. Gabor filters as well as histogram extricate attributes from MRIs and the attributes are sorted through SVM RBF and SVM with suggested kernel optimizations. Outcomes reveal that the suggested method increases classification accuracies in a significant manner. The SCSDS-SVM (RBF) method increased classification accuracy by 195.78%, 195.92%, 8.39%, 3.35% & 0.96% when compared with various number of CSGabor – SVM (Poly), CSGabor – SVM

(RBF), SDSGabor- SVM (Poly), SDSGabor- SVM (RBF) and CSSDS-SVM (Poly) methods. Discovering most optimal C and variables are NP-hard.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Zhang D, Wang Y, Zhou L, Yuan H, Shen D.[ 2011] Alzheimer's Disease Neuroimaging Initiative. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3): 856-867.
- [2] Tóth L, Gosztolya G, Vincze V, Hoffmann I, Szatlóczki G. [2015] Automatic detection of mild cognitive impairment from spontaneous speech using ASR. *ISCA*.
- [3] Duchesne S, Caroli A, Geroldi C, Collins DL, Frisoni GB. [2009] Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *NeuroImage*, 47(4): 1363-1370.
- [4] Guyon I, Elisseeff A. [2003] An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3: 1157-1182.
- [5] Roman G, Pascual B. [2012] Contribution of neuroimaging to the diagnosis of Alzheimer's disease and vascular dementia. *Archives of medical research*, 43(8): 671-676.
- [6] Zhou X, Wang ., Xu W, Ji G, Phillips P, Sun P, Zhang Y. [2015] Detection of pathological brain in MRI scanning based on wavelet-entropy and naive bayes classifier. *In Bioinformatics and biomedical engineering* (pp. 201-209)
- [7] Zhuang L, Wen W, Zhu W, Trollor J, Kochan N, Crawford J, Sachdev P. [2010] White matter integrity in mild cognitive impairment: a tract-based spatial statistics study. *Neuroimage*, 53(1):16-25.
- [8] Liu F, Wee CY, Chen H, Shen D. [2014]. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification. *NeuroImage*, 84: 466-475.
- [9] Zhang Y, Wang S, Ji G, Dong Z. [2013]An MR brain images classifier system via particle swarm optimization and kernel support vector machine. *The Scientific World Journal*.
- [10] Kekre HB, Bharadi VA. [2010]Gabor filter based feature vector for dynamic signature recognition. *International Journal of Computer Applications*,2(3): 74-80.
- [11] Derpanis KG. [2007] Gabor filter. York University, April, 23
- [12] Nasuto S, Bishop M.[1999]Convergence analysis of stochastic diffusion search. *PARALLEL ALGORITHMS AND APPLICATION*, 14(2): 89-107. Chicago
- [13] Myatt DR, Bishop JM, Nasuto SJ. [2004]Minimum stable convergence criteria for stochastic diffusion search. *ELECTRONICS LETTERS-IEE*, 40(2):112-112.
- [14] DURGESH K.S, Lekha B. [2010] Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*, 12(1): 1-7.
- [15] Renukadevi N, Thangaraj P. [2013] Optimizing of SVM for CT Classification. *Journal of Theoretical & Applied Information Technology*, 55(2):203-208.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and checked/reviewed by guest editor.

# CLOUD WORKFLOW SCHEDULING ALGORITHMS USING CUCKOO SEARCH (CS) WITH NOVEL FITNESS FUNCTION

Sivakumar<sup>1\*</sup>, Chitra<sup>2</sup>, Madhusudhanan<sup>3</sup>

<sup>1</sup>Dept. of Computer Science and Engineering, Mookambigai College of Engineering, Kalamavur, TN, INDIA

<sup>2,3</sup>Dept. of Computer Science and Engineering, Er. PerumalManimekalai College of Engineering, Hosur, TN, INDIA

## ABSTRACT

Cloud Computing is an emerging technique. Certain parallel applications show a reduction in utilization of CPU resources when there is a rise in parallelism if jobs are not schedules correctly then it decreases the computer performance. Task scheduling is a valuable tool which influences performance of cloud service providers to a great extent. Conventional approach that is used in optimizations is deterministic, fast, and gives perfect answers but frequently bog down in local optimum. In this paper, Cuckoo Search (CS) is proposed for the optimization. Experiments conducted and the results revealed that the proposed method outperformed other methods.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Cloud Computing, Scheduling, Cuckoo Search (CS).

\*Corresponding author: Email: [sivakumarm.clo@gmail.com](mailto:sivakumarm.clo@gmail.com)

## INTRODUCTION

Cloud computing offers resources to users as per their demands. The users demand for available services as per their desired Quality of Service, and pay for them on a pay per use basis. A challenging problem in Cloud computing is workflow scheduling. The processing units in cloud environment are known as virtual machines. It is to make sure that tasks are not loaded heavily on a single VM and other VMs do not remain idle and/or under loaded [1].

There is great need for the cloud services and their scheduling. Scheduling will be followed by the task or job scheduling within the resources. There might be more instances of a single resource which can be run at the same instant. It is necessary to check availability and reliability while also load ought to be equalized amongst the resources of the same kind. For the above variables there needs to be a function or procedure which could check them and allotment should be done in an optimum way.

The best way is combine together the computability of network strategies with scheduling algorithms [2]. Typically when tasks are scheduled they are done as per user's requirements as well as requests but when looking into all the features the computation requires to be done. Application scalability is the primary aim for cloud services to attain. In cloud scalability of resources permits real time provisioning of resources. Cloud has complicated execution environment however it has to offer QoS to its users. Virtual resources are utilized effectively for the entirely customizable configuration environment for application.

Workflows enable arranging apps in a directed acyclic graph form, wherein every node denotes the constituent job and edges denote inter job dependencies of apps. One workflow comprises a set of jobs that may interact with other jobs in the workflow. Hence, workflow scheduling is important in managing workflow implementation.

For taking cloud computing, scientific workflow gains more utilizations. But, we face several new challenges, wherein data as well as task scheduling are one. How to effectively schedule all tasks in an application is the most crucial problem. In order to reduce the executing time, schedule the computing intensive tasks to the high performance computer [3]. As the task scheduling is an NP-complete problem, certain heuristic algorithms have been utilized to resolve it.

For solving the NP complete as well as NP hard problems heuristic methods can be used. The several algorithms that may be utilized for the scheduling are evolutionary protocols which are based on the biological evolution of species. The evolutionary protocols are Particle Swarm Optimization as well as Genetic algorithm [4].

Optimization means finding the best solution for a given problem. The field of optimization algorithms studies algorithms derived from the observations and these algorithms are a source of inspiration for designing novel protocols for solving optimization as well as distributed control issues. Conventional methods need a lot of computational efforts which tend to fail when problem size rises [5].

Usage of bio-inspired algorithm is a motivational method for employing computationally effective alternative systems for deterministic method. Swarm intelligence denotes the group intelligence of social insects because of their efficacy in resolving complicated issues like discovering shortest route from nests to food sources or the organization of their nest. Despite the fact that the insects are not sophisticated as an individual unit, as a swarm through interactions with one another and the environment make them collectively intelligent.

In recent years, this activity of swarms is abstracted as numerical optimization techniques [6]. This collective intelligence arises out of a process of self-organizing of units that evolve automatically as per a set of rules denoting movement patterns as well as interactions with environment and other agents so that intelligent activity rises out of simplistic individual activity.

One of such nature inspired algorithms is Cuckoo Search (CS) algorithm. It owes its inspiration to the obligate brood parasitism of certain cuckoo species which lay their eggs in the nests of host birds [7]. CS is a recent meta-heuristic algorithm and is utilized in solving complex optimization problems. The optimum solutions acquired by CS are far better than those obtained by other swarm intelligence algorithms.

In the current paper we use CS algorithm to optimize the scheduling process. Remaining sections are formed as: Section 2 discusses the related work in literature. Section 3 describes the method used in the proposed work. Section 4 reveals the experiment results and Section 5 concludes the proposed work.

## RELATED WORKS

Nandhakumar and Ranjithprabhu [8] compared and analyzed the performance of different heuristic workflow scheduling algorithms with various QoS parameters and scheduling factors.

Saxena and Saxena [9] proposed a new scheduling algorithm for workflow, which are having dependencies among tasks, taking into consideration important parameters of transfer time and bandwidth along with basic requirements of optimizing the execution time and cost. The simulation was experimented using cloudSim toolkit. The proposed algorithm provided better results over other existing algorithm like PSO and CSO and is more closely related to real world scenario.

Wang et al., [10] proposed a competitive, dynamic and multiple DAG scheduling algorithm which takes link communication processor into consideration (CCRH). The algorithm used communication competition model to describe the communication between the processors. Simulation results showed that under the premise of ensuring reliable scheduling, the algorithm not only can improve the fairness of the multiple DAGs scheduling, but also effectively short the average multiple DAGs scheduling time, and make the robustness of algorithm is better.

Verma and Kaushal [11] proposed Bi-Criteria Priority based Particle Swarm Optimization (BPSO) for scheduling workflow tasks across the present cloud resources which minimized the execution costs as well as the execution time under the specified deadline as well as budget restrictions. The suggested algorithm was tested using simulations with four separate real world workflow applications and they are compared with Budget Constrained Heterogeneous Earliest Finish Time (BHEFT) as well as standard PSO. The simulation outcomes showed that the proposed scheduling algorithm considerably decreased the execution cost of schedule when compared with BHEFT as well as PSO with Deadline, Budget Constraints as well as using same pricing model.

Thanh et al., [12] proposed a metaheuristic algorithm called PSO<sub>i</sub> which based on the Particle Swarm Optimization method. The experiments that were arranged by utilizing simulation tool CloudSim reveal that PSO<sub>i</sub>



is better than the generic Random as well as RoundRobin, furthermore the deviation between the solutions found by PSOi as well as the optimal solution is negligible.

Tang et al., [13] proposed a DVFS-enabled efficient energy workflow task scheduling algorithm: DEWTS. Through merging the relatively inefficient processors by reclaiming the slack time, DEWTS may exploit the useful slack time once more employing DVFS method after servers combined. On the basis of the amount of arbitrarily generated DAGs workflows, experimental outcomes showed that DEWTS is capable of reducing the total power utilization with various parallel applications and balancing the scheduling performance.

Bittencourt et al., [14] analyzed the problems in the scheduling the workflows in the hybrid clouds and surveyed few scheduling algorithms used for the cloud systems. To impact of the communication channels on the allocation of the jobs was compared and evaluated for various scheduling algorithms.

Lu and Gu [15] proposed a load-adaptive cloud resource scheduling model based on ant colony algorithm. By analyzed an example result, the model could meet the goals and requirements of self-adaptive cloud resources scheduling and improved the efficiency of the resource utilization.

A scheduling algorithm based on Genetic Algorithm (GA) was proposed by Wei and Tian [16]. Scheduling scheme is encoded in integer sequence, as well as a fitness function on the basis of influence degree is formulated. The genetic operations are selection, crossover, mutations as well as elitist selection. An optimal method was proposed according to the practical application. Finally resources scheduling problem in a cloud design platform proved the validity of the scheduling algorithm and the effectiveness of the optimization method.

Ge and Wei [17] presented a scheduling algorithm to make a scheduling decision by comparing and judging about the entire group of tasks in the job queue. For the optimization of the parameters of the scheduler, a GA was designed. Simulations were conducted and the results were proved that the proposed scheduling policy balanced the load among the nodes in better than the First in First Out (FIFO) and delay scheduling.

Raghavan et al., [18] utilized a metaheuristic method known as bat algorithm. It is specifically formulated for optimizing hard problems. Bat algorithm with the help of binary bat algorithm was utilized for scheduling workflow in a cloud. Particularly, the mapping of tasks and resources is performed through this method. The optimum resources were selected such that the overall cost of the workflow is minimal.

## METHODOLOGY

Cuckoo Optimization Algorithm (COA) is improved to select the resources and schedule the tasks to minimize the overall completion time of tasks optimally by combining random local search and basic CS. Mapping of the real world problem to the meta-heuristic algorithm is very important and explained in detail. For jobs to be executed in cloud resources it compute the time taken by to compute and so on. A time to execute matrix is created as shown in [Table -1].

**Table: 1. Mapping of resource to jobs**

	CR1	CR2	CR3
j1	3.54	0.27	2.1
j2	5.28	1.12	9.07
j3	0.81	0.83	3.05
j4	8.12	5.33	5.22
j5	1.06	9.19	2.76
j6	0.29	0.18	1.79
j7	2.27	7.93	9.98
j8	4.92	8.9	3.14
j9	3.47	2.3	2.18
j10	4.39	1.58	8.04

j11	0.3	0.77	8.85
j12	8.35	4.71	3.03
j13	5.98	8.69	3.57
j14	8.35	4.77	8.35
j15	4.77	9.55	1.5

**Generating initial cuckoo habitat:** To resolve optimizing issues, it is required for the values of the issue variables be grouped as arrays. In GA as well as PSO terms, the array is known as 'Chromosome' and 'Particle Position' respectively. But here, in COA is known as 'habitat'. To begin the optimization algorithm, candidate habitat matrices are created. Few, arbitrarily produced quantity of eggs are assumed for every original cuckoo habitat. In the real world, cuckoos lay around five to twenty eggs [19]. These numbers are utilized as upper and lower limits of  $eff$  designated to every cuckoo at various iterations. Another habit of cuckoos is that they lay an egg at the farthest distance from their own habitat. This is known as 'egg laying radius' (ELR). Every cuckoo possesses an ELR that is suitable for the overall quantity of eggs, number of cuckoo eggs and also differing limits of  $var_{hi}$  and  $var_{low}$ . So ELR is given by:

$$ELR = \alpha \times \frac{\text{Number of current cuckoo's eggs}}{\text{Total number of eggs}} \times (Var_{hi} - Var_{low})$$

Which  $\alpha$  is an integer, supposed to handle the maximum value of ELR.

**Immigration of cuckoos:** Once young cuckoos grow old, they fly to live in their own region and when the season for laying eggs rolls around, they shift to fresh habitats with most similar host eggs and more food for the young birds. Then the cuckoo groups are created in several regions, the society with greatest fitness value is chosen as target point and all cuckoos move toward it. When mature cuckoos that live in those environments identify cuckoos belonging to other groups is a tough task. Most benefit is defined by target group and subsequently the group's most optimal habitat is the new destination habitat for moving cuckoos. When moving to the target point, cuckoos do not fly directly straight to destination. They cross partial distance and deviate. Pseudo code of Cuckoo Optimization Algorithm

1. Initialize cuckoo habitats with random points
2. Define ELR for each cuckoo
3. Let cuckoo to lay eggs inside their corresponding ELR
4. Kill those eggs that are identified by host birds
5. Eggs hatch and chicks grow
6. Evaluate the habitat of each newly grown cuckoo
7. Limit cuckoos maximum number in environment and kill those that live in worst habitats
8. Cuckoos find best group and select goal habitat
9. Let new cuckoo population move toward goal habitat
10. If stop condition is satisfied end, if not go to 2

### CUCKOO SEARCH (CS)

CS optimization algorithm is one of evolutionary algorithms and it was introduced by Yang and Deb in the year in 2009 [20]. The lifestyle and behavior of a bird called the Cuckoo was inspired by the developers of this algorithm. The brooding nature of this bird is different from the other birds. Cuckoo bird does not use its nest for laying the eggs and use other bird's nest for laying eggs. If the host bird finds that the eggs are not belongs to other bird, it will throw away or leave the nest. The grown cuckoo bird becomes a mature bird, and then it continues the mother's life instinctively [21].

**Cuckoo Behavior:** Certain cuckoo species have evolved such that female parasitic cuckoos are typically specialized in mimicry in colour as well as pattern of the eggs of a few particular host species. This decreases the probability of eggs being discarded and increases their reproductively. Cuckoos often select nests wherein host birds have just laid their own eggs. Generally, cuckoo eggs hatch a little earlier than the host eggs. When first cuckoo chick hatches, the instinctive action will evicting the host eggs by blindly pushing the eggs out of the nest, thereby increasing the cuckoo chick's share of food given by the host bird. Cuckoo characteristics could be described, as a model for good behavior other animals have extensive use in computing Intelligence Systems.

**Levy Flights:** The activity of animals to scour for food is quasi-random in practice. In recent research, it has been proven that flight activity of several creatures demonstrate generic features of Levy flights. Typically, foraging routes of creatures is technically an arbitrary walk as the next move is on the basis of current position and transition probability to the next locale [22]. Selecting the direction relies on a certain probability that may be abstracted mathematically. Several researches have employed these activities in optimizations, optimal searches and initial outcomes reveal its promise.

Levy flight is the most popular technique used and handles [23]:

- The generation of how each step should be
- The random direction of flight which is given by equation:

$$L = \frac{u}{|v|^{1/\beta}}$$

Where  $\beta$  is the scaling value with a range of [1, 2].  $u$  and  $v$  are generated from normal distribution and shown in equation:

$$u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$$

Where  $\sigma_u$  and  $\sigma_v$  are calculated using equation:

$$\sigma_u = \left\{ \frac{\Gamma(1+\beta) \sin(\pi\beta/2)}{\Gamma[(1+\beta)/2] \beta 2^{(\beta-1)/2}} \right\}^{1/\beta}, \sigma_v = 1$$

Where  $\Gamma$  is the gamma function.

**Cuckoo Rules and Parameters:** To simplify the principles of the new CS algorithm, three exemplary rules can be used.

- Every cuckoo lays a single egg at a time, and deposits it in an arbitrarily chosen nest,
- The best nests with excellent quality of eggs (solutions) will be carried over to the next generation.
- The quantity of available host nests is static, and hosts can discover alien eggs with a probability of  $pa \in [0, 1]$ . In such a case, host birds can either discard the egg or abandon nest and build an entirely new nest in a fresh location.

In this overall, minimizing the overall task completion time is taken as the fitness function. The parameters updation and usage are given by:

**Step 1:** Assign the nests randomly. This indicates selection of random solution

**Step 2:** Select one random nest and replace it by a best solution. Best solution is found by a levy flight operation.

When generating new solutions,  $x_i^{(t+1)}$  for the  $i$  th Cuckoo, a Lévy flight is performed using the equation:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \cdot S$$

Here  $\alpha > 0$ , the parameter  $\alpha$  is used as the step size parameter. It must be selected based on the problem scale. Most of the times it is set to unity in the CS [24] and reduced in the improved CS algorithm. The solutions or nests in the current positions are then used for finding best solution so far as the origin of the Lévy flight. The step size also increases the efficiency and performance of the CS algorithm. The parameter value  $S$  represents the length of random walk with the Lévy flights.

**Step 3:** In the next step, the fraction value  $pa$  is used to discover the worst nest, so that they can be and replaced by the best ones. This parameter  $pa$  is considered as the probability of a solution's that will be discovered. Therefore, a probability matrix is created by the equation:

$$P_{ij} = \begin{cases} 1 & \text{if } rand < pa \\ 0 & \text{if } rand \geq pa \end{cases}$$

Where  $pa$  represents the discovering probability and maximum number of analyses is the stopping criterion.  $rand$  is an arbitrary number within the range [0, 1] and  $P_{ij}$  is the probability of discovering  $j$ th variable of  $i$ th nest. It assigns either 0 or 1 to each variable on the nest. Using random walk, the point wise multiplication of random step sizes with probability matrix from their current positions according to quality.

Pseudo code for The CS algorithm

*Input :*  
*function to optimize, fit*  
*Population of n host nests  $x_i = i(1, 2, \dots, n)$*   
*Output :*  
*best solutions (nests with quality solutions),*  
*Initialize :*  
*While ( $t < \text{MaxGeneration}$ )*  
*Get a cuckoo randomly by Levy flights,*  
*Evaluate :*  
*fit*  
*Randomly choose nest among n available nests*  
*If ( $\text{fit}_i < \text{fit}_j$ )*  
*Re place j by the new solutions;*  
*End if*  
*bandona fraction (pa) of worse nests and build*  
*Re peat*  
*new nest*  
*New locations via Levy flights;*  
*Keep the best solutions;*  
*Rank the solutions and find the current best;*  
*end while*  
*Post process results and visualization;*  
*END*

**Fitness Value:** Fitness or quality value reveals how fit a solution is, i.e. how well it can adapt to its environment. For maximization problems, the fitness of solutions are proportional to values of objective functions. For simplicity, suppose every egg in a nest denotes a solution, and cuckoo egg denotes a new solution. The aim is to utilize the new and potentially improved solutions (cuckoos) to substitute a not-so-good solution in the nests. Here, it use the simplest approach where each nest has only a single egg [22].

## EXPERIMENTAL SETUP

Thirty tasks are assigned to Cloud with 5 resources and 10 resources. CloudSim simulator is used for conducting the experiments. The resources are located at two data centers. Each resource has 1 CPU with 512 Mb RAM. Each task is of size 1, 2, 3 or 4 units. The simulations were conducted using random local search and CS, the overall task completion time or makespan is used for comparing the performance.

## RESULTS

The overall task completion time is shown in [Table -2].

Table: 2. Overall completion Time

Technique used	10 resource	5 resource
Random Local Search	7.96 second	15.36 second

Cuckoo Search	7.42 second	15.18 second
---------------	-------------	--------------

From the numerical results, it is noted that the overall task completion time of the proposed CS optimization drastically reduces the overall completion time.

## CONCLUSION

Task scheduling problem concerns about the dynamic distribution of the tasks over the Cloud resources to achieve the best results. In the current paper, a task scheduling algorithm has been suggested to the independent task over Cloud Computing. The suggested algorithm is the CS algorithm. CS algorithm is based on the obligate brood parasitic behavior of some cuckoo species in combination with the Levy flight behavior of some birds and fruit flies. In the proposed CS algorithm, all the nests are ranked then a random local search is initiated with the average value of the top three nests (solutions). The best solution obtained by CS and the best solution obtained by Random Local Search are sent to the next iteration. For the simulation 30 tasks are taken with number of resources as 5 and 10.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Lovesum SJ, Krishnamoorthy K, Prince P. [2014] An optimized qos based cost effective resource scheduling in cloud. *Journal of Theoretical and Applied Information Technology*, 66(1).
- [2] Manimaran V, Prabhu S. [2010] A Survey on Resource Scheduling and Allocation Policy in a Cloud Environment.
- [3] Guo L, Zhao S, Shen S, Jiang C. [2012] Task scheduling optimization in cloud computing based on heuristic algorithm. *Journal of Networks*, 7(3): 547-553.
- [4] Jacob L, Jeyakrishnan V, Sengottuvelan P. [2014] Resource Scheduling in Cloud using Bacterial Foraging Optimization Algorithm. *International Journal of Computer Applications*, 92(1).
- [5] Singh G, Kaur A. [2015] Bio Inspired Algorithms: An Efficient Approach for Resource Scheduling in Cloud Computing. *International Journal of Computer Applications*, 116(10).
- [6] Milani FS, Navin AH. [2015] Multi-Objective Task Scheduling in the Cloud Computing based on the Patrice Swarm Optimization.
- [7] Soneji H, Sanghvi, RC.[2012, October] Towards the improvement of cuckoo search algorithm. In *Information and Communication Technologies (WICT), IEEE* 878-883.
- [8] Nandhakumar C, Ranjithprabhu K. [2015] Heuristic and meta-heuristic workflow scheduling algorithms in multi-cloud environments—A survey. In *Advanced Computing and Communication Systems*, 2015 International Conference on (pp. 1-5). *IEEE*.
- [9] Saxena S, Saxena D.[2015] EWSA: An enriched workflow scheduling algorithm in cloud computing. In *Computing, Communication and Security (ICCCS)*, 2015 International Conference on (pp. 1-5). *IEEE*.
- [10] Wang Y, Jia C, Xu Y. (2014) Multiple DAGs dynamic workflow scheduling based on the primary backup algorithm in cloud computing system. In *Broadband and Wireless Computing, Communication and Applications (BWCCA)*, 2014 Ninth International Conference on (pp. 177-182). *IEEE*.
- [11] Verma A, Kaushal S. [2014] Bi-criteria priority based particle swarm optimization workflow scheduling algorithm for cloud. In *Engineering and Computational Sciences (RAECS)*, 2014 Recent Advances in (pp. 1-6). *IEEE*.
- [12] Thanh TP, LN, Doan CN.[2015] A novel workflow scheduling algorithm in cloud environment. In *Information and Computer Science (NICS)*, 2015 2nd National Foundation for Science and Technology Development Conference on (pp. 125-129). *IEEE*.
- [13] Tang Z, Cheng Z, Li K, Li K. [2014]An Efficient Energy Scheduling Algorithm for Workflow Tasks in Hybrids and DVFS-enabled Cloud Environment. In *Parallel Architectures, Algorithms and Programming (PAAP)*, 2014 Sixth International Symposium on (pp. 255-261). *IEEE*.
- [14] Bittencourt LF, Madeira ER, Da Fonseca, NL. [2012] Scheduling in hybrid clouds. *Communications Magazine, IEEE*, 50(9):42-47.

- [15] Lu X., Gu Z. [2011] A load-adaptive cloud resource scheduling model based on ant colony algorithm. In Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on (pp. 296-300). IEEE.
- [16] Wei Y, Tian L. [2012] Research on cloud design resources scheduling based on genetic algorithm. In Systems and Informatics (ICSAI), 2012 International Conference on (pp. 2651-2656). IEEE.
- [17] Ge Y, Wei G. [2010] GA-based task scheduler for the cloud computing systems. In Web Information Systems and Mining (WISM), 2010 International Conference on , 2: 181-186). IEEE.
- [18] Raghavan S, Marimuthu C, Sarwesh P, Chandrasekaran K. [2015] Bat algorithm for scheduling workflow applications in cloud. In Electronic Design, Computer Networks and Automated Verification (EDCAV), 2015 International Conference on (pp. 139-144). IEEE.
- [19] Rabiee M, Sajedi H. [2013] Job scheduling in grid computing with cuckoo optimization algorithm. *International Journal of Computer Applications*, 62(16): 38-44.
- [20] Yang XS, Deb S. [2009] Engineering optimization by cuckoo search. *International Journal of Mathematical Modelling and Numerical Optimization*, 1: 330-343.
- [21] Al-maamari A, Omara FA. [2015] Task Scheduling using Hybrid Algorithm in Cloud Computing Environments. *Journal of Computer Engineering (IOSR-JCE)*, 17: 96-106
- [22] Navimipour NJ, Milani FS. [2015] Task scheduling in the cloud computing based on the cuckoo search algorithm. *International Journal of Modeling and Optimization*, 5(1): 44.
- [23] Baskan, O. [2013] Determining optimal link capacity expansions in road networks using Cuckoo Search algorithm with Lévy Flights. *Journal of Applied Mathematics*, 2013.
- [24] Yang XS, Deb S. [2009] Cuckoo search via Lévy flights. In Nature and Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on (pp. 210-214). IEEE.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and checked/reviewed by guest editor.

# PERFORMANCE EVALUATION OF DSR ROUTING PROTOCOL UNDER DIVERGING AND CONVERGING NODES

Umapathi<sup>1\*</sup> and Ramaraj<sup>2</sup>

<sup>1</sup>Dept. of Electronics and Communication, GKM college of Engg and Tech, Chennai, INDIA

<sup>2</sup>Dept. of Computer Science Engineering, Thangavelu Engg. College, Chennai, INDIA

## ABSTRACT

Structure less wireless mobile networks or Ad hoc networks under the MANET scheme are created on a limited temporary basis. But its importance can never be over-emphasized as it has found applications in so many fields of human endeavour. Thus, it is important to consider effective routing procedures which assists the proper functioning and deployment of MANET. In this paper, the overall performance of the Dynamic Source Routing (DSR) protocol, which is basically on-demand, under diverging and converging nodes is investigated. Detailed simulations were carried out, using OPNET Modeller.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

MANET, Routing Protocols,  
Dynamic Source Routing (DSR),  
Performance analysis.

## INTRODUCTION

MANET (Mobile Ad hoc Networks) are unstructured wireless networks which are temporarily used in timely projects [1]. A cost effective project management technique suitable for technologically arid regions, wireless networks of this category has been researched for a few decades, and after dedicated efforts put into research, a variable format of MANET was formed comprising of several levels of applications. MANETs are well suited in a situation in which the deployment of an infrastructure is not feasible and cost effective. Over last two decades, MANET became a very interesting research area. Many institutes and corporations have sponsored MANET. It is commonly used in daily communication technologies, such as, conferencing and also a host of emergency services; local home networks; embedded computing applications; and personal networks, etc. [2].

A major problem concerning MANET networks are proper routing methods, because neither does it have specialized infrastructure to counter mobile network needs nor does it have an assurance of stable positions. Thus, the goals of MANET is to enable tackling mobility issues through their limited resource capacity, heterogeneity, etc. [3]. This technique enables multiple alternatives to resolve the above challenges. So in interest among researchers is to supply suitable ad hoc routing networks for assisting academic and industrial spheres.

Several routing protocols have been designed for multi-hop ad hoc networks. The routing protocols obtained here needs a range of design choices and approaches, from simple modifications of internet protocols, to more complex multi-level hierarchical schemes. Although the ultimate end goal of a protocol may be operation in large networks, most protocols are typically designed for moderately sized networks of 10 to 100 nodes [4].

The Dynamic Source Routing protocol (DSR) [5], an basic and well-functioning protocol designed for wireless multiple hop ad hoc networks utilizes DSR to self-configure and process data, without the use of existent administrative structures. Communication is processed through several “hops” transmitted between each other, but not within the direct range of existent networks. Since routing networks are automatically formatted and maintained through routine DSR checkups, nodes are often made to join or exit wireless transmission networks to prevent any

source of interference. But the above process can be rapidly changing and due to this issue it becomes difficult to estimate the number of intermediary hops.

DSR sequential protocols depend on two main techniques which function together to propagate encounters and procedural maintenance of ad hoc source network routes in:

A method technique under Route Discovery, uses package systems to transfer between nodes S to its destination to gain a proper route. It is only used to send packages to the destination without the need to know a stable route.

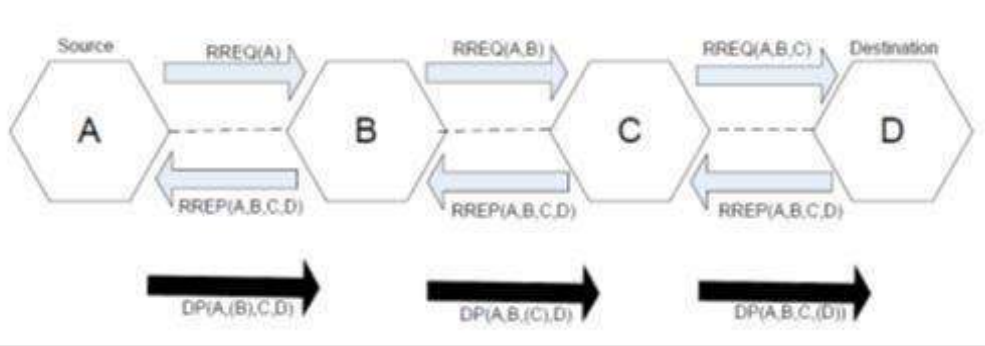


Fig: 1.Route Discovery in DSR

This is a detectable mechanism used to source linkages to D, if the topology of the networks have been modified and is no longer capable of of creating functional connections. Broken links are usually replaced by other routes to D, or the nodules of S attempt to create new links to contacted destinations D. This method is only used for transferring packages between S and D.

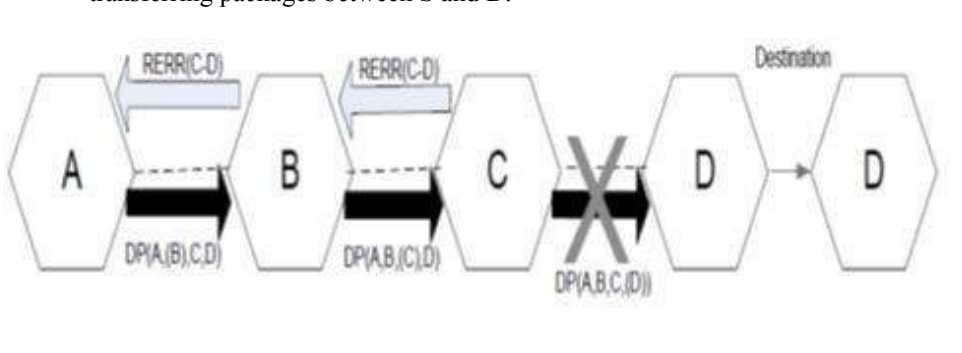


Fig: 2. Route Maintenance in DSR

A demand based system, DSR technology completely operate on Route Discovery and Route Maintenance. Unlike other protocols, DSR does not need any form of periodic packages within any layer of the network, such as, the lack of necessity to depend on periodic routing for advertizing, status link sensing or detecting neighbor packages, without relying on unwritten protocols of the system.

ROUTE ERRORS in DSRs occur when there is a disconnection in linkages, and the package is sent back to its source through another route discovery operation. Adding to this procedure, every broken link is removed through its intermediate cache nodules transmitted to the source. Thus, the increase in overhead traffic is the aftermath of complete routing procedures and knowledge acquisition among DSR systems. The size of the network is taken to be 5 to 10 nodes diametrically, even though the network might be small, it is applicable only to a relatively minute number of nodes, usually less than 100 nodes. This is done to maintain equilibrium in the system and not to cause problems in the system later.

In this paper, the overall performance of the Dynamic Source Routing (DSR) protocol, which is basically on-demand, under diverging and converging nodes is investigated. Detailed simulations were carried out, using



OPNET Modeller. The paper is organized as follows: this section introduces MANET, routing protocols and explains in the routing protocol DSR.

## RELATED WORKS

Functional ad hoc networks or MANET, researched through works by Valentina Timcenko et al., [6] proposed routing protocols under group mobilities and entity measuring models. The most commonly used and researched routing systems are: Destination Sequence Distance Vector (DSDV), Ad-hoc On-demand Distance Vector (AODV), and Dynamic Sourcing Route (DSR). These models include mobility structures like the Reference Point Group Mobility, Gauss-Markov (GM) and Manhattan Grid (MG) models.

Simulator models are usually merged with Bonn Motion scenario tools to process any simulation under the Network Simulator Version 2 (NS2). Successful results of the above simulator process have assisted in creating a set of specific simulated scripts which are applicable for usage on a wider range of MANET scenarios. The procured results indicate a relative ranking among protocols, which vary based on the level of mobility. Depending on nodular speeds, mobility presence can indicate failed linkages, which reacts depending on routing levels. Entity models enhance performances due to its low level of GM and MG randomness, under which the AODV model performs better with RPGM group models. Although some models experience stipulated delays, AODV models experience high levels of overhead routing speed, whereas DSR has a lower overhead speed with higher standard delays, especially under a MG and GM model. The above method performs the best with RW models. Hence, researchers need to consider energy efficient consumption patterns by allocating varied propagation and MAC models in the future.

Mobile ad hoc networks, based on “random walk” unrestricted mobile simulations, recreate an unrealistic vision of a world in which individuals try to surpass walls, no need for cars on the road and people drive on waterbodies. A new graph model introduced by Jing Tian et al., [7], provides better movement which are realistic unlike random walk models. This graph model portrays real-world spatial constraints. DSDV, DSR and AODV, the most common techniques use both a random walk-based and the proposed graph-based mobility model are analyzed. The simulation results show that the spatial constraints have a strong impact on the performance of ad hoc routing protocols. A graph from external spatial data is extracted to represent the realistic movement constraints of pedestrians walking in the city. As the result showed, routing protocols performed quite differently in this graph walk model from the random walk model. Moreover, comprehensive simulations are made with short radio ranges considering the energy constraint of handheld devices.

The last years has seen a surge in the popularity of Mobile Ad hoc Networks (MANETs), among researchers, especially in military and civilian applications due to its rapid deployment abilities. Due to its capacity to function albeit proper infrastructure and durability to sustain itself through rapid network changes. MANET systems are mainly evaluated through simulations, although there have been instances where variable graphs have been employed to formally research it. Anuj K. Gupta et al., [8] attempts to make a comprehensive performance evaluation of three commonly used mobile ad hoc routing protocols (DSR, TORA and AODV) which contain identical capabilities and working conditions with identical loads and conditions in environment which help in evaluating its relative performance with respect to the other two performance metrics: average End-to-End delay and packet delivery ratio. Over the past few years, new standards have been introduced to enhance the capabilities of ad hoc routing protocols. The latest simulation environment NS 2 is used to evaluate the protocols using packet-level simulation. Various simulation scenarios with varying pause times were investigated. From the detailed simulation results and analysis, a suitable routing protocol can be chosen for a specified network and goal. The experimental results obtained can be concluded as follows:

- Increase in the density of nodes yields to an increase in the mean End-to-End time deficiencies.
- The ability to delay means in end-to-end delays through adding more pauses in time
- The mean time loop detections are recognized through the spiraling in the number of nodes.

The above pointers indicate the steady performance of AODV. Unlike TORA systems which are suitable for moderately steady and mobile networks, DSR systems suitable for low bandwidth power usages is also suitable for moderate mobility rates. The major benefit of TORA systems is its excellent support for multiple routes and multicasting.

A collection of wireless mobile nodes or an Ad Hoc Network can dynamically form temporary networks without using any older network infrastructure or forms of centralized administration. There are a number of routing protocols like Dynamic Source Routings (DSR), Ad Hoc On-Demand Distance Vector Routings (AODV) and Destination-Sequenced Distance-Vectors (DSDV) which can be implemented. Samyak Shah et al., [9] attempted to compare the performance of two prominent on-demand reactive routing protocols for mobile ad hoc networks: DSR and AODV using ns-2 simulations, alongside traditionally proactive DSDV protocols. Thus, simulation models using MACs and physical layer models are used to investigate interlayer interactions and possible implications in performance. It is to be noted that on-demand protocols like AODV and DSR can perform better than any table-driven DSDV protocols. Eventhough DSR and AODV do share mutual on-demand behavioral patterns, differences in the protocol mechanics leads to significantly different level of performances. There is a variety of workload and scenarios patterns which are characterized by mobility, load and size of an ad hoc network. Performance differentials are then analyzed via differential network load, mobility, and network size. And the above simulations are carried out through the Rice Monarch Project which has managed to procure substantial extensions to the ns -2 network simulator to run ad hoc simulations. General observations from the above simulations show that for application-oriented metrics like packet delivery fractions and AODV delays can outperforms DSR in intensely “stressful” situations (i.e., smaller number of nodes and lower load and/or mobility), with wider gaps in performance gaps through each increasing stress level (e.g., more load, the higher the mobility rate). However, DSR, can consistently produce less routing load than standard AODV levels. Poorly performing DSRs can mainly be attributed to a growing role of aggressively using caching techniques, and lack of proper mechanisms to shut down non-functional routes or even predict the age of given routes especially when there exists multiple choices. But, Aggressive caching seems to assist DSR in low load situations and it also keeps the load levels in routing down.

Md. Anisur Rahman et al., [10] made a comprehensive attempt to study and compare performancea ofprominently existent on-demand routing protocols especially for mobile ad hoc networks like DSRs and AODVs, alongside traditional proactive methods like the protocols for DSDV. Simulation models with MAC and physical layer models have been used to study interlayer interactions and their performance implications. The On-demand protocols, AODV and DSR perform better than the table-driven DSDV protocol. Even though both DSR and AODV seem to share the same on-demand behavior, differences in the protocol mechanics lead to significant differentials in performance. Performance differentials are analyzed through the varying network loads, network size, and mobility. DSRs have a remarkably low packet dropping rate when compared to DSDV and AODV which indicates its efficiency level. But it is to be noted tha both models of AODVs and DSRs can fare better in high mobility situations unlike DSDVs. High mobility situations can be caused due to frequency in linkage failures and any overhead cost which is incurred while updating all newly routed information node as DSDVsare more involved than in the case of AODVs and DSRs. Particularly, DSRs utilise source routes and caches, and it does not purely depend on periodically involved activities. Thus, DSRs exploits caches for the purpose of route storage and it maintains a set of multiple routes per every destination. Unlike the above case, AODVs, on the other hand, utilize routing tables, at one route per destination, and proposed sequence of destination numbers, a mechanism to prevent loops and to determine freshness of routes. General observations made from the above simulation is based on application-oriented metrics such as packet delivery fraction and delay, DSR performs higher than the DSDV and AODV. DSR consistently generates less routing load than AODV.

Bai, et al., [11] proposed framework aims to evaluate the impact of different mobility models on the performance of MANET routing protocols. Interesting characteristics on mobility are captured through various independent protocol metrics, which include both spatial and temporal dependence and eventhe proposal of geographic restriction. Additionally, a set of richly parameterized mobility models will be introduced through mechanisms like the Group Mobilities, Random Waypoints, Group, and Manhattan and Freeway models. And based on the above ‘test suite’ models severalscenarios are carefully chosen to expand metric space. The utility of the proposed test-suite is demonstrated by evaluating various MANET routing protocols, including DSR, AODV and DSDV. Results from the above test showcase how performanes of protocol can drastically vary across various models and this can affect the ranking of performance protocols alongside each model used. This has been effectively explained through the interaction ofmobility characteristics alongside connectivity graph properties. Finally, decomposing the reactive routing protocols into mechanistic “building blocks” to gain a deeper insight into the performance variations across protocols in the face of mobility is attempted.

## METHODS

Performances of different protocols are determined by various interrelated metrics. Most important parameters to be regarded are the End to end delays, routing traffic data received, routing traffic data transferred and Throughputs which have been considered altogether to draw analytical observations. The throughput is generally taken as the key parameter. Throughput is the measure of how soon an end user is able to receive data. It is determined as the ratio of the total data received to required propagation time. A higher throughput will directly impact the user's perception of the quality of service (QoS).

Experiments carried out in a structured set up using an OPNET where the topology structure of the network and the motion mode of the nodes, to configure the service source and the receiver, to create the statistical data track file and so on is defined. Traffic models are often used as continuous bit rate (CBR) sources where the source-destination pairs are spread randomly over the network of area 4 Sqkm. Only 512-byte data packets are used. The number of source-destination pairs and the packet sending rate in each pair is constant. End to end delay includes all possible delays caused by buffering during route discovery latency, queuing at the interface queue, retransmission delays at the MAC, and propagation and transfer times of data packets.

Mobility models thus, attempt to understand movements of real mobile nodes. These are based from settling different parameters which can be related to nodular movements. Basic parameters are the starting location of mobile nodes, their movement direction, velocity range, speed changes over time. Mobility models can be classified to entity and group models [12]. Each entity models covers situations in which mobile nodes can shift independently from each other, while on the other hand in-group models nodes are highly dependent on each other or on a predefined leader node. In this study, Mobility Model used is diverging and converging nodes. It is noted that each packet is bound to start its journey from a randomly selected location to another randomly elected destination with an unspecified speed (uniformly distributed between 0–20 m/s). Simulations are run for 1000 simulated seconds

## RESULTS AND DISCUSSION

The following [Figures- 3-6] show the simulation results of end to end delay, routing traffic received, routing traffic sent and throughput.

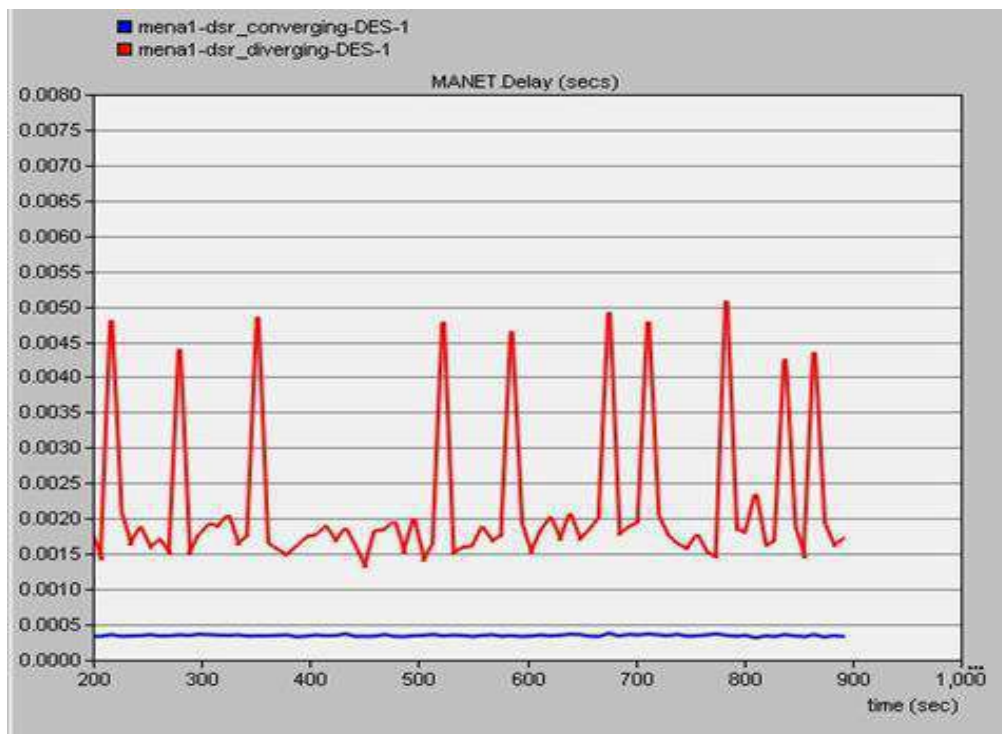


Fig:3. End to end delay for the converging and diverging nodes

It is observed from [Figure- 3] that the end to end delay increases substantially due to diverging nodes while the end to end delay is constant for converging nodes.

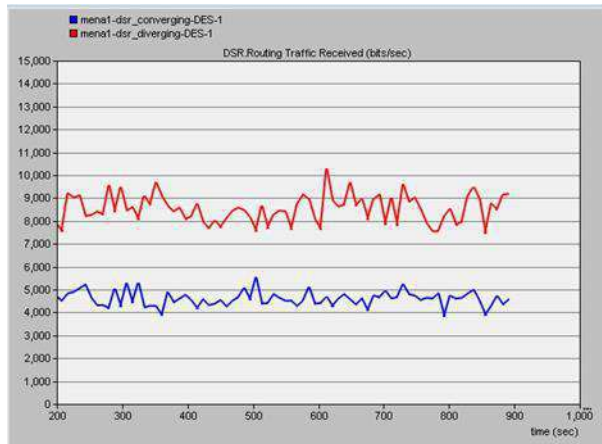


Fig. 4. Routing traffic received for the converging and diverging nodes

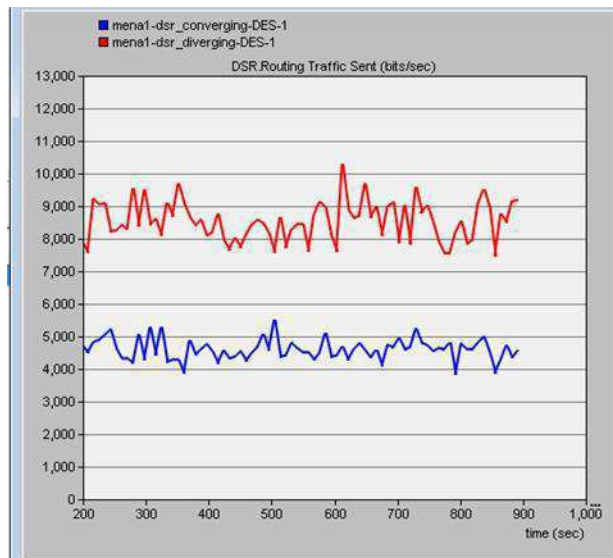


Fig. 5. Routing traffic sent for the converging and diverging nodes

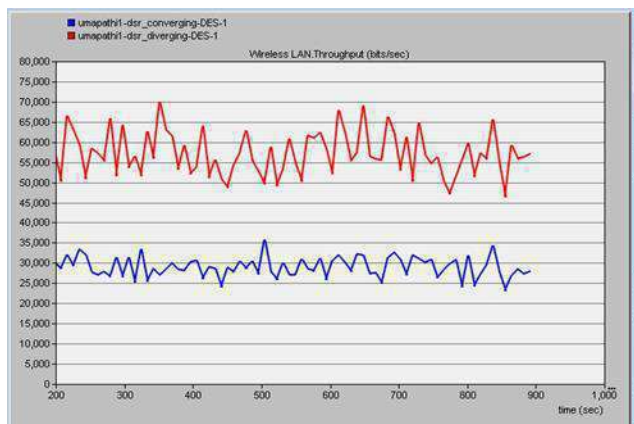


Fig. 6. Throughput in bits/sec

[Figure -6] shows the throughput for the diverging nodes and converging nodes. It is seen from the graph that the throughput falls drastically for the diverging nodes.

## CONCLUSION

In this paper, the overall performance of the Dynamic Source Routing (DSR) protocol, which is basically on-demand, under diverging and converging nodes is investigated. Detailed simulations were carried out, using OPNET Modeller. Simulation results show that the end to end delay increases and throughput reduces drastically in diverging scenario. Since throughput is lower in diverging scenario, further work can be done to improve the parameters by optimizing techniques.

## REFERENCES

- [1] Moussaoui A, Boukeream A. [2015] A survey of routing protocols based on link-stability in mobile ad hoc networks. *Journal of Network and Computer Applications*, 47:1-10.
- [2] Sari A. [2015] Security Issues in Mobile Wireless Ad Hoc Networks: A Comparative Survey of Methods. *New Threats and Countermeasures in Digital Crime and Cyber Terrorism*, 66.
- [3] Moussaoui A, Boukeream A. [2015] A survey of routing protocols based on link-stability in mobile ad hoc networks. *Journal of Network and Computer Applications*, 47: 1-10.
- [4] de M Cordeiro C, Gossain, H, P Agrawal, D [2015]. A directional antenna medium access control protocol for wireless ad hoc networks. *Journal of Communication and Information Systems*, 19(3).
- [5] Ilyas B, Fedoua D. [2015] A Novel Proactive Routing Method for Mobile Ad Hoc Networks. *American Journal of Applied Sciences*, 12(6):382.
- [6] Valentina Timcenko, MirjanaStojanovic, SlavicaBostjancicRakas, MANET Routing Protocols vs. Mobility Models: Performance Analysis and Comparison, Proceedings of the 9th WSEAS International Conference on Applied Informatics And Communications (AIC '09)
- [7] J Tian, J Hähner C, Becker I, Stepanov, and K Rothermel.[ 2002] Graph based mobility model for mobile ad hoc network simulation,” presented at the 35th Annual Simulation Symp., San Diego, CA, Apr
- [8] Anuj K Gupta, Member, IACSIT, Harsh Sadawarti, Anil K Verma.[ 2010] Performance analysis of AODV, DSR & TORA Routing Protocols, *IACSIT International Journal of Engineering and Technology*, 2(2)
- [9] Samyak Shah1, Amit Khandre2, Mahesh Shirole3 and GirishBhole. Performance Evaluation of Ad Hoc Routing Protocols Using NS2 Simulation,
- [10] M Anisur Rahman, Md Shohidul Islam, Alex Talevski.[ 2009] Performance Measurement of Various Routing Protocols in Ad-hoc Network, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong
- [11] Fan Bai a, Narayanan Sadagopan b, Ahmed Helmy.[2003] ] The IMPORTANT framework for analyzing the Impact of Mobility on Performance Of Routing protocols for AdhocNetworks, *AdHoc Networks* 1 :383–403.
- [12] Natarajan K, Mahadevan G. [2015] A Comparative Analysis and Performance Evaluation of TCP over MANET Routing Protocols. *Journal of Wireless Network and Microsystems*, 4:1-2

\*\*DISCLAIMER: This published version is uncorrected proof. plagiarisms and references are not checked by IIOABJ. the article is published as provided by author and checked/reviewed by guest editor.

# EFFICIENT SCALABLE AND ROBUST ZONE STRUCTURED MULTICAST ROUTING PROTOCOL FOR MANET

Umarani<sup>1\*</sup> and Thangaraj<sup>2</sup><sup>1</sup>Department of Computer Science and Engineering, SSM College of Engineering, INDIA<sup>2</sup>Department of Computer Science and Engineering, Bannari Amman Institute of Technology, INDIA

## ABSTRACT

Routing problems have become highly challenging because of the popularity of mobile devices. A rising amount of interest and importance has sparked among groups which support and communicate through Mobile Ad Hoc networks or MANET. Exchanging messages among a set of army soldiers on duty and communications between firemen during a disaster are examples of the above technology. With a one-to-many or many-to-many transmission pattern, multicast is an efficient method to realize group communications. Group communications are important in MANETs. Multicast is an efficient method to implement group communications. It is challenging to implement efficient and scalable multicast in MANET due to the difficulty in group membership management and multicast packet forwarding over a dynamic topology. A novel Efficient Geographic Multicast Protocol (EGMP) is proposed. EGMP uses a virtual-zone-based structure to implement scalable and efficient group membership management. A network wide zone-based bidirectional tree is constructed to achieve more efficient membership management and multicast delivery. The position of information guides the zone structure building, multicast tree construction, and multicast packet forwarding, which efficiently reduces the overhead for route searching and tree structure maintenance.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Geographic Routing, Wireless Networks, Mobile Ad Hoc Networks, Multicasting, Protocol.

\*Corresponding author: Email: [ctptr@yahoo.com](mailto:ctptr@yahoo.com)

## INTRODUCTION

The main aim of MANET is to extend mobility into the area of autonomous, mobile and wireless domains where set of nodes form network routing infrastructure in an adhoc fashion [1]. In a MANET, a group of mobile terminals work together to perform a particular task and hence it plays an important role in such networks [2]. Through wireless hosts MANET communicates with each other in absence of a fixed infrastructure. Hops are connected to each other through designed routes among two hosts within a network. With the rapid growth of demand in group communications the multicast technology in MANET has attracted a lot of attention [3]. Here mobile nodes are bounded to any centralized control like base stations or mobile switching centers. Limited transmission ranges affect wireless technologies and related network interfaces, and this is a reason why multi hops are needed for a single node to exchange data with another across a network. In this case, each mobile node operates as a host and as a router, forwarding packets to other mobile networks through its nodes which do not exist directly within wireless transmission range. Every nodal member participates in an ad hoc routing function that allows it to discover multihop paths through network to any other node.

Multicasting is proposed for group-oriented computing where the member of a host group is dynamic that means hosts may join or leave groups at any time. Members of host groups do not face any form of location or number restriction. A host may be a member of more than one group at a time and it doesn't have to be a member of group to send packets to the members in the group [4].

## RELATED WORK

Multicast routing protocol can be classified based on the type of routing structure they construct that are: First tree-based multicast routing protocol builds a tree-type multicast delivery structure for a multicast request. Second mesh based routing protocol builds a mesh structure for performing a multicast task. In general, these types of

protocols are resilient to network dynamics with certain sacrifice in forwarding efficiency [3]. It is a well-organized way of delivering content to a large group of receivers by using a tree structure embedded in the network. Building a multicast tree is called multicast routing that is used extensively to support many multicast application like teleconferencing and remote diagnosis [5,13]. It is an efficient way to transmit packets from one point or multi-points to multi-points that utilizes wireless channel bandwidth reasonably and reduces consumption of power [6].

Xia Deng et al., [1] have proposed a multicast routing protocol called CMRP (A Combined Multicast Routing Protocol) that considers three factors while selecting the routing path such as path's expiration time, the number of non-forwarding nodes, and the hop path. In order to meet user requirements, each component has an adjustable weight. They have showed that combination of these three factors achieve good performance in terms of data delivery ratio and energy consumption efficiency. The drawback of this paper is that the proposed protocol is not efficient to control the routing overhead and unable to handle traffic load.

Shigang Chena et al., [5] have proposed a scalable QoS multicast routing protocol (SoMR) that supports all three types of QoS requirement. SoMR is scalable due to small communication overhead. It achieves satisfactory tradeoff between routing performance and routing overhead by carefully selecting subgraphs in network aimed to search for path which are capable of supporting needs of QoS. This automatically tunes the scope based on the current network conditions. An early-warning mechanism helps to detect and route around the long-delay paths in the network. The operations of SoMR are completely decentralized. They rely only on the local state are stored at each router. The drawback of this paper is that they have not considered the throughput metrics and energy efficiency.

Neng-Chung Wang et al., [10] have proposed a power-aware dual-tree-based multicast routing protocol (PDTMRP) for mobile ad hoc networks (MANETs). Nodes described in the above scheme are classified randomly into two kinds: group-0 and group-1. In order to achieve the required load balance, two multicast trees (tree-0 for group-0 and tree-1 for group-1) are constructed. In this scheme, load balance is used to improve the lifetime of a network. In the route discovery, this scheme not only solves the stability routing problem, but also achieves the load balance of data transmission. Thus, controlling overhead route constructions and the required number of route reconstructions are proportionately decreased. The packet delivery ratio and the control overhead of the proposed scheme outperform that of MAODV and RMAODV. Moreover, the traffic load can be balanced and the network lifetime can be prolonged. This proposed scheme has a major drawback which is that they have not considered about the delay metrics in their proposed protocol.

Chia-Cheng Hu et al., [14] have proposed distinct strategy to select stable Backbone Hosts (BH). Extra or remaining connection time period between two neighbors is calculated through Global Positioning Systems (GPS), and with its aid a long lasting and stable BHs from a selected set of hosts. Additionally, new multicast protocols are proposed according to selected set of stable BHs to select stable multicast routes. A stable route is a route that is available for a longer time. Simulation results show that the proposed protocol has shorter transmission latency, shorter more stable multicast routes, lower overhead, more stable attachment of multicast members to BHs, and higher receiving data packet ratios than other existing two-tier multicast protocols. The future work includes study of the link quality for different application of ad hoc networks.

X. Xiang et al., [15] have proposed a novel Efficient Geographic Multicast Protocol (EGMP) which utilize virtually enhanced base structures to add scalable and efficient group management membership. Network wide bidirectional trees are constructed to achieve mre through its zone-based efficiency in managing membership and systems for multicast deliveries. Positional information is often used as a guideline for building zone structures, constructing multicast trees, and forwarding multicast packet, all of which effectively reduces overhead route searches and tree structure maintenance. Several strategies was proposed to improve the efficiency of the protocol, for example, introducing the concept of zone depth for building an optimal tree structure and integrating the location search of group members with the hierarchical group membership management. Finally, they have designed a scheme to handle empty zone problem faced by most routing protocols using a zone structure. The scalability and the efficiency of EGMP are evaluated through simulations and quantitative analysis. The results demonstrate that EGMP achieves a high packet delivery ratio, and low control overhead and multicast group joining delay under all test scenarios, and is scalable to both group size and network size. The proposed scheme has a major drawback which is that they have not considered about efficient utilization of bandwidth to improve QoS in the multicast routing.

## EXISTING PROTOCOL AND ITS PERFORMANCE

The following sections will try to illustrate some of the basic EGMP protocols briefly. Section A will provide an overview of all the protocols and required definitions to be used in the rest of the paper. Sections B and C, present the designs for construction of zone structure and the zone-based geographic forwarding.

### OVERVIEW OF PROTOCOLS

EGMP is a method which promotes scalability, reliable forms of managing membership and multicast forwarding established through using a structural two-tier virtual zone. At the lower layer, in reference to a predetermined virtual origin, nodes in the network are self-organizes by themselves into a set of zones, from which the elected leader of the zone manages local group memberships. The upper layers of leadership serves as a representative to join or leave multicast groups in its zone. Due to this model of functioning, a multicast tree of network-wide zone was created. For efficient and reliable management, location information is integrated with the design and used to guide the zone construction, group membership management, Maintenance, multicast tree constructions, and forwarding packets. This zone-based tree is sharable among groups of multicast sources.

Some notations used are:

- Zone: In this the network terrain is divided into square.
- Zone size, the length of a side of the zone square transmission range of the mobile nodes.
- To reduce intra zone management overhead, the intra zone nodes can communicate directly with each other without any intermediate relays.

• Zone ID: This is used for the identification of a zone. A node can calculate its zone ID (a, b) from its position

• Zone Forwarder. A zone forwarder is elected in each zone for managing the local zone group membership and taking part in the upper tier multicast routing.

The tree zones are responsible for multicast packet forwarding. A tree zone may have group members or just help forward the multicast packets for zones with members.

### NEIGHBOR TABLE GENERATION AND ZONE LEADER ELECTION

Neighboring tables are constructed through nodes without the need for extra signaling. When receiving a beacon from a neighbor, a node records the node ID, position, and flag contained in the message in its neighbor table. The zone ID of sending node can be calculated from its position. Failures in routing can be avoided by updating topology information and removing entries which have not been refreshed for a long period of time. Corresponding neighbors or TimeoutNT are unreachably detected by MAC layer protocols. The election of zone leaders through cooperative nodal systems is responsible for maintaining the consistency of the zone. A node sends a beacon announcing its existence once a node appears in the network. Then, it waits for an Intvalmax period for the beacons from other nodes. Every Intvalmin node checks its neighbor table and determine its zone leader under different cases: 1) the neighbor table contains no other nodes in the same zone; it will announce itself as a potential leader. 2) Flags of every node existent in one zone are unset, which indicates that no node can announce leadership of the zone. If the node is closer to the zone center than other nodes, it will announce its leadership role through signal messages between leadership flagsets. 3) When in more than a single node in one zone has selected a leader flags set, the highest node membership ID is selected. 4) But only one node in a zone can have a flag set, and this node is selected as the leader.

### CONSTRUCTING MULTICAST TREES

The section presents creation of multicast trees along with its maintenance schemes. Thus, instead of connecting each group present in EGMP, the member directly connect to the tree, and this tree is formed after guided location information in the granularity of zone, which can significantly reduce overhead the tree management. A control message can be sent immediately based on setting destination locations, without having to incur high overhead charges and delay in finding paths, which enables quick group joining and leaving. In the following description, except when explicitly indicated, we use G, S, and M, respectively, to represent a multicast group, a source of G and a member of G.



## MULTICASTING ROUTE OPTIMIZATION AND MAINTENANCE

It is crucial to maintain connection modes in dynamic networks and this requires adjusting the structure of the tree based on topological changes which optimize multicast routing. In the zone structure, some zones are empty due to the movement of nodes between different zones which is critical to handle the empty zone problem. Comparing the connections of individual nodes, however, there is a much lower rate of zone membership change and hence a much lower overhead in maintaining trees which are zone-based. Disconnected zones can effectively establish reconstructions to the tree due to the guided location constructions. Additionally, zones can be partitioned among multiple clusters based on the effects of fading and signal blocking.

## PERFORMANCE EVALUATION

Periodically, multicast sources broadcast Join-Query messages to an entire network. Intermediary nodes store source ID and sequence numbers, after updating the routing table with the required node ID (i.e., backward learning) and from the received details messages can be traced back to the source. A receiver creates and broadcasts a Join Reply to its neighbors, with the next hop node ID field that are filled by extracting information from its routing table. Matching ID neighbor nodes of the message realize the paths to the source and become a part of the forwarding group. It then broadcasts its own Join Table built upon matched entries. This whole process constructs (or updates) the routes from sources to receivers and builds a mesh of nodes, the forwarding group. [Table -1] lists the simulation parameters of EGMP with beacon interval 200sec. The simulations for ODMRP are based on the codes carried with the simulator, with the parameters set as in [9].

**Table: 1. Parameter Values for EGMP Simulations**

Parameter	Value
r(zone size)	75 m
Intval min	2 sec
Intval max	4 sec
Intval active	3 sec
Timeout NT	3 sec

Several bugs in the GloMoSim codes were fixed to prevent forwarding group node from sending any form of JOIN TABLES. This impacts and improves the delivery ratio by doubling its capabilities and reducing controlling overhead ODMRP. Additionally, we implemented SPBM in GloMoSim according to [20] and then two codes can provide the authors with similarly required parameter settings but it should be noted that square sizes is set as 150 m to assist nodes in a square which are in between each transmission range. Quad-trees transform according to the number of levels which is based accordingly to the square size and the network size used. For packet forwarding in SPBM [20], the square center is used as the destination position, which improves the delivery ratio. Also improves the stateless multicast protocol which allows it a better scalability to group size. Contrastly, EGMP uses a location-aware approach for more reliable membership management and packet transmissions, and supports a scalability of both group size and network size.

## AN EFFICIENT SCALABLE AND ROBUST ZONE STRUCTURED MULTICAST PROTOCOL FOR MANET

Work on a Receiver-Based Multicast protocol, RBMulticast, which is a stateless cross layer multicast protocol where packet routing is extended, splitting packets into multiple routes, and the medium access of individual nodes rely solely on the location information of multicast destination nodes. Multicast members are included in a list of RBM multicast packet header locations and this can prevent building and maintaining overhead multicast trees set at intermediate sensor nodes and due to the above important routing information, the packet is included within the packet header. Additionally, the medium access method employed does not require any state information such as neighbor wake-up time or any a priori operations such as time synchronization. Tree creation,

maintenance or neighbor table maintenance is not required. It makes RBMulticast as the least state of any multicast routing protocol. It is ideally suited for dynamic networks. In RBMulticast the following two techniques instead of two tier zone structures are proposed.

### NODE LIFETIME PREDICTION ALGORITHM

Consider, two nodes having the same residual energy level. Among that an active node quickly consumes energy that is used in many data-forwarding paths which shortens its lifespan than when a node remains in inactive node. The lifetime of node is based on its current residual energy and its past activity solution that does not need to calculate the predicted node lifetime from each data packet.  $E_i$ , represents current residual energy  $i$ , is exponentially used to weigh moving average and estimate energy draining rates  $e_{vi}$ , and this is the rate of energy depletion.  $E_i$  which is obtained easily online from instrumental battery management techniques, and  $e_{vi}$  is thus a statistical value which is obtained through recent history. The estimated energy drain rate in the  $n$ th period is  $e_{vin}$ , and  $e_{vi(n-1)}$  is the estimated energy drain rate in the previous  $(n - 1)$ th period,  $\alpha$  denotes the coefficient that reflects the relation between  $e_{vin}$  and  $e_{vin-1}$ , from which its constant value is estimated to be within the range of  $[0, 1]$ .

### LINK LIFETIME- PREDICTION ALGORITHM

When the minimum node lifetime in a route from 2 nodes of stable connection within the communication range of each other, then connection lifetime may last longer, and they cannot be a bottleneck in the route to which they belong. Unstable connections can also have the capability to model the flexibility and mobility of nodes which exist in shorter periods of its unstable nature. Nodes can move at a constant speed towards the same direction in a short period. It is easy to measure the distance between nodes  $N_i$  and  $N_{i-1}$  by applying Global- Positioning-System- based location information. Transmitted packages are forwarded with the same power level owned by a receiver and can measure the strength of a received signal power especially when receiving packages. Then the distance is calculated by directly applying the radio propagation model to it. If the received signal power strength is lower than a threshold value, then this link as an unstable state and then calculate the connection time.

LLT prediction algorithm requires only two sample packets, and implements piggyback information on route-request (RREQ) and route-reply (RREP) packets during a route-discovery procedure with no other control message overhead, and thus, it does not increase time complexity.

## RESULTS AND DISCUSSIONS

After the above subject of variability in moving speed of EGMP and its needed node density, the paper proceeds to investigate scalability of three protocols by modifying group and network size. We focus on the studies of the scalability and efficiency of the protocol under the dynamic environment and also in consideration with the energy and power utilization of nodes.. After evaluating the proposed algorithm, performance metrics are utilized in the simulations for performance comparison.

Packet arrival rate: The ratio of the number of received data packets to the number of total data packets sent by the source.

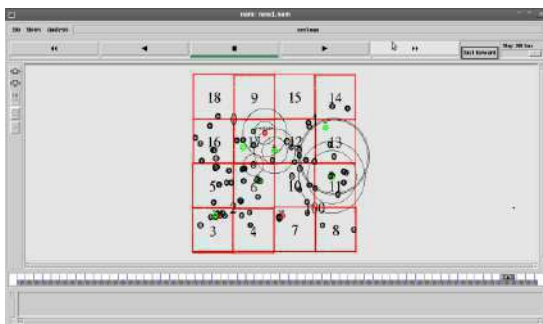


Fig:1.Packet from source is delivered to the destinations.

Xgraph: The xgraph shows the packet delivery ratio.

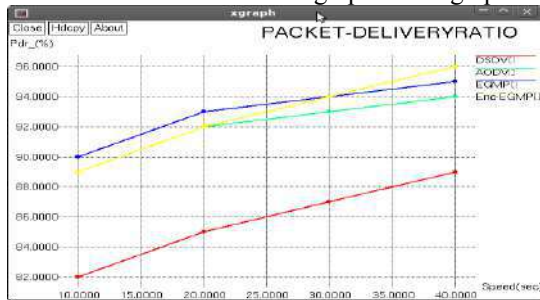


Fig. 2. The packet delivery ratio.

Average end-to-end delay: The average time elapsed for delivering a data packet within a successful transmission.

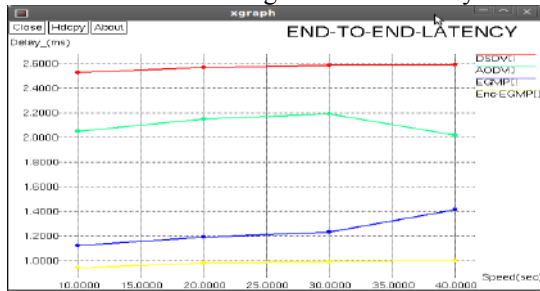


Fig. 3. End to End latency

Energy consumption: The calculation of energy consumption for the entire network includes the transmission energy consumption for both the data and control packets.

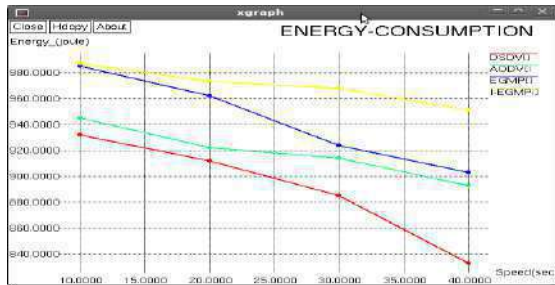


Fig.4. Energy Consumption

Throughput: The throughput for the entire data transmission from source to destination is increased when compared to the existing protocol.

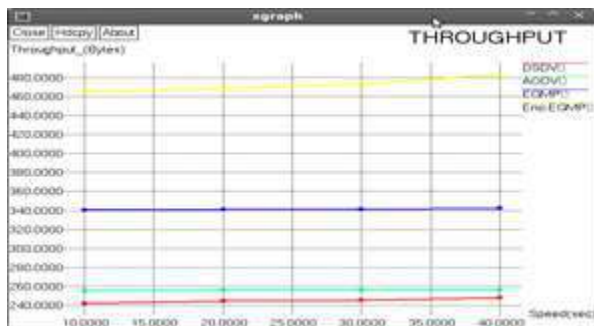


Fig. 5. Increase Throughput

Collision rate: The average Collision rate for the entire data transmission from source to destination is much controlled and reduced when compared to the existing protocol.

Communication overhead: The average number of transmitted control bytes per second, including both the data packet header and the control packets.

## CONCLUSION

Achieving stability of zone structure relies on underneath geographic unicast routing for reliable packet transmissions. We build a zone-based bidirectional multicast tree at the upper tier to achieve more efficient multicast membership management and delivery, and use a zone structure at the lower tier to realize the local management of membership. The research has also created a scheme which can handle problems of encountering empty zones which can challenge zone-based protocols. The position information is used in the protocol to guide the zone structure building, multicast tree construction and multicast packet forwarding. As compared to traditional multicast protocols, our scheme allows the use of location information to reduce the overhead in tree structure maintenance and can adapt to the topology change more quickly. Results shows that the throughput for the entire data transmission from source to destination is increased and the average time elapsed for delivering a data packet within a successful transmission when compared to the existing protocol. Future work should involve multicast routing protocols that aim at providing reliability, QoS guarantees, security, and so on. Hence, transmitting Multicast systems are more effective when compared to supported groups unicast from group communication applications and thus this aspect is important for development of future networks.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Deng Xia, Jianxin Wang, Yao Liu. [2011] A Combined Multicast Routing Protocol in mobile ad hoc networks." Communications, Computers and Signal Processing (PacRim), 2011 IEEE Pacific Rim Conference on. *IEEE*,
- [2] Junhai, Luo, Xue Liu, and Ye Danxia. [2008]Research on multicast routing protocols for mobile ad-hoc networks, *Computer Networks* 52(5): 988-997.
- [3] Yan, Yan, et al. [2012] D-ODMRP: a destination-driven on-demand multicast routing protocol for mobile ad hoc networks. Communications, *IET* 6(9):1025-1031.
- [4] Biradar RC, and SS Manvi. [2011 ]Agent-driven backbone ring-based reliable multicast routing in mobile ad hoc networks.Communications, *IET* 5(2) :172-189.
- [5] ShigangChena,andYuvalShavittb, SoMR.[2008 ]Ascalable distributed QoS multicast routing protocol, *J Parallel Distrib Comput* 68: 137 – 149.
- [6] Tang Yan, Xu Li, and Mingqiang Yang. [2011]Improvement of multicast routing supporting mobile ad hoc networks with unidirectional links. Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on. *IEEE*.,
- [7] Hong Xiaoyan, KaixinXu, and Mario Gerla. [2002 ]Scalable routing protocols for mobile ad hoc networks. Network, *IEEE* 16(4):11-21.
- [8] deMoraisCordeiro, Carlos, HrishikeshGossain, and Dbarma P.[2003] Agrawal. Multicast over wireless mobile ad hoc networks: present and future directions. Network, *IEEE* 17(1): 52-59.
- [9] Nguyen HaiTrung. [2011]An efficient and message-optimal multicast routing protocol in mobile ad-hoc networks. Advanced Technologies for Communications (ATC), 2011 International Conference on. *IEEE*
- [10] Wang, Neng-Chung, et al. [2011]A Dual-Tree-Based On- Demand Multicast Routing Protocol for Mobile Ad Hoc Networks. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2011 12th ACIS International Conference on. *IEEE*.
- [11] Garcia-Luna-Aceves, JJ. and Rolando Menchaca-Mendez[2011] PRIME: an interest-driven approach to integrated unicast and multicast routing in MANETs. *IEEE/ACM Transactions on Networking* (TON) 19(6): 1573-1586.
- [12] Sun, Jun, et al.[ 2011]QoS multicast routing using a quantum-behaved particle swarm optimization

algorithm. *Engineering Applications of Artificial Intelligence* 24(1): 123-131.

- [13] B Madhusudhanan, S Chitra, C Rajan, Mobility Based Key Management Technique for Multicast Security in Mobile Ad Hoc Networks, *The Scientific World Journal*, Hindawi Publishing Corporation, 2015
- [14] Hu Chia-Cheng, EH-K. Wu, and Gen-Huey Chen. [(2009)]Stable Backbone Hosts and Stable Multicast

Routes in Two-Tier Mobile Ad Hoc Networks." *IEEE Transactions on Vehicular Technology*, 58(9):5020-5036.

- [15] Xiang, Xiaojing, Xin Wang, and Yuanyuan Yang. [2011]Supporting efficient and scalable multicasting over mobile ad hoc networks. *Mobile Computing, IEEE Transactions on* 10(4): 544-559.

# DETECTION OF CALCIFICATION IN MAMMOGRAM USING NEAREST NEIGHBOUR ALGORITHM

Usha<sup>1\*</sup> and Arumugam<sup>2</sup>

<sup>1</sup>PARK College of Engineering and Technology, Coimbatore, TN, INDIA

<sup>2</sup>Dept of Computer Science and Engineering, Nandha Engineering College, Erode, TN, INDIA

## ABSTRACT

Breast cancer is a dangerous and it increases the death rate among women cancer detection in early stage is not an easy task. The reason of the cancer is uncontrollable cells growth. In this paper an automatic mammogram classification techniques using symlet wavelet, gabor filter and nearest neighbour algorithm are used for getting better result.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Breast Cancer, Digital mammogram Wavelets, gabor, Feature Selection, nearest neighbour algorithm.

\*Corresponding author: Email: [usha.samiappan@gmail.com](mailto:usha.samiappan@gmail.com), [arumugamdote@yahoo.co.in](mailto:arumugamdote@yahoo.co.in)

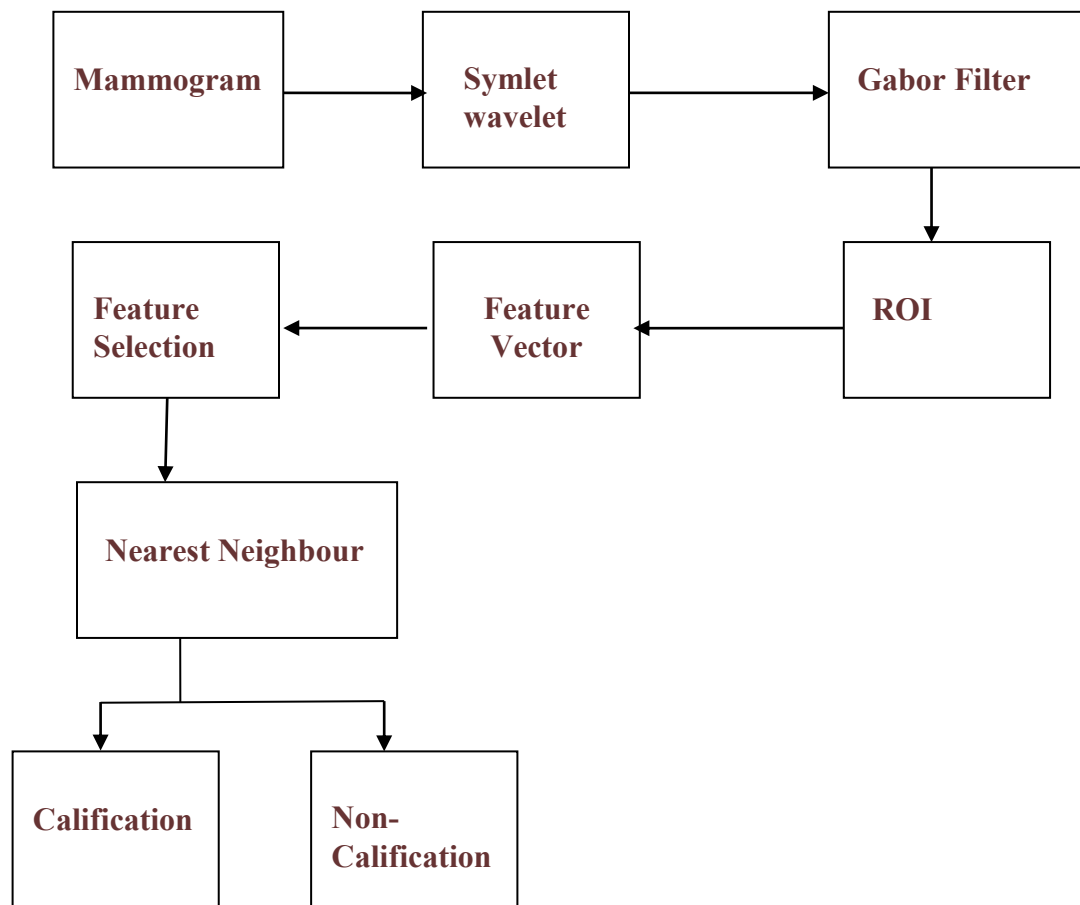
## INTRODUCTION

Mammography is a useful method for early breast carcinomas detection [1]. Diagnosis of Breast cancer using xray mammography is very good method. But many studies reveal that radiologists can be done misdetection of abnormalities in addition to having higher rates of false positive. 75% is the estimated sensitivity of radiologist in Breast cancer screening. To avoid this double reading was suggested to be effective to improve sensitivity but it is costly and it takes time to read [2]. cost effectiveness is important for mass screening programs to succeed. The diagnosis may be affected due to human factor of the abnormality indicators usually varied in shape, size and brightness. Difficult task to find out tumours and ordinary calcification [3,4]. In this paper, an automatic mammogram classification techniques using wavelets, gabor, feature selection is used to know the relationship between the features and classifier used to classify benign and malignant.

## METHODOLOGY

MIAS is a digital mammogram database where films from U.K.'s National Breast Screening Programme have been digitized to 50 micron pixel edge with a Joyce-Loebl scanning microdensitometer, a linear device in a 0-3.2 optical density range representing pixels with an 8-bit word [5]. The database includes 322 digitized films and radiologist "truth"-markings on locations of abnormalities detected. The database was reduced to 200 micron pixel edge with padding/clipping to ensure that all images are 1024x1024 pixels at 8 bits per pixel. Erosion followed by dilatation has a similar structuring element, completing the opening function.

The MIAS database [6] though no longer supported, is old used much in literature. MIAS annotations are insufficient for some studies as all circumscribed/speculated lesions are to be manually segmented [7]. Another drawback is its digitized resolution which renders it unsuitable for micro-calcification detection experiments. Regarding calcifications, healthier tissue is found in the ground truth region which is justified through calcifications shape as the latter are small lesions spread over a large area with all this being included in the annotation. Cross validation ensures higher formalism in entry data division considered necessary due to limited images with calcifications available in the database. The database Mini MIAS, prevents excessive network training and so a better system generalization.



**Fig. 1. Proposed system for calcification of Mammogram**

### ANALYSIS USING MULTI RESOLUTION APPROACH

Multi resolution is a scaling function. It is used to create a series of approximation of a function or image [8]. Each differing by a factor of 2 in resolution from its nearest neighbouring approximation [Gonzalez et al 2004]. In this system image scaling is used in order to reduce computing time. The wavelet transform have been used for micro calcification detection using symlet wavelet, a function of a continuous variable into a sequence of coefficient like sequence of number usually called as detail and approximation [9]. The wavelets are functions and those functions are basis for other functions called mother wavelet. The series of function generated by translation and dilation of mother function.

The four different coefficient are produced in each level of decomposition. The decomposition takes place while applying 2D wavelet transform on image. Those are horizontal, vertical, diagonal and approximation coefficient. DWT analysis the image by decomposing into coarse approximation via lowpass filtering and detailed information via highpass filtering. The decomposition is performed recursively on lowpass approximation coefficient at each level until the necessary iteration are reached. Micro calcification appear in the mammogram image as fine and bright grains in the tissue. so the detailed coefficient having micro calcification in the mammogram image using wavelet decomposition. In this research three level discrete wavelet decomposition by using symmetric daubechies of order 2 [10]. The preprocessing time can be reduced using region of interest (ROI). The ROI extraction ignored the dark areas at particular frequency and orientation gabor filter can be viewed as a sinusoidal plane [11]. Symlet wavelet along with gabor filter for detecting mass easily. Gabor filter bank applied in different frequency and orientation on HH high frequency subband image obtained using symlet and extracted statistical features like standard deviation and mean in the form of feature vector.

ROI selected the centers of the abnormality. After getting the feature vector, the feature can be selected using searching method [12]. The number of features obtained after applying the feature selection method [13]. The classifier used to classify the mass into calcification and non-calcification. The classification can be done using Euclidian distance as a measurement between the coefficient [14][15]. In a class, there is a N of images. These images are used to create class core vector and the core vector calculated by

$$CCV^i = \frac{1}{n} \sum_{j=1}^{j=n} CCV_j^i$$

Where  $j = 1, 2, 3, \dots, m$

The Euclidian distance can be calculated by

$$D = \sqrt{\sum_{i=1}^m (CCV^i - V_{Test}^i)^2}$$

It is used to calculate the distance between the tested image and class core vector.

### DISTRIBUTION OF THE MIAS DATA BASE

In this research MIAS (Mammographic Image Analysis Society) data set is used. This data set was investigated & labeled by the experts. This data set having 322 mammogram image of right and left breast. From 161 patients, 51 diagnosed as malignant, 64 has benign and 207 as normal. This result shown in [Table -1].

Table: 1. Distribution of the Mias data base

Cases	B	M
Circ	19	04
Ill	06	08
Spic	11	08
Arch	09	10
Ass	06	09
Norm	-----	----

- Abnormality Class
- Norm - normal tissue
- Calc - microcalcification clusters
- Circ - circumscribed masses
- Ill – ill-defined masses
- Spic – spiculated lesions
- Arch – architectural
- Asym – asymmetry
- Type of cancer
- B – benign
- M – malignant

### RESULT AND ANALYSIS

To distinguish between the types of tumors based on the physical properties and level of risk. The six abnormality cases are used as important classes. Those are microcalcification, circumscribed masses, ill-defined masses, spiculated lesions, architectural, asymmetry. Next, classifying whether those cases are benign or malignant tumors. The following table shows the classification rate with accuracy based on 10 fold cross validation. In each fold, the average rate calculated and total average of 10 fold can be calculated (Table-1).

Table-1: Successful rate of calcification and non-calcification.

Class	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	AVG
B	100	90.20	84.33	92.35	80.20	94.3	85.60	90.00	86.78	93.50	89.72
M	100	100	98.50	98.40	98.04	96.20	100	97.30	100	96.30	98.47
Avg	100	95.1	91.41	95.37	89.12	95.25	92.8	93.65	93.39	94.9	94.09

B : Bennign, M : Malignant, F: Fold



## CONCLUSION

The diagnosis of breast cancer in a digital mammogram is a practical field of investigation. In this study the concept of using wavelet coefficients and gabor coefficients are used to form future vector and multi resolution analysis used in future extraction. Experiment which applied on real data shows the above results. The accuracy rate of classification achieved to distinguish between benign and malignant is 94.09 %. The results show a special concentration on the wavelet coefficient that gives high percentage of success to find the tumour in each class.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Curtis C, Shah SP, et al. [2012] The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346-352.
- [2] Tang J, Rangayyan RM, Xu J, El Naqa I, Yang Y. [2009] Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *Information Technology in Biomedicine, IEEE Transactions on*, 13(2):236-251.
- [3] Virnig BA, Tuttle TM, Shamliyan T, Kane RL. [2010] Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes. *Journal of the National Cancer Institute*, 102(3): 170-178.
- [4] Wasif N, Maggard MA, Ko CY, Giuliano AE. [2010] Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes. *Annals of surgical oncology*, 17(7):1862-1869.
- [5] Suckling J et al. [1994] The mammographic image analysis society digital mammogram database, *International Congress Series* 1069:375-378.
- [6] HD Cheng, X cia, X chen, LH lou. C. [2003] Computer-aided detection and classification of microcalcification in mammogram: a survey. *Pattern Recognition Letters* 36:2967-2991.
- [7] Dean J, Ilvento V. [2006] Improved cancer detection using computer-aided detection with diagnostic and screening mammography: Prospective study of 104 cancers. *American Journal of Roentology*, 187:20-28.
- [8] S Liu, CF Babbs, E Delp. [2001] Multiresolution detection of spiculated lesions in digital mammograms, *IEEE Transactions on Image Processing* 10 (6):874-884.
- [9] EA Rashed, IA, Ismail SI Zaki. [2007] Multiresolution mammogram analysis in multilevel decomposition. *Pattern Recognition Letters* 28:286-292.
- [10] Hajar Moradmand, Saeed Setayeshi, Ali Reza Karimian, Mehri Siros, Mohammad Esmail Akbari "Comparing the performance of image enhancement methods to detect microcalcification clusters in digital Mammography", *Spring*, 5(2)
- [11] Yufeng Zheng [2010] Breast cancer detection with gabor features from digital mammograms. *Algorithms*, 3:44-62.
- [12] Wei CH, Li Y, Li CT. [2007] Effective extraction of Gabor features for adaptive mammogram retrieval. In *Multimedia and Expo, 2007 IEEE International Conference on* (pp. 1503-1506). *IEEE*
- [13] Mohammed Meselhy Eltoukhy, Ibrahim Faye, Brahim Belhaouari Samir "Curvelet based feature extraction method for breast cancer diagnosis in digital mammogram.
- [14] Min-Ling Zhang; National Lab. for Novel Software Technol., Nanjing Univ., China; Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification.
- [15] Keller JM. Gray, M.R.; Givens, J.A.. A fuzzy K-nearest neighbor algorithm

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and approved by guest editor.

# PERFORMANCE & ANALYSIS AUTOMATION OF MEDICAL REIMBURSEMENT FOR BSNL

Anto Bennet<sup>1\*</sup>, Ramesh Kandasamy<sup>2</sup>, Sankaranarayanan<sup>1</sup>, Nandhini<sup>1</sup>, Meena<sup>1</sup>, Alamelu<sup>1</sup>

<sup>1</sup> Dept. of ECE, Vel Tech, Chennai, INDIA

<sup>2</sup> Dept. of IT, Nandha Engineering College, Erode, Chennai, INDIA

## ABSTRACT

The fundamental goal of this paper is to automate the Medical Reimbursement strategies for BSNL with the Seamless Integration of procedure. This delineates the Medical Reimbursement components of the association. Our essential objective is to plan the device in a manner that the End-User and the Administrator need not battle to search for any data he/she needs. The Automation of Medical Reimbursement System has been utilized to beat the issues brought about by the present manual preparing. In the proposed framework, the information sustained by the workers is kept in a secured database which thus has made the recuperation of the information more straightforward. Then again it likewise gives attention to the workers about the way in which it can be effortlessly associated with the database in the back end furthermore altered. Further, information reflection is done to guarantee that fitting coding at the server side is kept a long way from the ordinary web clients. Our Application is adaptable for different upgrades and improvements in not so distant future. All in all, the proposed framework has been utilized to give an undeniable support of the client and the head in their own particular terms.

Published on: 10<sup>th</sup>– August-2016

## KEY WORDS

Medical Reimbursement,  
Seamless Integration, Data  
Abstraction.

\*Corresponding author: Email: [bennetmab@gmail.com](mailto:bennetmab@gmail.com) Tel.: +91 9965576501

## INTRODUCTION

The fundamental reason for this paper is to computerize the handling of the Medical bills by the representatives for their Reimbursement. In this task we have taken a shot at making a Medical bill preparing web application with the goal that it ought to be an effectively justifiable application. The application that we have grown in this is for the representatives of Bharat Sanchar Nigam Limited (BSNL). Any web application without the cutting edge methodology is futile. It is additionally intended to convey substance to client's requests, along these lines lessening the clients' work significantly expanding solace and effectiveness. The substance of the web application are chosen and gathered organized appropriately such that any non-specialized and new client can scan through the data he/she needs. Bharat Sanchar Nigam Limited (curtailed BSNL) is an Indian state-possessed information transfers organization headquartered in New Delhi, India.

It was consolidated on 15 September 2000. It assumed control over the matter of giving of telecom administrations and system administration from the recent Central Government Departments of Telecom Services (DTS) and Telecom Operations (DTO), with impact from 1 October 2000 on going concern premise. It is the biggest supplier of settled telephony and fourth biggest telephony supplier in India, and is additionally a supplier of broadband administrations. Be that as it may, as of late the organization's income and piece of the pie dove into overwhelming misfortunes because of exceptional rivalry in the Indian information transfers sector. BSNL is India's most established and biggest correspondence administration supplier (CSP). It had a client base of 117 million as of Jan 2014. It has foot shaped impressions all through India with the exception of the metropolitan urban communities of Mumbai and New Delhi, which are overseen by Mahanagar Telephone Nigam (MTNL).

## MATERIALS AND METHODS

### LITERATURE SURVEY

BSNL is an extremely enormous system which is overseen by a gathering of individuals. Preparing and sorting out the entire system is a dreary procedure. In BSNL every single Medical case are handled physically, which devours part of time and staff assets. At first, the Employee needs to fill the Application shape and submit it to the individual officer alongside the Medical Bills for preparing. The officer thus sits tight for a chunk of such claims to get gathered and afterward passes the gathered cases to the following level of preparing. The structure is finally gone to the Account officer who checks the subtle elements for sum freedom. At last, the case is authorized which might take couple of months or more.

### PROPOSED SYSTEM

The Proposed framework has been worked for the advantage of BSNL to beat the troubles confronted in manual handling of Medical Reimbursement via robotizing the whole process as a Web based application. In Proposed system [1-4], at the client level the client sign in through their one of a kind id's to fill in the Online Reimbursement shape and submit it and gets their status redesigned occasionally through email alarms. At the administrator level, the submitted structure is gotten by the separate authorities in different levels, for example, Dispatch, Receive, Amount Clearance and Sanction. Along these lines, the reengineered framework will indicate most extreme capacity and similarity to be changed over to a client justifiable organization [5-8].

### SYSTEM DESIGN

The User applies for therapeutic repayment by filling the online application structure. The structure submitted is put away in the server and after that gets redesigned into the database. At the Admin level, the dispatcher dispatches the application structure to the collector.

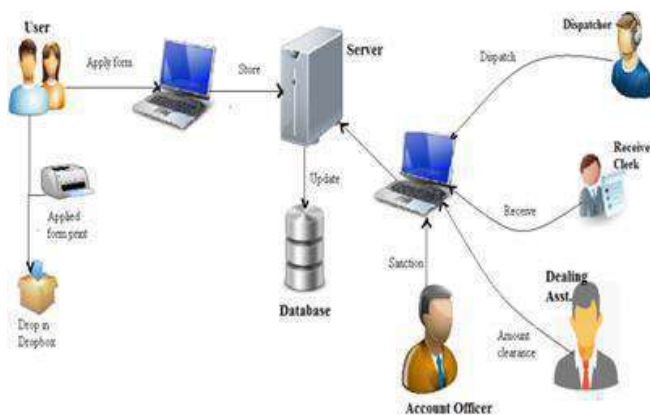


Fig. 1: System Architecture

The recipient on confirming it passes it to the managing collaborator for sum leeway. At last, the record officer authorizes the case of the client and the client is insinuated about the same through email correspondence appeared in [Figure 1].

### DATA FLOW DIAGRAM

Data Flow Diagram (DFD) is a method for demonstrating a framework's abnormal state point of interest by indicating the sequence of transformations. DFD uncovers connections among and between the different segments in a project or framework. DFD comprises of four noteworthy segments: Entities, Processes, Data stores and Data streams

### LEVEL 0

Level 0 Data stream graph will speak to the info, process and the yield of the framework. A DFD might seem to be like a stream diagram. Be that as it may, there is a noteworthy distinction with the information stream graph. The bolts in DFDs demonstrate that there is a stream of information between the two parts and not that the segment is sending the information that should be executed in the accompanying segment appeared in [Figure -2].



Fig. 2: Level 0 DFD

A segment in DFD may not proceed with execution when sending information and amid execution of the segment getting the information. The segment sending information can send numerous arrangements of information along a few associations. Indeed, a DFD hub can be a segment that never closes.

**LEVEL 1**

The Level 1 DFD indicates how the framework is separated into sub-frameworks (forms), each of which manages one or a greater amount of the information streams to or from an outer operators, and which together give the majority of the usefulness of the framework all in all. It likewise distinguishes interior information stores that should be available all together for the framework to carry out its employment, and demonstrates the stream of information between the different parts of the framework appeared in [Figure -3]

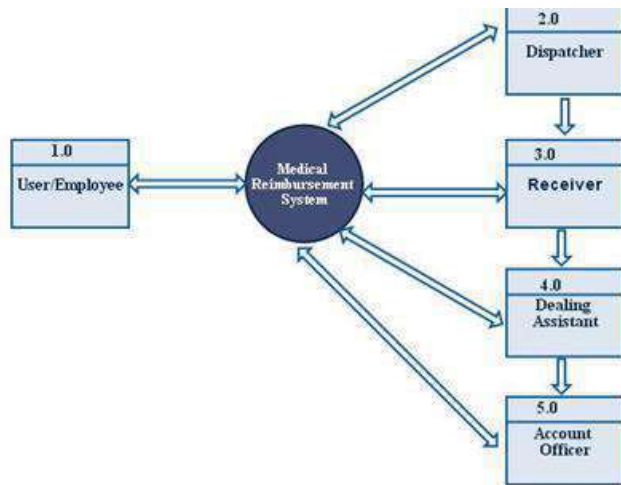


Figure: 3 Level 1 DFD

**LEVEL 2**

In the level 2 DFD, there is number of intermediate nodes which are used in each module is elaborated and the exact flow of the processing system is shown in detailed structure shown in [Figure -4].

**UML DESIGN**

Unified Modeling language (UML) is a standardized modeling language enabling builders to specify, visualize, construct and document artifacts of a software procedure. Therefore, UML makes these artifacts scalable, relaxed and effective in execution. UML is an main aspect concerned in object-oriented software progress. It uses photo notation to create visible models of software systems.

**USE CASE DIAGRAM**

The utilization case graph is alert in nature there ought to be some inward or outside elements for making the communication. These inside and outer operators are known as on-screen characters. So utilize case outlines are comprises of performing artists, use cases and their connections. The outline is utilized to display the framework/subsystem of an application. A solitary use case chart catches a specific usefulness of a framework. So to display the whole framework quantities of utilization case graphs are utilized appeared as a part of [Figure 5].

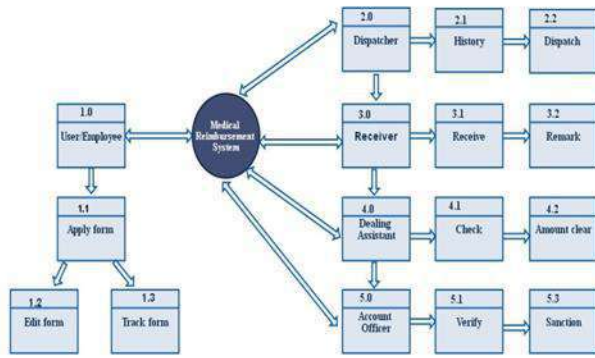


Fig. 4: Level 2 DFD

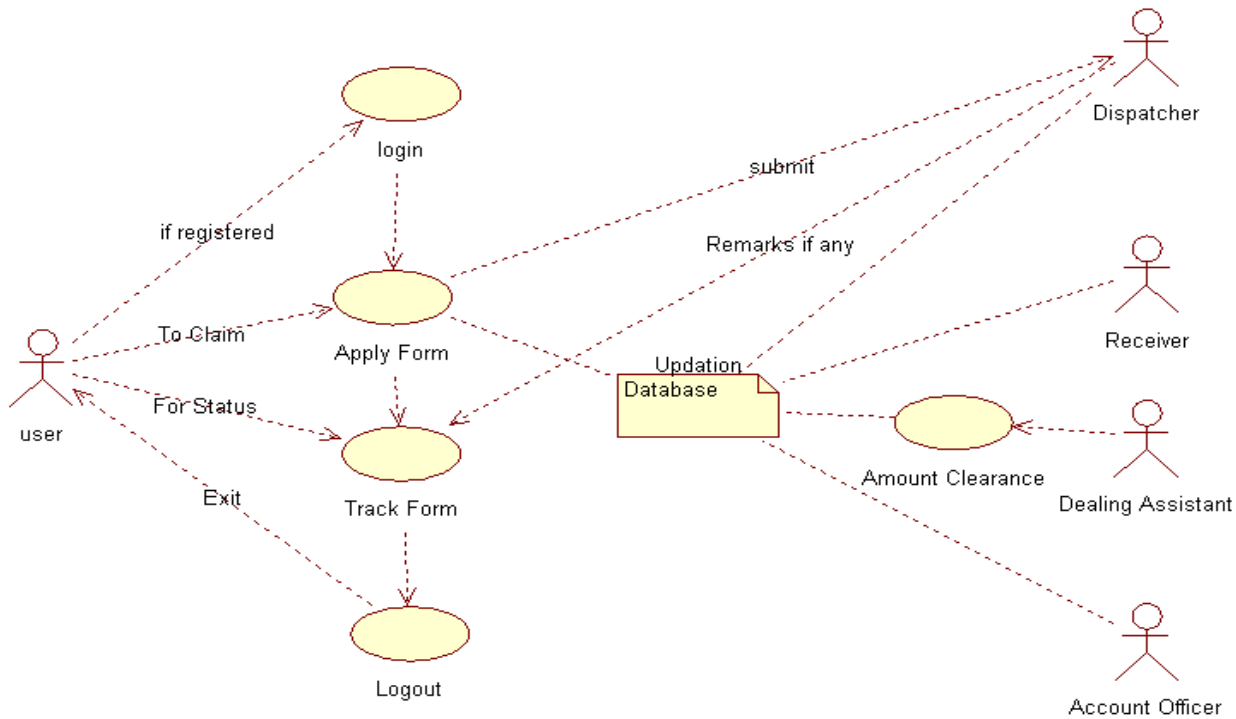


Fig. 5: Use Case Diagram

**STATE CHART DIAGRAM**

The name of the chart itself elucidates the motivation behind the graph and different subtle elements. It depicts distinctive conditions of a part in a framework. The states are particular to a segment/object of a framework. A State graph outline depicts a state machine. Presently to clear up it state machine can be characterized as a machine which characterizes diverse conditions of an article and these states are controlled by outside or inward occasion appeared in [Figure -6].

**CLASS DIAGRAM**

The class layout is a static chart. It identifies with the static point of view of an application. Class chart is not simply used for envisioning, portraying and reporting unmistakable parts of a system also to develop executable code of the item application. The class outline delineates the qualities and operations of a class moreover the restrictions constrained on the system. The class

blueprints are extensively used as a part of the showing of thing arranged structures since they are the fundamental UML diagrams which can be mapped clearly with article organized tongues. The class plot exhibits a gathering of classes, interfaces, affiliations, facilitated endeavors and goals. It is generally called an assistant outline showed up in [Figure 7].

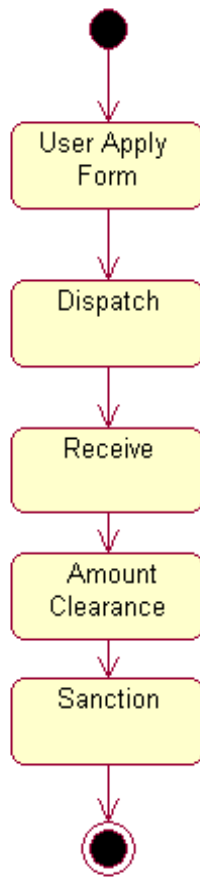


Fig. 6: State Chart Diagram

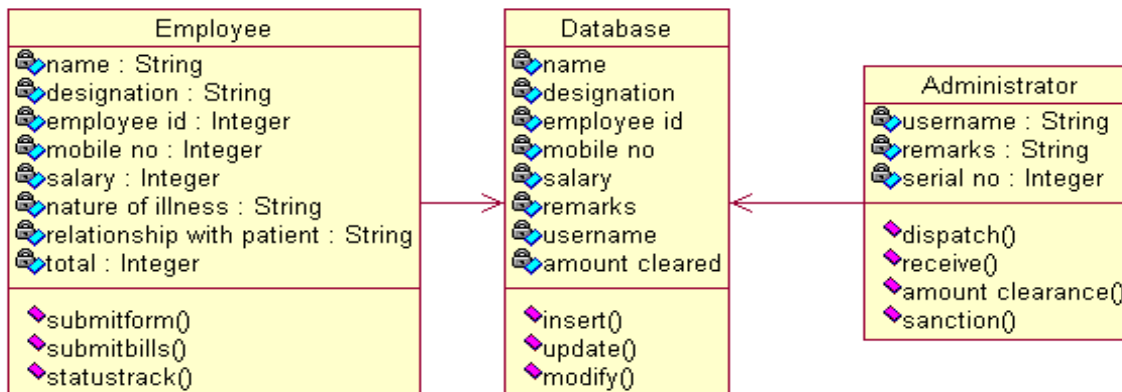


Fig. 7: Class Diagram

### SEQUENCE DIAGRAM

UML grouping charts are utilized to show how questions interface in a given circumstance. A vital normal for a grouping graph is that time goes through and through: The connection begins close to the highest point of the outline and closures at the bottom. A prevalent use for them is to archive the progress in an item situated framework. For every key joint effort, outlines are made that show how questions cooperate in different delegate situations for that coordinated effort appeared in [Figure 8].

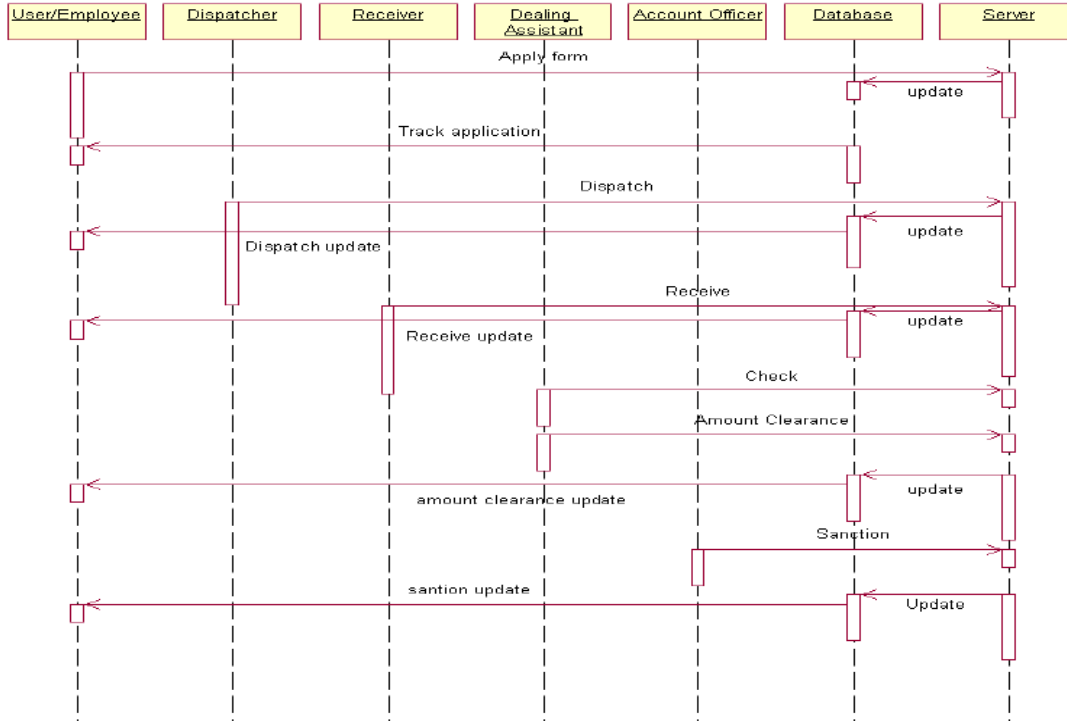


Fig. 8: Sequence Diagram

### COLLABORATION DIAGRAM

Collaboration diagram demonstrates the item association as demonstrated as follows. Here in Collaboration diagram the strategy call succession is demonstrated by some numbering system as demonstrated as follows. The number demonstrates how the techniques are called in a steady progression. We have taken the same request administration framework to portray the joint effort graph. The technique calls are like that of a succession outline. In any case, the distinction is that the arrangement chart does not portray the item association where as the joint effort graph demonstrates the article association appeared in [Figure -9].

### RESULTS

#### MODULES

##### USER LOGIN

- Apply for Medical Reimbursement
- Track user application

##### ADMIN LOGIN

- Dispatch and Receive user application
- Dealing Assistant check and amount clearance
- Sanction by Account officer.

#### MODULES DESCRIPTION:

##### USER LOGIN:

Apply for Medical Reimbursement

The User logs in through his/her employee id and fills in the Medical reimbursement form online and submits it. The hard copy of the submitted Application form, along with the original medical bills of the employee is submitted to the respective official for further processing of the claim.

Track user application

The Proposed system has an additional advantage for the user in a way of tracking his/her application. The User gets updated about the position of their application.

**ADMIN LOGIN**

Dispatch and Receive User Application

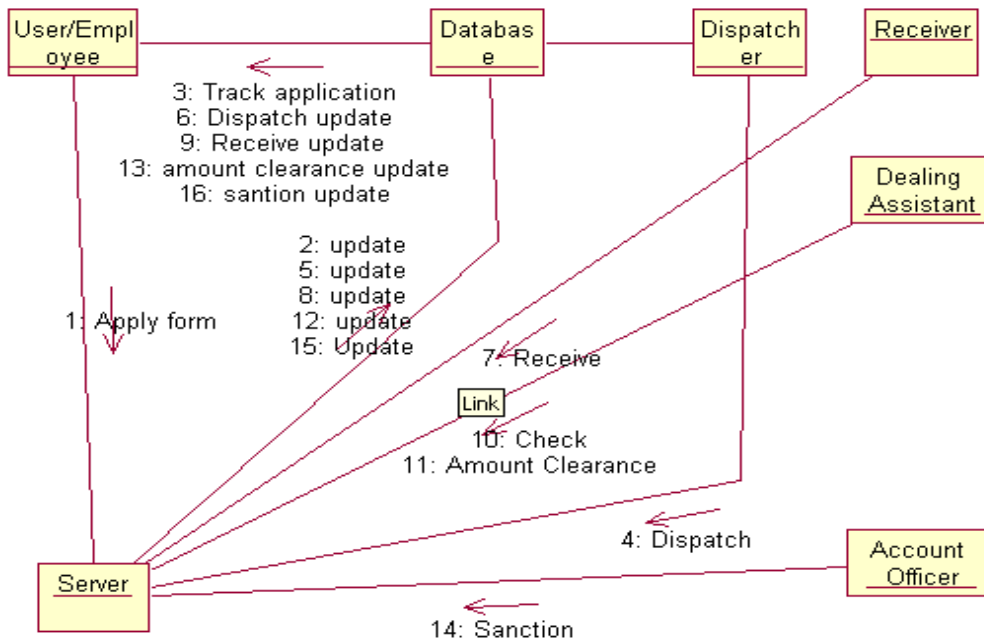
The final application form of the employee is checked along with the original medical bills by the Dispatcher for verification and remarks, if any are written and passed to the Receiver. The Receiver on receiving the form cross verifies the form and checks whether the imposed remarks of the dispatcher are met and forwarded to the Dealing assistant.

Dealing Assistant check and amount clearance

The Dealing assistant initially looks for the designation, salary and his/her relationship with the patient, based on the above mentioned details he calculates eligible amount of money to be cleared. The Dealing assistant passes the employee's application form with the amount eligible for sanctioning the claim.

Sanction by Account officer

The Account Officer checks the cleared amount quoted by the dealing assistant with the employee's medical bill amount. If all the requirements are met for sanctioning, the amount is sanctioned. Finally, the employee is reimbursed with the amount.



**Fig. 9: Collaboration Diagram**



## CONCLUSION

In this paper, we have tended to the issue of Medical Reimbursement preparing for BSNL. By method for devices, for example, Microsoft Front Page, XAMPP server and PHP, we have made a web application to apply for the case, dispense the case to the dispatcher and to redesign the status of every client. It likewise gives an office to store the repaid archive in the database, which can be recovered at whatever time for reference. Based upon the learning accumulated, we have additionally given an alternative to the client to check the repayment status on the web. Subsequently the proposed model has been created under different conditions and the outcomes are contrasted and the current framework. Along these lines this framework is produced to be more viable and effective.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] AntoBennet M, Thamilvalluvan B, Ashokram S, Sankarnarayanan S [2014] Efficient Energy Conservation Algorithm For Mobile Sensor Nodes in Wireless Sensor Networks, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 3 (8): 2650-2655.
- [2] AntoBennet M, Sankaranarayanan S, Ashokram S, Dinesh Kumar TR [2014] Testing of Error Containment Capability in can Network, *International Journal of Applied Engineering Research*, 9( 19): 6045-6054.
- [3] AntoBennet M, Nehru V [2014] An Architectural-Model for Mobile Based E-Learning Algorithm, *International Journal of Computer Science & Mobile Applications*, 2( 11):41-48.
- [4] AntoBennet M, Sankaranarayanan S [2015] Performance& Analysis of storage node in wireless networks, *International Journal of Computer & Modern Technology*, 2(2): 87-93.
- [5] R Agrawal and J Kiernan [2002] Watermarking Relational Databases, Proc. 28th Int'l Conf. Very Large Data Bases (VLDB '02), *VLDB Endowment*, pp. 155-166.
- [6] P Bonatti, SDC di Vimercati, and P Samarati [2002] An Algebra for Composing Access Control Policies, *ACM Trans. Information and System Security*, 5( 1): 1-35.
- [7] P Buneman, S Khanna, and WC Tan [2001] Why and Where: A Characterization of Data Provenance, Proc. Eighth Int'l Conf. Database Theory (ICDT '01), J.V. den Bussche and V. Vianu, eds, 316-330.
- [8] Panagiotis Papadimitriou and Garcia-Molina [2011] Data Leakage Detection, *Transaction Knowledge and Data Engineering*, 23(1) : 13-24.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and approved by guest editor.

# IMAGE FEATURE EXTRACTION OF K-MEANS CLUSTERING IMAGE SEGMENTATION TECHNIQUE FOR EARLY DETECTION OF DISEASES

Anto Bennet\*, Sankaranarayanan, Deepa, Banu, Priya

Dept. of ECE, Vel Tech, Chennai, INDIA

## ABSTRACT

Crop development is a center component of the field administration. Suitable assessment and determination of yield infection in the field is exceptionally basic for the expanded generation. Early malady discovery is a noteworthy test in agribusiness field. Henceforth legitimate measures must be taken to battle bioaggressors of yields while minimizing the utilization of pesticides. This proposed work depends on Image Segmentation methods utilizing K-means clustering as a part of which, the caught pictures are handled for advancement first. The K-Means grouping method is a surely understood methodology that has been connected to understand low-level picture division assignments. This bunching calculation is merged and its point is to improve the dividing choices in light of a client characterized introductory arrangement of groups that is redesigned after every emphasis. In the initial step we recognize the for the most part green shaded pixels. Next, these pixels are veiled in view of particular edge values that are figured utilizing Otsu's technique, then those for the most part green pixels are conceal. The other extra step is that the pixels with zeros red, green and blue qualities and the pixels on the limits of the tainted bunch (item) were totally uprooted. The test results exhibit that the proposed strategy is a vigorous system for the recognition of plant leaves ailments.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Image Segmentation, K-Means clustering, Crop diseases

\*Corresponding author: Email: [bennetmab@gmail.com](mailto:bennetmab@gmail.com) Tel.: +91 9965576501

## INTRODUCTION

A considerable measure of exploration has been done on nursery agro frameworks and all the more for the most part on secured harvests to control vermin and ailments by natural means rather than pesticides. Research in horticulture is pointed towards increment of efficiency and nourishment quality at decreased use and with expanded benefit, which has gotten significance in late time. A solid request now exists in numerous nations for non-concoction control techniques for vermin or illnesses. Nurseries are considered as biophysical frameworks with inputs, yields and control process circles. A large portion of these control circles are automatized (e.g., atmosphere and fertirrigation control). The administration of lasting organic product crops requires close checking particularly for the administration of illnesses that can influence generation altogether and consequently the post-harvest life. In the event of plant the illness is characterized as any disability of ordinary physiological capacity of plants, delivering trademark manifestations [1-4].

An indication is a wonder going with something and is viewed as proof of its presence. Sickness is brought about by pathogen which is any operators bringing on malady. In the greater part of the cases vermin or illnesses are seen on the leaves or stems of the plant. Thusly distinguishing proof of plants, leaves, stems and discovering the bug or infections, rate of the irritation or sickness frequency, manifestations of the nuisance or illness assault, assumes a key part in effective development of products. In natural science, in some cases a huge number of pictures are created in a solitary examination. These pictures can be required for further studies like characterizing injury, scoring quantitative characteristics, ascertaining territory eaten by creepy crawlies, and so forth. All of these assignments are prepared physically or with particular programming bundles. It is colossal measure of work as well as experiences two noteworthy issues: extreme handling time and subjectiveness ascending from various people [5-8].

## METHODS

Clustering is the process of partitioning a group of data points into a small number of clusters. Image analysis can be applied for the following purposes:

- To detect diseased leaf, stem, fruit

- To quantify affected area by disease.
- To find the boundaries of the affected area.
- To determine the color of the affected area.

The K-Means clustering algorithm is proposed by Mac Queen in 1967 which is a segment based group examination strategy. It is utilized generally as a part of group examination for that the K-implies calculation has higher effectiveness and adaptability and merges quick when managing huge information sets [9]. . In the initial step we recognize the for the most part green shaded pixels. Next, these pixels are covered in light of particular limit values that are figured utilizing Otsu's strategy, then those generally green pixels are veiled. The other extra step is that the pixels with zeros red, green and blue qualities and the pixels on the limits of the contaminated bunch (item) were totally evacuated. Be that as it may it additionally has numerous efficiencies: the quantity of groups K should be instated, the introductory bunch focuses are self-assertively chose, and the calculation is affected by the commotion focuses. In perspective of the deficiencies of the customary K-Means bunching calculation, this paper exhibits an enhanced K-implies calculation utilizing cluster information channel. The calculation created thickness construct identification strategies based with respect to attributes of commotion information where the disclosure and handling ventures of the cluster information are added to the first calculation. By preprocessing the information to prohibit these cluster information before grouping information set the clustering of the grouping results is enhanced fundamentally and the effect of commotion information on K-means calculation is diminished viably and the clustering results are more exact[11].

### Steps for disease detection

- RGB image acquisition.
- Create the colour transformation structure.
- Convert the colour values in RGB to the space specified in the colour transformation structure.
- Apply K-means clustering.
- Masking green-pixels.
- 6.Remove the masked cells inside the boundaries of the infected clusters.
- Convert the infected (cluster / clusters) from RGB to HSI Translation.
- SGDM Matrix Generation for H and S.
- Calling the GLCM function to calculate the features.
- Texture Statistics Computation.

The proposed approach step - by - venture of the picture division and acknowledgment procedures is represented in Algorithm 1. In the beginning step, the RGB pictures of all the leaf tests were grabbed. Some genuine examples of those maladies are appeared in [Figure -2]. It is evident from [Figure-2] that leaves fitting in with ahead of schedule sear, cottony mold, powder-colored mold and late singe have noteworthy contrasts structure oily spot leaves as far as shading and surface. Likewise, Figure indicates two pictures; the left picture is contaminated with modest whiteness sickness, and the right picture is a typical picture. Notwithstanding, the leaves identified with these six classes (early burn, cottony mold, colorless mold, late singe, minor whiteness and ordinary) had little contrasts as recognizable to the human eye, which might legitimize the misclassifications in light of bare eye [9].

In points of interest, in step 2 a shading change structure for the RGB leaf picture is made, and afterward, a gadget autonomous shading space change for the shading change structure is connected in step 3. Steps 2 and 3 are inescapable for completing step 4. In this stride the current pictures are fragmented utilizing the K-Means bunching method. These four stages constitute stage 1 though, the tainted article is/are resolved. In step 5, we distinguish the for the most part green shaded pixels. After that, taking into account determined and shifting edge esteem that is processed for these pixels utilizing Otsu's strategy, these for the most part green pixels are covered as takes after: if the green segment of pixel intensities is not exactly the pre-registered edge esteem, the red, green and blue segments of the this pixel is relegated to an estimation of zero. This is done in sense that these pixels have no significant weight to the sickness ID and grouping steps, and most likely those pixels speak to solid zones in the leave. Besides, the picture handling time ought to wind up essentially decreased. In step 6 the pixels with zeros red, green and blue qualities and the pixels on the limits of the contaminated bunch (item) were totally evacuated. Steps 5 and 6 structure stage 2, and this stage is useful as it gives more exact ailment arrangement and distinguishing proof results with fulfilled execution and the general calculation time ought to wind up essentially less. The perceptions behind steps 5 and 6 were tentatively accepted. Next, in step 7 the tainted group was then changed over from RGB configuration to HSI position. In the following step, the SGDM grids were then produced for every pixel guide of the picture for just H and S pictures. The SGDM is a measure of the likelihood that a given pixel at one specific dark level will happen at an unmistakable separation and introduction edge from another pixel, given that pixel has a second specific dim level. From the SGDM grids, the surface insights for every picture were created. Briefly, the components set were figured just to pixels inside the limit of the contaminated zones of the leaf. As such, solid territories inside the tainted zones were likewise uprooted. Steps 7 – 10 structure stage 3 in which the surface elements for the portioned tainted items in this stage are figured. At long last, the acknowledgment process in the fourth stage was performed to the separated components through a pre-prepared neural system. For every picture in the information set the resulting ventures in Algorithm 1 were rehashed. The picture information of the leaves chose for this study would be gathered. Calculations in light of picture handling strategies for highlight extraction and order would be outlined. Manual bolstering of the datasets, as digitized RGB shading photos would be ruined element extraction and preparing the SAS factual classifier. Subsequent to preparing the SAS classifier, the test information sets would be utilized to dissect the execution of precise arrangement. The entire methodology of investigation would be duplicated for three substitute arrangement ways to deal with incorporate; measurable classifier utilizing the ahalanobis least separation technique, neural system based classifier utilizing the back engendering calculation and neural system based classifier utilizing spiral premise capacities.

Correlation of the outcomes acquired from the three methodologies would be finished and the best approach for the current issue would be resolved

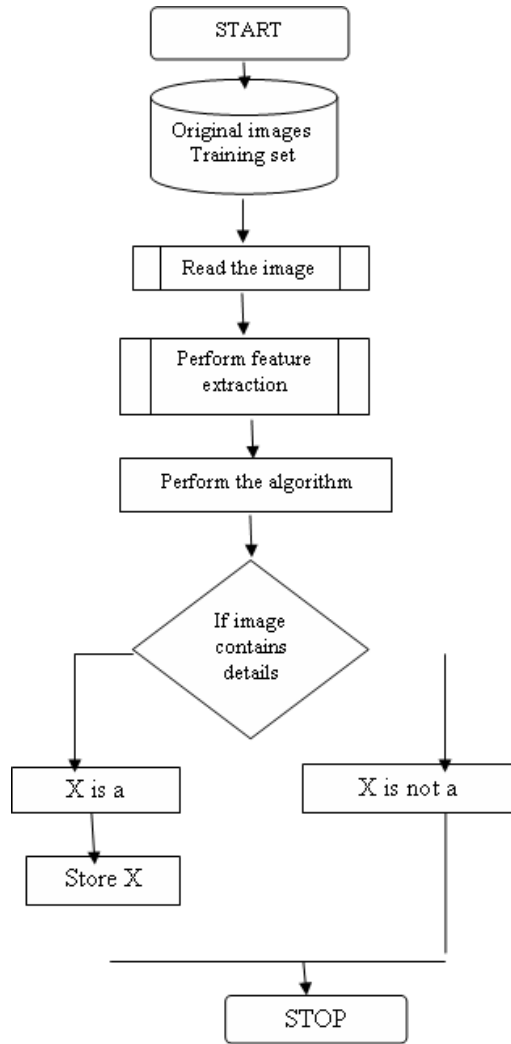


Fig: 1. Algorithm 1- Basic steps describing the proposed algorithm

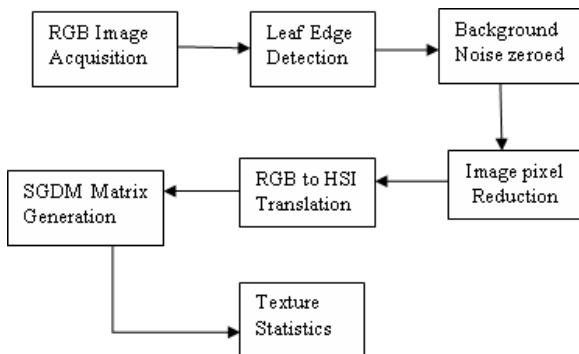


Fig: 2. Image acquisition and image Classification chart

## K-MEANS CLUSTERING ALGORITHM

K-means clustering algorithm is simply described as follows:

**Input:** N objects to be cluster ( $x_1, x_2, \dots, x_n$ ), the number of clusters  $k$ ;

**Output:**  $k$  clusters and the sum of dissimilarity between each object and its nearest cluster centre is the smallest;

- Arbitrarily select  $k$  objects as initial cluster centres ( $m_1, m_2 \dots m_k$ );
- Calculate the distance between each object  $X_i$  and each cluster centre, and then assign each object to the nearest cluster, formula for calculating distance as:  $d(X_i, m_j)$  is the distance between data  $i$  and cluster  $j$ ;
- Calculate the mean of objects in each cluster as the new cluster centres,  $N_i$  is the number of samples of current cluster  $i$ ;
- Repeat 2 & 3 until the criterion function  $E$  converged, return ( $m_1, m_2 \dots m_k$ ).

### Advantages of K-Means Clustering

- This algorithm is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is  $O(nkt)$ , where  $n$  is the number of objects,  $k$  is the number of clusters, and  $t$  is the number of iterations.
- It works well when the clusters are compact clouds that are rather well separated from one another.
- The algorithm is not only simple, but also the results are easily understandable and it can be easily modelled to deal with streaming data.
- Continual improvements and generalizations of the algorithm have ensured its continued relevance and gradually increased its effectiveness as well.

## RESULTS

Results snapshots are shown in [Figure -3], [Figure -4], [Figure -5], [Figure -6], [Figure -7], [Figure -8], [Figure -9] and [Figure -10].

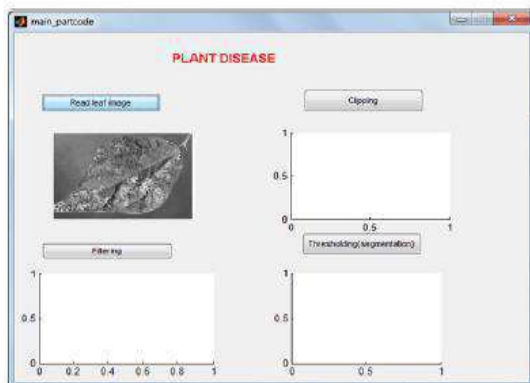


Fig. 3. Taking infected image as input

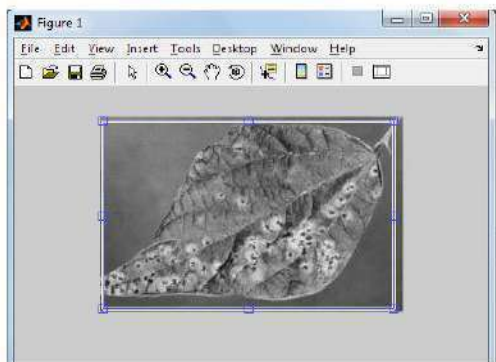


Fig. 4. Selected Crop section.

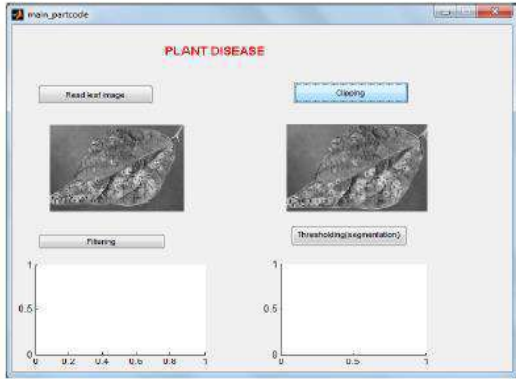


Fig: 5.Clipping section of diseased leaf

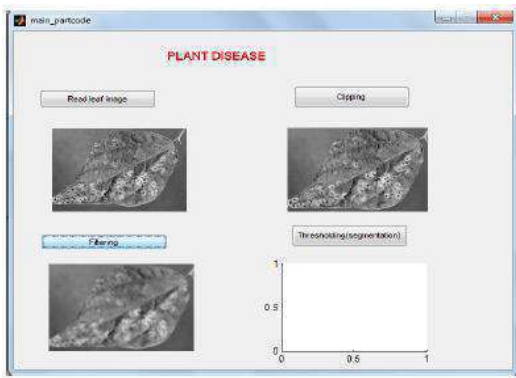


Fig: 6.Filtering of diseased leaf

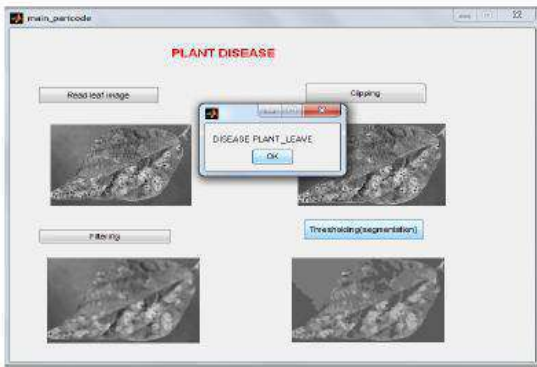


Fig: 7.Segmented output & infected part of leaf is detected

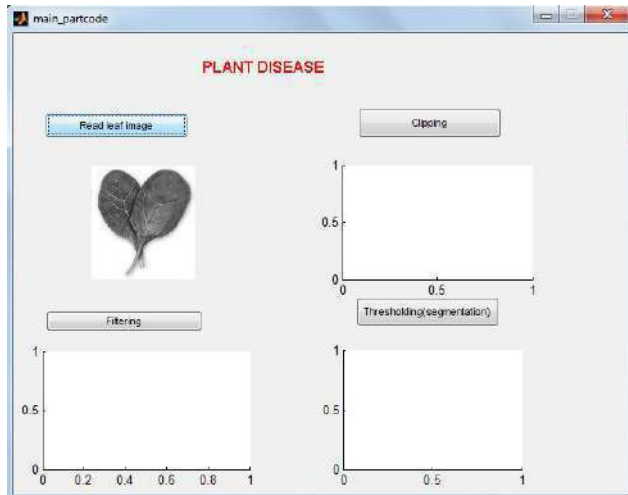


Fig: 8. Reading image of normal leaf

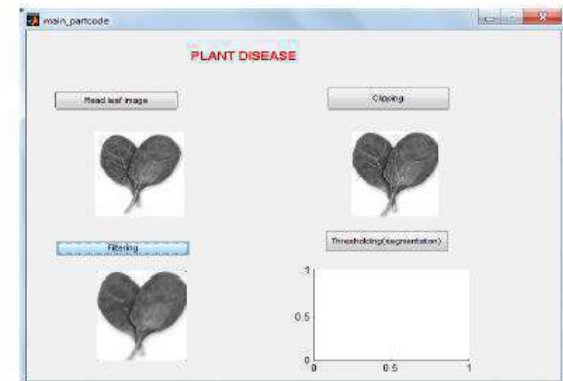


Fig: 9. Normal leaf image after filtering

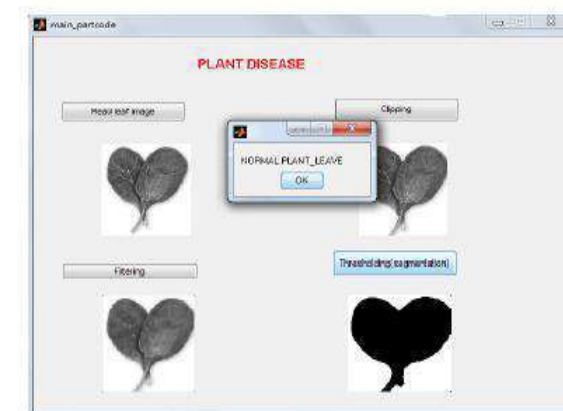


Fig: 10. Image of normal leaf after segmentation

### CONCLUSION

We exhibit a general k-means based clustering calculation that can distinguish common groups in datasets, whether they are inserted in the first space or subspaces. Like conventional k-means calculation, the time many-

sided quality of the calculation is straight with the quantity of the information focuses, the dimensionality of the information, and the quantity of groups in the dataset. The test results demonstrate that our calculation is a proficient calculation with high bunching precision. Bunching investigation strategy is one of the primary systematic techniques in information mining; the strategy for grouping calculation will impact the bunching comes about straightforwardly. Standard renditions of k-means calculations appear be better in discovering high wellness arrangements. In the same time results acquired in standard and hereditary forms of k-means calculations with respect to legitimacy files are additionally equivalent.

The outcomes exhibited in this paper are promising however a few enhancements in both material and techniques can be completed to achieve the necessities of an Integrated Pest Management framework. In future the component extraction of picture will be done. From this outcomes sort, shape, shading, surface of irritation will be distinguished. From these measures what preventive activity against irritation ought to be taken will be chosen through which the creation of products can be expanded. Amid broad inquiry of arrangement space, Genetic adaptations of k-means calculations frequently discover arrangements with somewhat more regrettable wellness values yet in the meantime with incredibly great estimations of individual legitimacy files. Further examination concerning this matter could display beginning stage into change of k-means based picture bunching procedures.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Digital image processing Using MATLAB codes by Dhananjay Theckedath Tech-Max Publication.
- [2] Ali SA, Sulaiman N, Mustapha A and Mustapha N, [2009] K-means clustering to improve the accuracy of decision tree response classification. *Inform. Technol J*, 1256-1262.
- [3] Hillnhuetter C. and AK Mahlein. [2008] Early detection and localisation of sugar beet diseases: new approaches, *Gesunde Pflanz* 60(4)
- [4] B Cunha. [2003] Application of Image Processing in Characterisation of Plants, *IEEE Conference on Industrial Electronics*.
- [5] AntoBennet M, JacobRaglend.[2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, *Journal of Computer Science*, 8( 9): 1447-1454.
- [6] AntoBennet, M & JacobRaglend (2011) Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering, *European Journal of Scientific Research*, 66 (1):120-129.
- [7] Anto Bennet M, Mohan babu G, Rajasekar C, Prakash P. [2015] Performance and Analysis of Hybrid Algorithm for Blocking and Ringing Artifact Reduction, *Journal of Computational and Theoretical nanoscience* 12(1):141-149.
- [8] AntoBennet M, Jacob Raglend. [2013] Performance and Analysis of Compression Artifacts Reduction for MPEQ-4 Moving Pictures Using TV Regularization Method, *Life Science Journal*, 10(2): 102-110.
- [9] P. Thangaraj K. Geetha , FGT2- ABR: Fuzzy Game Theory Trust Associativity Based Routing to Mitigate Network Attacks in Pervasive Health Monitoring Systems, *Journal Of Pure And Applied Microbiology*, Volume 9, Pages 161-168, 2015.
- [10] Banuchandar J, S Deepa N Tamilarasi, and J Parkavi.,Eye for the visually impaired." *International Journal of Modern Engineering Research*, 2(2) :368-372
- [11] AntoBennet M, JacobRaglend (2012) A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach, *Journal of Applied Sciences Research*.8(5): 2429-2438.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and approved by guest editor.



# PERFORMANCE EVALUATION OF ENHANCING THE CAPACITY OF SPECTRUM SHARING COGNITIVE RADIO NETWORKS

Anto Bennet<sup>1,\*</sup>, Sankaranarayanan<sup>1</sup>, Ramesh Kandasamy<sup>2</sup>, Aruna<sup>1</sup>, Kavitha<sup>1</sup>, Thamizhoviya<sup>1</sup>

<sup>1</sup>Dept. of ECE, VEL Tech, Chennai, TN, INDIA

<sup>2</sup>Dept. of IT, Nandha Engg College, Erode, Chennai, TN, INDIA

## ABSTRACT

In spectrum-sharing cognitive radio systems, the transmit power of secondary users (SU) has to be very low due to the limitations on the interference power dictated by primary users (PU). In order to enlarge the coverage area of secondary transmission and reduce the corresponding interference region, multi-hop amplify-and-forward (AF) relaying can be implemented for the communication between secondary transmitters and receivers. Monte Carlo simulation is a method proposed in this project for iteratively evaluating a deterministic model using sets of random numbers as inputs. The optimal power allocation is employed to allocate the transmit power of secondary users (SU) to avoid the interference at the primary user (PU). The performance can be calculated for different number of hops in terms of probability and interference power at the primary user with the signal to noise ratio using an amplify and forward (AF) relay protocol.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Amplify and forward (AF),  
Cognitive radio (CR),  
Cooperative relaying, multi-hop  
relaying, power allocation,  
spectrum sharing

\*Corresponding author: Email: [bennetmab@gmail.com](mailto:bennetmab@gmail.com) Tel.: +91 9965576501

## INTRODUCTION

Lately, the field of remote correspondence frameworks has demonstrated a huge measure of improvement concerning research and practice. Applications range from the day by day needs like mobiles, Wi-Fi, to business utilizes like satellite correspondences. With the guide of current innovation, it is conceivable to correspond with any side of the world. These innovations require a dependable and web framework for better execution. Right now clients are being locked in by the administrations of various accessible remote access frameworks. Especially, a number of new systems are capable of using not only the 800 MHz to 6000 MHz band which is suitable for broadband wireless access systems and for cellular communications but also the frequency bands such as the very high frequency (VHF) and ultra high frequency (UHF) bands. It seems that after around ten years, the majority of frequency bands, suitable for mobile communication systems, are entirely engaged and new solutions are compulsory. One of the possible solutions is to use the "Cognitive Radio" technology which is a radio or system, that is able to sense and that is fully aware of its functioning situation and can regulate its radio operating parameters autonomously according to collaborating wireless and wired networks. In order to more efficiently use the available spectrum on the frequency band, this technology is expected as a key technology.

Cognitive radio (CR) is a promising remote innovation to determine the developing lack of the essential electromagnetic range assets. By utilization of CR, secondary users (SUs) without expressly allotted range assets can exist together with primary users (PUs) authorized with specific range. By and by, some real correspondences controllers like the Federal Communications Committee (FCC) in U.S. what's more, the Office of Communications (OFCOM) in U.K. have permitted secondary access for unlicensed gadgets to the physical TV telecast groups.

Among different types of CR usage, range sharing CR is particularly engaging for down to earth arrangement since it doesn't include complex range detecting instruments. All the more particularly, range sharing CR restricts just the transmit force of SUs such that their hurtful obstruction onto PUs stays beneath recommended mediocre levels. As a result of the obstruction power limitation directed by PUs, the transmit force of SUs in range sharing frameworks must be low, which constrains the scope territory of secondary transmission. To amplify the scope range of secondary transmission and certification solid correspondence, helpful transferring procedures can be abused.

Utilizing handing-off strategies, a solitary or different unmoving users can be included in sending messages between a secondary source and its destination.

## MATERIALS AND METHODS

### SYSTEM MODEL

The last decade has witnessed the increasing popularity of wireless services. In fact, recent measurements by Federal Communications Commission (FCC) have shown that 70% of the allocated spectrum in US is not utilized. CR is a kind of intelligent wireless device, which is able to adjust its transmission parameters, such as transmit power and transmission frequency band, based on the environment. In a CR network, ordinary wireless devices are referred to as primary users (PUs), and CRs are referred to as secondary users (SUs). CR is defined as an intelligent wireless communication system that provides more efficient communication by allowing secondary users to utilize the unused spectrum segments.

A K-hop cooperative relaying system operating in a spectrum sharing cognitive radio (CR) environment is considered. The secondary users and primary users exchange data with some consecutive Amplify and Forward (AF) relay. All nodes are equipped with a single half duplex omni directional antenna. For the secondary multihop AF relaying link, all SUs work in a time division multiple access (TDMA) fashion. Only one SU transmits to its next node along the multiple path during each time slot. The [Figure - 1] describes the system model for spectrum sharing with the distance and the fading coefficient for the desired link and the interference link where(  $d_k, f_k$  ) – Distance and the channel fast fading between SU $k-1$  and SU $k$  (desired link) ( $l_k, h_k$  ) – Distance and the channel coefficient between SU $k-1$  and PU1(interference link)

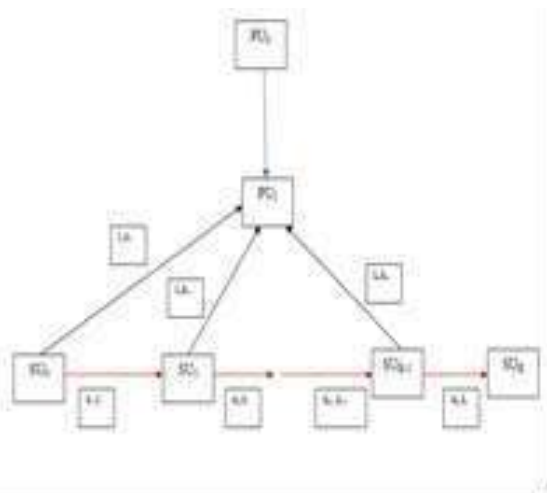


Fig: 1. System model

For the secondary multi-hop AF relaying link, all SUs work in a time division multiple access (TDMA) fashion. Only one SU transmits to its next node along the multiple path during each time slot. The received signal to noise ratio (SNR) at the Kth secondary node is defined as:

$$\gamma_k = (P_{k-1} - \sigma^2) d_k^{-\alpha} f_k^2 \quad (1)$$

where  $\gamma_k$  is the received SNR,  $P_{k-1}$  is the transmit power of secondary users,  $\sigma^2$  is the additive white Gaussian noise (AWGN) variance,  $d_k$  is the distance,  $f_k$  is the channel fast fading coefficient. The received SNR for the secondary users are calculated based on the different parameters. The interference coming from the primary transmitter is treated as noise.

### OPTIMAL POWER ALLOCATION

The criterion for optimal power allocation at the secondary nodes is established. The average tolerable interference power at the primary user is  $W$  dB. The optimal power is constrained and this tolerable interference power is based on the interference power or average peak power. The power allocation parameter is determined based on the average interference power and it is expressed as

$$E[\lambda_{k-1} - (\sigma^2 / \eta_k)(h_k^2 / f_k^2)] = 10W / 10 \quad (2)$$

Where  $\lambda_{k-1}$  is the optimal power,  $\sigma^2$  is the variance,  $\eta_k$  is the path loss ratio,  $h_k$  is the fading coefficient of the interference link,  $f_k$  is the fading coefficient of the desired link,  $W$  is the tolerable interference power at the primary user.

### CHANNEL ESTIMATION

The channel estimation depends on detecting the primary and secondary users. The quantity of channels are allotted for the transmission of the information outlines. The channel state data is given to every client. The detecting time depends on the discovery likelihood.

The quantity of diverts allotted in this task is 8. The clamor and the force level can be evaluated for every channel.

**PERFORMANCE ANALYSIS**

Consider the quantity of hops as a significantly number with K=4. The multihop connection is opposite by utilizing an open up and forward transfer convention. In this handing-off plan, the hand-off sends an opened up rendition of the got signal in the last time-space. For helpful correspondence, AF plans give spatial assorted qualities to fight against blurring; for limit estimation of transfer systems, such plans give achievable lower limits that are known not ideal in some correspondence situations and for simple system coding, given the telecast way of the remote medium that permits the blending of the signs noticeable all around, these plans give a correspondence procedure that accomplishes high throughput with low computational unpredictability at interior nodes. The way path loss worth is taken as 10 and it is utilized to ascertain the separation parameters for the coveted connection and the obstruction join. The way path loss is given as

$$\eta = dk^{-\epsilon} / lk^{-\epsilon} \tag{3}$$

where, dk and lk are the distance parameters, ε is the path loss exponent and the value is 4. If the multihop link is perpendicular to the interference link then the value is normalized to unity. For K=4 hops the distance between the PU1 and SU2 is normalized to unity. The various distance parameters are calculated using the path loss ratio. The monte carlo simulation method is used in which the gain of the channel is subject to Rayleigh distribution with unit mean and the variance of AWGN at all nodes is set to unity.

A Monte Carlo method is a technique that involves using random numbers and probability to solve problems. Monte Carlo simulation is a method for iteratively evaluating a deterministic model using sets of random numbers as inputs. This method is often used when the model is complex, nonlinear, or involves more than just a couple uncertain parameters. The [Table -1] shows the value for different distance parameters and it is calculated based on the path loss ratio.

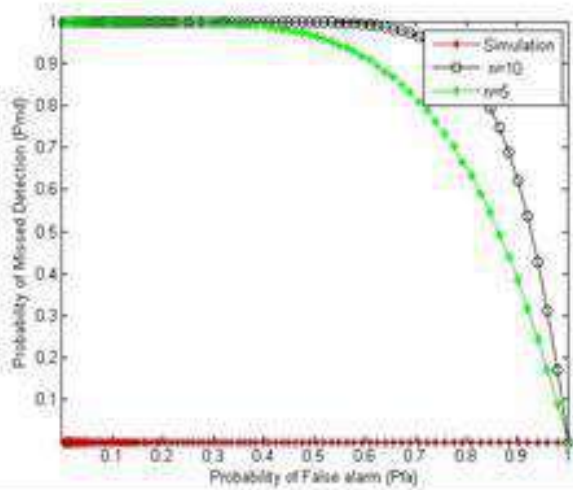
**Table: 1. Distance Parameters**

Parameters	Value
d <sub>1</sub>	1.6
d <sub>2</sub>	0.6
d <sub>3</sub>	0.56
d <sub>4</sub>	0.04
l <sub>1</sub>	2.12
l <sub>2</sub>	1.06
l <sub>3</sub>	1
l <sub>4</sub>	1.14

The SNR of the received PU signal at the sensor depends on the PU transmitted power and the propagation environment. The two error probabilities are linked to each other through sensing time, SNR, and detection threshold. The detection performance improves with an increase in the SNR. After determining the SNR value as 15 dB then the optimal power is allocated to each secondary user under some constraints. The probability is determined based on probability density function.

The tolerable interference power is measured with respect to the optimal power and the distance parameters. The total transmit power at the secondary user is 10 and the interference power is found as W=30 dB. It indicates the increase in tolerable interference power leads to the better performance for different hops in multipath.

**RESULTS**



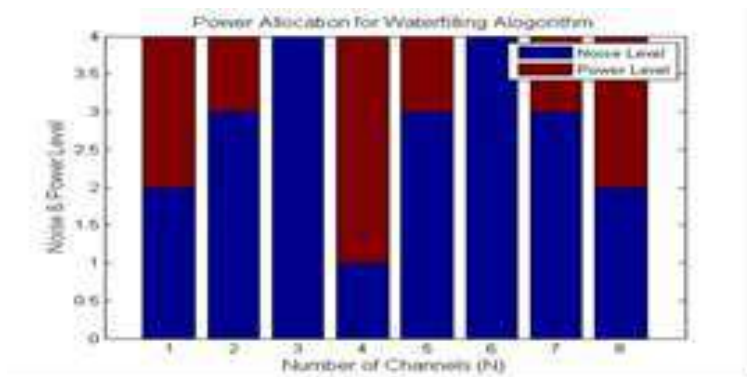
**Outputs**

In this paper, the performance of cooperative spectrum sharing is evaluated for different number of hops using amplify and forward (AF) relaying protocol.

**Fig: 2. Performance of Spectrum sharing**

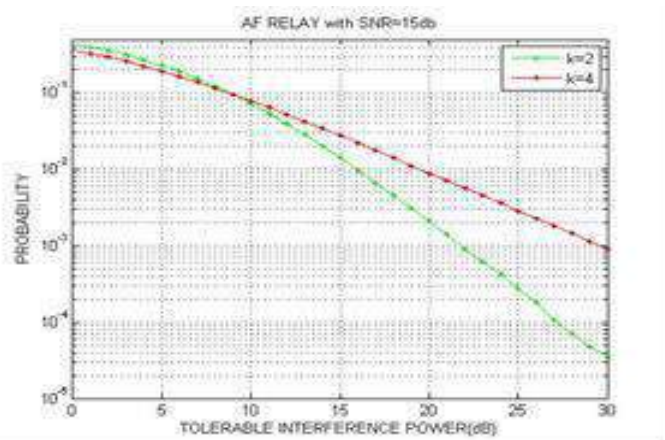
The [Figure -2] indicates the performance of spectrum sharing for the different number of cognitive radio users. From the above graph, it is very clear that the alarm increases with the detection probability and it indicates that the channel can be reused when it is available. The probability of detection is the time during which the PU (licensed) is detected. If the sensing time is increased then PU can make better use of its spectrum and the limit is decided that SU can't interfere during that much of time. More the spectrum sensing more PUs will be detected and lesser will be the interference because PU can make best use of their priority right.

To avoid the interference at the primary user the transmit power of secondary users has to be very low, so that the optimal power allocation algorithm is used to allocate the power optimally under maximum and minimum conditions. The optimal power is constrained and this tolerable interference power is based on the interference power or average peak power. The number of channels allocated is 8. The noise and the power level can be estimated for each channel. The below [Figure -3] represents the power allocation using the water filling for 8 channels and the various noise and power levels can be estimated.



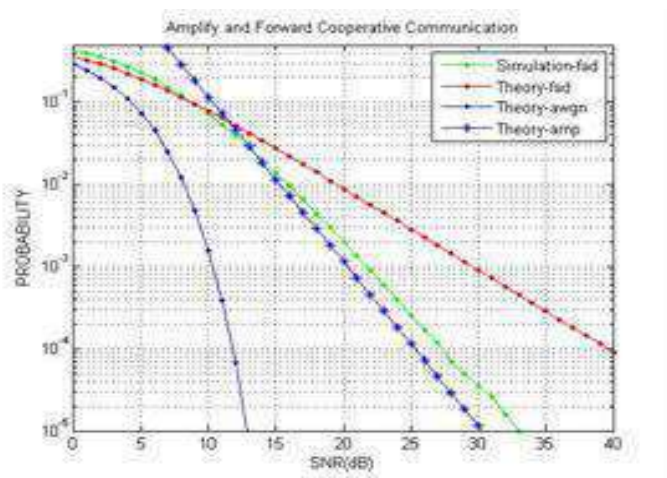
**Fig: 3. Optimal power allocation**

After the power assignment is done, the execution can be resolved concerning the likelihood and the normal impedance power. The amplify and forward (AF) transfer convention is utilized for helpful transferring as a part of multihop networks .The SNR worth is measured as 15 dB and it is utilized to gauge the interference power or normal top force which is termed as middle of the road impedance power at the primary user. The beneath [Figure -4] demonstrates the likelihood debases when there is an expansion in decent obstruction power for various number of jumps as  $k=2, 4$ . As the bounces increments in the system, the scope zone for transmission can be amplified.



**Fig: 4. Performance of interference power with probability**

The below [Figure -5] shows the performance of AF relay under different channels with SNR and probability.



**Fig: 5. Performance of Amplify and Forward relay**

## CONCLUSION

In this project, the performance of cooperative spectrum sharing in cognitive radio is analyzed using multihop relay networks. The coverage area can be extended using the multihop cooperative relaying. The amplify and forward relaying protocol improves the performance of the multihop network and it is simple when compared with the other techniques. To avoid the interference at the primary user, the transmit power of secondary users has to be low, so optimal power allocation is done at the secondary user. The signal to noise ratio (SNR) value is 15dB, and the results are analyzed for different hops in terms of tolerable interference power and probability.

In future, the analysis of cooperative spectrum sharing in multihop networks can be done for different relaying protocols. The optimal power allocation can be used to limit the transmit power of secondary users for the tolerable interference at the primary user with the different number of hops.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Ernesto Zimmermann, Patrick Herhold, Gerhard Fettweis, On the Performance of Cooperative Relaying Protocols in Wireless Networks.
- [2] Hong J, Hong B, Ban T, and Choi W. [2012] On the cooperative diversity gain in underlay cognitive radio systems, *IEEE Trans. Commun.*, 60(1): 209–219.
- [3] Lee J, Wang H, Andrews JG, and Hong D. [2011] Outage probability of cognitive relay networks with interference constraints, *IEEE Trans Wireless Commun*, 10( 2): 390–395.
- [4] AntoBennet M, JacobRaglend. [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, *Journal of Computer Science*, 8(9): 1447-1454.
- [5] AntoBennet M, JacobRaglend .[2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering, *European Journal of Scientific Research*, 66(1):120-129.
- [6] Anto Bennet, M, Mohan babu, G, Rajasekar, C & Prakash, P [2015] Performance and Analysis of Hybrid Algorithm for Blocking and Ringing Artifact Reduction, *Journal of Computational and Theoretical nanoscience*,12(1):141-149.
- [7] AntoBennet, M & JacobRaglend [2013] Performance and Analysis of Compression Artifacts Reduction for MPEQ-4 Moving Pictures Using TV Regularization Method, *Life Science Journal* , 10( 2): 102-110.
- [8] AntoBennet M, JacobRaglend. [2012] A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach, *Journal of Applied Sciences Research*,8(5): 2429-2438.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# INVESTIGATION OF AUDIO SIGNALS WATERMARKING USING EMPIRICAL MODE DECOMPOSITION

Anto Bennet<sup>1,\*</sup>, Sankaranarayanan<sup>1</sup>, Ramesh Kandasamy<sup>2</sup>, Rathna<sup>1</sup>, Jaishreebai<sup>1</sup>, Nalini<sup>1</sup>

<sup>1</sup>Dept. of ECE, VEL TECH, Chennai, INDIA

<sup>2</sup>Dept. of IT, Nandha Engg College, Erode, Chennai, INDIA

## ABSTRACT

In this paper a new adaptive audio watermarking algorithm based on Empirical Mode Decomposition (EMD) is proposed. The audio signal is separated into frames and each one is decomposed adaptively, by EMD, into intrinsic oscillatory components called Intrinsic Mode Functions (IMFs). The watermark and the synchronization codes are embedded into the extrema of the last IMF, a low frequency mode stable under different attacks and preserving audio perceptual quality of the host signal. The data embedding rate of the proposed algorithm is 46.9–50.3 b/s. Relying on exhaustive simulations, we show the robustness of the hidden watermark for additive noise, MP3 compression, re-quantization, filtering, cropping and resampling. The comparison analysis elucidates that our method has better performance than watermarking schemes reported recently.

Published on: 10<sup>th</sup>– August-2016

### KEY WORDS

Empirical Mode Decomposition,  
Intrinsic Mode Functions,  
Synchronization Codes.

\*Corresponding author: Email: [bennetmab@gmail.com](mailto:bennetmab@gmail.com) Tel.: +91 9965576501

## INTRODUCTION

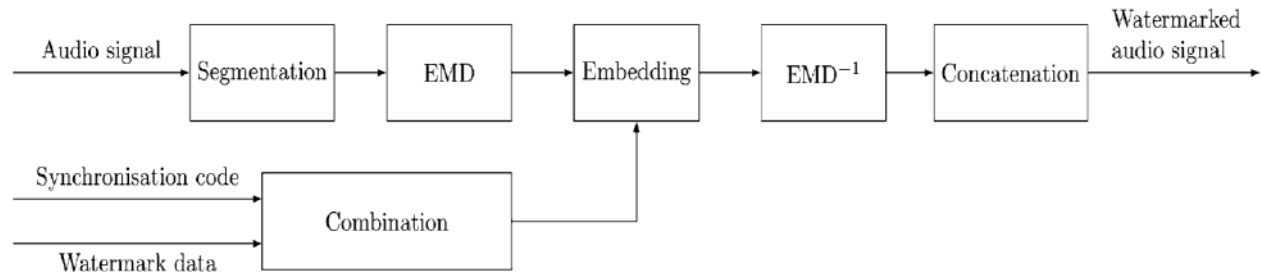
Digital audio watermarking has received a great deal of attention in the literature to afford efficient solutions for copyright protection of digital media by embedding a watermark in the original audio signal. Main necessities of digital audio watermarking are imperceptibility, robustness and data capacity. More precisely, the watermark must be inaudible within the host audio data to maintain audio quality and robust to signal distortions applied to the host data. Finally, the watermark must be easy to extract to prove ownership. To achieve these requirements, seeking new watermarking schemes is a very challenging problem. Different watermarking techniques of varying complexities have been proposed in a robust watermarking scheme to different attacks is proposed but with a limited transmission bit rate. To advance the bit rate, watermarked schemes performed in the wavelets domain have been proposed [1]. A limit of wavelet approach is that the basis functions are fixed, and thus they do not necessarily match all real signals. To overcome this limitation, recently, a new signal decomposition method referred to as Empirical Mode Decomposition (EMD) has been introduced for analyzing non-stationary signals derived or not from linear systems in totally adaptive way [2-4]. A major advantage of EMD relies on no a priori choice of filters or basis functions. Compared to classical kernel based approaches, EMD is fully data-driven method that recursively breaks down any signal into a reduced number of zero-mean with symmetric envelopes AM-FM components called Intrinsic Mode Functions (IMFs). The decomposition starts from finer scales to coarser ones. Any signal is expanded by EMD as follows:

$$x(t) = \sum_{j=1}^C IMF_j(t) + r(t)$$

where C is the number of IMFs and  $r(t)$  denotes the final residual. The IMFs are nearly orthogonal to each other, and all have nearly zero means [5]. The number of extrema is decreased when going from one mode to the next, and the whole decomposition is guaranteed to be completed with a finite number of modes. The IMFs are fully described by their local extrema and thus can be recovered using these extrema. Low frequency components such as higher order IMFs are signal dominated and thus their alteration can lead to degradation of the signal. As result, these modes can be considered to be good locations for watermark placement. Some preliminary results have appeared recently in showing the interest of EMD for audio watermarking. In [5], the EMD is combined with Pulse Code

Modulation (PCM) and the watermark is inserted in the final residual of the subbands in the transform domain. This method supposes that mean value of PCM audio signal may no longer be zero. As stated by the authors, the method is not robust to attacks such as band-pass filtering and cropping, and no comparison to watermarking schemes reported recently in literature is presented. Another strategy is presented in [6-8] where the EMD is associated with Hilbert transform and the watermark is embedded into the IMF containing highest energy. However, why the IMF carrying the highest amount of energy is the best candidate mode to hide the watermark has not been addressed. Further, in practice an IMF with highest energy can be a high frequency mode and thus it is not robust to attacks.

Watermarks inserted into lower order IMFs (high frequency) are most vulnerable to attacks. It has been argued that for watermarking robustness, the watermark bits are usually embedded in the perceptually components, mostly, the low frequency components of the host signal.



**Fig: 1. Watermark embedding**

It concurrently has better resistance against attacks and imperceptibility, we embed the watermark in the extrema of the last IMF. Further, unlike the schemes introduced in, the proposed watermarking is only based on EMD and without domain transform. We choose in our method a watermarking technique in the category of Quantization Index Modulation (QIM) due to its good robustness and blind nature. Parameters of QIM are chosen to guarantee that the embedded watermark in the last IMF is inaudible. The watermark is associated with a synchronization code to facilitate its location [9]. An advantage to use the time domain approach, based on EMD, is the low cost in searching synchronization codes. Audio signal is first segmented into frames where each one is decomposed adaptively into IMFs. Bits are inserted into the extrema of the last IMF such that the watermarked signal inaudibility is guaranteed. Experimental results demonstrate that the hidden data are robust against attacks such as additive noise, MP3 compression, requantization, cropping and filtering. Our method has high data payload and performance against MP3 compression.

## MATERIALS AND METHODS

### PROPOSED WATERMARKING ALGORITHM

The thought of the proposed watermarking technique is to cover up into the first sound flag a watermark together with a Synchronized Code (SC) in the time area. The information sign is initially sectioned into edges and EMD is directed on each casing to extricate the related IMFs [Figure -2]. At that point a paired information arrangement comprised of SCs and enlightening watermark bits [Figure -3] is installed in the extrema of an arrangement of successive last-IMFs. A bit (0 or 1) is embedded per extrema.

Since the number of IMFs and then their number of extrema depend on the amount of data of each frame, the number of bits to be embedded varies from one frame to the following. Watermark and SCs are not all embedded in extrema of last-IMF of only one frame. In general the number of extrema per last-IMF (one frame) is very small compared to length of the binary sequence to be embedded.

This also depends on the length of the frame. If we design by  $N_1$  and  $N_2$  the number of bits of SC and watermark respectively, the length of binary sequence to be embedded is equal to  $2N_1 + N_2$ . Thus, these  $2N_1 + N_2$  bits are spread out on several last IMFs (extrema) of the consecutive frames. Further, this sequence of  $2N_1 + N_2$  bits is embedded  $P$  times. Finally, inverse transformation (EMD-1) is applied to the modified extrema to recover the watermarked audio signal by superposition of the IMFs of each frame followed by the concatenation of the frames [Figure -1]. For data extraction, the watermarked audio signal is split into frames and EMD applied to each frame [Figure -4]. Binary data sequences are extracted from each last IMF by searching for SCs [Figure -5]. We show in [Figure -6] the last IMF before and after watermarking. This figure shows that there is little



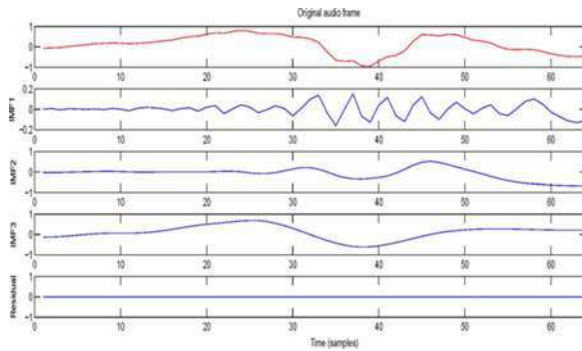


Fig: 2. Decomposition of an audio frame by EMD

Sync-code	Watermark bits	Sync-code
-----------	----------------	-----------

Fig:3. Data structure (mi)

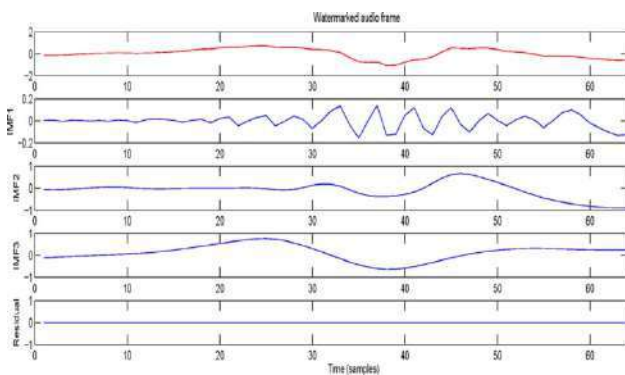


Fig: 4. Decomposition of watermarked audio frame by EMD

difference in terms of amplitudes between the two modes. EMD being fully data adaptive, thus it is important to guarantee that the number of IMFs will be same before and after embedding the watermark [Figure. -2, 4]. In fact, if the numbers of IMFs are different, there is no guarantee that the last IMF always contains the watermark information to be extracted. To overcome this problem, the sifting of the watermarked signal is forced to extract the same number of IMFs as before watermarking. The proposed watermarking scheme is blind, that is, the host signal is not required for watermark extraction. Overview of the proposed method is detailed as follows:

### SYNCHRONIZATION CODE

To locate the embedding position of the hidden watermark bits in the host signal a SC is used. This code is unaffected by cropping and shifting attacks.

Let  $U$  be the original SC and  $V$  be an unknown sequence of the same length. Sequence  $V$  is considered as a SC if only the number of different bits between  $U$  and  $V$ , when compared bit by bit, is less or equal than to a predefined threshold.

### WATERMARK EMBEDDING

Before embedding, SCs are combined with watermark bits to form a binary sequence denoted by  $m_i \in \{0, 1\}$ .  $i$ -th bit of watermark [Figure -3]. Basic of our watermark embedding are shown in [Figure -1] and detailed as follows:

Step 1: Split original audio signal into frames.

Step 2: Decompose each frame into IMFs.

Step 3: Embed P times the binary sequence {mi} into extrema of the last IMF (IMFo) by QIM:

Step 4: Reconstruct the frame (EMD-1) using modified IMFc and concatenate the watermarked frames to retrieve the watermarked signal.

### WATERMARK EXTRACTION

For watermark extraction, host signal is splitted into frames and EMD is performed on each one as in embedding. We extract binary data . We then search for SCs in the extracted data

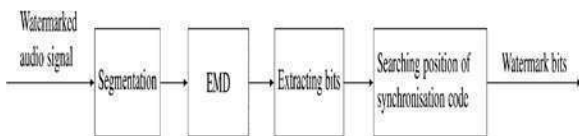


Fig: 5. Watermark extraction

This procedure is repeated by shifting the selected segment (window) one sample at time until a SC is found. With the position of SC determined, we can then extract the hidden information bits, which follows the SC. Let  $y = \{m_{i^*}\}$  denote the binary data to be extracted and U denote the original SC. To locate the embedded watermark we search the SCs in the sequence  $\{m_{i^*}\}$  bit by bit. The extraction is performed without using the original audio signal. Basic steps involved in the watermarking extraction, shown in [Figure -5], are given as follows:

Step 1: Split the watermarked signal into frames.

Step 2: Decompose each frame into IMFs.

Step 3: Extract the extrema  $\{e_{i^*}\}$  of IMFc.

Step 4: Extract the P watermark and make comparison bit by bit between these marks, for correction, and finally extract the desired watermark.

We evaluate the performance of our method in terms of data payload, error probability of SC, Signal to Noise Ratio (SNR) between original and the watermarked audio signals, Bit Error Rate (BER) and Normalized cross-Correlation (NC). According to International Federation of the Photographic Industry (IFPI) recommendations, a watermark audio signal should maintain more than 20 dB SNR. To evaluate the watermark detection accuracy after attacks, we used the BER and the NC defined as follows :

$$BER(w, \hat{w}) = \frac{\sum_{i=1}^M \sum_{j=1}^N W(i,j) \oplus \hat{W}(i,j)}{M \times N}$$

Where  $\oplus$  is the XOR operator and  $M \times N$  are the binary watermark images sizes. w and  $\hat{w}$  are the original and the recovered watermark respectively. BER is used to evaluate the watermark detection accuracy after signal processing operations. To evaluate the similarity between the original watermark and the extracted one we use the NC measure defined as follows :

$$NC(w, \hat{w}) = \frac{\sum_{i=1}^M \sum_{j=1}^N w(i,j) \hat{w}(i,j)}{\sqrt{\sum_{i=1}^M \sum_{j=1}^N w^2(i,j)} \sqrt{\sum_{i=1}^M \sum_{j=1}^N \hat{w}^2(i,j)}}$$

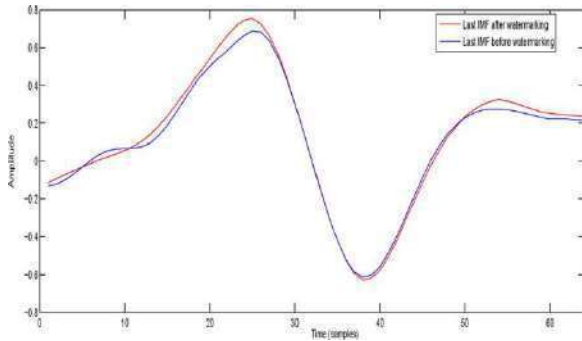
A large NC indicates the presence of watermark while a low value suggests the lack of watermark. Two types of errors may occur while searching the SCs: the False Positive Error (FPE) and the False Negative Error (FNE). These errors are very harmful because they impair the credibility of the watermarking system.

## RESULTS

### OUTPUT

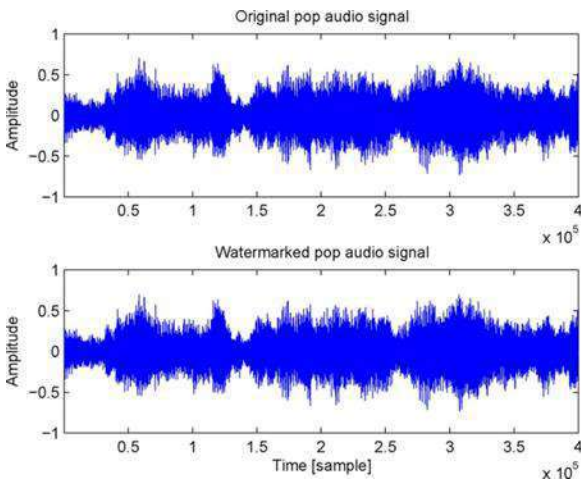
To show the effectiveness of the scheme , simulations are performed on audio signals including pop, jazz , rock, and classic sampled at 44.1 kHz. The embedded watermark, W , is a binary logo image of size  $M \times N = 34 \times 48 = 1632$ bits . We convert this 2D binary image into 1D sequence in order to embed it into the audio signal. The SC used is a 16 bit Barker sequence 1111100110101110. Each audio signal is divided into frames of size 64 samples and the threshold  $T$  is set to 4. The S value is fixed to 0.98. . These parameters have been chosen to have a good compromise between imperceptibility of the watermarked signal, payload and robustness. [Figure -9]shows a

portion of the pop signal and its watermarked version. This figure shows that the watermarked signal is visually indistinguishable from the original one.



**Fig: 6. Last IMF of an audio frame before and after watermarking**

Perceptual quality assessment can be performed using subjective listening tests by human acoustic perception or using objective evaluation tests by measuring the SNR and Objective Difference Grade (ODG). In this work we use the second approach. ODG and SNR values of the four watermarked signals are reported. The SNR values are above 20 dB showing the good choice of  $\alpha$  value and confirming to IFPI standard. All ODG values of the watermarked audio signals are between -1 and 0 which demonstrates their good quality.



**Fig:7. A portion of the pop audio signal and its watermarked version**

**ROBUSTNESS TEST**

To assess the robustness of our approach, different attacks are performed:

Noise: White Gaussian Noise (WGN) is added to the watermarked signal until the resulting signal has an SNR of 20 dB.

Filtering: Filter the watermarked audio signal using Wiener filter.

Cropping: Segments of 512 samples are removed from the watermarked signal at thirteen positions and subsequently replaced by segments of the watermarked signal contaminated with WGN.

Resampling: The watermarked signal, originally sampled at 44.1 kHz, is re-sampled at 22.05 kHz and restored back by sampling again at 44.1 kHz.

Requantization: The watermarked signal is re-quantized down to 8 bits/sample and then back to 16 bits/sample.

## CONCLUSION

In this paper another versatile watermarking plan in view of the EMD is proposed. Watermark is installed in low recurrence mode (last IMF), subsequently accomplishing great execution against different assaults. Watermark is connected with synchronization codes and accordingly the syn-chronized watermark can oppose moving and trimming. Information bits of the synchronized watermark are inserted in the extrema of the last IMF of the sound sign in view of QIM. Broad reenactments over various sound signs show that the proposed watermarking plan has more prominent power against normal assaults than nine re-cently proposed calculations. This plan has higher payload and better execution against MP3 pressure contrasted with these before sound watermarking strategies. In all sound test flags, the watermark introduction duced no discernable mutilation. Tests exhibit that the water-checked sound signs are vague from unique ones. These exhibitions exploit the self-versatile decay of the sound sign gave by the EMD. The proposed plan accomplishes low false positive and false negative blunder likelihood rates. Our watermarking strategy includes simple estimations and does not utilize the first sound sign. In the directed analyses the inserting quality S is kept consistent for all sound documents. To assist enhance the execution of the strategy, the S parameter ought to be adjusted to the sort and plan of an answer technique for versatile installing issue. Additionally as future examination we plan to incorporate the attributes of the human sound-related and psychoacoustic model in our watermarking plan for a great deal more change of the execution of the watermarking technique. At long last, it ought to be fascinating to research if the proposed strategy underpins different inspecting rates with the same payload and heartiness furthermore if in genuine applications the technique can deal with D/An A/D transformation issues. Additionally , the execution of sound watermarking utilizing EMD is being finished by ARM processor.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] IJ Cox and ML Miller. [2002] The first 50 years of electronic water- marking, *J Appl Signal Process*, 2: 126–132.
- [2] MD Swanson, B Zhu, and AH Tewfik. [1998] Robust audio water- marking using perceptual masking, *Signal Process*, 66( 3): 337–355.
- [3] S Wu, J Huang, D Huang, and YQ Shi. [2005] Efficiently self-synchro- nized audio watermarking for assured audio data transmission, *IEEE Trans. Broadcasting*, 51(1): 69–76.
- [4] V Bhat, KI Sengupta, A Das. [2010] An adaptive audio watermarking based on the singular value decomposition in the wavelet domain, *Digital Signal Proces*, 20: 1547–1558.
- [5] AntoBennet M, JacobRaglend. [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, *Journal of Computer Science*, ( 9): 1447-1454.
- [6] AntoBennet M, JacobRaglend. [2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering, *European Journal of Scientific Research*, 66(1):120-129.
- [7] Anto Bennet, M, Mohan babu, G, Rajasekar, C & Prakash, P [2015] Performance and Analysis of Hybrid Algorithm for Blocking and Ringing Artifact Reduction, *Journal of Computational and Theoretical nanoscience*, 12(1):141-149.
- [8] AntoBennet M, JacobRaglend. [2013] Performance and Analysis of Compression Artifacts Reduction for MPEQ-4 Moving Pictures Using TV Regularization Method, *Life Science Journal* , 10(2): 102-110.
- [9] AntoBennet M, JacobRaglend. [2012] A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach, *Journal of Applied Sciences Research*, 8(5): 2429-2438.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and approved by guest editor.

# PERFORMANCE EVALUATION OF OPTIMIZATION ALGORITHM USING SCHEDULING CONCEPT IN GRID ENVIRONMENT

Krishnamoorthy, Karthikeyan, Sangeetha

Dept. of Computer Science and Engineering, Kongu Engineering College, Perundurai, INDIA

## ABSTRACT

Grid computing has been an important technology by a noteworthy span in the fields of scientific and engineering. This boundless enactment of Grid computing paradigm has taken place very promptly even faster than the case for the web. Optimization is the process of choosing the desirable event in the "finest" way. The perception of 'different factors' means that there are different feasible solutions, and a perception of 'achieving enticing outcomes' means that there is an objective of searching progress on how to find the best result. In this paper, comparisons of three optimization approaches are proposed for grid Scheduling problem. As in other generic techniques of optimization such as genetic algorithm, ant colony optimization, etc., Swallow Swarm Optimization (SSO) is the new algorithm in optimization with high convergence rate is compared with the two existing standard other optimization techniques namely Particle Swarm Optimization (PSO) which is very difficult to optimize in a highly discontinuous data surface features and Fish Swarm Optimization (FSO) algorithm may give local minimum results in occurrence of stagnation condition and eventually converges at global minimum points. Simulation results show that the SSO algorithm performs better than existing methods, and performance improvement is especially significant in large-scale applications. We analyze here the use of a Grid computing systems to cope up with the limits of performance metrics. It is obtained from results that FSO gives the next higher execution time in large values. SSO on contrary executes all jobs in minimum time interval, thus obtained to be the optimal algorithm from three methods. They produce good results with the large scale applications.

Published on: 14<sup>th</sup>– August-2016

### KEY WORDS

Grid computing, Resource management, Scheduling, Performance metrics.

## INTRODUCTION

Grid computing is mostly widely used in computational science: bridging the gap between Grid computing and workflow management is the further progress.

Over the last decade we have gathered an experience in process modeling, analysis and enactment in a great accord. One of the most eloquent and sophisticated open-source workflow systems available today is YAWL (Yet Another Workflow Language). Furthermore, the analysis process is specialized in the management of workflow. With the help of Petri nets as a foundation of theoretical, variety of real-life process models have been analyzed and ranging from BPEL (Business Process Execution Language) and workflow stipulation to the entire SAP model.

In recent years, on focusing the analysis of processes based on system logs. The ProM framework has been developed at TU/e provides an adaptable toolset for mining process that is notably useful in a Grid environment. The server on the web has direct contact to the individuals to talk independently in large numbers: the collection of servers and clients often working together to solve a problem with the help of Grids. Grids are viewed by the user as a virtual environment with uniform access to resources but actually they are intrinsically distributed and heterogeneous. Major issues in Grid software addresses the security, scheduling of resources along with quality of services and many, which allows Grid to be anticipated as a single virtual platform by the users. The Grid is viewed as the backbone of the Internet, for all users. Since now the Grid computing is primarily focused on framework. Users can submit their problems to the Grid using Grid software. Most application helps the user in finding the solutions to the problem efficiently.

Optimization is the process of finding an alternative with the most cost efficient or best feasible performance for some constraints, by maximizing desired aspects and minimizing the excluded ones, which must be correct in regardless of a solution. In other words, the optimization finds the solution for a function within an addicted domain. On comparing, maximization is trying to attain the best solution or highest result or better outcome no matter what the cost is. Due to the lack of full information and time to evaluate all the available information is restricted.

The design of optimization is simple, to reduce the cost of production and improve the efficiency of production. The procedure of optimization algorithm is that the process is executed iteratively and compared the results until the optimal solution is achieved. Optimization has taken its role in computer aided design activities. The optimization algorithm has been divided into two distinct types, they are, Deterministic and Stochastic algorithms. The main contribution of this paper is Scheduling. A Grid computing provides both hardware and software framework that provides true, persistent, ubiquitous, and low-cost access to high-end computational capabilities [1]. Grid is a shared environment created through the distribution of a constant, based on standard service framework that is used for creation and resource sharing within communities. Resources can be any computer, storage media, instruments, software applications or data, all connected through the web. The middleware software provides services mainly for security and resource management. Resources under various organizations are being shared by locally defined policies that indicate what is actually shared to whom is the access allowed and under what conditions it is transferred [2]. Resource sharing and problem solving in dynamics are the two major problems that underlie the Grid concept [3]. From the scheduling point of view, a higher level abstraction for the Grid can be applied by eliminating some framework parameters such as authentication, authorization, and resource discovery and access control. To facilitate the discussion, the following frequently used terms are defined:

- The job properties are parameters like memory requirement, targets, priorities, etc.
- A job is a set of tasks that is executed on a various resources. In this paper, each jobs minimal completion time is carried for scheduling.
- A resource is something that is used to perform some operation.
- A job scheduling is the mapping of jobs to selected resources which is distributed in multiple domains.

## RELATED WORK

In Swarm Intelligence, Self-organization plays a major role with fewer restrictions and interactions with agents. Swarm intelligence came up with many famous examples from the world of wildlife, such as birds flock, fish school and insects swarm. The social interactions with social by individual help it to adapt to the situations more effectively because more information are collected from the entire swarm.

James Kennedy and Russell Eberhart [4] has reviewed on the relationship between PSO and both artificial life and metamorphic computation. Proposed the neural network and training and robot task learning and tested using Benchmark functions. They use a three layered network design to solve the XOR problem, as a demonstration of the particles swarm optimization concept. The network consists of two inputs, three hidden processing elements and one output processing element. The goal of this concept is to obtain the simplicity and robustness with the frequency which models cycle interminably around a non-global optimum.

Zainal et al., [5] presented an overview of Artificial Fish Swarm (AFSA) algorithm by describing the evolution of the algorithm along with all the improvements and its combinations with various algorithms and methods as well as its applications in solving industrial problems.

Zhehuang Huang and Yidong Chen [6] proposed an improved artificial fish swarm algorithm based on hybrid behavior selection to select behavior of fishes. First, they proposed an improved algorithm based swallowed behavior to speed up the convergence. Second it deals with the problems of easy fall into local optimum value. The experiment shows that the proposed algorithm has more powerful global exploration ability and faster convergence speed.

Revathi and Krishnamoorthy [7] made a comparison of PSO, FSO and SSO algorithms with different parameters. The swallow swarm optimization algorithm has been proven to have faster convergence speed of getting the optimal result at lower number of iterations. The design and performance evaluation of SSO were presented.

Farzi and Saeed [8] presented an Efficient Job Scheduling in Grid Computing with Modified Artificial Fish Swarm Algorithm. Job scheduling was the NP complete problem and an important issue in grid computing. To overcome the difficulties a new algorithm called modified artificial fish swarm algorithm (MAFSA) was proposed. In AFSA algorithm, leaping behavior was added, and adaptive step was used.

Biao Zhang, [9] developed Homogeneous Ant Colony Optimization (HACO) Algorithm to overcome the convergence of the Basic Ant Colony Optimization (BACO) algorithm for continuous domain problems. The proposed algorithm was demonstrated to be effective and robust, that has the potential to be implemented in various inverse heat transfer problems that are treated as the solving model for the coupled radiation and conduction of heat transfer. This is simulated by the Finite Volume Method (FVM) were served as an input for the inverse analysis. Fine-tuning of the algorithm and practical application of ACO algorithms in heat transfer.

## SYSTEM DESIGN

The workflow of the proposed system includes the following steps

- Initialize: The particles that are to be evaluated are formed a population.
- Leader Selection: From the population initialized the particle with lower convergence to the optimal solution is predicted as a leader. Always the leader particles guide the other particles in process.
- Update: After each iteration the position and velocity of every particle is updated to the predicted new value.
- Global best particle: By using distinct optimization method the unique global best value that is nearer or exact to the optimal value is obtained.
- Benchmark functions: Earlier they are tested with the standard 19 benchmark functions, result shows that SSO gives best result than the other two methods.
- Scheduling: Scheduling the jobs to the optimized resources in order to save time.
- Performance metrics: The various parameters are included for comparing the three algorithms say Time, Speed, etc., [Figure- 1]. Shows the Workflow of the proposed system.

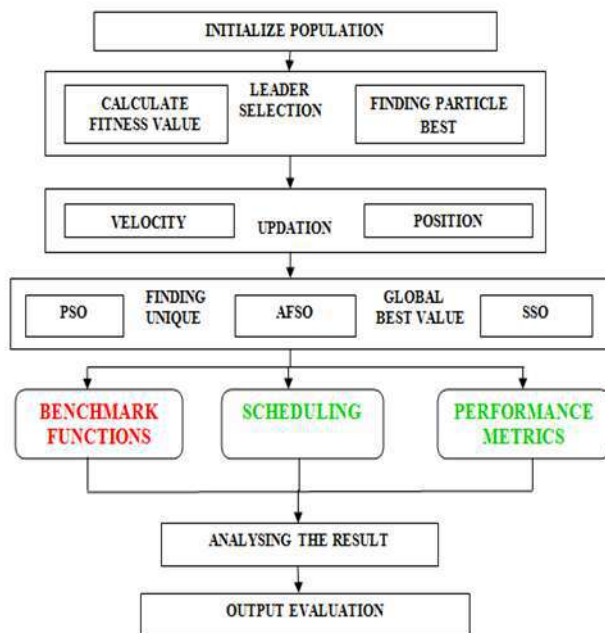


Fig: 1. Workflow of the proposed system

## FEATURES OF SWARM OPTIMIZATION

Some of the Swarm optimization is Ant Colony Optimization (ACO) where Pheromone segregation by ant easily gets evaporated which is the advantage of lowering the convergence rate at local optimal solution. Bee Colony Optimization (BCO) where the scope of exploration at local points is constantly focused on the best results. Particle Swarm Optimization (PSO) is a method based on searching the pattern that is not used to the elevation of

the problem to be optimized. Artificial Fish Swarm Optimization (AFSO) which achieve fast convergence rate and some parameters to be adjusted to not get stuck at local points. Swallow Swarm Optimization (SSO) has proved to have high efficiency and high convergence speed and not get trapped in local minima values.

## PARTICLE SWARM OPTIMIZATION

Particle Swarm Optimization (PSO) is a probing technique appropriate for finding of the optimal solution. The Particle Swarm Optimization algorithm is a biologically-encouraged algorithm excite by a social living. PSO has the ability to face the database classification occurrence is inspected. PSO algorithms using population contour or sections, which are small, fractional subsets of the best value. The formula (1) gives the updating of velocity for every particle in PSO algorithm.

$$V_i^d = \epsilon [v_i^d + c_1 \text{rand}_1^d(\text{pBest}_i^d - x_i^d) + c_2 \text{rand}_2^d(\text{gBest}^d - x_i^d)] \quad (1)$$

where,  $V_i^d$  indicates the velocity updation,  $\epsilon$  is Constriction factor,  $\text{gBest}$  is Particle's best position (global),  $x_i$  is the Current position of the particle,  $\text{pBest}$  is the particle's best position (local),  $w_{\max}$  and  $w_{\min}$  represents the maximum and Minimum weight (0.9 and 0.4),  $c_1$  and  $c_2$  is the Constant value. Sum of both constants is 4.1

## ARTIFICIAL FISH SWARM OPTIMIZATION

The next behavior of artificial fish mainly depends on its present state and ecological conditions. Aimless action indicates the initialization phase of the algorithm. The next step is a 'visual scope' of the Artificial Fish Swarm optimization algorithm. An elemental organic attitude of any creatures is to search of food, either by their eyes or through sensibility [10].

- Whenever the visual scope of fish is find empty, and there are no other fish to lead, current fish takes the random tour in searching of better position.
- Whenever the visual scope is hammed the fish failed to follow any single leader and takes random movement and searches for a better region.
- Whenever the visual scope is not hammed the fist chooses among two choices: to swarm energetic towards the best position.

$$X_i^y = X_i + \text{Visual. rand}() \quad (2)$$

$$X_{\text{next}} = X + \frac{X_v - X}{\|X_v - X\|} \cdot \text{Step. rand}() \quad (3)$$

The formula 2 and 3 gives the fish current position velocity and next step position and velocity updating respectively. Where,  $\text{rand}()$  indicates the random number between 0 and 1,  $X_i$  is fish Current position of fish,  $X_{\text{next}}$  represents the next before position of fish,  $\text{Step}$  indicates the Step length of the fish.

## SWALLOW SWARM OPTIMIZATION

Extensive design of this new optimization technique is motivated by the swallow swarm intelligence. There are three types of particles in this algorithm, they are

- Explorer particle ( $e_i$ )
- Aimless particle ( $o_i$ )
- Leader particle ( $l_i$ )

Those particles moves in parallel to each other and keep interaction with one another. Every particle in the colony (each colony can be consisted of some sub colonies) actively participate in obtaining the better situation all the time. Each particles behavior is briefly explained below.

- **Explorer particle**  
 These particles beset the major population of the colony and their main responsibility is to analyze the space problems. When the swallow reaches an extreme point that is best solution, using a contrasting sound it pays the attention of the group toward there. Suppose that place is the leading solution for the entire problem they act as a Head Leader (HL). On the contrary, if the particle is in a favorable position in parallel with its neighboring particles, it is enforced as a local leader (LL) or else, each explorer particle  $e_i$  respecting their velocities of both



leaders  $V_{HL}$  (velocity vector of particle toward HL),  $V_{LL}$  (velocity vector of particle toward LL), and competence of reaction of these two manner makes an extensive move.

- **Aimless particle**

The aimless particles do not follow their leaders. They do not have good position with the other particles. They take random movement in search of food. In case of food founded, they make different calls and pay the attention of other particles towards them.

- **Leader particle**

The Leader particles always guide the explorer particles. There exists two leaders in swallow swarm algorithm, one guides the local particles within the colony is the Local Leader. Another one guides the entire particle in progress is the Head Leader. The Head leader is the best leader. The forward movement of the velocity of leaders are given below

$$V_{HLi+1} = V_{HLi} + \alpha_{HL} \text{rand}() (e_{\text{best}} - e_i) + \beta_{HL} \text{rand}() (HL_i - e_i) \quad (4)$$

$$V_{LLi+1} = V_{LLi} + \alpha_{LL} \text{rand}() (e_{\text{best}} - e_i) + \beta_{LL} \text{rand}() (LL_i - e_i) \quad (5)$$

where,  $V_{HL}$  represents the Velocity of Head leader whereas  $V_{LL}$  represents the velocity of Local Leader and  $e_{\text{best}}$  represents the best position of the explorer particle and  $e_i$  indicates current position of the explorer particle.

Update the velocity using the formula,

$$V_{i+1} = V_{HLi+1} + V_{LLi+1} \quad (6)$$

The particle value is updated as,

$$e_{i+1} = e_i + V_{i+1} \quad (7)$$

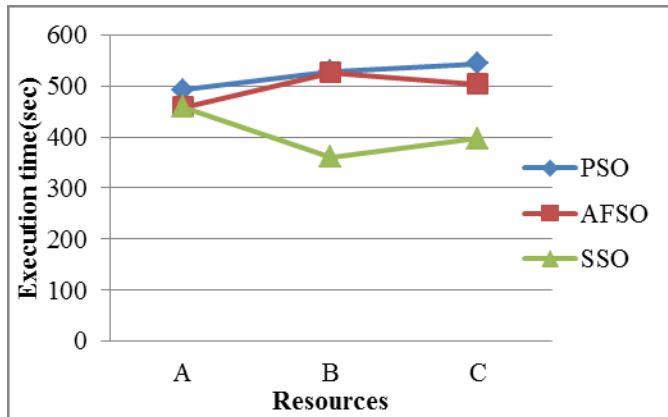
## SCHEDULING AND PERFORMANCE METRICS

Resource management and job scheduling are very important and complex problems in grid computing environment. It is necessary to do resource state prediction to get proper job scheduling. Swallow Swarm algorithm is a new heuristic algorithm. The inherent parallelism and scalability makes the algorithm very suitable to be used in grid computing resource optimization whose structure is dynamic changed almost all the time. Here scheduling process is done as, the jobs execution time is given as an input to all three optimization algorithms. Every job is scheduled to the optimized resources to get minimum execution time in a resource.

In order to produce a good schedule, estimating the performance of tasks on resources is crucial, especially for constructing a preliminary workflow schedule. By using performance estimation techniques, it is possible for workflow schedulers to predict how jobs in a work flow will be have on distributed heterogeneous resources and thus make decisions on how and where to run them. As indicated above, there are several performance estimation approaches: Time consumption, Delay rate, Transmission speed, Energy consumed and Accuracy predicted by each optimization technique.

## RESULTS

In our experiments, Particle Swarm Optimization (PSO) and Fish Swarm Optimization (FSO) algorithms along were used to compare with a new method Swallow Swarm Optimization (SSO). Specific parameter settings of all the considered algorithms are described in [Table-1](#). Each experiment (for every algorithm) was repeated 100 times with different jobs completion time as input value. The input file is converted into the comma separated value file and feed into each optimization methods for evaluating their performance.



**Fig: 2. Minimum execution time for job scheduling**

In a grid environment, the main emphasis was to generate the schedule as fast as possible. So the completion time for 100 trials was used as one of the criteria to improve their performance. First from the given 7 resources say 3 resources is selected using optimization algorithm to execute jobs in minimum time. The total time for executing the job in three resources A, B and C is given in the [Figure- 2] the PSO has higher execution time since they have lower convergence rate in large scale applications. FSO gives the next higher execution time in large values. SSO on contrary executes all jobs in minimum time interval, thus obtained to be the optimal algorithm from three methods. They produce good results with the large scale applications.

**Table: 1. Minimum execution time for job scheduling**

#	A (sec)	B (sec)	C (sec)
PSO	493	529	544
FSO	459	526	504
SSO	459	361	397

### Performance Estimation

- **Time consumption**

The jobs are distributed according to the time frame assigned to the jobs. Increasing time or decreasing time algorithm may be one of the examples of time based scheduling. Algorithm with higher convergence rate consumes more time. On the other hand algorithm with lower convergence rate algorithm consumes less time to find the optimal solution. [Figure-3] shows that PSO method consumes more time than FSO and SSO consumes very less time which is considered to be the best result.

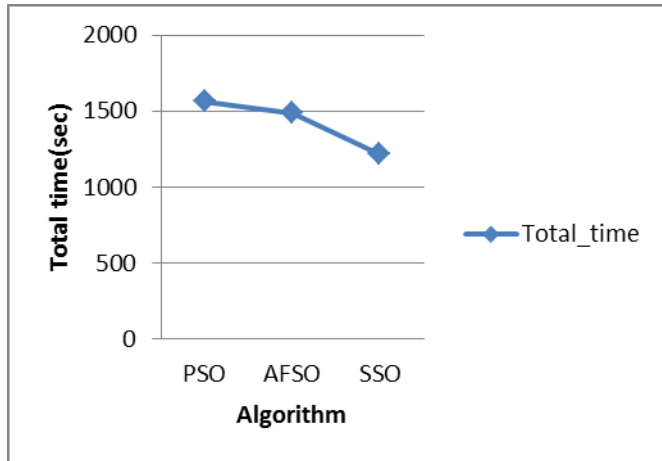


Fig.3. Time consumed by three optimization algorithms

- **Delay**

A delay is a period of time by which getting the optimal solution is postponed. [Figure -4] shows that PSO algorithm with less delay indicates that the particles gets stuck at the local minima points very easily. The high delay value in SSO algorithm indicates that SSO do not get stuck at the local minima points easily because they have two leaders head leader and the local leaders to guide the explorer particles.

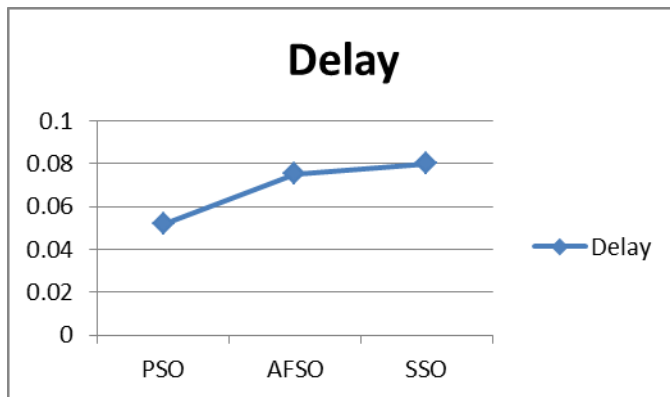


Fig.4. Delay time of optimization algorithms

- **Transmission Speed**

The rate at which the algorithm finds the optimal solution that is, one best particle from the given number of particles. [Figure -5] shows the transmission speed of each algorithm to get the best result. Lower the transmission speed higher the standard of the algorithm. Here PSO and FSO methods require more speed to converge the optimal solution than to the SSO method which executes in less transmission speed.

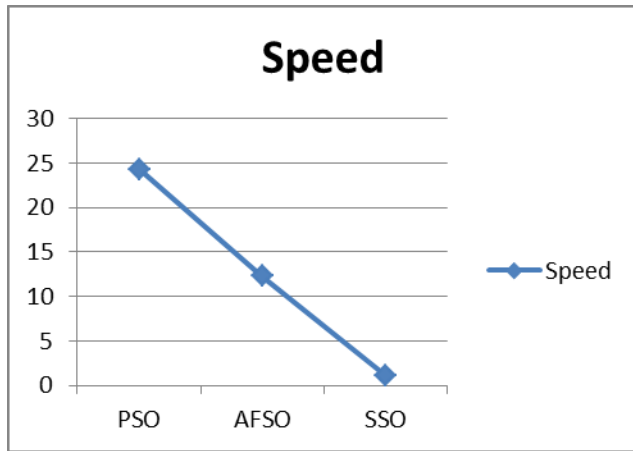


Fig: 5. Transmission speeds of optimization algorithms

- **Energy**

Virtue of the particles position to obtain the optimal solution. [Figure -6] shows that the new optimization method SSO requires less energy than to the other optimization methods.

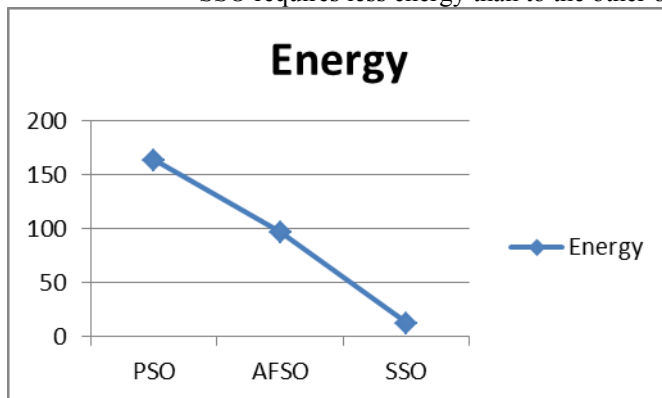


Fig: 6. Energy consumed by three optimization algorithms

## CONCLUSION AND FUTURE WORK

The proposed work is designed for comparing the behavior of optimization techniques in two forms. Firstly, Scheduling in which the jobs are executed in the optimized resources in order to save the time. Finally, the results are compared to estimate the best optimization technique that have high convergence rate and do not easily get stuck at the local minimum points. Secondly, Performance measure of all techniques is compared which includes various characteristics in terms of Time consumption, Delay rate, Transmission Speed, Energy consumed and Accuracy. On comparing results of three optimization methods, Swallow Swarm optimization algorithm proves to be even faster than the other two algorithms in having higher convergence rate and particles do not get stuck at the local minimum points easily, because there is more number of particles that follow their leaders. Unlike other optimization methods that have a one head leader here we have two leaders, one for local and other for global guidance to achieve the optimal solution even faster. The future work of this paper can be hybridization of SSO with other techniques gives better results and implementing SSO algorithm in other grid computing areas.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Venayagamoorthy, GK. [2015]Future Grids Will Not Be Controllable Without Thinking Machines. Newsletter, 2015.
- [2] Olaiifa M, Mapayi T, Van Der Merwe R. [2015]. Multi Ant LA: An adaptive multi agent resource discovery for peer to peer grid systems. In Science and Information Conference (SAI), *IEEE*, (pp. 447-451).
- [3] Rubab S, Hassan MF, Mahmood AK, Shah NM. [2015] Grid Computing in Light of Resource Management Systems: A Review.
- [4] Eberhart Russ C, and James Kennedy. [1995] A new optimizer using particle swarm theory, *Proceedings of the sixth international symposium on micro machine and human science*, Vol. 1
- [5] Zainal N, Zain AM, Sharif S. [2015] Overview of Artificial Fish Swarm Algorithm and Its Applications in Industrial Problems. In *Applied Mechanics and Materials*, 815: 253-257.
- [6] Huang Zhehuang and Yidong Chen. [2013] An improved artificial fish swarm algorithm based on hybrid behavior selection, *International Journal of Control and Automation*, 6: 103-116.
- [7] Revathi K, Krishnamoorthy N. [2015] The performance analysis of swallow swarm optimization algorithm. In *Electronics and Communication Systems (ICECS)*, 2015 2nd International Conference on (pp. 558-562). *IEEE*.
- [8] FarziSaeed. [2009] Efficient job scheduling in grid computing with modified artificial fish swarm algorithm, *International Journal of computer theory and engineering*, 1(1): 13-18.
- [9] Zhang Biao. [2013] Application of homogenous continuous Ant Colony Optimization algorithm to inverse problem of one-dimensional coupled radiation and conduction heat transfer, *International Journal of Heat and Mass Transfer*, 66: 507-516.
- [10] Huang Z, Chen Y. [2015] Log-linear model based behavior selection method for artificial fish swarm algorithm. *Computational intelligence and neuroscience*,

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

## A STUDY OF NATURE INSPIRED OPTIMIZATION ALGORITHMS

Amuthadevi<sup>1\*</sup>, Gayathri Monicka<sup>2</sup>, Madhusudhanan<sup>3</sup><sup>1</sup>Department of Computer Science & Engineering, Adhi College of Engineering & Technology, Chennai, INDIA<sup>2</sup>Department of EEE, Adhi College of Engineering & Technology, Chennai, INDIA<sup>3</sup>Department of Computer Science & Engineering, Perumal Manimekalai College of Engineering, Hosur, INDIA

## ABSTRACT

Nature has the abilities of balancing the 'eco-system', diversity maintenance and adaptation to changing environment which educated many strategies to the human beings and can be adaptable in the technologies. The generation of human beings and the behavior of many social agents and animals gave inspiration to design a set of meta-heuristic algorithms which are used to find optimal or best solutions for large number of complex problems. Most of these algorithms are independent of the nature of the problems to be solved. As many algorithms are being implemented for various applications, no one is proved as best among all the optimization problems. This paper surveys some of the nature inspired algorithms, their adaptability to real world problems and concludes with limitations and improvisation required in these algorithms.

Published on: 14<sup>th</sup>– August-2016

## KEY WORDS

Optimization algorithms, Bio-inspired algorithms, Natural Evolution.

\*Corresponding author: Email: [a.bcamuthadevi@gmail.com](mailto:a.bcamuthadevi@gmail.com), [drkssece@gmail.com](mailto:drkssece@gmail.com)

## INTRODUCTION

Charles Darwin analyzed the evolution of natural components and defined the "Theory of Natural Evolution". This theory describes about the "Survival of the Fittest" of the natural elements by taking on of the changing/dynamic environments. The natural elements have the ability of self-processing and self-learning. Best example for this is the 'generation of the human beings'. All the search/optimization problems go along with the "Survival of the Fittest". The optimization plays a major role in most of the engineering applications.

The problem solving methods are categorized into two types. They are Classical methods/traditional methods and Heuristic methods. Classical methods use either simple logical or mathematical steps and have clearly defined ways to get a solution. But heuristic methods are useful to solve NP-hard problems and need some optimization algorithms. These optimization algorithms mimic the behavior that inspired from the natural components [1].

The major reasons which makes trouble to solve a problem are [2,3]:

- Solution space has large number of possible solutions and this creates the need of exhaustive search to find the best answer.
- As the problem is more complicated, simple search is not useful.
- The evaluation strategy may vary with time or it may give a noisy solution.
- Constraint on the solution is so weighty and sometimes finding a single solution is so difficult.
- Wrong assumption about the problem/constraints may create barrier that prevents to find out a solution.

Some search algorithms like Gradient search are mostly problem dependent and convergence also depends on the selection of starting solutions. The algorithm cannot be parallelized. But nature inspired optimization algorithms can be used in wide class of applications and can be parallelized [2]. When adapting the nature inspired algorithms the following should be considered:

- The problem should be represented properly.

- The solution must be evaluated by a strategy. This will be useful to qualify the solution by using a fitness function.
- The operators should be designed to give the next set of solutions.
- This following contents of this paper is having 3 sections. First one explains about evolutionary algorithms and second one is swarm intelligence based optimization algorithms. Then the conclusion gives the merits and limitations of these algorithms.

## EVOLUTIONARY ALGORITHMS

### GENETIC ALGORITHM (GA)

GAs are powerful as they apply natural selection/natural evaluation concepts based stochastic search and optimization methods [4]. GAs work on individual populations, representing candidate solutions for optimization problems. Individuals comprise gene strings (chromosomes). GAs apply the survival of the fittest, selection, reproduction, crossover (recombining), and mutation principles on individuals to ensure better individuals (new solutions). GA's disadvantage is its inability to locate an exact global optimum, as there is no best solution guarantee.

A control parameters (optimal or near-optimal) set for a GA or GA application does not generalize all cases. The GA is defined by control parameter set  $II = \{P, C, U, M\}$ , where:

- P is population size.
- C is crossover rate. It decides convergence to pull a population to local maximum or minimum. Values range from 0 to 1. Taking higher value ensures faster convergence.
- U is the probability of an allele involved in a crossover. The probability specifies how often a crossover is allowed. 100% probability makes all offsprings and 0% makes new generation an exact copy of the earlier generation.
- M is mutation rate. It is a divergence operator to break one/more population members out of local maximum/minimum to get better maximum/minimum. Mutation rate ranges from 0 to 1. It is less frequent than cross over, so small values are taken for it.

GA operations are shown in the [Figure -1].

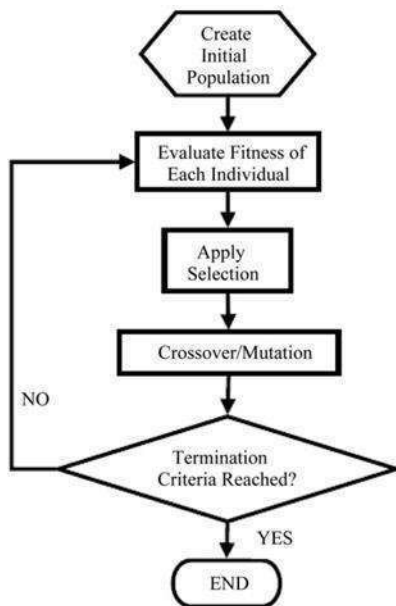


Fig: 1: Flow chart of GA

GA operations are:

- Initial population generation: This is generated randomly in a range of each parameter.
- Evaluation of fitness: Once initial population generation is over, each individual's fitness is determined. Fitness is a numeric index, measuring each individual's effectiveness as a solution. This is utilized from the population to choose members for reproduction.
- Selection Operation: An individual pair selected from current population using selection method.
- Crossover Operation: One/multipoint crossover is applied to newly selected (parents) individuals to generate two offspring. In detail, numbers of crossover points/parameters to be optimized are equal.
- Mutation Operation: Mutation operator applied randomly to newly generated offspring to prevent premature convergence to local minima [5].

- The population count is usually set as 50 for small dataset and 100 to 200 is the standard value for large data set. For mutation rate  $1/L$  is set, where  $L$  is the length of the encoding. Number of iterations is related to fitness function. Fitness function shows major improvements in earlier generations and then asymptotically reaches optimum.

## EVALUATION STRATEGIES

These use a simple variation in the process of GA [6,7]. Some of the ES schemes are:

- **(1+1)ES**: Single solution is selected and mutation is performed. The new one is compared with the solution that was before mutation. The best will be used as parent in further iterations.
- **( $\mu+\lambda$ )-ES**:  $\mu$  number of parent are selected from current iteration and  $\lambda$  off-strings are generated. From  $\mu$  and  $\lambda$  offstrings, best  $\mu$  number of offstrings will survive for next iteration.
- **( $\mu,\lambda$ )-ES**:  $\mu$  number of parent are selected from current iteration and  $\lambda$  off-strings are generated (with the condition that  $\lambda > \mu$ ). From the offstrings, best  $\mu$  number of offstrings will survive for next iteration and previously participated parents are discarded completely.

## SWARM INTELLIGENCE BASED ALGORITHMS

These algorithms are inspired based on the food searching behavior of social agents like insects and fishes. Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and Artificial Bee Colony Optimization (ABC), Cuckoo Search, Bat algorithm, Firefly algorithm come under this category.

### PARTICLE SWARM OPTIMIZATION (PSO)

PSO is based on rapid change in the movements and communications among the birds/fishes. Each particle finds its best solution from current position called as 'pbest' and all the particles have best among its group in nearby distances called as 'gbest' [9,12]. The purpose of PSO is to speed up the particles towards the gbest locations with a biased speeding factor in all the iterations. The algorithm of PSO is,

- Initialize a set of particles.
- Fitness value is calculated for each particle called as pbest.
- Compare pbest with previous iterations pbest and assign the best value for particle best.
- Compare all the nearby particles pbest and assign the best one as gbest for the group.
- For each particle, calculate particle velocity and the position according to equation based on current position, velocity and direction of particle that has gbest.
- Repeat the steps 2 to 5 until stopping criteria is reached.

The particle velocity is updated by the following equations:

- $v = v + c1 * \text{rand}() * (\text{pbest} - \text{present}) + c2 * \text{rand}() * (\text{gbest} - \text{present})$  (1)
- $\text{present} = \text{percent} + v$  (2)

where  $v$  is the particle velocity,  $\text{percent}$  is the current particle (position or solution).  $\text{pbest}$  and  $\text{gbest}$  are defined as stated before.  $\text{rand}()$  is a random number between (0,1).

PSO shares some of the properties like initial population generation and fitness function with GA. As PSO has no evolution operation, particles update the solution by their velocities towards the solution and have some memory utilization. PSO has only few parameters to set [12].

Even 10 particles can give good solutions. Standard number of particles ranges from 10 to 20. For some specific problems and wider solution space, particle range can be set as 100 to 200. Dimension of the particles depends on the problem to be solved.  $V_{\max}$  is the parameter used to define the maximum change of the particle in each iteration. In a particle  $x_1$ , if  $V_{\max}=20$ , then the changes in the solution can be [-10, 10]. There are 2 learning factors  $c_1$  and  $c_2$ , and it depends on the problem. Stop condition may be based on either the fitness value or the maximum number of iterations (Standard number of iterations is set as 1000 to 2000 for extremely complex problems)

### ANT COLONY OPTIMIZATION (ACO)

Real ants deposit a pheromone trail in the path of forward and return journey while food searching and nest building etc., This idea was used in ACO and originally implemented for Travelling salesperson problem (TSP) to find an optimal path in the weighted graphs. In ACO, artificial ants are used to find an optimal solution and then only best solutions are updated by increasing pheromone trail values and bad solutions are discarded by decreasing pheromone values [9]. ACO algorithm has the following steps.

- Parameter Setting: Number of artificial Ants and Pheromone trail, Pheromone evaporation rate and amount of reinforcement
- For each ant construct set of possible solutions.
- Daemon action is optional and depends on the problem before updating pheromone trails.
- For good solutions increase the pheromone value and decrease for others.



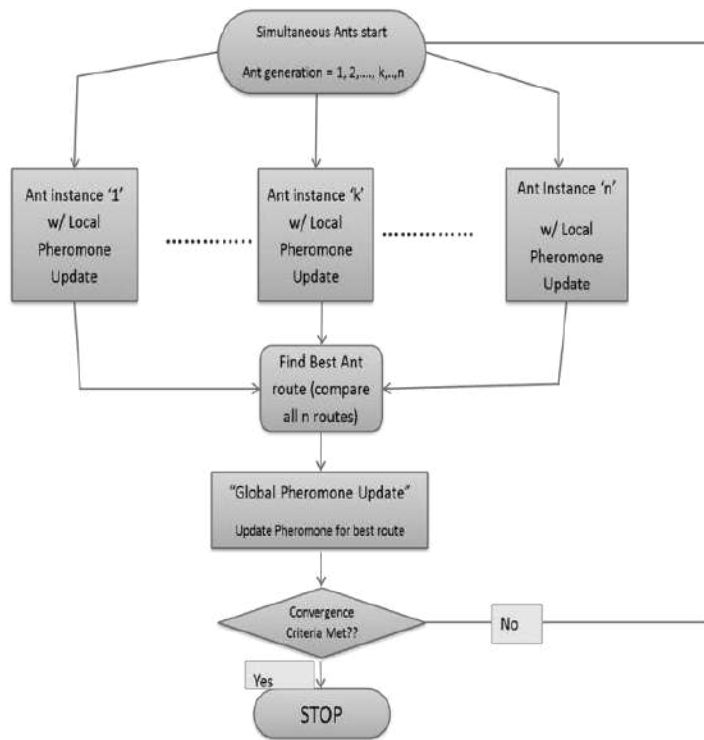


Fig: 2. Flow chart of ACO algorithm

When parameter setting, the number of ants depends on the optimization algorithm. M number of ants move from one solution x to y (in graphs one node x to another node y). In that tour, Pheromone relative importance  $\alpha$  and heuristic importance  $\beta$  (some prior information about movement from x to y). This denotes the strength of movement of solution x to y (in graphs selection of path from node x to node y).

An ant k moves from x to y with the probability,

$$P_{xy}^k = \frac{(\tau_{xy}^\alpha)(\eta_{xy}^\beta)}{\sum_{s \in \text{allowed}_x} (\tau_{xs}^\alpha)(\eta_{xs}^\beta)} \tag{3}$$

Where  $\tau_{xy}$  is the amount of pheromone deposited in the movement from x to y, and  $\eta_{xy}$  is the prior information. (In graphs this is based on the distance from node x to node y).

All the m ants complete their search then the trails are updated by,

$$\tau_{xy} \leftarrow (1 - \rho)\tau_{xy} + \sum_k \Delta\tau_{xy}^k \tag{4}$$

$$\Delta\tau_{xy}^k = \begin{cases} Q/L_k & \text{if ant } k \text{ uses curve } xy \text{ in its tour} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Where  $\rho$  is the pheromone evaporation coefficient ((1- $\rho$ ) indicates the pheromone persistence factor) and Q is the amount of pheromone ants release [14].

### ARTIFICIAL BEE COLONY OPTIMIZATION(ABC)

ABC algorithm is based on the food searching behavior of honey bees. Bees colony has three types of bees called as scout bees, onlooker bee and employee bees.

The initial step in this algorithm is to send the bees in different directions to search for 'best quality food'. When the bees found the location and the quality of the desired food, they come back to the colony and inform to the remaining bees using a communication mechanism called as 'waggle dance'.

This informs about the distance of food resource from colony, the way to reach the food and the quality of source of food. Onlooker bee is responsible for selection of best source. Then all the bees are attracted with the bee that brought information about the best quality food source [10, 11].

Actual Bee colonies behavior is given in the following pseudo code.

- Send the scouts onto the initial food sources  
REPEAT
- Send employee bees to find food sources & nectar Amounts
- Calculate probability of food sources to prefer by onlooker bees
- Send the onlooker bees onto food sources and calculate the nectar amounts.
- Stop the Exploitation process of the sources exhausted by the bees
- Send the scout bees in the search area for discovering new food sources
- Memorize the food source found so far  
UNTIL (termination conditions met)

In ABC optimization algorithm, the original concept is simply modified. Each food source is considered as a solution and the nectar amount qualifies the solution.

- Initialization: Assign an initial set of solutions in which 'employee bees' can search to find a set of possible solutions.
- Fitness Evaluation: Bees which searched the solutions are evaluated based on the fitness function using identified solutions and their visited locations.
- Evaluating Best Value: The bees having higher fitness value will be selected and visited locations are used for 'neighborhood search'.
- Iteration: If the solution is not optimal solution, then other employee bees are sent for new search. If a solution representing a food source is not improved by a predetermined number of trials, then that solution (food source) is abandoned.

In every search, quality of solution is better than previous, artificial bees forgot the previous solution and location and saves the better as new one. For better parameter setting, the number of employee and onlooker bees must be equal to the number of solutions to be searched. Maximum iteration count depends on the problem.

## CONCLUSION

This paper gives a review of few popular optimization algorithms. In real applications, many modified versions of these are used based on the nature of the problem and the size of the solution space.

GA uses many parameters to control the evolutionary search for a problem's solution. These include rates of crossover and mutation, maximum generations and number of individuals in a population. There are no hard and fast rules to choose appropriate values for parameters. PSO, ACO and ABC have few parameters to adjust when comparing to GA. As swarm intelligence algorithms does not have evolutionary process the stability and convergence is high. These also use some memory to remember good solutions.

GA can be used in wide range of applications including classification, data mining, bio-informatics and defect identification systems. ACO algorithms are suitable for scheduling, Routing and Graph designed problems. ABC is widely used in scheduling, classification and clustering algorithms.

In optimization problems the scope of the field is very vast. If the problem is not well-formulated for optimization, poor performance will be the result. There is no guarantee to get an optimal solution in finite amount of time. Expensive computation is also needed. Result mainly depends on the parameter setting. Scalability and performance evaluation are the major difficulties. Even though, some algorithms are stated as suitable for particular class of problems, no one optimization algorithm was proved as best one for particular problem.

Proper design of problems, self-adaption of parameters and hybrid optimization will improve the results with fast convergence that can reduce the computational needs for hard optimization problems.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Binitha S, S Siva Sathya.[2012] A Survey of Bio inspired Optimization Algorithms, *International Journal of Soft Computing and Engineering (IJSCE)*, 2(2): 2231-2307.
- [2] KarthikSindhya.[ 2014]An Introduction to Nature Inspired Algorithms,.
- [3] Michalewicz Z, Fogel DB.[ 2004] How to Solve It: Modern Heuristics, Springer.
- [4] Goldberg DE.[1989] Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley.
- [5] Acharya DP, Panda G, Mishra S, Lakshmi YVS.[2007] ‘Bacteria foraging based independent component analysis’, In *IEEE International Conference on Computational Intelligence and Multimedia Applications*, 2: 527-531.
- [6] Miettinen K, Neittaanmaki P.[1999] Evolutionary Algorithms in Engineering and Computer Science: Recent Advances in Genetic Algorithms”, Evolution Strategies, Evolutionary Programming, GE ,John Wiley & Sons, Inc. New York, NY, USA, 1999, ISBN:0471999024
- [7] Beyer HG, Schwefel HP. [2002] Evolution Strategies A Comprehensive Introduction, *Natural Computing*, 1:3-52, Kluwer Academic Publishers.
- [8] Schaffer JD, Caruana RA, Eshelman, LJ, Rajarshi Das. [1989] A study of control parameters affecting online performance of genetic algorithms for function optimization”, In Proceedings of the third international conference on Genetic algorithms, , pp 51–60. Morgan Kaufmann Publishers Inc., San Francisco.
- [9] Ant Colony optimization. <[http://en.wikipedia.org/wiki/Ant\\_colony\\_optimization\\_algorithms](http://en.wikipedia.org/wiki/Ant_colony_optimization_algorithms)>[Online][2015].
- [10] Karaboga D, Basturk B.[ 2007] A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, *Journal of Global Optimization*, 39: 459–471.
- [11] Karaboga D, Akay B. [2009] A comparative study Of Artificial Bee Colony algorithm, *Applied Mathematics and Computation*, 214: 108–132.
- [12] C. Rajan, N. Shanthi, Genetic based Optimization for multicast Routing algorithm for Manet’ Sadhana - Academy Proceedings in Engineering Science, 40 (7): 2341-2352, 2015.
- [13] PSO Tutorial’, Available in [www.swarmintelligence.org/tutorials.php](http://www.swarmintelligence.org/tutorials.php)

# FRAMEWORK FOR FDG-PET AND MRI FOR AUTOMATED EXTRASKELETAL BONE SARCOMA WITH CANCER DETECTION

Baskaran<sup>1</sup>, Malathi<sup>2</sup>, Thirusakthimurugan<sup>3</sup>

<sup>1</sup>\* Dept. of ECE/DrPEC and Research Scholar, Annamalai University, INDIA

<sup>2</sup> Dept. of EIE, Annamalai University, Chidambaram, INDIA

<sup>3</sup> Dept. of EIE, Pondicherry Engineering College, Pondicherry, INDIA

## ABSTRACT

Soft Tissue Sarcoma (STS) is a heterogeneous set of connective tissue malignancies that emerge from tissues of mesenchymal origin. They comprise lesser than 1% of all adult malignancies. Morphologic imaging modalities like CTs as well as MRIs may be utilized for assessing tumour localizations, size as well as infiltrations of the surrounding tissues and presence of stellate metastases. FDG PET imaging is complementary to radiological tomography as well as histological grading. Features are extracted through Wavelets as well as Wavelet and Grey Level Co-occurrence Matrix (GLCM). Correlation-based features selection methods are utilized for features selection. Structural optimization is suggested through usage of binary particle swarm optimization for classification. Neural networks are utilized as classifiers. Experimental evaluation confirmed the efficacy of the suggested technique

Published on: 14<sup>th</sup>– August-2016

### KEY WORDS

Soft Tissue Sarcoma (STS),  
Grey Level Co-occurrence  
Matrix (GLCM), Wavelets,  
Neural Networks, Levenberg  
Marquardt

\*Corresponding author: Email: [baskaranpec@gmail.com](mailto:baskaranpec@gmail.com); [vsmalu@gmail.com](mailto:vsmalu@gmail.com), [thirusakthimurugan@pec.edu](mailto:thirusakthimurugan@pec.edu)

## INTRODUCTION

Sarcoma refers to a heterogeneous set of tumours which emerge from tissues of mesenchymal origin. They are a rare kind of tumours and make up only around 1.5% of all cancers. Classifications by the World Health Organization (WHO) pinpoint around fifty distinct subkinds. Molecular study has revealed that the various subkinds are in fact biologically different which implies that sarcomas are basically a set of very rare diseases, more so than others. Clinically speaking, sarcomas are split into soft-tissue sarcomas (STS), bone sarcomas, and gastrointestinal stromal tumours (GIST) [1]. STS represents both the biggest as well as most heterogeneous set of tumours amongst sarcomas. Sarcoma originates basically from elements in the mesodermal embryonic layers. [2].

F-fluorodeoxyglucose positron emission tomography (FDG PET) is capable of identifying primary, recurrent as well as metastatic cancers in the breasts, colon, lungs as well as lymphoma in a successful manner. It is also capable of detecting STS as well as providing an indicator of grade. There is very little data available on its usage in earlier detection of local recurrence and metastases after primary surgical treatment of soft-tissue sarcomas. FDG-PET also has the benefit of detecting both problems in one procedure. Two individuals had whole-body FDG-PET initially during the evaluation stage and were hence incorporated only in the assessment of the value of FDG PET in the identification of distant metastases.

FDG-PET/CT both refer to imaging tools which are capable of measuring a tumour's metabolic activity and might be more helpful than Response Evaluation Criteria in Solid Tumors (RECIST) for the assessment of the efficacy of therapy as well as the prediction of survival of the afflicted individual. The tool exploits the fact that aggressive tumours utilize great levels of glucose for fuelling their growth and utilize radio-labelled glucose for measuring tumour's metabolic activities. Although FDG-PET/CT scans are now a fundamental part of cancer follow-up care, two UCLA works offer proof that the scans are efficient for direction of treatment schemes earlier in the care of the patients. FDG-PET/CT is capable of detecting the stage of sarcoma in children and research is being conducted to observe if it is capable of determining earlier response to therapies like in the case of adults [3].

Contrast improved Magnetic Resonance Images (MRI) displayed a  $5.1 \times 4.2 \times 11.0$  cm multilobulated forearm mass with prominent peripheral improvement as well as a central non-improving element. The mass was centred at the flexor digitorum profundus and did not appear to be involved with the base osseous structure. FDG PET scans were consequently obtained. Mild, diffused as well as increased activities all through the muscles secondary to insulin administration was present. Administration of insulin was carried out for lowering serum glucose to large right forearm soft tissue masses or within several lung nodules. The individual was chemoradiotherapy naïve when PET scan was taken [4]. Follow-up MRIs of upper extremities revealed extreme distribution of tumours by original extensions and it also infiltrated into the axilla, scapular musculature as well as underlying rib cage. CT scans of the chest revealed that there was considerable worsening of pulmonary as well as parenchymal metastatic diseases.

Hence, it is considerably significant to detect individuals who will potentially have benefits from chemotherapies or other molecule targeted agents. Morphological imaging modalities such as CTs or MRIs may be utilized for assessing tumour localization, size as well as infiltration of the surrounding tissue apart from the presence of satellite metastases. But it is not conclusively proved that considerable changes in the size of tumours are a useful tool of result of individuals suffering from soft-tissue sarcoma. Standard radiographic responses have not correlated in a consistent manner with histological responses or with illness-free survival. Other techniques for identifying individuals who will probably find chemotherapy or other agents beneficial would be of great use. Hence, PET with F-FDG is increasingly finding great usage in oncology as it permits functional imaging of viable tumorous tissue. But not every tumour is PET avidin spite of being viable.

Structural optimization is proposed through usage of binary particle swarm optimization for the classification. The remaining sections organized as: Section 2 presents the related work in literature. Section 3 details the methods which are utilized in the proposal. Section 4 discusses the experiment results and section 5 gives the conclusion of the proposed work.

## LITERATURE REVIEWS

Farhidzadehet al., [5] suggested a new model for the classification of STSs that had a focus on radiologically-defined sub-areas known as habitats. The important habitats are areas wherein the evolution of tumours can be seen. The investigators measured T1 post- as well as pre-contrast gadolinium as well as T2 non-contrast MRIs of 36 patients before treatment. The suggested method took into consideration spatially separate habitats that might be useful in clinical treatments, particularly chemotherapy as well as radiation.

Karakatsaniset al., [6] suggested the facilitation of transitions from static to dynamic multi-bed FDG PET/CT imaging wherein considering the trouble of sparse temporal sampling at all beds, new dynamic acquisition strategies are to be utilized for yielding quantitative whole-body imaging of FDG uptakes. A group of new dynamic multi-bed PET images acquisition strategies have been formulated through usage of Monte Carlo simulation, for quantitative evaluations of the clinical feasibility of the technique as well as optimization of the quantity of passes per bed as well as the total study time period. In the end, clinical whole-body patient information has been obtained in a dynamic manner and the outcomes revealed the potential of the suggested technique in the enhancement of treatment response monitoring capacities of clinical PETs researched.

O'Sullivan et al., [7] presented new methods of characterizing the total profile of the tumours, as well as a method for measuring the phase of development. Phase metric is capable of distinguishing between the earlier phase tumours wherein uptake is greatest at the core and latter phase masses wherein frequently there may be central voids in FDG uptakes. A set of FDG-PETs examined from around 185 individuals is utilized for the formal evaluation of the prognostic benefits. The study proved that more detailed quantitative appraisals of the spatial patterns of PET image data of tumour masses, beyond the maximal FDG uptakes (SUVmax) as well as earlier regarded metrics of heterogeneity, offer enhanced data for possible inputs to treatment decisions for future cases.

Wanget al., [8] suggested an automatic protocol for the detection of occurrences as well as changes of hotspots in intra-subject FDG-PET scans from fused PET-CT scanners. In this protocol, several CT scans of one subject were aligned through the usage of affine transformations, while the predicted transformations are then utilized for aligning the related PET scans into the same coordinate systems. Hotspots were detected through thresholding as

well as regions growing with variables defined particularly for various body parts. The alterations of the identified hotspots with time are examined and provided. The outcomes in nineteen clinical PET-CT studies proved that the suggested method yielded excellent performance.

Zhong and Kundu [9] suggested an optimized compartment framework that is capable of concurrently correcting for spillovers as well as partial volume impacts for both blood as well as tissue, calculate kinetic rate variables as well as create model corrected blood input functions (MCBIF) from OS EM-MAP cardio-respiratory gated 18F-FDG PET scans of mouse heart with attenuation corrections in vivo, with no invasive blood sampling. The method improves quantitation however it is iterative.

Mabrouk et al., [10] obtained rat cardiac scans in list-mode with 16 ECG-gates with PET as well as FDG. The investigators suggested a custom coupled active contour framework for reducing contamination from blood to tissue as well as from tissue to blood that are because of organ movement as well as spillovers. The novel discoveries included the fact that the investigators appended external energies to internal contours for considering the contrast blood-to-tissue as significant as the contrast tissue-to-outside myocardium. For correcting blood as well as tissue areas for spillovers, the investigators disintegrated the two dynamic ROIs in the blood as well as tissue component through usage of Bayesian probability. The outcomes revealed an excellent distinction of blood as well as tissue component in images as opposed to external blood sampling.

Gray et al., [11] suggested the initial usage of multi-region FDG-PET information for classifying subjects from Alzheimer's Disease Neuroimaging Initiative. Image information was acquired from 69 normal subjects, 71 Alzheimer's afflicted individuals as well as 147 individuals with base diagnosis of mild cognitive impairment (MCI). Anatomical segmentation was automatically created in the native MRI-space of all subjects while the mean signal intensity per cubic millimetre in all regions were extricated from the FDG-PET scans. Through usage of FDG-PET, a method that is frequently utilized clinically in the workup of dementia patients, the investigators attained outcomes that are equivalent to those got through data from research-quality MRIs or biomarkers got in an invasive way from cerebrospinal fluid.

Zhenget al., [12] suggested a new model for the derivation of generalized optimum quantitative index (QI) as well as its related optimum range of imaging protocols for more enhancement of performance of dual-time FDG-PET imaging in diagnosing lung cancers.

Tafstet et al., [13] suggested a novel technique for segmenting (BTV) in 18F-FDG-PET scans through usage of an automated Gaussian mixture model (GMM) on the basis of Akaike information criteria (AIC). The protocol was confirmed as valid on two patients out of seven who had laryngeal tumour. The volumes predicted were contrasted with macroscopic laryngeal specimens wherein a 3-D biological tumour volume (BTV) specified by histology was utilized as reference. Outcomes from experiments revealed that the technique was capable of segmenting BTV in a more accurate fashion than other threshold-based techniques.

## METHODOLOGY

### DATASET

A dataset comprising fifty one patients with histologically confirmed primary soft tissue sarcomas of the extremities was obtained in a retrospective fashion. Patients with metastatic and/or recurrent soft tissue sarcomas at presentation were discarded from the research. The individual were split into two broad groups [14]:

- 32 who did not develop lung metastases (represented as 'NoLungMets'); and
- 19 who developed lung metastases (represented as 'LungMets') within the follow-up period.

Individuals from the first with follow-up time lesser than a year were discarded from the research. Lung metastases were confirmed either through biopsies or through diagnoses by healthcare professionals through the presence of common pulmonary lesions in CT and/or FDG-PET scans.

### WAVELET

Features extraction is the initial phase of classification wherein features of all images are distinctly extricated from MRIs through wavelets which is regarded as the optimal technique for extracting most emphasizing pixels apparent in images for improving outcomes. For decomposition of data into distinct frequencies, wavelet mathematical functions are utilized and all components

are examined with resolutions matched to their degrees. For analyzing complicated dataset, wavelets are now regarded as the most powerful mathematical tool available.

Wavelet refers to mathematical functions that disintegrate data into distinct frequency components with resolutions matched as per its state. It has several advantages at the time of analysis of physical situations with discontinuities as well as sharp edges. Wavelet transforms are similar to hierarchical sub-band filtering systems. Mostly all practical DWTs utilize discrete time filter banks. These are known as wavelets as well as scaling coefficient in wavelet terminology.

Wavelet transforms (WT) are a comparatively novel kind of transform. A significant benefit is that the transform possesses the capacity to offer data regarding the time-frequency abstractions of the signal. For almost all practical applications, there are two types of wavelets present which are CWT as well as DWT. Wavelet coefficients at all scales generate large quantities of information. Because of the large quantities of information created via CWT, training classifiers on the basis of the coefficients at various scales may frequently become hard.

DWT refers to an implementation technique for wavelets following certain specified rules as well as discrete set of wavelet translations. It is required for practical computations to make wavelets discrete. Scale limits are then discredited with regard to translation limit ( $\tau$ ). The formula below reveals the scale as well as translation of wavelets:

$$s = 2^{-m}$$

$$\tau = n2^{-m}$$

The typical format of transform kind of image combination protocols is the wavelet fusion protocol due to its simplicity as well as its capacity to preserve time as well as frequency details of image scans to be combined. Wavelet transfers of wavelet fusion protocols of two registered scans  $P_1(x_1, x_2)$  and  $P_2(x_1, x_2)$ . It may be denoted by [15]:

$$I(x_1, x_2) = W^{-1}(\psi(W(P_1(x_1, x_2)), W(P_2(x_1, x_2))))$$

Wherein  $W$ ,  $W^{-1}$  as well as  $\psi$  represent wavelet transform operators, inverse wavelet transform operators as well as fusion rule, correspondingly.

### GREY LEVEL CO-OCCURRENCE MATRIXES (GLCM)

GLCM refers to a statistical technique of examination of textures which consider the spatial relations of pixels. GLCM function characterizes the texture of images through calculation of how frequently pairs of pixels with particular values as well as in particular spatial relation occur in a particular image. The features are created through calculation of features for all co-occurrence matrices acquired through the usage of directions  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ , and  $135^\circ$ , and then average of the four values.

The noted co-occurrence matrix features equation that calculated from image  $(i, j)$  are angular second moment, energy, contrast, dissimilarity as well as GLCM correlation.

$$ASM = \sum_{i,j=0}^{N-1} P_{i,j}^2$$

$$Energy = \sqrt{\sum_{i,j=0}^{N-1} P_{i,j}^2}$$

$$Contrast = \sum_{i,j=0}^{N-1} P_{i,j} (i-j)^2$$

$$Dissimilarity(DIS) = \sum_{i,j=0}^{N-1} P_{i,j} |i-j|$$

$$GLCM \text{ Correlation} = \sum_{i,j=0}^{N-1} P_{i,j} \left[ \frac{(i-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right]$$

GLCM correlations indicate the linear dependency between

gray levels as well as neighbor pixels wherein  $\mu_i$  denotes horizontal mean in matrix,  $\mu_j$  denotes vertical mean in matrix,  $\sigma_i^2$  as well as  $\sigma_j^2$  denotes dispersion around the mean of combinations of target as well as neighboring pixel [16].

### CORRELATION-BASED FEATURE SELECTION (CFS)

Like most features selection protocols, CFS utilizes a search protocol alongside functions for evaluating merits of features subset. Heuristics through which CFS assesses excellent of features subset considers the utility of individual features for prediction of class labels along with level of inter-correlation amongst them.

Correlation coefficient is utilized for estimation of correlation between subsets of features as well as classes, and inter-correlations amongst features. Relevance of a set of attributes rises with correlation between attributes as well as classes, and reduces with increasing inter-correlation [17]. CFS is utilized for determining optimal features subset and is typically fused with search schemes like forward selection, backward elimination, bi-directional searches, best-first search as well as genetic search.

$$r_{zc} = \frac{k\bar{r}_{zi}}{\sqrt{k + k(k-1)\bar{r}_{ii}}}$$

where  $r_{z_c}$  refers to the correlation between summed features subset as well as the class parameter,  $k$  refers to the quantity of subset features,  $r_z$  refers to the average of correlations between the subset features as well as the class parameters, while  $r_r$  refers to the average inter-correlation between subset features.

### NEURAL NETWORK LANGUAGE MODEL (NNLM)

Neural Network (NN) is exhaustively utilized for patterns classification, because of the fact that they do not need any details about probability distributions as well as the a priori probabilities of various classes. Neural network classifications systems imitate the human approach to thinking and in particular scenarios, they provide the decision for several classes to indicate the possibility of few other infections. In the event of brain MRI classification as either normal or anomalous, the technique utilizes back propagation protocol LM with Artificial Neural Network (ANNLM) for categorizing inputs into the set of target categories (normal or anomalous) as per features extraction variables.

Levenberg-Marquardt (LM) strategy is a high order adaptive method and significantly reduces Mean Square Error of neural networks. In the suggested method, optimized LM strategy is employed for diminishing errors at the time of classification of tumours in brain MRIs. In the current study, MRIs are subjected to ANN strategy through LM training for discovering as well as categorizing occurrences of tumours in MRIs and experimental evaluation is carried out for deciding the sensitivity, specificity as well as accuracy of optimized LM strategy. The total procedure utilized in tumour in MRIs classification is given.

LM method as well as Network infrastructure,

Assume a nonlinear model of the generic form in MRI brain images classification,

$$x_i = p(y_i, \alpha) + \varepsilon_i \text{ where, } (i = 1, 2, 3, \dots, m) \text{ Wherein } \alpha \text{ refers to a vector comprising } n \text{ variables while } m > n.$$

It is also to be assumed that  $g$  is non-linear in  $\alpha^T = [\alpha_1, \alpha_2, \dots, \alpha_n]$ . The method of least squares is used for estimation of indefinite variables in the event of non-linear regression functions. Proportional to the strategy, estimates of  $\alpha_1, \alpha_2, \dots, \alpha_n$  are acquired through minimization of quantity,

$$\sum g_i^2(\alpha)$$

Sum of the squares of the errors of the estimations of the classification of normal or anomalous brain images is inferred through the previously mentioned, wherein:  $g_i(\alpha) = x_i - p(y_i, \alpha)$

As given by non-linear regressions, the strategy of non-linear least-squares data fitting possesses excellent form for gradient as well as Hessian.

Standard LM training procedure is given by the pseudo code below, [18]

- Initialize weights as well as variable  $\mu$  ( $\mu = .01$  is adequate).
- Calculate sum of squared errors over inputs  $F(w)$ .
- Solve (2) for obtaining increment of weights  $\Delta w$
- Recalculate sum of squared errors  $F(w)$

Utilizing  $w = w + \Delta w$  as the trial  $w$ , and assess

IF trial  $F(w) < F(w)$  in step 2 THEN

IF trial  $F(w) < F(w)$  in step 2 THEN

$$w = w + \Delta w$$

$$\mu = \mu \cdot \beta (\beta = .1)$$

Go back to step 2

ELSE

$$\mu = \mu / \beta$$

go back to step 4

END IF

### NEURAL NETWORK BACK PROPAGATION (NN\_BPP)

Neural networks comprise of sets of nodes as well as connections between them. Typically nodes are grouped in layers with connections that go from one layer to the next. Input layers of nodes are present that are activated by inputted image data. Output layers of nodes represent output classes to train for. There are one or more hidden layers in the middle. Nodes in a layer are linked to all nodes in the next. Nodes in hidden layers obtain inputs from all nodes in the earlier layer. Output values from hidden layers are spread to output layers which comprise one node for every output class. All node connections possess weights that



multiply the signal traversing the connection. Nodes in the hidden as well as output layers sum the weighted signals they obtain and employ functions for producing output values. During learning phases, example spectral pattern is passed through network in a set of iterations. The second stage in training is a backward pass through the network for reducing errors between real as well as anticipated output.

Artificial neural networks (ANN) take into consideration classification as a significant research as well as application area. The primary shortcoming in utilizing ANN is the finding of accurate grouping of training, learning as well as transfer functions for classification of datasets with rising quantity of features as well as classified sets. The various combinations of functions as well as their impact when utilizing ANN as a classifier is examined and the accuracy of the functions are studied for several types of datasets. The real-life issues that are denoted by multi-dimensional data sets are obtained from medical backgrounds.

Classifying as well as clustering the datasets is important. Datasets are split into training as well as testing sets and have no utilization in the training procedure. The outcomes are got with the assistance of the data sets and are utilized for testing. Training sets are obtained from 2/3<sup>rd</sup> of the datasets and the remainder is taken up as test dataset. This is made via the measurement of accuracy attained through testing against the datasets. Then networks are simulated with the same information. Back propagation protocols train the neural networks. Gradient descent methods (GDM) were utilized for decreasing mean squared errors between network output as well as actual error rates. The variables given below are regarded for measuring efficacy of the network: rate of convergence, number of epochs for converging network, computed MSE.

With the adequate combinations of training, learning as well as transfer function, data set classification utilizes the very successful tool known as back propagation neural networks [19].

### STRUCTURE OPTIMIZED BINARY PARTICLE SWARM OPTIMIZATION (PSO)

Kennedy and Eberharts suggested a discrete binary variant of PSO for binary issues. Binary values may be representations of real values in binary search space. In bPSO, particles' personal bests as well as global bests are updated like in the continuous variant. The primary variation is that velocity of the particles is specified with regard to probabilities that a bit will modify to 1. Utilizing this definition, velocities are restricted to [0,1]. Hence maps are suggested for mapping all real valued numbers of velocity to [0,1]. Normalization functions utilized here are sigmoid functions given by:

$$v'_{ij}(t) = sig(v_{ij}(t)) = \frac{1}{1 + e^{-v_{ij}(t)}}$$

Furthermore the formula given above is utilized for updating velocity vectors of the particles. New positions of the particles are got through the formula given below:

$$x_{ij}(t+1) = \begin{cases} 1 & \text{if } r_{ij} < sig(v_{ij}(t+1)) \\ 0 & \text{otherwise} \end{cases}$$

Wherein  $r_{ij}$  refers to a uniform arbitrary number within [0,1] [20].

### RESULT AND DISCUSSION

The experiments conducted using 85 Normal, 35 sarcoma images from the dataset. [Figure -1 & 2] shows the sample images. [Table - 1] listed the experiment results.

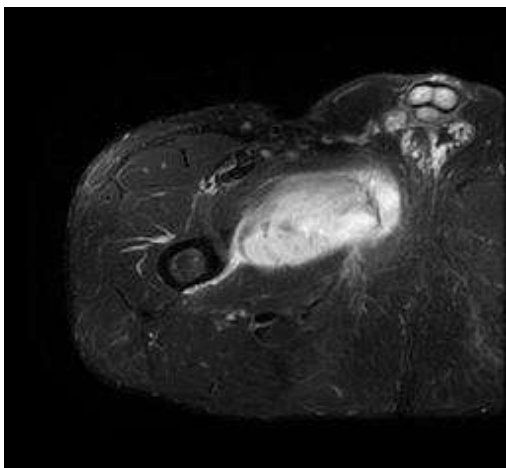


Fig: 1. Sample image 1

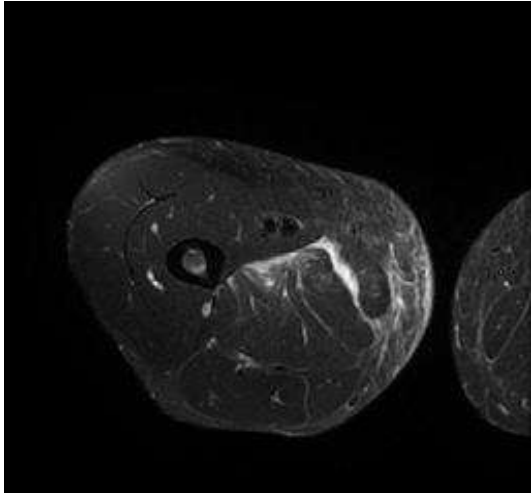


Fig: 2. Sample image 2

Table: 1. Results

Techniques	NN-BP	NN-LM	NN-structure optimized
classification accuracy	0.875	0.8898	0.925
Positive Prdictive Value for Normal	0.7632	0.8056	0.8421
Positive Prdictive Value for Bone Sarcoma	0.9268	0.9268	0.9634
Sensitivity for Normal	0.8286	0.8286	0.9143
Sensitivity for Bone Sarcoma	0.8941	0.9157	0.9294
F measure for Normal	0.9102	0.9212	0.9461
F measure for Bone Sarcoma	0.7946	0.8169	0.8767

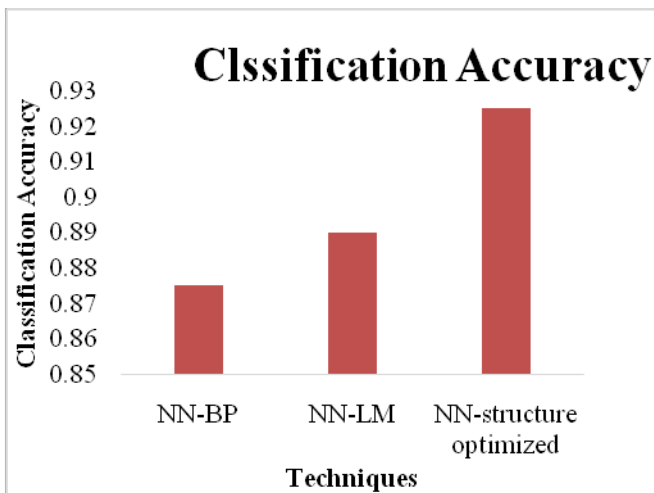
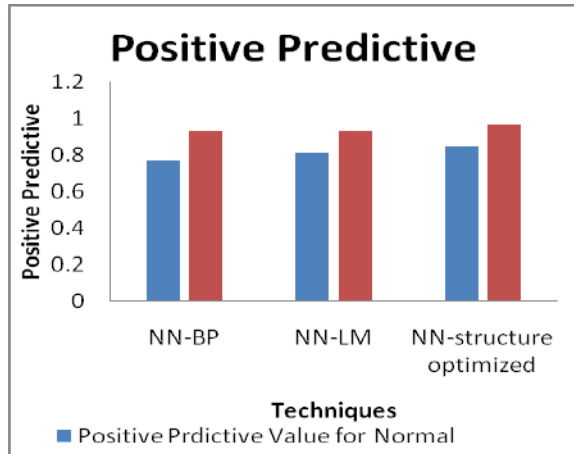


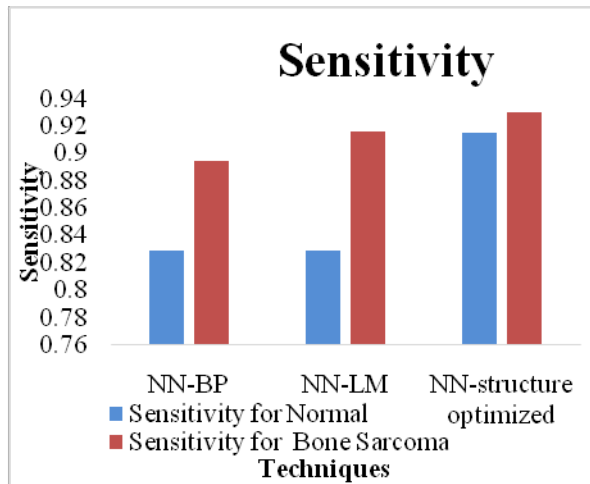
Fig: 3. Classification Accuracy

It can be observed from the [Figure -3] that the proposed NN-structure optimized with binary PSO method improved accuracy by 5.56% and 3.88% when compared with NN-BP and NN-LM approaches respectively.



**Fig: 4. Positive Predictive**

It can be observed from the [Figure -4] that the proposed NN-structure optimized with binary PSO method improved positive predictive value by 9.83% and 3.87% when compared with NN-BP for normal and bone sarcoma images respectively.



**Fig: 5. Sensitivity**

It can be observed from the [Figure -5] that the proposed NN-structure optimized with binary PSO method improved sensitivity value by 9.83% and 1.49% when compared with NN-LM for normal and bone sarcoma images respectively.

From [Figure -6] it is observed that the proposed NN-structure optimized with binary PSO method improved F measure value by 2.67% and 7.06% when compared with NN-LM for normal and bone sarcoma images respectively.

## CONCLUSION

STs typically form in the body's muscle, fat, nerve, deep skin tissue as well as blood vessels. Taken in tandem, FDG-PET definitely plays a growingly significant prognostic as well as predicting role in managing sarcoma. It can be utilized for assessing aggressiveness of tumours for making earlier clinical decision regarding the utility of treatment option for patients. In this paper, features extracted using wavelet and GLCM methods and CFS is used

for feature selection. And we proposed the neural network classifier optimized with the binary particle swarm optimization for the classification of sarcoma images. The results proved that the proposed optimization improved the accuracy, positive predictive and sensitivity as well as f measure.

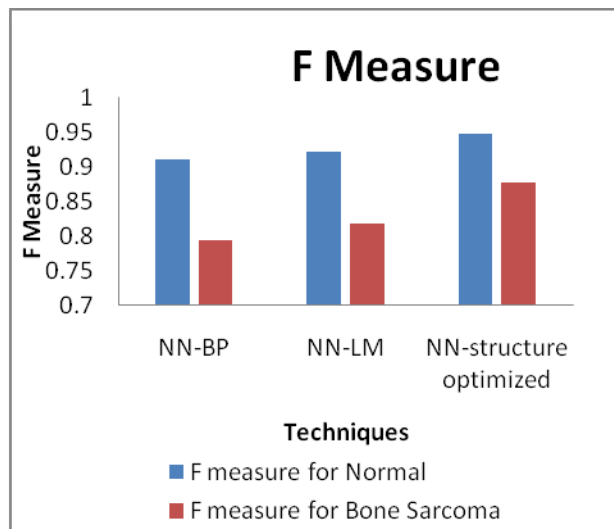


Fig: 6. F Measure

#### CONFLICT OF INTEREST

The authors declare no conflict of interests.

#### ACKNOWLEDGEMENT

None

#### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

#### REFERENCES

- [1] Levard A, Tassy L, Cassier PA. [2013] Emerging therapies for soft-tissue sarcomas. *Hematology/oncology clinics of North America*, 27(5): 1063-1078.
- [2] Mankin HJ, Hornicek FJ. [2005] Diagnosis, classification, and management of soft tissue sarcomas. *Cancer Control*, 12(1): 5-21.
- [3] Benz MR, Czernin J, Allen-Auerbach MS, Tap WD, Dry SM, Elashoff D, Weber WA. [2009] FDG-PET/CT imaging predicts histopathologic treatment responses after the initial cycle of neoadjuvant chemotherapy in high-grade soft-tissue sarcomas. *Clinical cancer research*, 15(8): 2856-2863.
- [4] Musana KA, Raja S, Cangelosi CJ, Lin YG. [2006] FDG PET scan in a primitive neuroectodermaltumor. *Annals of nuclear medicine*, 20(3): 221-225.
- [5] Farhidzadeh H, Goldgof DB, Hall LO, Gatenby RA, Gillies RJ, Raghavan M. [2015] Texture feature analysis to predict metastatic and necrotic soft tissue sarcomas.
- [6] Karakatsanis NA, Lodge MA, Zhou Y, Mhlanga J, Chaudhry MA, Tahari AK, Rahmim A. [2011] Dynamic multi-bed FDG PET imaging: feasibility and optimization. In *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, IEEE (pp. 3863-3870).
- [7] O'Sullivan F, Wolsztynski E, O'Sullivan J, Richards T, Conrad EU, Eary JF. [2011] A statistical modeling approach to the analysis of spatial patterns of FDG-PET uptake in human sarcoma. *Medical Imaging, IEEE Transactions on*, 30(12):059-2071.
- [8] Wang J, Feng DD, Xia Y. [2010] Automated Detection of the Occurrence and Changes of Hot-Spots in Intro-subject FDG-PET Images from Combined PET-CT Scanners. In *Digital Image Computing: Techniques and Applications (DICTA)*, 2010 International Conference on, IEEE, (pp. 63-68).
- [9] Zhong M, Kundu BK. [2013] Optimization of a model corrected blood input function from dynamic FDG-PET

- images of small animal heart in vivo. *Nuclear Science, IEEE Transactions on*, 60(5):3417-3422.
- [10] Mabrouk R, Bentabet L, Dubeau F, Bentourkia M. [2010] Input functions extraction from gated 18 F-FDG PET images. In *Nuclear Science Symposium Conference Record (NSS/MIC)*, IEEE (pp. 2982-2986).
- [11] Gray KR, Wolz R, Keihaninejad S, Heckemann RA, Aljabar P, Hammers A, Rueckert D. [2011] Regional analysis of FDG-PET for use in the classification of Alzheimer's disease. In *Biomedical Imaging: From Nano to Macro, IEEE International Symposium on* (pp. 1082-1085). IEEE.
- [12] Zheng X, Tian G, Wen L, Feng DD. [2010] Generalized optimal quantitative index of dual-time FDG-PET imaging in lung cancer diagnosis. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, 201-204).
- [13] Tafast A, Hadjili ML, Hafdaoui H, Bouakaz A, Benoudjit N. [2015] Automatic Gaussian Mixture Model (GMM) for segmenting 18 F-FDG-PET images based on Akaike Information Criteria.
- [14] Freeman C R, Skamene SR, El Naqa, I. [2015] A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in medicine and biology*, 60(14):5471.
- [15] Korchiyne R, Farssi SM, Sbihi A, Touahni R, Alaoui MT. [2014] A Combined Method Of Fractal And GLCM Features For MRI And CT Scan Images Classification. arXiv preprint arXiv:1409.4559.
- [16] Kaur N, Bahl M, Kaur H, [2014] Review On: Image Fusion Using Wavelet and Curvelet Transform. *International Journal of Computer Science and Information Technologies*, 5 (2):2467-2470
- [17] Karegowda AG, Manjunath AS, Jayaram MA. [2010] Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271-277.
- [18] Palanivelu L, Vivayakumar P. [2014] IMPROVED NEURAL NETWORK TRAINING ALGORITHM FOR CLASSIFICATION OF COMPRESSED AND UNCOMPRESSED IMAGES. *Journal of Theoretical & Applied Information Technology*, 62(1).
- [19] Saravanan K, Sasithra S. [2014] Review On Classification Based On Artificial Neural Networks. *International Journal of Ambient Systems and Applications (IJASA)* 2(4)
- [20] Khanesar MA, Teshnehlab M, Shoorehdeli MA. [2007] A novel binary particle swarm optimization. In *Control & Automation, 2007. MED'07. Mediterranean Conference on* (pp. 1-6). IEEE.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as it is provided by author and approved by guest editor.

# GENERALIZED REGRESSION NEURAL NETWORK FOR SOFTWARE DEFECT ESTIMATION

Sankara Rao<sup>1\*</sup> and ReddiKiran Kumar<sup>2</sup>

<sup>1</sup>JNTU, Kakinara, Andhra Pradesh, INDIA

<sup>2</sup>Krishna University, Machilipatnam, Andhra Pradesh, INDIA

## ABSTRACT

Software development estimation is important task in managing huge software projects. It is well known that software industry is unable to properly estimate effort, time and development cost. Many estimation models exist for effort prediction but there is a need for a new model to get more accurate estimates. This paper proposes a Generalized Regression Neural Network (GRNN) to use improved software estimation effort for COCOMO dataset. This paper uses Mean Magnitude Relative Error (MMRE) and Median Magnitude Relative Error (MdMRE) as evaluation criteria. The new GRNN is compared to varied techniques like linear regression, M5, RBF kernel and Sequential Minimal Optimization (SMO) Poly kernel.

Published on: 14<sup>th</sup>– August-2016

### KEY WORDS

Effort estimation, Constructive Cost Model (COCOMO), Magnitude Relative Error (MMRE) and Median Magnitude Relative Error (MdMRE)

\*Corresponding author: Email: [psankararao.cse@gmail.com](mailto:psankararao.cse@gmail.com)

## INTRODUCTION

A software defect is a condition which fails to meet software requirements or end user expectations in software products. A defect is an error/bug in coding/logic causing a program malfunction or produces incorrect or unanticipated results. Software defect prediction locates defective modules in software. To ensure high quality software, the final product should have very few defects. Early defects detection leads to reduced time, development cost and rework and reliable software. So, defect prediction is for good software quality. Software defect prediction metrics have a big role in constructing a statistical defect prediction model for by software organizations during early software development to identify defect modules [1].

Software cost estimation predicts effort to develop a software system. Many estimation models were used over three decades. Computing power is a subordinate resource for software developing companies as it doubles every 18 months, costing a fraction compared to late 60's. Personnel costs are an important expense in a software company's budget. In view of this, proper planning is a key aspect for companies. Software community developed tools/techniques like effort, size and cost estimation to offset challenges facing software development projects management. These tools/techniques are used for software development phases starting with software requirements specification. As demand for software applications goes up continually and software scope/complexity go higher software companies need accurate estimates of under development projects. Good software effort estimates are critical to companies/clients [2].

Software effort estimation methods are categorized as algorithmic/non-algorithmic methods. The former are mainly COCOMO, Function Points and Software Life-Cycle Model (SLIM). They are also called parametric methods as they predict software development effort using fixed mathematical formula parameterized from historical data. But, preliminary stage estimates of a project are difficult to get as primary estimate effort source is from a SRS document. They also face problems in modelling inherent complex relationships. Algorithmic methods limitations opt for non-algorithmic methods based on soft computing. These methods learn from previous data and model complex relationship between dependent (effort) and independent variables [3].

Costs and efforts are predicted by using mathematical formulae in algorithmic cost estimation. Formulae are historical data based. A known algorithmic cost model called COCOMO published by Barry Boehm in 1981 was developed from analyzing 63 software projects. Boehm proposed three model levels called Basic COCOMO, Intermediate COCOMO and Detailed COCOMO. Intermediate COCOMO is described as follows [4]:

Intermediate COCOMO: Basic COCOMO is based on relationship: Development Effort (DE),

$$DE = a*(SIZE)^b \quad (1)$$

Where, SIZE is measured in 10000 delivered source instructions. Constants a, b are dependent upon 'mode' of projects development. DE is measured in man-months. Boehm proposed 3 project modes:

- Organic mode – simple projects engaging small teams working in known/stable environments.
- Semi-detached mode – projects engaging teams with a mixture of experience. It is between organic and embedded modes.
- Embedded mode – complex projects developed under tight constraints with changing requirements.

Basic COCOMO accuracy is limited as it does not consider hardware, personnel, modern tools use and attributes affecting project cost. Also, Boehm proposed an Intermediate COCOMO that adds accuracy to Basic COCOMO by multiplying 'Cost Drivers' into an equation with a new variable: Effort Adjustment Factor (EAF) seen in [Table -1] [5].

**Table: 1. DE FOR THE INTERMEDIATE COCOMO**

Organic	
Semi-detached	$DE = EAF * 3.0 * (SIZE)^{1.12}$
Embedded	$DE = EAF * 2.8 * (SIZE)^{1.2}$
Semi-detached	$DE = EAF * 3.0 * (SIZE)^{1.12}$

An Artificial NN (ANN) is an information processing system resembling a biological neural network in characteristics. ANN's have many highly interconnected processing elements named neurons which are connected to others through a connection link. Each link is associated with weights have input signal information. Neuron net uses the information to solve specific problems. A neuron has its own internal state called neuron activation level which is a function of inputs received by a neuron. There are many activation functions applied over net input like Gaussian, Sigmoid, Linear and Tanh. Neural nets frequently use sigmoid function [6].

This paper proposes to investigate MMRE/MdMRE using techniques like M5, SMOPolykernel, Linear regression, RBF kernel and the new GRNN. The COCOMO dataset is used for investigations. The paper is organized as follows: section 2 deals with related work, section 3 details materials and methods used. Section 4 provides experiments results and discussion of the same and section V concludes the paper.

## RELATED WORKS

Kernel principal component analysis (KPCA) with Kernel Density Estimation (KDE) approach was applied to Tennessee Eastman process to detect faults by Samuel & Cao [7]. Results confirmed that associating KPCA with kernel density estimated control limits ensured better monitoring than using normal probability density function based control limits.

An innovative idea of the working of Principal Component Analysis (PCA) with ANN by keeping base of Constructive Cost Model II (COCOMO II) was presented by Patil et al., [8]. Feed forward ANN used delta rule learning method train a network. ANN training was PCA and COCOMO II sample dataset repository based. PCA was a classification method which filters multiple input values into certain values. It reduces the gap between actual/estimated effort. Test results from the hybrid model were compared to COCOMO II and ANN.

Use of KPCA as an approximation technique for nonlinear thermodynamics or kinetic functions parameterized using available plant archived data was explored by Mukhopadhyay et al., [9]. Simulation on a complex binary distillation column proved the new approach's applicability.

An adaptation mechanism for a mixture of Gaussian process regression models based soft sensor model was proposed by Grbic et al., [10]. Also presented was a procedure for input variable selection based on mutual information. This selects most important input variables for output variable prediction, simplifying the model for development/adaptation. This soft sensor is used for adaptive process monitoring in addition to online prediction of difficult-to-measure variables. The proposed method's efficiency was benchmarked with commonly applied recursive PLS and recursive PCA method on Tennessee Eastman process and on two real industrial examples.

A popular component analysis methodology, i.e., PCA, in complex reproducing kernel Hilbert spaces (CRKHS) formulated by Papaioannou&Zafeiriou [11] defined a widely linear complex kernel PCA framework. Also it shows how to efficiently perform linear PCA in small sample sized problems. Finally, it shows the new framework's usefulness in robust reconstruction using Euler data representation.

Performance of the aforementioned feature selection methods on LR and  $\ell_1$ -regularized logistic regression using different statistical measures was assessed by Musa [12]. Varied performance metrics like sensitivity, precision, specificity, accuracy, area under receiver operating characteristic curve and receiver operating characteristic analysis was used. This study included a comprehensive statistical analysis.

A new method to assign weights to features by considering their specific importance on cost was proposed by Tosun et al., [13]. Two weight assignment heuristics inspired by a popular statistical technique called PCA was used.

Potential/accuracy of MART as a new software effort estimation model compared to recently published models like neural networks, Radial Basis Function (RBF) linear regression, and Support Vector regression models with linear and RBF kernels was evaluated by Elish [14]. Comparison was based on a NASA software project dataset.

A new model using COCOMO II, 5 Scale factors and 17 Effort multipliers used as input was proposed by Attarzadeh&Ow [15]. A sigmoid activation function created a network to accomplish post architecture COCOMOII model. Results regarding MMRE, and Pred (0.25) were compared to traditional COCOMO.

Kalichanin-Balich&& Lopez-Martin [16] used Feed forward NN to estimate software development effort on short-scale projects. Totally 132 projects verified the new mechanism. Accuracy was measured regarding MER, i.e., MMER was 0.26, LRM 0.26 and NN 0.25.

A Modified MMRE proposed Dave &Dutta [17] used NASA dataset of 60 projects. They undertook experiments with three differing evaluation methods, i.e., MMRE, Modified MMRE, and Relative Standard Deviation (RSD). Three estimation modes were used, i.e., FFNN, Regression analysis and RBFNN. RBFNN was found to be better for effort estimation based on RSD and Modified MMRE according to the authors.

To ensure good results for problems with noise inputs, complex relationships between inputs and outputs and where inputs had high noise levels RBFN was proposed by Srichandan [18]. COCOMO 81 and Tukutuku were datasets. Clustering algorithm configured RBFN hidden layer. After using widths for models, it was found RBF accuracy using minimum width was better than using maximum width.

Various parameters affecting software development effort studied by Park &Baek [19] identified six variables other than software size for accurate effort estimation by using NN. Authors compared NN model with the two current regression models and human expert judgments. It was revealed that NN model was more accurate than other estimation procedures.

A NN based model and stepwise regression model for software development effort was implemented by de BarcelosTronto et al., [20]. Author reported results restate that NN model estimated software development effort more accurately. Authors compared results with multiple regression, COCOMO and SLIM models, showing that NN model suited effort estimation.

A multilayer feedforwardNN to accommodate the COCOMO model was proposed by Reddy &Raju [21]. COCOMO database consisting 63 projects was the dataset. Data set was divided into training and validation sets in a 80 %: 20 % ratio. Training set had 50 randomly chosen projects while the validation set had 13 projects.



## MATERIALS AND METHODS

This section describes COCOMO dataset, RBFN, KPCA, MMRE, Linear regression, GRNN methods.

### PROMISE EFFORT ESTIMATION DATASET COCOMO

COCOMO dataset has details of 63 software projects each described by 16 cost drivers or effort multipliers. Of 16 attributes, 15 are measured on a scale of six categories: *very low*, *low*, *nominal high*, *very high*, and *extra high*. A numeric values represents categories. Kilo Delivered Source Instructions (KDSI) is a numeric attribute. COCOMO dataset assesses new techniques comparative accuracy. [Figure -1] shows a COCOMO dataset's effort histogram [22].

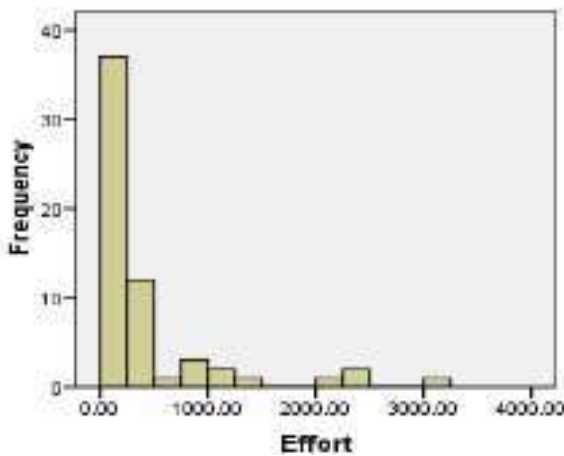


Fig: 1. Effort histogram of COCOMO81

### REGRESSION ANALYSIS WITH MEAN MAGNITUDE RELATIVE ERROR (MMRE) AND MEDIAN MAGNITUDE RELATIVE ERROR (MDMRE) FORMULA

An effort predictor's value is reported in many ways including MMRE and probability of a project having relative error less than or equal to L (PRED (L)). MMRE and PRED (L) are the accepted evaluation criteria to evaluate different software effort estimations [23].

MMRE and PRED are computed from relative error, or RE, the relative size of difference between actual/estimated value of individual effort  $i$

$$RE_i = (\text{predicated effort}_i - \text{actual effort}_i) / (\text{actual effort}_i) \quad (2)$$

Magnitude of Relative Error (MRE) was calculated by taking absolute value of relative error that is,

$$MRE_i = \text{abs}(RE_i) \quad (3)$$

MRE value is calculated for every observation  $i$  of actual/predicted effort. MRE aggregation over multiple observations ( $N$ ) is achieved through Mean MRE (MMRE) as follows:

$$MMRE = \frac{1}{N} \sum_i^N MRE_i \quad (4)$$

A complementary criterion is a prediction at level L, Pred (L) =  $k/N$ , where  $k$  is number of observations where MRE is less than/equal to L and  $N$  is total observations. Thus, Pred (25) gives projects percentage predicted with a MRE less than or equal to 0.25.

MMRE computes average of MREs over reference projects. As MMRE is susceptible to an individual outlying prediction, MdMRE is adopted when many observations are available. MREs median for  $n$  projects is MdMRE less sensitive to extreme values of MRE adopted. Despite this, MMRE is used in estimation accuracy. MMRE was criticized as it is unbalanced for validation procedures and often results in overestimation [24].

$$MdMRE = \text{median}(MRE_i) \quad (5)$$

### KERNEL PRINCIPAL COMPONENT ANALYSIS (KPCA)

In KPCA, this is crucial at two levels. From a practical point of view, this connection allows reduction of Eigen decomposition of (infinite dimensional) empirical kernel covariance operator to Eigen decomposition of kernel Gram matrix, which makes an algorithm feasible. From theory's view, it is a bridge between kernel covariance's spectral properties and those of the kernel integral operator [25].

So, KPCA's properties theoretical insight goes beyond this algorithm with direct consequences for understanding the kernel matrix/kernel operator's spectral properties. This makes a study of KPCA interesting: kernel Gram matrix is a central object in kernel methods and its spectrum has a major role in kernel algorithms; this was shown in Support Vector Machines (SVM). Understanding the kernel matrices eigenvalues behaviour, their stability and how they relate to eigenvalues of corresponding kernel integral operator is crucial to understand kernel-based algorithms statistical properties.

KPCA is a PCA's functional generalization similar to Locally Linear Embedding, Isomap or spectral clustering methods. It allows as many principal components as data samples in a training set, with directions being nonlinear [26].

The approach's rationale is considering straight lines to recover from n-dimension space are data's principal components. Though linear, PCA cannot be used as:

- directions looked for are not orthogonal;
- There are more directions to find than space dimension (under determined context).

Though Kernel PCA was defined in RKHSs, same framework applies to reproducing kernel space, provided it has a representer theorem.

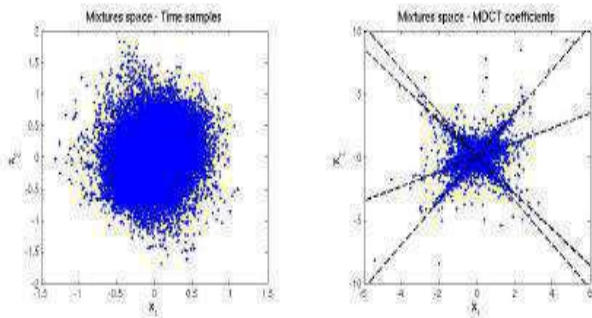


Fig: 2(a). Scatter plot of x1 w.r.t x2 in the time-domain. (b) Scatter plot of x1 w.r.t x2 in the MDCT domain - the dashed lines represent the directions of the mixing matrix

It refers to Algorithm 1 for precise descriptions of algorithm involving KPCA in underdetermined contexts. For a given (reproducing) kernel  $k : X \times X \rightarrow \mathbb{R}$  and data set  $\{x_t\}_{t=1, \dots, N}$ , one performs Eigen decomposition of a kernel matrix defined as:

$$K = [k_{tt'}]_{t,t'=1, \dots, N}$$

with :  $k(x_t, x_{t'})$  (6)

(in PCA, this is performed over correlation matrix  $E[xx^T]$ ). Directions corresponding to n largest eigenvalues are scatter data line estimates.

It has to assure that K's eigenvectors are recoverable, i.e. that are expressed as a finite linear kernels combination evaluated on training data. This is got through a representer theorem that extends Wahba and Kimeldorf's to Krein reproducing spaces, thereby ensuring that one expresses the eigenvectors as:

$$v_i(X) = \sum_{t=1}^N \alpha_t^i k(x, x_t) \tag{7}$$

With  $\{\alpha_1^i, \dots, \alpha_N^i\}$  in  $\mathbb{R}^2$ , for any x in  $\mathbb{R}^2$ , and  $v_i$  eigenvector corresponding to ith largest eigenvalue. This

Equation proves that KPCA yields linear directions if the kernel  $k(\cdot, \cdot)$  is linear. Next paragraph is dedicated to presenting kernel in Algorithm 1.

Algorithm 1: KPCA for mixing matrix estimation.

- Step 0: Initialization  
 Set  $\eta, \theta_0$  ( $m, N$  and  $n$  are known);  
 Define reproducing kernel  $k(\cdot, \cdot)$  as in equation (8);  
 Define  $x_i = [x_{i1}, x_{i2}]^T$ ; perform shrinkage with threshold  $\eta$  on the  $\|x_i\|$ 's; map (by symmetry) all remaining  $x_i$ 's to the positive half - plane.
- Step 1: Krein kernel principal component analysis  
 Define matrix  $K = [k_{ij}]_{i,j=1,\dots,N}$  as in equation (6);  
 Find the eigenvalues and eigenvectors for  $K$ .
- Step 2: Estimation of the mixing matrix  $A$   
 Keep the eigenvectors corresponding to the  $n$  largest eigen - values;  
 Express the corresponding  $n$  directions in  $X$  using equation (7).

Fig: 3. Reproducing kernel  $k(\cdot, \cdot)$  plotted as a function (left); eigenvalues for  $k(\cdot, \cdot)$  (right)

The KPCA kernel needs to be:

- Linear, as it looks for lines in  $R^m$ ;
- Parameterized so that KPCA can be tuned according to problem difficulty (sparsity of the sources? noisy data?);
- Designed in a Krein framework (obvious linear kernel  $\langle \cdot, \cdot \rangle_X$  which is a Hilbert rk that cannot be parameterized regarding angles).

Denoting  $\hat{\langle x_i, x_i \rangle}$  angle between  $x_i$  and  $x_i$ , kernel

$$k_{\theta_0}(x_i, x_i) = \begin{cases} \frac{\cos(\hat{\langle x_i, x_i \rangle}) - \cos \theta_0}{1 - \cos \theta_0} & \text{if } (\hat{\langle x_i, x_i \rangle}) \leq \theta_0 \\ 0 & \text{else} \end{cases}$$

$$(x, x') \rightarrow \frac{k(x, x')}{\sqrt{k(x, x)k(x', x')}} \tag{8}$$

Satisfies all the required conditions (it recall that if  $k(\cdot, \cdot)$  is a RKKS kernel, then so is

that  $\cos(\hat{\langle x_i, x_i \rangle}) = \langle x_i, x_i \rangle_X$  as the  $x_i$ 's have unit norm). This kernel is plotted in [Figure -3], with eigenvalues.

Computing dot products in feature space: From  $\tilde{K} : \tilde{K}_{ij} = (\tilde{\phi}(y_i) \cdot \tilde{\phi}(y_j))$  Equation, it is seen that to compute kernel matrix, only vectors dot products in feature space  $F$  are required, while explicit calculation of map  $U(y)$  need not be known. A dot product is computed through use of a kernel function. This is a "kernel trick". The Mercer kernels alone are used as a kernel function [27].

Kernel function  $k(y_i, y_j)$  calculates dot product in space  $F$  directly from input space vectors  $R^M$ :

$$k(y_i, y_j) = (f(y_i) \cdot f(y_j)) \tag{9}$$

Common kernel functions are Gaussian kernel and polynomial kernel.

Kernel for linear PCA: If the kernel function is chosen as polynomial kernel of order one

$$k(y_i, y_j) = (y_i \cdot y_j) \tag{10}$$

Then linear PCA is performed on sample realizations. Using a kernel matrix to perform linear PCA is the same as "method of snapshots" known in reduced-order modeling. This method is computationally efficient than K-L expansion's standard implementation. Using a kernel matrix, an eigenvalue problem of size  $N \times N$  is needed, whereas in size of eigenvalue problem is  $M \times M$ . In most cases, available experimental data is smaller than data dimensionality.

Kernel for nonlinear PCA: Choosing a nonlinear kernel function results in nonlinear PCA performance. A common kernel function is Gaussian kernel:

$$k(y_i, y_j) = \exp\left(-\frac{\|y_i - y_j\|^2}{2\sigma^2}\right) \quad (11)$$

Where  $\|y_i - y_j\|$  is squared L2-distance between two realizations. A kernel width parameter  $\sigma$  controls kernel flexibility. A larger value of  $\sigma$  allows more "mixing" between realizations elements, whereas smaller value of  $\sigma$  uses few significant realizations. A choice for  $\sigma$  is average minimum distance between two realizations in input space:

$$\sigma^2 = c \frac{1}{N} \sum_{i=1}^N \min_{j \neq i} \|y_i - y_j\|^2, \quad j = 1, \dots, N, \quad (12)$$

Where c is a user-controlled parameter.

Why KPCA is better than PCA: 1) PCA does not support mean for multilayer NN. 2) Large dataset like 0.000009 assume non-linear value PCA does not take nonlinear space value [28]. KPCA identifies the kernel's principal directions where data varies largely. 3) PCA supports mapping and so PCA works with single layer NN. A huge data set in practice leads to huge K, storing which is an issue. A way to handle this is to perform clustering on the huge dataset, populating the kernel with the clusters means. 4) KPCA supports implicit mapping and so it works with multilayer NN.

## LINEAR REGRESSION

Regressions techniques predict software evaluate accuracy in evaluation/validation. A Regression Analysis views effect of independent variables on a dependent variable. It aims to see how much dependent variable are independent variables based. Linear regression is a statistical technique used for prediction or evaluating a linear interrelationship between two numerical variables.

A linear regression model having exponential transformation predicts relations between variables involving software size and effort to raise reliability in software effort estimation. It changes the independent variable value and sees resulting dependent variable change. The aim is to locate to what extent a dependent variable is described using an independent variable. Simple Linear Regressions have one dependent and independent variable each [29].

$$Y = a + bX + C \quad (13)$$

Where

Y: Dependent variable X: Independent variable

b: Coefficient of variable X

a: Y intercept

C: Constant

More than one independent variable describes dependent variable change in Multiple Linear Regression Analysis [30].

C: Constant

Consider the problem of approximating set of data,

$$D = \left\{ (x^1, y^1), \dots, (x^l, y^l) \right\}, \quad x \in \mathfrak{R}^n, y \in \mathfrak{R} \quad (14)$$

With a linear function,

$$f(x) = \langle w, x \rangle + b \quad (15)$$

The optimal regression function is given by the minimum of the functional,

$$\Phi(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_i (\xi_i^- + \xi_i^+) \quad (16)$$

Where C - a pre-specified value, and  $\xi^-$ ,  $\xi^+$  - slack variables that represent system outputs upper and lower constraints.

SMOPolykernel and Radial Basis Neural Networks (RBFN) kernel

The RBFN NN involves three layers. An input layer of sources nodes (cost drivers); a hidden layer where neurons compute output using RBF, which is a Gaussian function, and an output layer that constructs a linear hidden neuron outputs weighted sum and supplies the network's response (effort). A RBF NN configured for software effort estimation has an output neuron. Hence, it implements an output-input relation in the Equation which is composition of nonlinear mapping realized by hidden layer realized by the output layer's linear mapping [31]

$$F(x) = \sum_{j=1}^M \beta_j \phi_j(x) \quad (17)$$

Where M is number of hidden neurons,  $x \in \mathbb{R}^p$  is input,  $\beta_j$  are RBFN networks output layer weights and  $\phi(x)$  is Gaussian RBF given by:

$$\phi_j(x) = e^{-\left(\frac{\|x-c_j\|^2}{\sigma_j^2}\right)} \tag{18}$$

Where  $c_j \in \mathbb{R}^p$  and  $\sigma_j$  are center and width of jth hidden neuron and  $\|\cdot\|$  denotes Euclidean distance. RBFN are powerful alternatives approximating classifying a pattern set some times better than Multi-Layer Perceptron (MLP) NN [32].

RBFs differ from MLPs as an overall input-output map is constructed from local Gaussian axons contributions and need fewer training samples training faster than a MLP. A popular method to estimate centers and widths includes using an unsupervised technique called k-nearest neighbour rule. The clusters centers give RBF's centers and distance between clusters is the Gaussians width. Centers computation, used in RBF NN kernels function is main focus to achieve efficient algorithms in the pattern set's learning process. Adequate centers choice implies high performance, concerning convergence, learning times and generalization.

SVMs based methods are for classification [33]. For a training data  $(x_i, y_i), i = 1, \dots, n$ , where  $x_i \in \mathbb{R}^d$  is a feature vector and  $y_i \in \{+1, -1\}$  indicates class value of  $x_i$  solves the optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \tag{19}$$

Subject to

$$\begin{aligned} y_i (w^T \Phi(x_i) + b) &\geq 1 - \xi_i \text{ for } i=1 \dots n \\ \xi_i &\geq 0 \end{aligned} \tag{20}$$

Where  $\Phi: \mathbb{R} \rightarrow H$ , H being a high dimensional space  $w \in H$ , and  $b \in \mathbb{R}, C \geq 0$  is a parameter controlling margin errors minimization and margins maximization.  $\Phi$  is chosen that an efficient kernel function K exists. In practice, this optimization problem is solved using Lagrange Multiplier. SMO [34] is an to solve SVM QP problem. Its advantage is its ability to solve Lagrange multipliers without numerical QP optimization. The Lagrangian form is as follows :

$$\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^n \alpha_i \tag{21}$$

Subject to

$$\begin{aligned} 0 &\leq \alpha_i \leq C \text{ for } i=1 \dots n \\ \sum_{i=1}^n \alpha_i y_i &= 0 \end{aligned} \tag{22}$$

On solving optimization problem, w is computed as [35]:

$$w = \sum_{i=1}^n \alpha_i y_i \Phi(x_i) \tag{23}$$

$x_i$  is a support vector if  $\alpha_i \neq 0$ . New instance x is computed by:

$$f(x) = \sum_{i=1}^{n_s} \alpha_i y_i K(s_i, x) + b \tag{24}$$

Where  $s_i$  are support vectors and  $n_s$  is number of vectors. The polynomial kernel function is given by:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \text{ where } \gamma > 0 \tag{25}$$

And the Radial basis function (RBF) kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \text{ where } \gamma > 0 \tag{26}$$

M5 algorithm

Tree-building algorithms like C4.5 determine which attributes best classify remaining data, followed by iterative tree construction. Decision trees immediate conversion to rules easily interpreted by decision-makers is their advantage. For numeric data mining

prediction, it is common to use regression or model trees [36]. Both build decision tree structures where leaves are responsible for a specific input space local regression. The difference is that while regression trees generate constant output values for input data subsets (zero-order models), model trees generate linear (first-order) models for every subset.

M5 algorithm builds trees with leaves being linked to multivariate linear models and tree nodes are chosen over an attribute maximizing expected error reduction as a standard deviation function of output parameter. M5 algorithm builds decision trees that divide attribute space in orthohedric clusters, with border paralleling the axis. Their advantage is their being converted o rules easily; each tree branch has the following condition: attribute  $\leq$  value or attribute  $>$  value.

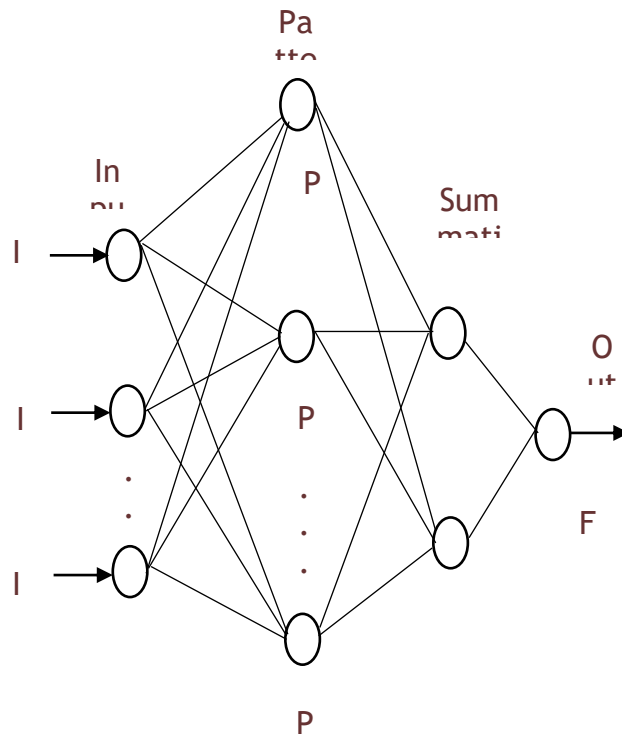
**GENERALIZED REGRESSION NEURAL NETWORKS (GRNN)**

ANN is used for cost estimation as it learns from earlier data. Three factors defining ANN are: i) interconnection pattern between different neuron layers, ii) learning process to update interconnections weights and iii. activation function converting a neuron's weighted input to output activation.

Network nodes are split into input layer which is linked to weights having information on input signals and output layer which goes through network nodes in hidden layer [37].

A basic NN has inputs applied by weights combined to give an output. Different NN learning algorithms like Delta rule leaning, Perceptron learning and back propagation learning are used. It uses Delta rule learning algorithm to train a NN and solve different problems. Delta rule learning algorithm uses sigmoid activation function in which every neuron has continuous activation function rather than threshold activation function.

GRNN is a RBF based, supervised learning model used for classification, regression and time series predictions. GRNN architecture is seen in [Figure -4] [38].



**Fig: 4. GRNN Architecture**

GRNN comprises four layers named input layer, pattern layer, summation layer and output layer. Input units numbers depend on total observation parameters i.e. input vector 'I' (feature matrix  $F_i$ ). Input layer connected to pattern layer has neurons to ensure training patterns and output to summation layer for performing normalization of resultant output set. Each pattern layer is connected to summation neurons which calculate weight vector using these equations [39].

$$W_i = e^{-\left[\frac{\|I - I_i\|^2}{2h^2}\right]}$$

$$F(I) = \frac{\sum_{i=1}^n T_i W_i}{\sum_{i=1}^n W_i} \tag{27}$$

Where output F (I) is weighted average of target values Ti of training cases Ii close to input case I. GRNN is one-pass learning algorithm based having a highly parallel structure. GRNN is a powerful memory based network that estimates continuous variables and converges to an underlying regression surface. GRNN's strength is its ability to deal with sparse data effectively. GRNNs feature fast training times, model non-linear functions and are known to do well in noisy environments if provided enough data. The GRNN algorithm can ensure smooth transition from one observed value to another, even with sparse data in a multidimensional measurement space. GRNN applications produce continuous valued outputs.

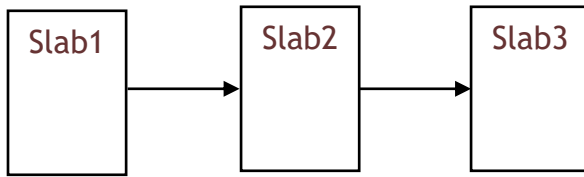


Fig: 5.GRNN Network

For GRNN networks, the number of hidden layer (Slab2) neurons is the number of patterns in a training set as a neuron represents each pattern. Input layer neurons (Slab1) are the inputs, and output layer neurons (Slab3) correspond to number of outputs.

GRNN's advantage is the speed with which a network is trained. There are no training parameters like learning rate and backpropagation network momentum, but a smoothing factor is applied after training a network. Smoothing factor is an GRNN adjustable parameter, so overtraining is not likely in GRNN. Smoothing allows GRNN interpolate between patterns/spectra in training sets. Smoothing determines how tightly a network matches predictions to training patterns data. Smoothing factor for GRNN networks must be greater than 0 ranging from 0.01 to 1 with good results. The proposed GRNN parameters are tabulated in [Table -2].

Table: 2.PARAMETERS FOR THE PROPOSED GRNN

<b>Number of input</b>	<b>17</b>
Number of output	1
Number of hidden layer	2
Number of neurons in hidden layer	6
Number of cluster centers	85
Competitive rule	conscience full metric – Euclidean
Activation function	tanh
Momentum	0.5
Learning rate	0.1

## RESULTS AND DISCUSSION

The attributes of the COCOMO dataset is used as it is and feature transformation of the attributes using KPCA). In the first set of experiment without data transformation using PCA, the MMRE and MdmRE is evaluated using various techniques such as M5, Linear regression, SMOPolykernel and RBF kernel and the proposed Generalized Regression Neural Network. In the second set of experiments, the same is evaluated with feature transformation of the attributes using KPCA.

Table: 3. PARAMETERS FOR THE PROPOSED GRNN

M5 Actual	M5 - Kernel PCA predicted	MMRE
387	251.58	0.35
18	-61.69	4.43
15	-103.45	7.9
237	259.65	0.1
958	628.78	0.34

14	49.01	2.5
57	107.34	0.88
33	31.3	0.05
98	463.89	3.73
605	1021.06	0.69
423	655.96	0.55
702	2917.59	3.16
724	1126.56	0.56
70	255.04	2.64
20	52.31	1.62
523	242.14	0.54
7.3	8.69	0.19
1272	589.25	0.54
88	171.16	0.95
55	135.16	1.46
8	53.42	5.68
45	376.78	7.37
1075	874.58	0.19
243	2134.84	7.79
38	82.67	1.18
106	239.03	1.26
321	1213.44	2.78
1063	1093.85	0.03
201	133.11	0.34
126	576.56	3.58
240	1259.74	4.25
6600	11253.02	0.71
87	502.26	4.77
61	324.14	4.31
122	316.15	1.59
8	-227.09	29.39
40	643.72	15.09
1600	1664.47	0.04
11400	2364.59	0.79
6400	2361.72	0.63
79	-206.75	3.62
2455	1586.89	0.35
156	63.17	0.6
41	424.56	9.36
8	-77.01	10.63
130	1267.25	8.75
6	-361.3	61.22
82	468.67	4.72
12	-218.62	19.22
73	952.22	12.04
36	65.26	0.81
176	445.23	1.53
83	250.66	2.02
218	1048.4	3.81
453	320.4	0.29
539	523.01	0.03
5.9	-338.05	58.3
9	220.04	23.45
43	13.26	0.69
230	142.14	0.38
47	173.46	2.69
2040	270.99	0.87
50	-4014.22	81.28



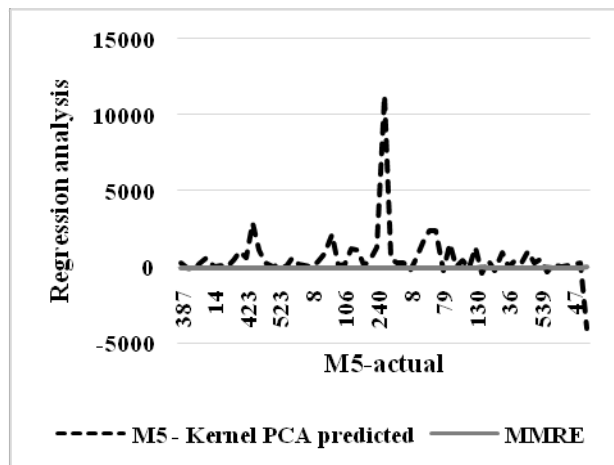


Fig. 6. M5-actual

From the [Table -3] and [Figure -6], it can be observed that the MMRE method averagely decreased by 195.62% when compared with M5 - Kernel PCA predicted method with M5-actual.

Table: 4.LINEAR REGRESSION-ACTUAL

Linear regression actual	linear regression Kernel PCA-Predicted	MMRE
387	469.65	0.21
18	182.52	9.14
15	205.73	12.72
237	255.78	0.08
958	327.05	0.66
14	175.42	11.53
57	187.25	2.29
33	270.92	7.21
98	251.77	1.57
605	463.41	0.23
423	380.06	0.1
702	3096.17	3.41
724	1050.5	0.45
70	329.5	3.71
20	176.36	7.82
523	427.19	0.18
7.3	159.62	20.87
1272	3698.57	1.91
88	219.79	1.5
55	261.59	3.76
8	169.59	20.2
45	466.34	9.36
1075	306.92	0.71
243	1099.36	3.52
38	211.47	4.56
106	395.59	2.73
321	349.34	0.09
1063	621.22	0.42
201	362.85	0.81
126	734.3	4.83
240	598.59	1.49
6600	11830.69	0.79
87	299.23	2.44
61	113.51	0.86
122	149.44	0.22

8	127.66	14.96
40	126.16	2.15
1600	1775.59	0.11
11400	1963.6	0.83
6400	2037.19	0.68
79	284.55	2.6
2455	1724.29	0.3
156	704.66	3.52
41	173.58	3.23
8	163.35	19.42
130	334.35	1.57
6	196.49	31.75
82	244.74	1.98
12	259.87	20.66
73	177.68	1.43
36	320.99	7.92
176	323.42	0.84
83	352.96	3.25
218	379.3	0.74
453	799.12	0.76
539	696.17	0.29
5.9	170.66	27.93
9	165.67	17.41
43	165.9	2.86
230	298.88	0.3
47	563.01	10.98
2040	940	0.54
50	338.73	5.77

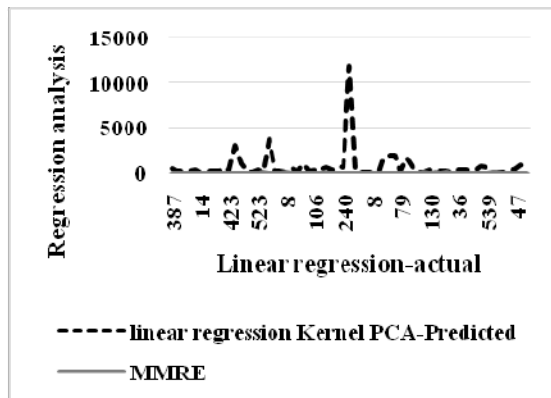


Fig:7. Linear regression-actual

From the [Table -4] and [Figure -7], it can be observed that the MMRE method averagely reduced by 197.18% when compared with linear regression Kernel PCA-Predicted method with Linear regression-actual.

Table: 5: M5-ACTUAL

M5 Actual	M5 predicted	MMRE
387	270.89	0.3
18	-134.35	8.46
15	-27.83	2.86
237	32.02	0.86
958	528.87	0.45
14	-77.98	6.57
57	80.88	0.42
33	-413.87	13.54
98	892.3	8.11
605	2070.09	2.42
423	479.76	0.13
702	3028.23	3.31
724	1408.59	0.95
70	-208.06	3.97

20	-82.82	5.14
523	406.23	0.22
7.3	-26.33	4.61
1272	4881.49	2.84
88	224.52	1.55
55	119.59	1.17
8	-88.36	12.04
45	538.44	10.97
1075	1114.52	0.04
243	1954.62	7.04
38	-181.48	5.78
106	474.68	3.48
321	-60.63	1.19
1063	1749	0.65
201	237.37	0.18
126	276.43	1.19
240	130.22	0.46
6600	12754.64	0.93
87	-0.89	1.01
61	147.99	1.43
122	176.05	0.44
8	-58.91	8.36
40	113.44	1.84
1600	1388.09	0.13
11400	6500	0.43
6400	2907.02	0.55
79	-530.49	7.72
2455	2046.46	0.17
156	-110.05	1.71
41	214.05	4.22
8	-359.67	45.96
130	1130.62	7.7
6	-118.21	20.7
82	-100.9	2.23
12	171.41	13.28
73	-123.38	2.69
36	-30.83	1.86
176	788.17	3.48
83	355.92	3.29
218	-3.75	1.02
453	620.92	0.37
539	588.57	0.09
5.9	-65.98	12.18
9	250.13	26.79
43	27.17	0.37
230	897.51	2.9
47	-326.15	7.94
2040	901.71	0.56
50	-12.54	1.25

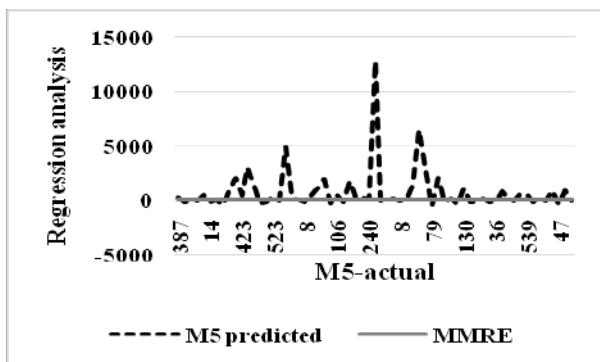


Fig: 8. M5-actual

From the [Table -5] and [Figure -8], it can be observed that the MMRE method averagely decreased by 197.64% when compared with M5 Predicted method with M5-actual.

Table: 6. LINEAR REGRESSION-ACTUAL

Linear regression actual	linear regression Predicted	MMRE
387	478.42	0.24
18	182.61	9.14
15	208.85	12.92
237	258.51	0.09
958	329.45	0.66
14	175.51	11.54
57	185.44	2.25
33	273.74	7.3
98	251.19	1.56
605	469.18	0.22
423	378.98	0.1
702	3085.02	3.39
724	1040.45	0.44
70	326.36	3.66
20	179.67	7.98
523	435.07	0.17
7.3	162.49	21.26
1272	3673.21	1.89
88	217.67	1.47
55	260.24	3.73
8	169.03	20.13
45	473.4	9.52
1075	305.79	0.72
243	1090.32	3.49
38	213.79	4.63
106	398.51	2.76
321	355.72	0.11
1063	624.37	0.41
201	368.3	0.83
126	737.04	4.85
240	603.88	1.52
6600	11768.41	0.78
87	302.75	2.48
61	115.12	0.89
122	150.58	0.23
8	128.49	15.06
40	127.91	2.2
1600	1795.77	0.12
11400	1952.73	0.83
6400	2055.82	0.68
79	282	2.57
2455	1751.55	0.29
156	709.27	3.55
41	173.25	3.23
8	164.83	19.6
130	333.87	1.57
6	194.74	31.46
82	249.12	2.04
12	263.25	20.94
73	178.5	1.45
36	319.74	7.88
176	326.66	0.86
83	359.24	3.33
218	383.32	0.76
453	794.14	0.75
539	702.06	0.3
5.9	170.69	27.93
9	165.08	17.34

43	167.9	2.9
230	301.94	0.31
47	562.96	10.98
2040	936.85	0.54
50	337.21	5.74

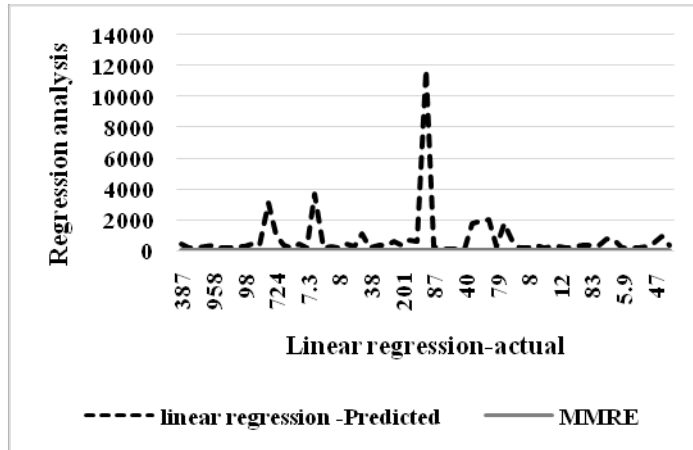


Fig:9. Linear regression-actual

From the [Table -6] and [Figure -9], it can be observed that the MMRE method averagely reduced by 197.17% when compared with linear regression Predicted method with Linear regression-actual.

## CONCLUSION

This paper proposes a Generalized Regression NN to use improved software effort estimation for COCOMO dataset. This paper uses Mean Magnitude Relative Error (MMRE) and Median Magnitude Relative Error (MdMRE) as evaluation criteria. The new method performance was compared to that of three regression algorithms including M5, linear regression and modified SVM to avoid quadratic problem. Two kernels were used for SVM with the first being a polykernel and second using RBF. It is found that the new method outperformed classical regression algorithms in experiments.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Paramshetti P, Phalke DA. [2015] Software Defect Prediction for Quality Improvement Using Hybrid Approach. *International Journal of Application or Innovation in Engineering & Management (IJAIEEM)*, 4(6)
- [2] Shubitsha P, Rajan JK.[2014] Artificial Neural Network Models For Software Effort Estimation.
- [3] Santani D, Bunde M, Rijwani, P. Artificial Neural Networks for Software Effort Estimation: A Review.
- [4] Prasad Reddy PVGD, Sudha KR, Rama Sree P. [2011] Prediction of Software Development Effort Using RBNN And GRNN. *International Journal of Computer Science Engineering and Technology*, 1(4):185-190.
- [5] Reddy PVGD, Sudha KR, Sree PR, Ramesh SNSVSC. [2010] Software effort estimation using radial basis and generalized regression neural networks. arXiv preprint arXiv:1005.4021.
- [6] Kaushik A, Soni AK, Soni R. [1969] A Simple Neural Network Approach to Software Cost Estimation. *Global Journal of Computer Science and Technology*, 13(1).
- [7] Samuel RT, Cao Y. [2014] Fault detection in a multivariate process based on kernel PCA and kernel density estimation. In *Automation and Computing (ICAC), 2014 20th International Conference on* (pp. 146-151). *IEEE*.

- [8] Patil LV, Waghmode RM, Joshi SD, Khanna V. [2014] Generic model of software cost estimation: A hybrid approach. In Advance Computing Conference (IACC), 2014 IEEE International (pp. 1379-1384). *IEEE*.
- [9] Mukhopadhyay S, Gundappa M, Srinivasan R, Narasimhan S. [2013] A Novel attempt to reduce engineering effort in modeling non-linear chemical systems for Operator Training Simulators. In *American Control Conference (ACC)*, 2013 (pp. 1902-1907). *IEEE*.
- [10] Grbic R, Sliskovic D, Kadlec, P. [2013] Adaptive soft sensor for online prediction and process monitoring based on a mixture of Gaussian process models. *Computers & chemical engineering*, 58:84-97.
- [11] Papaioannou A, Zafeiriou S. [2014] Principal component analysis with complex kernel: The widely linear model. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(9): 1719-1726.
- [12] Musa AB. [2014] A comparison of  $\ell_1$ -regularization, PCA, KPCA and ICA for dimensionality reduction in logistic regression. *International Journal of Machine Learning and Cybernetics*, 5(6):861-873.
- [13] Tosun A, Turhan B, Bener AB. [2009] Feature weighting heuristics for analogy-based effort estimation models. *Expert Systems with Applications*, 36(7): 10325-10333.
- [14] Elish MO. [2009] Improved estimation of software project effort using multiple additive regression trees. *Expert Systems with Applications*, 36(7):10774-10778.
- [15] Attarzadeh I, Ow SH. [2010] Proposing a new software cost estimation model based on artificial neural networks. In *Computer Engineering and Technology (ICCET)*, 2010 2nd International Conference on, *IEEE* 3: V3-487).
- [16] Kalichanin-Balich I, Lopez-Martin C. [2010] Applying a feedforward neural network for predicting software development effort of short-scale projects. In *Software Engineering Research, Management and Applications (SERA)*, 2010 Eighth ACIS International Conference on, *IEEE*, (pp.269-275).
- [17] Dave VS, Dutta K. [2011] Neural network based software effort estimation & evaluation criterion MMRE. In *Computer and Communication Technology (ICCCT)*, 2011 2nd International Conference on (pp. 347-351). *IEEE*.
- [18] Srichandan S. [2012] A new approach of Software Effort Estimation Using Radial Basis Function Neural Networks. *International Journal on Advanced Computer Theory and Engineering (IJACTE)*, 1(1):113-120.
- [19] Park H, Baek S. [2008] An empirical validation of a neural network model for software effort estimation. *Expert Systems with Applications*, 35(3):929-937.
- [20] deBarcelosTronto IF, da Silva, JDS, Sant'Anna, N. [2008] An investigation of artificial neural networks based prediction systems in software project management. *Journal of Systems and Software*, 81(3):356-367.
- [21] Reddy CS, Raju KVSVN. [2009] A concise neural network model for estimating software effort. *International Journal of Recent Trends in Engineering*, 1(1):188-193.
- [22] Boetticher G, Menzies T, Ostrand T. [2007] PROMISE Repository of empirical software engineering data <http://promisedata.org/> repository, West Virginia University, Department of Computer Science.
- [23] Singh BK, Misra AK. [2012] An Alternate Soft Computing Approach for Efforts Estimation by Enhancing Constructive Cost Model in Evaluation Method. *iji*, 10(1).
- [24] Shepperd MJ, Schofield C. [1997] Estimating Software Project Effort Using Analogies, *IEEE Transaction on Software Engineering* 23:736-743.
- [25] Blanchard G, Bousquet O, Zwald L. [2007] Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3): 259-294.
- [26] Desobry F, Févotte C. [2006] Kernel PCA based estimation of the mixing matrix in linear instantaneous mixtures of sparse sources. In *Acoustics, Speech and Signal Processing*, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, *IEEE*, 5: V-V).
- [27] Ma X, Zabarar N. [2011] Kernel principal component analysis for stochastic input model generation. *Journal of Computational Physics*, 230(19):7311-7331.
- [28] Waghmode S, Kolhe K. [2014] A Novel Way of Cost Estimation in Software Project Development Based on Clustering Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(4)
- [29] Bhatia S, Attri VK. [2015] MACHINE LEARNING TECHNIQUES IN SOFTWARE EFFORT ESTIMATION USING COCOMO DATASET. *International Journal of Research and Development organization (IJRD)*, 2 (6)
- [30] V Vapnik, S Golowich, A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems* 9, pages 281–287, Cambridge, MA, 1997. MIT Press.
- [31] Idri A, Zakrani A, Zahi, A. [2010] Design of radial basis function neural networks for software effort estimation. *International Journal of Computer Science*, 7(4).
- [32] Bautista AM, Castellanos A, San Feliu T. [1993] SOFTWARE EFFORT ESTIMATION USING RADIAL BASIS FUNCTION NEURAL NETWORKS. *INFORMATION THEORIES & APPLICATIONS*, 319.
- [33] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001.
- [34] Platt JC. [1999] Fast training of support vector machines using sequential minimal optimization". *Advances in kernel methods: Support vector machines*, B. Schölkopf et al. (ed.), MIT Press,
- [35] Breiman L, Friedman J, Olshen R, Stone CJ. [1984] *Classification and Regression Trees*, Chapman and Hall, New York, 1984.
- [36] Waghmode RM, Patil LV, Joshi SD. [2013] A Collective Study of PCA and Neural Network based on COCOMO for Software Cost Estimation. *International Journal of Computer Applications*, 74(16):25-30.
- [37] Ajay Prakash, BV Ashoka, DV Manjunath, Aradhya VN. [2015]. Estimating Software Development Effort using Neural Network Models. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(7)
- [38] Thwin MMT, Quah TS. [2005] Application of neural networks for software quality prediction using object-oriented metrics. *Journal of systems and software*, 76(2):147-156.

\*\*DISCLAIMER: This published version is uncorrected proof; plagiarisms and references are not checked by IIOABJ; the article is published as it is provided by author and approved by guest editor.

# SCHEDULING BASED ON HYBRID PARTICLE SWARM OPTIMIZATION WITH CUCKOO SEARCH ALGORITHM IN CLOUD ENVIRONMENT

Sumathi<sup>1\*</sup> and Poongodi<sup>2</sup>

<sup>1</sup> Dept. of Computer Science and Engineering, Jayam College of Engineering and Technology, Dharmapuri, TN, INDIA

<sup>2</sup> Dept. of ECE, Karpagam College of Engineering, Coimbatore, TN, INDIA

## ABSTRACT

Cloud computing provides a centralized pool of configurable computing resources and computing outsourcing mechanisms that enable different computing services to different people in a way similar to utility based systems. Job scheduling algorithm achieves a high performance computing and the best system throughput. The scheduling problem can be solved by enumeration method, heuristic method or approximation method. In this work, a heuristic method is proposed for optimizing scheduling in cloud. Particle Swarm Optimization (PSO) refers to a population-based meta-heuristic algorithm that is inspired by the social behavior of populations with collaborative properties. Cuckoo Search (CS) is a new and efficient population-based heuristic evolutionary algorithm for solving optimization problems. CS has the advantages of simple implementation and few control parameters. In this paper, scheduling is performed based on hybrid PSO with CS algorithm. Results show that the proposed method performs better in terms of average schedule length and ratio of successful execution.

Published on: 14<sup>th</sup>– August-2016

### KEY WORDS

Cloud computing, Job scheduling, hybrid Particle Swarm Optimization (PSO) with Cuckoo Search (CS) algorithm, average schedule length and ratio of successful execution.

\*Corresponding author: Email: [sumathid.clo@gmail.com](mailto:sumathid.clo@gmail.com)

## INTRODUCTION

Cloud computing refers to a model of processing information, storing it as well as delivery wherein physical resources are provided to clients on demand. Instead of purchasing actual physical devices servers, storage, or any networking equipment, clients lease these resources from a cloud provider as an outsourced service. It can also be defined as “management of resources, applications and information as services over the cloud (internet) on demand”. Cloud computing is a model for enabling convenient and on demand network access to a shared group of computing resources that can be rapidly released with minimal management effort or service provider interaction [1]. The goal of cloud computing is to provide on-demand computing service with high reliability, scalability, and availability [2]. Cloud Computing is rapidly spreading out at an excellent pace amongst IT enterprises because of the incredible saving of costs in infrastructure as well as decrease in the costs of IT management.

This is a vast domain and has several aspects on which efficacy is dependent on, one major aspect being scheduling. Scheduling is an important factor requiring attention within the domain of cloud computing. Quantity of energy utilized, costs incurred for providing services over the cloud, execution time are all important concerns and improving task scheduling assists in their minimization. Plenty of research has been carried out in this domain.

Job scheduling [3] is an important activity that is carried out in all computational environments. Cloud computing is emerging rapidly to become one of the most useful latest technologies. For effectively increasing the functioning of cloud computing environments, task scheduling is carried out for gaining most amount of profit. The aim of scheduling algorithms within distributed systems is the spreading of load on processors as well as the maximization of their usage and the minimization of total task implementation time. Task scheduling, a popular optimization issue, is key in improving flexible as well as dependable models. The major purpose is the scheduling of tasks to adaptable resources as per adaptable time, which includes discovering proper sequences in which tasks may be implemented

under transaction logic constraints. There are two major groups of scheduling algorithms: Static scheduling algorithm and Dynamic scheduling algorithms. Both of them have their own benefits as well as limitations. The latter possesses better performance over the former, although it also has more overheads in comparison.

Scheduling process in cloud can be generalized into three stages: 1. Resource discovering and filtering: Datacenter Broker discovers the resources present in the network system and collects status information related to them. 2. Resource selection: target resource is selected based on certain parameters of task and resource which is deciding stage. 3. Task submission: task is submitted to resource selected.

Optimization problems are in Class NP-hard [4]. These issues may be resolved through enumeration, heuristic or approximation methods. In the first approach, optimal solutions may be chosen if all feasible solutions are enumerated and contrasted one by one. When quantity of instances is huge, extensive enumeration is not possible for scheduling issues. In this case, the second approach, the heuristic is the next best way to obtain reasonably good solutions in a short amount of time. The last kind of algorithms is approximation algorithms that are utilized to discover approximate solutions for optimized solutions. The algorithms are later utilized for issues wherein exact polynomial time algorithms are known. Improving task data locality in huge scale data processing models is critical to complete the tasks. Almost all methods for improving data locality are greedy or ignore global optimization or even suffer from high computational complexities. This issue is handled through a heuristic task scheduling model.

Task scheduling in cloud computing is an NP-hard problem, PSO as one of the heuristic algorithms has been applied in solving scheduling problem and other NP-hard problems, it is relatively easy to implement compared with the ant colony optimization algorithm and genetic algorithm. PSO has been utilized in workflow scheduling issue in cloud computing environments. However, PSO has some disadvantages, such as poor local search ability and not suitable for problems in discrete areas. This study proposes a workflow scheduling strategy in cloud computing based on hybrid particle swarm algorithm with Cuckoo Search algorithm (HPSOCS) in order to solve these shortages.

PSO has several advantages, such as fast convergence speed, but it also has some defects, such as premature convergence, and it easily falls into local optima. CS has several advantages, such as few control parameters and high efficiency, but it also has some defects, such as slow convergence speed and low accuracy. A PSO and CS hybrid algorithm should be developed as a hybrid algorithm with an outstanding performance because of the complementation of PSO and CS. In this study, HPSO-CS is proposed to improve the performance of scheduling in cloud. Section 2 explains the literatures that are related to this study, section 3 explains the techniques and algorithms, section 4 discusses about the obtained results and section 5 explains the conclusion of this study.

## LITERATURE SURVEY

Masdari et al [5] presented a comprehensive survey and analysis of scheduling schemes in the cloud computing and provides a classification of the proposed schemes based on the type of scheduling algorithm applied in each scheme. Beside, each scheme was illustrated and a complete comparison of them was given for illustrating their aims, properties as well as their limitations. In the end, conclusion as well as indicators for future research is provided.

Tsai et al [6] presented a novel heuristic scheduling algorithm, called Hyper-Heuristic Scheduling Algorithm (HHSA), to find better scheduling solutions for cloud computing systems. The diversity detection and improvement detection operators were employed by the proposed algorithm to dynamically determine which low-level heuristic has to be used to find better candidate solutions. To evaluate the performance of the proposed method, this study compares the proposed method with several state-of-the-art scheduling algorithms, by having all of them implemented on CloudSim (a simulator) and Hadoop (a real system). The results show that HHSA can significantly reduce the makespan of task scheduling compared with the other scheduling algorithms, on both CloudSim and Hadoop.

Guo et al [7] formulated a model for task scheduling to minimize the cost and proposed a PSO model that has its basis in small position value rules. Through comparison of PSO with PSO embedded in crossovers as well as mutations and in local research, experimental evaluation reveals that PSO not only converges more rapidly but



also runs quicker than the other two on a large scale. Experimental results proved that PSO was more suitable to cloud computing.

The scheduling of dependent tasks is a NP-complete problem and has become as one of the most challenging problems in cloud environment. There is a need of specifying a sequence of execution of these tasks to satisfy the user requirements in terms of QoS parameters such as cost, execution time, etc. The workflow scheduling is considered to be difficult, when it becomes a multi-objective optimization problem. Dutta&Aggarwal [8] presented a comprehensive description of the existing approaches based on meta-heuristics for workflow scheduling. On the basis of the related works, it was found that the Genetic algorithm as the best method for scheduling. A GA searches the problem space globally and therefore, scholars have investigated combining GAs with other meta-heuristic methods to resolve the local search problem. There is a scope of using hybrid meta-heuristics approach that combines Artificial Bee Colony algorithm and Genetic Algorithm (ABC-GA) for scheduling workflows in Cloud computing. Cross-over and mutation operators of GA can be embedded into ABC to improve scheduling strategy.

Ramezani et al [9] developed a comprehensive multi-objective [9] model for optimizing task scheduling to minimize task execution time, task transferring time, and task execution cost. However, the objective functions in this model are in conflict with one another. Considering this fact and the supremacy of PSO algorithm in speed and accuracy, we design a multi-objective algorithm based on Multi-objective PSO (MOPSO) method to provide an optimal solution for the proposed model. To implement and evaluate the proposed model, Jswarm package was extended to Multi-objective Jswarm (MO-Jswarm) package. Also extend the Cloudsim toolkit applying MO-Jswarm as its task scheduling algorithm. MO-Jswarm in Cloudsim determines the optimal task arrangement among VMs according to MOPSO algorithm. The simulation results reveal that the suggested technique has the ability to find optimal trade-off solutions for multi-objective task scheduling problems that represent the best possible compromises among the conflicting objectives, and significantly increases the QoS.

Durgadevi&Srinivasan [10] focused on Meta-heuristic Swarm Optimization Algorithms (MSOA) which handles issue of VM placements as well as task scheduling in cloud environments. MSOA are simple, parallel algorithms which may be employed in several ways for resolving task scheduling issues. The suggested model is regarded as a combination of SO as well as CS algorithms called MSOACS. The proposed algorithm was tested with CloudSim simulator. The outcomes confirmed the decrease in makespan as well increase in utilization ratio of the suggested MSOAC as opposed to SOA or RA algorithms.

Xue& Wu [11] presented a QoS-based Genetic Hybrid Particle Swarm Optimization (GHPSO) for scheduling applications to cloud resources. Crossovers as well as mutations are embedded into PSO, in GHPSO, so that it can play a role in the discrete problem, in addition, variability index, changing with the number of iterations, was suggested for ensuring that populations can have better global search capacity in earlier stages of evolution, with no premature phenomena. Hill climbing algorithms are also introduced into PSO for improving local search capacity as well as for maintaining diversity of population. Simulation outcomes reveal that GHPSO attains greater performance than standard PSO used in minimizing costs in a specified execution time.

Scheduling refers to tasks that are carried out for obtaining maximum profit for increasing cloud computing workload efficacy. The objective is to utilize resources in an adequate manner as well as the management of loads between resources with minimal execution time. Extreme transmissions costs are incurred in clouds prevent task schedulers from being employed in huge scale distributed environments. Sridhar &Babu [12] proposed a hybrid Particle Swarm Optimization (PSO) which performs better in execution ratio and average schedule length.

Bittencourt et al [13] introduced the scheduling problem in hybrid clouds presenting the main characteristics to be considered when scheduling workflows, as well as a brief survey of some of the scheduling algorithms used in these systems. To assess the influence of communication channels on job allocation, proposed method compared and evaluated the impact of the available bandwidth on the performance of some of the scheduling algorithms.

Babukarthik et al [14] presented a Hybrid algorithm, on the basis of ACO as well as Cuckoo Search that effectively resolves the task scheduling issue that decreases total implementation time Within ACO, pheromones are chemical substances which are deposited by ants when they walk. For resolving optimization issues, it behaves as if it lures artificial ants. For performing local searches, proposed method use Cuckoo Search where there is essentially only a single parameter apart from the population size and it is also very easy to implement.

The issue of scheduling in a cloud is an NP-hard optimization problem. Maintaining a load balance between processing units in the system is of great significance in cloud technology. When a set of tasks arrive at the cloud, the system is supposed to respond to all of them as it manages to achieve the shortest possible time. Branch [15] used Cuckoo algorithm to perform such a management. The purpose of the proposed method achieved an order of processing units such that the time of responses to queries was minimized. The input to cuckoo algorithm is the number of virtual machines and the number of tasks. By examining various orders of these machines, the proposed method allocates hosts to tasks in a proper way. Simulation results show that using Cuckoo algorithm for the intention of reaching the best order of processing units leads to improve performance parameters. In addition, simulations reveal that if tasks are scheduled without any primary information about the resources, the results will not be satisfying enough.

## METHODS

In this work, scheduling is performed based on hybrid PSO with CS algorithm.

### PARTICLE SWARM OPTIMIZATION (PSO)

Particle swarm optimization (PSO) is a non-traditional, modern optimization method. It is population based, it is inspired by the natural behavior of animals or insects e.g., bird flocking, fish schooling. PSO was formulated for resolving non-linear optimization issues, but recently the algorithm has been utilized in several domains, including real world application issues. It is a significant tool of swarm intelligence and it owes its inspiration to the natural activity of birds as well as fish and their movements. Regard this scenario; if flocks of birds are looking for one from one location to another, there is no leader present for that flock. All birds follow that one bird that is nearest to food source and they transmit the data to one another. The flock attains its best position toward the food source through transmissions with the member nearest to it. The procedure is iterated till food source is identified. To find optimal solutions, PSO algorithm follows the exact same procedure [16].

Each individual in the swarm is represented as a particle in a D –dimensional space. Each particle is represented by its position ( $X_i$ ) and velocity ( $V_i$ ). The particle's personal best location is specified by  $P_i = \{p_{i1}, p_{i2}, \dots, p_{in}\}$  and the Global best of all particles is given by  $G = \{g_1, g_2, \dots, g_n\}$ . The algorithm begins with a set of particles whose location as well as velocity is set arbitrarily. All particles' fitness values are computed. Fitness values of all particles are recorded as personal best (pbest) values. Best fitness values as got so far by all particles is considered as the global best (g best). Once these two values are obtained, then each particle will updates its position and velocity using the two equations (1):

$$\begin{aligned} V_{id}(t) &= wV_{id}(t) + c_1 r_1 [X_{id}(t) - p_{id}(t)] + c_2 r_2 [X_{id}(t) - g_d(t)] \\ X_{id}(t) &= X_{id}(t) + V_{id}(t) \end{aligned} \quad (1)$$

Where  $c_1$  and  $c_2$  are the Cognitive factors  $r_1$  and  $r_2$  are the values randomly chosen between 0 and 1 and w is the cognitive weight factor.

### BINARY PARTICLE SWARM OPTIMIZATION (BPSO)

Binary Particle Swarm Optimization (BPSO) is possible with a modification of the PSO version. In a binary version, particle's personal best and global best are updated as in a typical version [3]. The difference between BPSO and PSO is that relevant variables (particles velocities and positions) are defined regarding the change of probabilities. Particles are formed by integers in

{0, 1}. A logistic Sigmoid transformation function  $s(v_{ij}^k)$  limits velocity in the interval [0, 1].

$$s(v_{ij}^k) = 1 / (1 + e^{-v_{ij}^k}) \quad (2)$$

Thus, real velocity is digitized (1/0) by logistic functions for binary space. The BPSO's update equation is done in 2 steps. First, the equation is used to update particle velocity. Second, the particle's new position is obtained using equation (3) [30]

$$X_{ij}^k = \begin{cases} 1: & \text{if } rand() \leq s(v_{ij}^k) \\ 0: & \text{otherwise} \end{cases} \quad (3)$$

Where,  $v_{ij}^k$  is a velocity of  $j$ th dimension in  $i$ th particle,  $X_{ij}^k$  is current position of  $j$ th dimension in  $i$ th particle at iteration  $k$ .  $\text{rand}()$  is a uniform random number in a range  $[0, 1]$ . The scheduling algorithm is given as:

1. Set particle dimension as equal to the size of ready tasks in  $\{t_i\} \in T$
2. Initialize particles position randomly from  $PC = 1, \dots, j$  and velocity  $v_i$  randomly.
3. For each particle, calculate its fitness value
4. If the fitness value is better than the previous best  $pbest$ , set the current fitness value as the new  $pbest$ .
5. After Steps 3 and 4 for all particles, select the best particle as  $gbest$ .
6. For all particles, calculate velocity and update their positions
7. If the stopping criteria or maximum iteration is not satisfied, repeat from Step 3.

### CUCKOO SEARCH (CS) ALGORITHM

The CS is an optimization algorithm proposed by Yang and Deb. Cuckoo search [17] is a search algorithm based on the natural behavior of brood parasitism of cuckoo birds. Cuckoo birds display aggressive breeding behavior. They do not build their own nests, rather they lay their eggs in the nests of other hosts. Host birds are typically not aware of cuckoo eggs because the color as well as pattern is imitated. Cuckoo eggs are hatched at an earlier stage than host eggs. When hosts discover that the eggs do not belong to it, they throw away the foreign eggs or destroy them and rebuild their nests elsewhere. This behavior of cuckoos helps in several optimization issues. CS has its basis in three assumptions: Every cuckoo lays only one egg at a time and the eggs are arbitrarily placed into any nest, the nest containing eggs of better quality will be passed to the subsequent iteration and the quality of host nest is static and cannot be changed.

### L'EVY FLIGHTS

As it is well-known, random searching [18] is crucial importance in meta-heuristic algorithms. The L'evy flight is a random process which consists of taking a series of consecutive random steps [36]. From the mathematical point of view, two consecutive steps need to be performed to generate random numbers with L'evy flights: [Figure-1]

- The generation of steps and
- The choice of a random direction. To do this, one of the most efficient methods is to use the so-called Mantegna algorithm where the step length  $L$  can be determined as in equation (4):

$$L = \frac{u}{|v|^{1/\beta}} \quad (4)$$

Where  $\beta$  the scale parameter and its recommended range is  $[1, 2]$ .

```

Step (0): Initialization
Objective function  $f(x)$ ,  $x = \{x_1, x_2, \dots, x_d\}$ 
Generate an initial population of  $n$  host nests  $x_i$ ,  $i=1, 2, \dots, n$ 
Step (1): Updation loop
While(Termination Criterion)
Choose a cuckoo bird ( $i$ ) arbitrarily using levy Flights
Find the Fitness function  $F_i$  Choose a nest ( $j$ ) arbitrarily among  $n$ 
If( $F_i > F_j$ ) Replace  $j$  by  $i$ 
End
A Probability ( $P_a$ ) of worst nest is removed.
Build the new nest
Record the best solutions
Sort these solutions and find current best
End While
Pass the Best solution to next iteration
End
  
```

Fig: 1. Pseudo code for Cuckoo Search (CS) Algorithm

### HYBRID PSO WITH CS

To improve the performance of CS, PSO [19] is introduced in the update process of CS. Thus, a PSOCS hybrid algorithm is developed. PSOCS first uses Lévy flights in the search space to search, and then it uses the position of the PSO update mode to accelerate the particles to the optimal solution convergence. At the same time, the random elimination mechanism of CS can successfully escape local optima, thereby improving the performance of searching for the optimal solution. [Figure-2] shows the flowchart for hybrid PSO-CS.

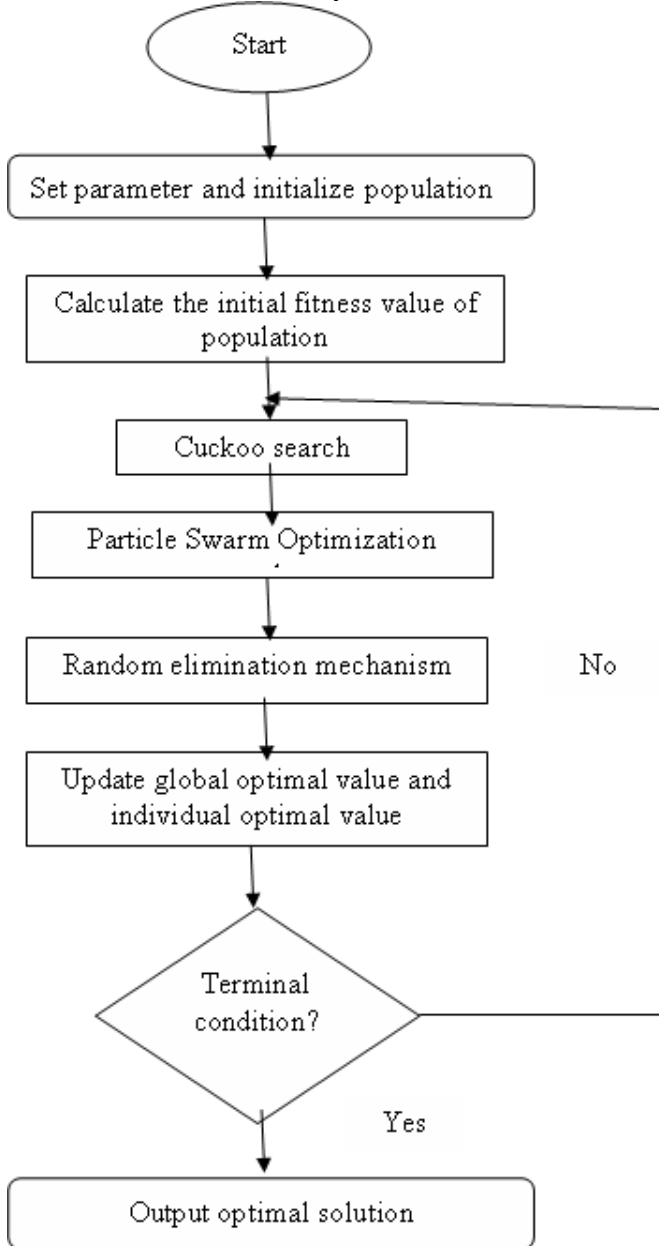


Fig: 2. Flowchart for Hybrid PSO-CS algorithm

## RESULTS AND DISCUSSION

For experiments, the numbers of tasks used are 100, 300, 500, 700 and 900. PSO and HPSO with CS is applied for scheduling. [Table -1], [Table -2] and [Figure-3], [Figure-4] shows the result table and graph for average schedule length and ratio of successful execution respectively.

Table: 1. Average Schedule Length

Number of tasks	PSO	HPSOCS
100	324	321
300	1014	989
500	1712	1665
700	2354	2339
900	3044	2953

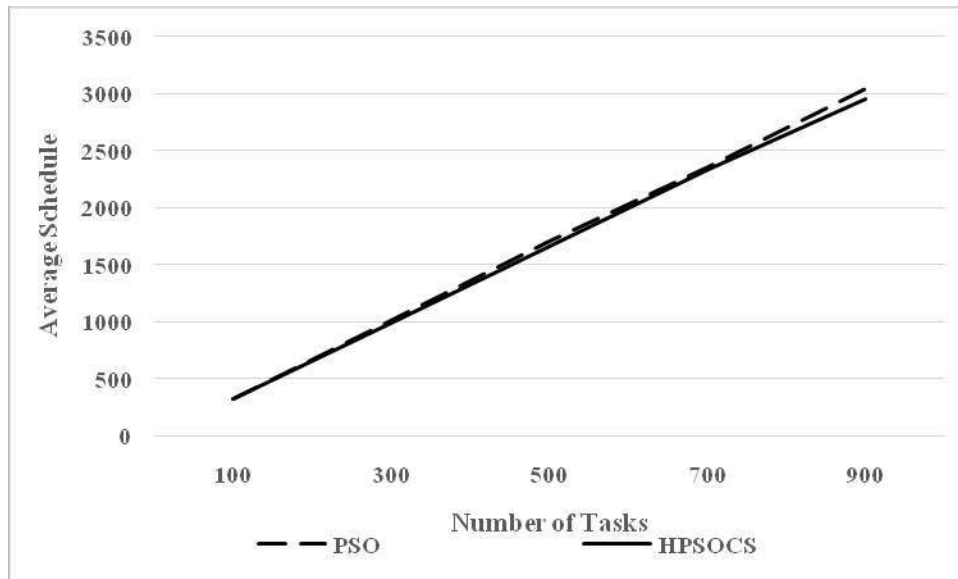


Fig: 3. Average Schedule Length

It is observed from [Table -1] and [Figure -3] that the average schedule length of proposed HPSOCS performs better by 0.93% than PSO at number of task is 100 and by 3.03% than PSO at number of task is 900. The average schedule length is increased when the number of tasks increases.

Table: 2. Ratio of Successful Execution

Number of tasks	PSO	HPSOCS
100	0.85	0.88
300	0.83	0.86
500	0.82	0.84
700	0.81	0.81
900	0.8	0.81

It is observed from [Table -2] and [Figure -4] that the ratio of successful execution of proposed HPSOCS performs better by 3.5% than PSO at number of task is 100 and by 1.24% than PSO at number of task is 900. The ratio of successful execution gets decreased when the number of tasks increases.

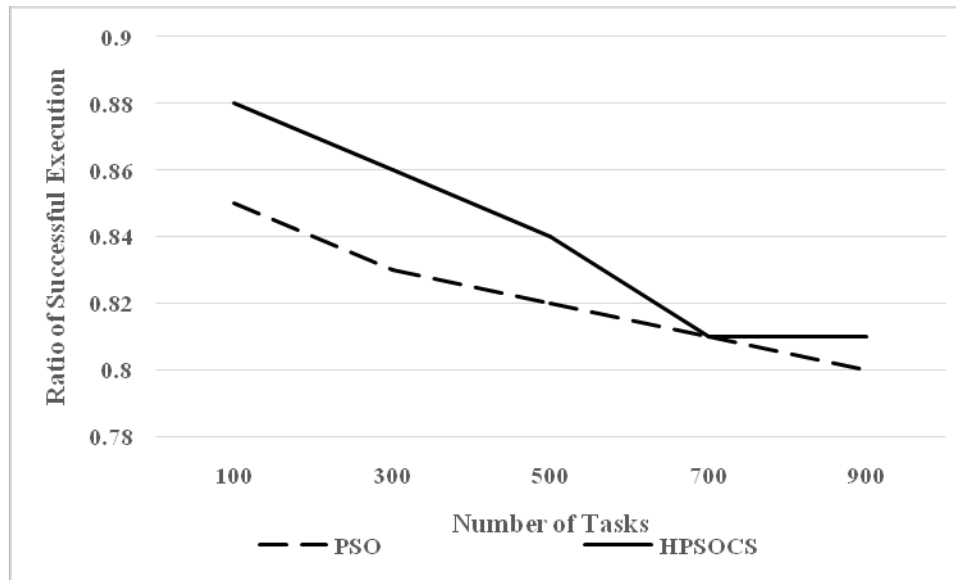


Fig. 4. Ratio of Successful Execution

## CONCLUSION

Cloud computing has become an important platform for companies to build their infrastructures upon. Scheduling is one of the most important task in cloud computing environment. Results shows that the average schedule length of proposed HPSOCS performs better by 0.93% than PSO at number of task is 100 and by 3.03% than PSO at number of task is 900. The average schedule length is increased when the number of tasks increases. Also the ratio of successful execution of proposed HPSOCS performs better by 3.5% than PSO at number of task is 100 and by 1.24% than PSO at number of task is 900. The ratio of successful execution gets decreased when the number of tasks increases.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Mell P, Grance T.[2011]The NIST definition of cloud computing.
- [2] Kumar P, Gopal K, Gupta JP. [2015] Scheduling algorithms for cloud: A survey and analysis. *Journal of Information Sciences and Computing Technologies*, 3(1):162-169.
- [3] Salot P. [2013] A survey of various scheduling algorithm in cloud computing environment. *International Journal of research and engineering Technology (IJRET)*, ISSN, 2319-1163.
- [4] Chawla Y, Bhonsle M. [2012] A study on scheduling methods in cloud computing. *International Journal of Emerging Trends & Technology in Computer Science (IJETCS)*, 1(3): 12-17.
- [5] Masdari M, ValiKardan S, Shahi Z, Azar SI. [2016] Towards workflow scheduling in cloud computing: A comprehensive analysis. *Journal of Network and Computer Applications*.
- [6] Tsai CW, Huang WC, Chiang MH, Chiang MC, Yang CS. [2014] A hyper-heuristic scheduling algorithm for cloud. *Cloud Computing, IEEE Transactions on*, 2(2):236-250.
- [7] Guo L, Zhao S, Shen S,,Jiang C. [2012] Task scheduling optimization in cloud computing based on heuristic algorithm. *Journal of Networks*, 7(3): 547-553.
- [8] Dutta M, Aggarwal N. [2016]Meta-Heuristics Based Approach for Workflow Scheduling in Cloud Computing: A Survey. In *Artificial Intelligence and Evolutionary Computations in Engineering Systems* (pp. 1331-1345). Springer India.

- [9] Ramezani F, Lu J, Hussain F. [2013] Task scheduling optimization in cloud computing applying multi-objective particle swarm optimization. In *Service-Oriented Computing* (pp. 237-251). Springer Berlin Heidelberg.
- [10] Durgadevi P, Srinivasan S. [2015] Task Scheduling using Amalgamation of Metaheuristics Swarm Optimization Algorithm and Cuckoo Search in Cloud Computing Environment. *Journal for Research*, 1(09).
- [11] Xue SJ, Wu W. [2012] Scheduling workflow in cloud computing based on hybrid particle swarm algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 10(7):1560-1566.
- [12] Sridhar M, Babu G. [2015] Hybrid Particle Swarm Optimization scheduling for cloud computing. In *Advance Computing Conference (IACC), 2015 IEEE International* (pp. 1196-1200). *IEEE*.
- [13] Bittencourt LF, Madeira ER, Da Fonseca NL. [2012] Scheduling in hybrid clouds. *Communications Magazine, IEEE*, 50(9): 42-47.
- [14] Babukarthik RG, Raju R, Dhavachelvan P. [2013] Hybrid algorithm for job scheduling: Combining the benefits of ACO and Cuckoo search. In *Advances in Computing and Information Technology* (pp. 479-490). Springer Berlin Heidelberg.
- [15] Branch K. [2015] A Novel Task Scheduling Method in Cloud Environment using Cuckoo Optimization Algorithm.
- [16] Sangale SY. [2016] SCHEDULING BASED ON PARTICLE SWARM OPTIMIZATION ALGORITHM: REVIEW PAPER. space, 29, 30th.
- [17] Banu AS, Helen W R. [2015] Scheduling Deadline Constrained Task in Hybrid IaaS Cloud using Cuckoo Driven Particle Swarm Optimization. *Indian Journal of Science and Technology*, 8(16).
- [18] Baskan O. [2013] Determining optimal link capacity expansions in road networks using cuckoo search algorithm with lévy flights. *Journal of Applied Mathematics*.
- [19] Guo J, Sun Z, Tang H, Jia X, Wang S, Yan X, Wu G. [2016] Hybrid Optimization Algorithm of Particle Swarm Optimization and Cuckoo Search for Preventive Maintenance Period Optimization. *Discrete Dynamics in Nature and Society*, 2016.

\*\*DISCLAIMER: This published version is uncorrected proof; plagiarisms and references are not checked by IIOABJ; the article is published as it is provided by author and approved by guest editor.

# MULTI-LEVEL CO-VARIANCE MEASURE BASED VIDEO COMPRESSION TECHNIQUE FOR EFFICIENT VIDEO TRANSMISSION

Tamil Selvi<sup>1</sup> and Rajiv Kannan<sup>2</sup>

<sup>1</sup>Research scholar, Anna University, Chennai, TN, INDIA

<sup>2</sup>Dept. of Computer Science and Engineering, K.S. Rangasamy College of Engineering, Thiruchengode, TN, INDIA

## ABSTRACT

The high volume data transmission has various issues like bandwidth occupation, latency and much more. To overcome the difficulty in data transmission, we discuss a frame level Co-Variance Measure based video compression technique. The method first splits the video into some frames and for each successive frame, the method computes the co-variance matrix. The method generates three different co-variance matrix, where each matrix maintains the variance of RGB layers of the frames. The method extracts the features of RGB image and at each layer the method computes the variance of feature values. Based on extracted features, the method computes the co-variance measure between all the three layer feature matrix. Based on the measure computed, the method identified variance of feature and based on identified feature variance the method sends the core frame or subsequent frame.

Published on: 14<sup>th</sup>– August-2016

### KEY WORDS

Data Transmission, Video Compression, Multi Level Covariance Matrix, Bandwidth Utilization

\*Corresponding author: Email: [tamilselvi.phd@rediffmail.com](mailto:tamilselvi.phd@rediffmail.com)

## INTRODUCTION

The modern information technology uses multimedia content for the various purpose of authentication in security systems. Not only that the video transmission being used in other applications like video conferencing where the user can perform his activities through video conferencing. In video conferencing the events happening at the remote site will be transferred to the web technology and the end user can vide the events from his own place. The video conferencing has many emerging applications in educational institutions. In educational society, the learner need not go to the place of a tutor but still he can learn from the tutor. The lecture given by the tutor is transferred through the network technology and provided to the learner.

In order to provide the e-learning to the learner, the huge video must be transferred to the user location. The video content occupies more amount of bandwidth and introduces more overhead in data transmission. While performing data transmission, the data occupies more bandwidth and introduces more overhead in different intermediate nodes. In order to overcome the problem of overhead in data transmission, the video content has to be compressed and transferred. There are many data compression techniques available, but suffers from the problem of accuracy. A video can be referred as the collection of frames and snapshot. Each snapshot has a number of frames and each frame occupies a certain amount of memory. If you consider the subsequent frames of any video, you can find only a minimum difference in the video content or visual features. By handling the visual feature variation a high-quality video compression is achieved.

Because of the video is transferred through the network and the network has fixed bandwidth. To improve the throughput of the network, the bandwidth must be utilized in an efficient manner. By utilizing the network bandwidth condition in data transmission, the efficiency of the video transmission can be achieved. Also, the video frame can be identified as an image. An image can be viewed as a container with multiple layers. The image has three different layers of red, green and blue. When you consider the subsequent images and verify the layer



content, there will be little change in any one of the layer feature. The multi layer covariance matrix can be used to maintain the difference in the layer features of the image. The multi layer covariance matrix used to store the difference in color features and by computing the difference between them, the feature difference between two images can be obtained.

## METHODS

There are some methods has been discussed for the video compression and this section discusses some of the methods.

Low-complexity depth map compression in HEVC-based 3D video coding [1], discusses a low-complexity procedure is proposed to reduce the complication of depth map density in the high-efficiency video coding (HEVC)-based 3D video coding (3D-HEVC). Since the complexity map and the corresponding texture video represent the same scene in a 3D video, there is a high correlation between the coding information from depth map and texture video. The experimental examination is done to study depth map and video texture correlation in the coding information such as the motion vector and prediction mode. Based on the correlation, they suggest three efficient low-complexity methods, including early finish mode decision, adaptive search variety motion estimation (ME), and fast disparity estimation (DE). Experimental results show that the future algorithm can decrease about 66% computational difficulty with a negligible rate-distortion (RD) performance loss in comparison with the original 3D-HEVC encoder.

Video Compression Algorithm Using Motion Compensation Technique [2], video compression using motion compensation technique that reduces video data based on motion estimation from one frame to another. Motion recompense is an algorithmic method employed for the brainwashing of video data for video density. Motion compensation describes a frame regarding the transformation of a reference frame on the current surround. The orientation edge may be preceding in time or smooth from the future. The proposed method reduces the candidate of the prediction modes based on the Sum of Absolute Hadamard-Transformed Difference (SATD) between the original block and the inter-predicted block. The motion of each block is obtained based on the SATD value. The current frames are further reduced by using the combination of motion and most probable displacement. The proposed method reduces the number of motion in frames to either one or two. When descriptions container is correctly synthesized from previously stored images, the compression efficiency can be improved. Temporal redundancy is exploited so that not every frame of the video needs to be coded independently as a new image.

Parametric Video Density Scheme Using AR Based Surface Synthesis [10,11], a video coding arrangement based on the parametric density of texture is future. Each macro block is branded either as an edge block, or as a non-edge block containing texture. The non-edge blocks are coded by modeling them as an auto-regressive process (AR). By applying the AR model in the spatiotemporal domain, we safeguard both spatial as healthy as chronological consistency. Edge blocks are programmed using the standard H.264/AVC. The proposed algorithm achieves up to 54.52% more compression as compared to the standard H.264/AVC at the similar visual quality.

Optimizing Motion Compensated Prediction For Error Resilient Video Coding [3], worried with optimization of the motion salaried prediction framework to recover the error resilience of video coding for broadcast over lossy networks. First, precise end-to-end misrepresentation estimation is working to optimize both motion approximation and prediction inside an overall rate-distortion outline. Low complexity practical variations are proposed: a technique to approximate the top motion via simple falsehood and source coding rate representations, and a source-channel forecast method that uses the predictable decoder orientation frame for prediction. Second, orientation frame generation is reentered as a problem of strainer design to optimize the error pliability versus coding competence tradeoff. The singular cases of leaky prediction and biased prediction (i.e., finite impulse response filtering), are examined. A novel reference surround generation method, called widespread source-channel forecast, is proposed, which involves immeasurable impulse reply filtering.

A Better Low Multifaceted Spatially Scalable Acc-Dct Based Video Density Method [4], suggest a low multifaceted Scalable ACC-DCT based video density approach which tends to adventure hard the pertinent chronological joblessness in the video edges to improve density efficiency with less dispensation complexity. The video signal has high temporal redundancies due to the high correlation between successive frames. This redundancy has not been exposed enough to current video compression techniques. Our model consists of 3D to 2D transformation of the video frames that allows exploring the temporal redundancy of the video using 2D transforms and avoiding the computationally demanding motion compensation step. This change turns the spatial-temporal association of the video into high latitudinal correlation. Indeed, this method converts each collection of pictures (GOP) to one image (Accordion Representation) ultimately with in height spatial relationship. This perfect is also combined with up/down sampling method (SVC) which is based on a combination of the forward and retrograde type discrete cosine transform (DCT) coefficients. As this grain has various regularities for efficient calculation, a debauched algorithm of DCT-based Scalability notion is also proposed.

Three-Dimensional Penetration Map Motion Approximation and Compensation for 3D Video Compression [5], propose a new method to 3D depth plan motion guesstimate and compensation for 3D video density. 3D video provides representative vision also will be a feature of forthcoming video displays. The numerous kinds of 3D formats include multiview 3D, a single interpretation of a depth map, time separation multiple 3D, and so on. 3D video requires huge amounts of data and needs a countless deal of storing space to store 3D material. Also, when 3D video is conveyed, the huge quantity of data should be compressed to decrease bandwidth usage. To resolve this problematic, we assume single view with a profundity map 3D format and enterprise a 3D depth map density arrangement for 3D video. We reflect depth map motion estimate and compensation to realize temporal compression of 3D video.

Video Compression by memetic algorithm [6], the position equation of Standard Particle Swarm Optimization is modified and used as step size comparison to find the best matching block in the present frame. To attain adaptive step size, time variable apathy weight is used in its place of constant inertia heaviness for getting true gesture vector dynamically. The period varying inertia weight is based up on preceding motion vectors. The step size reckoning is used to predict best corresponding macro block in the orientation frame on a macro chunk in the current frame for which motion vector is originated. The result of proposed technique is compared with existing block matching algorithms.

A New Video Density Method using DCT/DWT and SPIHT based on Accordion Representation [7], current a new video density method which tends to hard achievement the relevant progressive joblessness in the video to improve hardness efficiency with minimum dispensation complexity. It includes 3D (Three Dimension) to the 2D (Three Dimension) alteration of the video that allows traveling the temporal joblessness of the video using 2D transforms and evading the computationally difficult motion payment step. This alteration converts the three-dimensional and temporal correlation of the video signal into a high three-dimensional correlation. Indeed, this method transforms each assembly of movies into one picture finally with the high spatial association. SPIHT (Set Partitioning in Hierarchical Trees) exploits the possessions of the wavelet-transformed imageries to upsurge its efficiency. Thus, the De-correlation of the subsequent pictures by the DWT (Discrete Wavelet Transform) makes well-organized energy compaction and then produces a high video density ratio. Many untried tests had been conducted to prove the technique efficiency especially in high bit rate and with slow motion video.

Detection of Double Compression in MPEG-4 Videos Based on Markov Statistics [8], Markov founded features are accepted to detect double density artifacts, which suggest that the original video might have been interposed. The advantages and boundaries of double MPEG-4 compression finding are analyzed. Experimental consequences have demonstrated that our scheme outperforms most current methods.

All the methods discussed above has the problem of producing tampered video quality and produces poor video quality. Also, they could not achieve higher data compression ratio

## RESULTS

### Multilevel covariance video compression

The multi level covariance approach converts the video into some frames and for each frame, the method improves the quality by applying histogram equalization. The quality improved frames are converted into the matrix, and the method computes the deviation in feature to store them in the covariance matrix. Also, the method computes the texture variance to store them in the texture variance matrix. Based on both covariance matrix, the method decides whether the entire image has to be transferred or the covariance matrix has to be sent. The entire process can be split into three different stages namely preprocessing, covariance matrix generation and video compression.

The Figure 1, shows the architecture of multi level covariance measure based video compression. Also, the figure 1, depicts the stages of the proposed approach in detail.

### Preprocessing

At this stage, the input video is taken into processing and splits the entire video into some sub-sampling images. The generated image is applied with histogram equalization, which improves the quality image and removes the noise from the image. The generated image will be used to perform feature extraction in the next stage.

Pseudo Code of Preprocessing:

Input: Video v.

Output: Frame set Fs.

Start

Read Input Video V.

Split Video into Frames Fs.

$Fs = \int_{i=1}^{Video\ Length} (\sum Frames \in Fs) \cup Frames(i)$

For each video Fi from Fs

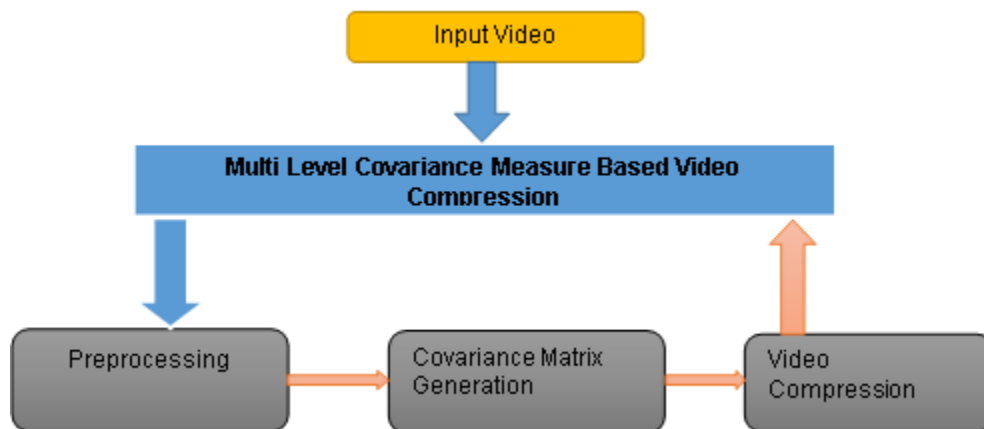
Perform Histogram Equalization.

$Fs = \int_{i=1}^{size(Fs)} HistogramEqualize(Fs(i))$

End

Stop.

The preprocessing algorithm discussed above converts the video into frame set and for each frame available, the method improves the quality of the image by applying histogram equalization.



**Fig1: Architecture of multi level covariance video compression**

### Co-variance Matrix Generation

For any two successive frames being generated, the method extracts the RGB features into three different matrices. From the matrix being generated, the method computes the matching between the pixels of two different images color values. The variance of the value is stored in the matrix. Finally the covariance of the color values is computed. Similarly the method generates the gray scale images and computes the texture variance between the frames.

Pseudo Code of Covariance Matrix Generation:

Input: Frame set Fs

Start

Initialize Covariance Matrix Rcm, Tcm

For each frame Fi from Frame set Fs

Read Previous Frame Fi-1

Generate Rgb matrix Rm = RGB(Fi).

Generate Rgb Matrix Rm1 = RGB(Fi-1).

Compute Rgb Covariance.

$$R_{cm} = \int_{i=1}^{\text{size}(R_{m1})} \text{Dist}(R_m(i), R_{m1}(i))$$

Convert Fi into Gray Scale.

Convert Fi-1 into Gray scale.

Compute Texture Covariance.

$$T_{cm} = \int_{i=1}^{\text{size}(F_i)} \text{Dist}(F_i(i), F_{i-1}(i))$$

End

Stop.

The covariance matrix generation algorithm computes the multi-layer rgb covariance and texture covariance between two different frames considered. The computed covariance value is stored in the concern matrix which will be used to perform video compression.

### Video Compression

The video compression approach performs a comparison of texture covariance and color covariance measures, if the texture covariance has more difference than the particular threshold then the method transmits the complete frame otherwise the method generates a frame with the new texture and color values which have been varying. This reduces the bandwidth occupancy and produces good compression ratio.

Pseudo Code of Video Compression:

Input: Frame Set  $F_s$ , RGB Covariance matrix  $R_{cm}$ , Texture Covariance Matrix  $T_{cm}$ .

Output; Compressed Video  $C_v$

Start

Compute Texture Covariance Similarity.

$$Tcs = \frac{\sum_{i=1}^{\text{size}(T_{cm})} T_{cm}(i) == T_{cm}(i)}{\text{size}(T_{cm})} \times 100$$

If  $Tcs > STh$  then

Compute RGB covariance Similarity.

$$Rcs = \frac{\sum_{i=1}^{\text{size}(R_{cm})} R_{cm}(i) == R_{cm}(i)}{\text{size}(R_{cm})} \times 100$$

If  $Rcs > RTh$  then

Transmit the covariance matrix.

End

Else

Transmit the frame.

End

Stop.

The video compression technique, computes the texture covariance similarity and RGB covariance similarity values. If the texture similarity is less than the threshold then the method transmits the frame otherwise the method transmits only the covariance matrix. On the other side, the receiver could reframe the frame using the previous frame and covariance matrix.

## DISCUSSION

The proposed multi level covariance measure based video compression has been implemented using Matlab and the performance of the methods has been evaluated using different videos. The methods have produced efficient result in compression ration and reduce the distortion ratio than other methods.

The Figure-2 shows the comparison of video compression ratio being achieved by different methods and it shows clearly that the proposed methods have produced more video compression ratio than other methods.

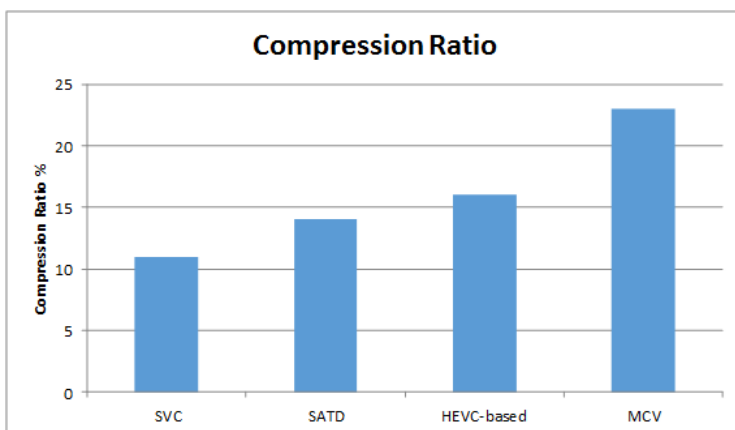
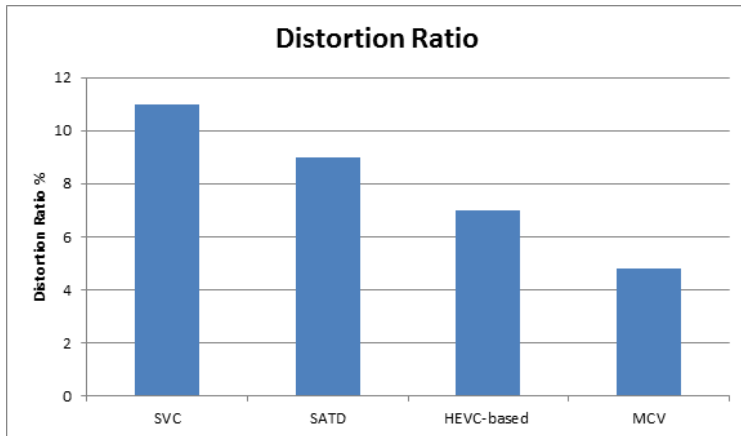


Fig. 2: Comparison of video compression ratio



**Fig. 3: Comparison of distortion ratio**

The Figure-3 shows the comparison of distortion ratio being produced by different methods and it shows clearly that the proposed method has produced less distortion ratio, and the methods reduce the distortion ratio rapidly.

**Table 1: Comparison of various video compression measures**

Method	Compression Ratio %	Distortion Ratio %	Time Complexity in seconds
SVC	11	11	87
SATD	14	9	81
HEVC-based	16	7	76
MCV	23	4.8	56

The Table-1, shows the comparison of different video compression measures produced, and it shows that the proposed method has produced an efficient result.

## CONCLUSION

The author proposed a multi level covariance measure based video compression. The method first splits the video into some frames and for each successive frame, the method computes the co-variance matrix. The method generates three different co-variance matrix, where each matrix maintains the variance of RGB layers of the frames. The method extracts the features of RGB image and at each layer the method computes the variance of feature values. Based on extracted features, the method computes the co-variance measure between all the three layer feature matrixes. Based on the measure computed, the method identified variance of feature and based on identified feature variance the method sends the core frame or subsequent frame.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] Queen Zhang , Ming Chen, Xinpeng Huang, Nana Li, Yong Gan, Low-complexity depth map compression in HEVC-based 3D video coding, EURASIP journal of image and video processing, 2015.
- [2] Video Compression Algorithm Using Motion Compensation Technique, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 3, Issue 6, June 2014.
- [3] Hua Yang, and Kenneth Rose, "Optimizing Motion Compensated Prediction For Error Resilient Video Coding," IEEE Transactions On Image Processing, Vol. 19, No. 1, January 2010 .
- [4] G. Suresh et al., "An Improved Low Complex Spatially Scalable Acc-Dct Based Video Compression Method," (IJCS) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010.
- [5] Yu-Cheng Fan, Shu-Fen Wu, and Bing-Lian Lin, "Three-Dimensional Depth Map Motion Estimation and Compensation for 3D Video Compression," IEEE Transactions on magnetic, VOL. 47, NO. 3, MARCH 2011.
- [6] Pooja Nagpal, Seema Baghla, "Video Compression by Memetic Algorithm," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
- [7] Jaya Krishna Sunkara, E Navaneethasagari,Atal, "A New Video Compression Method using DCT/DWT and SPIHT based on Accordion Representation, International journal of Image, Graphics and Signal Processing, 2012.
- [8] Xinghao Jiang, Wan Wang, Tanfeng Sun,Atal, "Detection of Double Compression in MPEG-4 Videos Based on Markov Statistics", IEEE Signal Processing Letters, VOL. 20, NO. 5, MAY 2013.
- [9] Ms. Bhavina Patel, Dr.R.V.Kshirsagar, "Review And Comparative Study Of Motion Estimation Techniques To Reduce Complexity In Video Compression," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Vol. 2, Issue 8, August 2013.
- [10] Rajan. C, Geetha. K.[2016, ]An Investigation Of Effective Medical Image Compression And Transmission In Wireless Ad Hoc Networks, International Journal on Concurrent Applied Research in Engineering and Management,vol.1, No.4, pp. 173-179
- [11] Khandelia, A. Parametric Video Compression Scheme Using AR Based Texture Synthesis ,Computer Vision, Graphics & Image Processing, 2008.

\*\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and checked/reviewed by guest editor.

## A SURVEY OF TWITTER SENTIMENT ANALYSIS

Anuprathibha. T<sup>1</sup> and C. S. Kanimozhi Selvib<sup>2</sup>

<sup>1</sup> Dept. of Master of Computer Applications, Chettinad College of Engineering and Technology, Karur, TN, INDIA

<sup>2</sup> Dept. of Computer Science and Engineering, Kongu Engineering College, Perundurai, TN, INDIA

### ABSTRACT

Published on: 18<sup>th</sup>– August-2016

#### KEY WORDS

Opinion mining, Twitter, tweets

Twitter gets to be a standout amongst the most mainstream social networking sites, which permits the clients to peruse and post messages (i.e. tweets). Among the immense assortments of subjects, individuals in Twitter tend to express their opinions for the brands, superstars, products and open occasions. Therefore, it draws in much thoughtfulness regarding assessment the swarm's sentiments in Twitter. Sentiment analysis is a sort of natural language processing for following the inclination of people in general around a specific product or subject. Sentiment analysis, which is additionally called, opinion mining, includes in building a framework to gather and examine opinions.

### INTRODUCTION

Opinion analysis is a fascinating exploration theme in both information extraction and knowledge discovery. Found opinions can be utilized as a part of numerous applications. Opinion mining pulls in interest in both the educated community and industry because of its potential relevance. A promising application is an analysis of social networks opinions. Numerous compose their opinions in forums, micro blogging or review websites. This data is helpful for organizations, governments, and people, who track attitudes and sentiments in such sites. Much data containing valuable information is accessible, for programmed analysis. Case in point, a customer who proposes to purchase a product ventures the Web to discover different customers/reviewers opinions about the product. Such reviews influence the customer's choices. Opinion mining is the computational investigation of content expressed opinions, sentiments and emotions. Opinions for any entity/object/product/individual exist on the Web and additionally for features/segments of items like, cell phone batteries, keyboards, touch screen displays, etc.

Twitter is a well-known micro blogging administration where clients make status messages (called "tweets"). These tweets a few times express opinions about different themes. A method was proposed to automatically extract sentiment (positive or negative) from a tweet. This is extremely helpful in light of the fact that it allows feedback to be accumulated without manual mediation. Customers can utilize sentiment analysis to research products or administrations before making a buy. Advertisers can utilize this to research popular opinion of their organization and products or to examine customer fulfillment. Associations can likewise utilize this to accumulate basic feedback about issues in recently discharged products.

Twitter gets to be a standout amongst the most well-known social networking sites, which permits the clients to peruse and post messages (i.e. tweets). Among the immense assortments of themes, individuals in Twitter tend to express their opinions for the brands, VIPs, products and open occasions. Thus, it draws in much regard for assessment the swarm's sentiments in Twitter.

Twitter messages have numerous special attributes, which separates our examination from previous exploration: Length: The greatest length of a Twitter message is 140 characters. From proposed training set, it is compute that the average length of a tweet is 14 words or 78 characters. This is exceptionally different from the previous

sentiment classification look into that concentrated on characterizing longer assemblages of work, for example, motion picture reviews.

Data availability: another difference is the greatness of data accessible. With the Twitter API, it is anything but difficult to gather a great many tweets for training. In past examination, tests just comprised of thousands of training things.

Language model: Twitter clients post messages from numerous different media, including their cell phones. The recurrence of incorrect spellings and slang in tweets is much higher than in different areas.

Domain: Twitter clients post short messages around an assortment of themes dissimilar to different sites which are custom-made to a specific point. This differs from a huge rate of past exploration, which concentrated on specific spaces, for example, film reviews.

The objective of feature selection is to pick a subset of features to diminish the length of feature vectors with the least information misfortune. Feature selection plans as per their subset evaluation routines, arranged into two groups: Filter and Wrapper. In channel routines, features assessed with their inborn impact on isolating classes while wrapper systems use accuracy of the learning strategies to assess subset of features.

Feature selection offers various favorable circumstances, eliminating so as to include all the more effective classification models unimportant or noisy features, more conservative and speedier models by building them utilizing just a little subset of the first arrangement of features, and the capacity to concentrate on a subset of pertinent features, which can be utilized for the discovery of new knowledge.

## LITERATURE REVIEW

Anjaria and Guddeti [1] presented the novel methodology of abusing the client influence factor with a specific end goal to predict the result of an election result. Exploratory results exhibited that SVMs outperformed every other classifier with most extreme effective prediction accuracy of 88% if there should arise an occurrence of US Presidential Elections held in November 2012 and greatest prediction accuracy of 58% in the event of Karnataka State Assembly Elections held in May 2013. Sindhura and Sandeep [2] exemplified Opinion mining the procedures and drew nearer that guarantee to precisely encourage the opinion-situated information looking for systems that included computational regimen of opinion or subjectivity in the content. Different data-driven strategies for opinion mining as Feature based Opinion Mining Technique, Machine learning based Opinion Mining Technique, and Ranking model with an opinionatedness feature were reviewed and their qualities and shortcoming are touched upon. Claster et al., [3] investigated motion picture sentiment expressed in Twitter microblogs which utilizes, Self-Organizing Maps and film knowledge keeping in mind the end goal to model opinion over a multi-dimensional sentiment space. The outcomes demonstrated the adequacy of the proposed visualization in mining sentiment in the space of Twitter tweets. Weitzel et al., [4] meant to investigate pitched stream of tweets from the Twitter microblogging webpage. In investigation comes about, the creators recognized more neutral passionate states than positive or negative (31%). The creators likewise connected statistical strategies with a specific end goal to derive if there exist correlation between client notoriety and enthusiastic substance. Selvan and Moh [5] concentrated on the computational framework for quick feedback opinion mining which requires a flexible platform to handle all the conceivable issues emerged from mining data streams of a social networking site. The structure made utilization of ongoing Twitter data stream. The system is based upon Apache Hadoop to manage tremendous volume of data streamed from Twitter. The investigations have demonstrated an 84% accuracy in the sentimental analysis and it is ready to give quick, profitable feedbacks to organizations. Aldahawi and Allen [6] broke down data gathered from Twitter and researched the fluctuation that emerges from utilizing an automated sentiment analysis instrument versus human characterization. The proposed results demonstrated that the two techniques yield altogether diverse positive, natural and negative arrangements depending on society and the relationship of the blurb to the two organizations, raising doubt about the reliability of automated sentiment analysis apparatuses for specific classes of clients. Bing and Chan [7] proposed a novel matrix-based fuzzy algorithm, called the FMM framework, to mine the characterized multi-layered Twitter data. Through sets of equivalent investigations connected on Twitter data, the proposed FMM framework accomplished an excellent performance, with both quick processing paces and high predictive accuracy. Spencer and Uchyigit [8] presented Sentimentor, an apparatus for sentiment analysis of Twitter data. Sentimentor used the gullible Bayes Classifier to order Tweets into positive, negative or objective sets. The creators presented exploratory evaluation of our dataset and characterization comes about, the proposed work found are not contradictory with



existing work. Pak and Paroubek [9] concentrated on utilizing Twitter, the most mainstream microblogging platform, for the undertaking of sentiment analysis. The creators demonstrated to consequently gather a corpus for sentiment analysis and opinion mining purposes. Utilizing the corpus, the creators assemble a sentiment classifier that can determine positive, negative and neutral sentiments for a report. Exploratory evaluations demonstrated that the proposed systems are efficient and performs superior to anything previously proposed routines. In the exploration, the creators worked with English, be that as it may, the proposed procedure could be utilized with some other language. Narahari et al., [10] gave a powerful instrument to perform opinion designing so as to mine an end to end pipeline with the help of Apache Flume, Apache HDFS, Apache Oozie and Apache Hive. To make this procedure close continuous the creators contemplated the workaround of overlooking Flume tmp records and expelling default hold up condition from Oozie work design. The proposed work investigated few of the utilization cases that could be produced into real working models. Gokulakrishnan et al., [11] talked about a methodology where an exposed stream of tweets from the Twitter microblogging webpage are pre-processed and grouped in view of their enthusiastic substance as positive, negative and immaterial; and broke down the performance of different arranging algorithms in view of their precision and recall in such cases. Further, the proposed work exemplified the uses of this examination and its constraints. Altrabsheh et al., [12] talked about how feedback could be gathered by means of social media, for example, Twitter and how utilizing sentiment analysis on educational data could enhance instructing. The proposed work likewise presented the proposed framework Sentiment Analysis for Education (SA-E). Mane et al., [13] gave a method for sentiment analysis utilizing hadoop which would prepare the tremendous measure of data on a hadoop cluster quicker progressively. Padmaja and Fatima [14] attempted to concentrate on the fundamental meanings of Opinion Mining, analysis of linguistic resources required for Opinion Mining, few machine learning systems on the premise of their use and significance for the analysis, evaluation of Sentiment orders and its different applications. Saif et al., [15] presented a review of eight freely accessible and physically annotated evaluation datasets for Twitter sentiment analysis. The proposed work likewise gave a relative investigation of the different datasets along a few measurements including: aggregate number of tweets, vocabulary size and sparsity. The creators likewise researched the pair-wise correlation among these measurements and additionally their correlations to the sentiment arrangement performance on diverse datasets. Khan et al., [16] concentrated on these issues and presented an algorithm for twitter bolsters order in view of a hybrid methodology. The proposed strategy included different pre-processing ventures before encouraging the content to the classifier. Test results demonstrated that the proposed method defeated the previous restrictions and accomplished higher accuracy when contrasted with comparative systems. Montejo-Ráez et al., [17] presented a novel way to deal with Sentiment Polarity Classification in Twitter posts, by separating a vector of weighted nodes from the graph of WordNet. These weights are utilized as a part of SentiWordNet to register a last estimation of the polarity. Technique proposed as a non-supervised arrangement is space independent. The evaluation of a produced corpus of tweets demonstrated that this strategy is promising. ElTayeby et al., [18] examined the media's influence on constructing so as to isolate opinions an aspect-based opinion mining structure. The fundamental errand is to distinguish the isolated groups of opinions by understanding the proposed model utilizing Expectation Maximization (EM) algorithm. The creators indicated fascinating perceptions on the sentiment utilized for specific points among the groups of opinions, and closed the rates of media influences among the isolated groups of opinions concerning these themes. Saif et al., [19] presented SentiCircle; a novel vocabulary based methodology that considered the contextual and reasonable semantics of words while figuring their sentiment orientation and quality in Twitter. Results are focused yet uncertain when contrasting with condition of-workmanship SentiStrength, and differ starting with one dataset then onto the next. SentiCircle outperformed SentiStrength in accuracy on average, yet falls insignificantly behind in F-measure. Shrivatava et al., [20] concentrated on tweets that would bring about dissecting the perspective of people in general on for the most part examined points. A tweets puller was created and the grouping depends on features separated and arranged into POSITIVE, NEGATIVE and NEUTRAL. The outcomes further assessed and finished up to gather the performance of the characterization through SVM. Saif et al., [21] proposed a novel approach that naturally catches patterns of words of comparable contextual semantics and sentiment in tweets. The creators utilized 9 Twitter datasets as a part of the evaluation and looked at the performance of the patterns against 6 cutting edge baselines. Results demonstrated that the patterns reliably outperformed every other benchmark on all datasets by 2.19% at the tweet-level and 7.5% at the entity-level in average F-measure. Table-1 demonstrates the written works reviewed.

Table 1: Comparison review of Literatures

S.no	Author	Techniques/Algorithm Used	Merits
1	Anjaria and Guddeti [1]	Support Vector Machines (Svms)	Maximum Successful Prediction Accuracy
2	Sindhura and Sandeep [2]	Opinion Mining	Perform a computational analysis of opinions
3	Claster et al., [3]	Movie Sentiment in Twitter Microblogs	Effective in visualization
4	Weitzel et al., [4]	Publicized Stream of Tweets	detected more neutral emotional states than positive or negative
5	Selvan and Moh [5]	Fast-Feedback Opinion Mining	provide fast, valuable feedbacks
6	Aldahawi and Allen [6]	Data Collected From Twitter	Better result
7	Bing and Chan [7]	Matrix-Based Fuzzy Algorithm	high predictive accuracy
8	Spencer and Uchyigit [8]	Sentimentor	Better result
9	Pak and Paroubek [9]	Microblogging Platform	efficient and better performance
10	Narahari et al., [10]	End To End Pipeline With The Help Of Apache Flume, Apache HDFS, Apache Oozie And Apache Hive	Better performance
11	Gokulakrishnan et al., [11]	Publicised Stream Of Tweets From The Twitter Microblogging Site	Better performance
12	Altrabsheh et al., [12]	Sentiment Analysis	Improve teaching
13	Padmaja and Fatima [14]	Opinion Mining, Analysis Of Linguistic Resources	Improved in usage and importance for the analysis
14	Saif et al., [15]	Twitter Sentiment Analysis	Better performance
15	Khan et al., [16]	Twitter Feeds Classification Based On A Hybrid Approach	achieved higher accuracy
16	Montejo-Ráez et al., [17]	Sentiment Polarity Classification In Twitter Posts	Promising techniques
17	EITayeby et al., [18]	Expectation Maximization (EM) Algorithm	detect the segregated groups of opinions
18	Saif et al., [19]	Senticircle	outperformed SentiStrength in accuracy on average
19	Shrivatava et al., [20]	Tweets Puller	Better performance
20	Saif et al., [21]	Contextual Semantics And Sentiment In Tweets	Out-performed all other baselines on all datasets by 2.19% at the tweet-level and 7.5% at the entity-level in average F-measure.

## CONCLUSION

Opinion mining attracts interest both in academia and industry due to its potential applicability. A promising application is an analysis of social networks opinions. This data is useful for businesses, governments, and individuals, who track attitudes and feelings in such sites. This study made a study on twitter sentiment analysis and correlation is made on the procedures utilized, results acquired. This review is utilized to get an idea on how to continue with this twitter sentiment classification to improve its accuracy, F-measures.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] Anjaria, M., and Guddeti, R. M. R. (2014, January). Influence factor based opinion mining of Twitter data using supervised learning. In *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on* (pp. 1-8). IEEE.
- [2] Sindhura, V., and Sandeep, Y. (2015, March). Medical data Opinion retrieval on Twitter streaming data. In *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on* (pp. 1-6). IEEE.
- [3] Claster, W. B., Hung, D. Q., and Shanmuganathan, S. (2010, July). Unsupervised artificial neural nets for modeling movie sentiment. In *Computational Intelligence, Communication Systems and Networks (CICSyN), 2010 Second International Conference on* (pp. 349-354). IEEE.

- [4] Weitzel, L., Aguiar, R. F., Rodriguez, W. F., and Heringer, M. G. (2014, June). How do medical authorities express their sentiment in Twitter messages?. In Information Systems and Technologies (CISTI), 2014 9th Iberian Conference on (pp. 1-6). IEEE.
- [5] Selvan, L. G. S., and Moh, T. S. (2015, June). A framework for fast-feedback opinion mining on Twitter data streams. In Collaboration Technologies and Systems (CTS), 2015 International Conference on (pp. 314-318). IEEE.
- [6] Aldahawi, H., and Allen, S. M. (2013, September). Twitter mining in the oil business: A sentiment analysis approach. In Cloud and Green Computing (CGC), 2013 Third International Conference on (pp. 581-586). IEEE.
- [7] Bing, L., and Chan, K. C. (2014, December). A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data. In Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing (pp. 652-657). IEEE Computer Society.
- [8] Spencer, J., and Uchyigit, G. (2012). Sentimentor: Sentiment analysis of twitter data. In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (pp. 56-66).
- [9] Pak, A., and Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).
- [10] Narahari, P R., Nitin Srinivas, S. and Prashanth C M. (2015). Real Time Opinion Mining of Twitter Data. International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2923-2927
- [11] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012, December). Opinion mining and sentiment analysis on a twitter data stream. In Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on (pp. 182-188). IEEE.
- [12] Altrabsheh, N., Gaber, M., and Cocea, M. (2013). SA-E: sentiment analysis for education. In 5th KES International Conference on Intelligent Decision Technologies.
- [13] Mane, S. B., Sawant, Y., Kazi, S., and Shinde, V. (2014). Real Time Sentiment Analysis of Twitter Data Using Hadoop. International Journal of Computer Science and Information Technologies, (3098-3100), 5(3).
- [14] Padmaja, S., and Fatima, S. S. (2013). Opinion Mining and Sentiment Analysis—An Assessment of Peoples’ Belief: A Survey. International Journal of Adhoc, Sensor & Uboquitos Computing, 4(1), 21.
- [15] Saif, H., Fernandez, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.
- [16] Khan, F. H., Bashir, S., and Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems, 57, 245-257.
- [17] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña-López, L. A. (2014). Ranked wordnet graph for sentiment polarity classification in twitter. Computer Speech & Language, 28(1), 93-107.
- [18] ElTayeb, O., Molnar, P., and George, R. (2014). Measuring the Influence of Mass Media on Opinion Segregation through Twitter. Procedia Computer Science, 36, 152-159.
- [19] Saif, H., Fernandez, M., He, Y., and Alani, H. (2014). Senticircles for contextual and conceptual semantic sentiment analysis of twitter. In The Semantic Web: Trends and Challenges (pp. 83-98). Springer International Publishing.
- [20] Shrivatava, A., Mayor, S., and Pant, B. (2014). Opinion Mining of Real Time Twitter Tweets. International Journal of Computer Applications, 100(19).
- [21] Saif, H., He, Y., Fernandez, M., and Alani, H. (2014). Semantic patterns for sentiment analysis of Twitter. In The Semantic Web—ISWC 2014 (pp. 324-340). Springer International Publishing.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

## ANALYZING AND REMOVING UNUSED ANDROID INTER-APP PERMISSIONS

M. Gowthami\*, S. Sriraj, G. Jitesh Kumar, G. Vishal

*Department of Computer Science and Engineering, Vel Tech High Tech Dr. Rangarajan Dr.Sakunthala Engineering College, Avadi, Chennai, INDIA*

## ABSTRACT

**Aims:** The aircrafts are playing vital role in Military and Commercial purposes. It is very difficult to communicate with the aircrafts when it is flying. The aircrafts requires more secure communication because it is used in defense. Radar communication has so many drawbacks if signal loss then we cannot track the aircrafts and also it does not provides secure communication. If any Eavesdropper wants to hacks our confidential aircrafts communication easily they can hack. **Materials and methods:** Android's enforcement of the permissions is at the level of individual apps, allowing multiple malicious apps to collude and combine their permissions or to trick vulnerable apps to perform actions on their behalf that are beyond their individual privileges. In this paper, we present COVERT, a tool for compositional analysis of Android inter-app vulnerabilities. COVERT's analysis is modular to enable incremental analysis of applications as they are installed, updated, and removed. It statically analyzes the reverse engineered source code of each individual app, and extracts relevant security specifications in a format suitable for formal verification. **Results:** Given a collection of specifications extracted in this way, a formal analysis engine (e.g., model checker) is then used to verify whether it is safe for a combination of applications—holding certain permissions and potentially interacting with each other—to be installed together. **Conclusion:** Our experience with using COVERT to examine over 200 real-world apps corroborates its ability to find inter-app vulnerabilities in bundles of some of the most popular apps on the market.

Published on: 18<sup>th</sup>– August-2016

## KEY WORDS

Formal verification, static analysis, Android, Inter-App vulnerabilities

\*Corresponding author: Email: [gowthamimailme@gmail.com](mailto:gowthamimailme@gmail.com) Tel.: +91-90474-91577; Fax: +91-44-26840 249

## INTRODUCTION

Mobile app markets are creating a fundamental model shift in the way software is delivered to the end users. The benefits of this software supply model are plenty, including the ability to rapidly and effectively acquire, introduce, maintain, and enhance software used by the consumers. By providing a medium for reaching a large consumer market at a nominal cost, app markets have leveled the software development industry, allowing small entrepreneurs to compete with prominent soft-ware development companies.

Application frameworks are the key enablers of these markets. An application framework, such as the one provided by Android, ensures apps developed by a wide variety of suppliers can interoperate and coexist together in a single system (e.g., a phone) as long as they conform to the rules and constraints imposed by the framework.

This paradigm shift, however, has given rise to a new set of security challenges. In parallel with the emergence of app markets, we are witnessing an increase in the security threats targeted at mobile platforms. This is nowhere more evident than in the Android market (i.e., Google Play), where many cases of apps infected with malwares and spywares have been reported [1]. Numerous culprits are at play here, and some are not even technical, such as the general lack of an overseeing authority in the case of open markets and inconsequential implication for those caught provisioning applications with vulnerabilities or malicious capabilities.

In this context, Android's security has been a thriving subject of research in the past few years. Leveraging program analysis techniques, these research efforts have investigated weaknesses from various perspectives, including detection of information leaks [2–4], analysis of the least-privilege principle [5,6], and enhancements to Android protection mechanisms [7–9]. The majority of these approaches, however, are subject to a common limitation: they are intended to detect and mitigate vulnerabilities in a single app, but fail to identify vulnerabilities that arise due to the interaction of multiple apps. Vulnerabilities due to the interaction of multiple

apps, such as collusion attacks and privilege escalation chaining [5], cannot be detected by techniques that analyze a single app in isolation. Thus, security analysis techniques in such domains need to become compositional in nature.

This paper contributes a novel approach, called COVERT, for compositional analysis of Android inter-app vulnerabilities. Unlike all prior techniques that focus on assessing the security of an individual app in isolation, our approach has the potential to greatly increase the scope of application analysis by inferring the security properties from individual apps and checking them as a whole by means of formal analysis. This, in turn, enables reasoning about the overall security posture of a system (e.g., a phone device) in terms of the security properties inferred from the individual apps.

COVERT combines static analysis with formal methods. At the heart of our approach is a modular static analysis technique for Android apps, designed to enable incremental and automated checking of apps as they are installed, removed, or updated on an Android device.

Through static analysis of each app, our approach extracts essential information and captures them in an analyzable formal specification language. These formal specifications are intentionally at the architectural level to ensure the technique remains scalable, yet represent the true behavior of the implemented software, as they are automatically extracted from the installation artifacts.

The set of models extracted in this way are then checked as a whole for vulnerabilities that occur due to the interaction of apps comprising a system. COVERT uses Alloy as a specification language [10], and the Alloy Analyzer as the analysis engine. Alloy is a formal specification language based on first order logic, optimized for automated analysis.

Since COVERT's analysis is compositional, it provides the analysts with information that is significantly more useful than what is provided by prior techniques. Our experiences with a prototype implementation of the approach and its evaluation against one of the most prominent inter-app vulnerabilities, i.e. privilege escalation, in the context of hundreds of real-world Android apps collected from variety of repositories have been very positive. The results, among other things, corroborate its ability to find vulnerabilities in bundles of some of the most popular apps on the market.

**Contributions:** This paper makes the following contributions:

**Formal model of Android framework:** We develop a formal specification representing the behavior of Android apps that is relevant for the detection of inter-app vulnerabilities. We construct this formal specification as a reusable Alloy module to which all extracted app models conform.

**Modular analysis:** We show how to exploit the power of our formal abstractions by building a modular model extractor that uses static analysis techniques to automatically extract formal specifications (models) of apps from their installation artifacts.

**Implementation:** We develop a prototype implementation on top of our formal framework for compositional security analysis of Android apps.

**Experiments:** We present results from experiments run on over 200 real-world apps, corroborating COVERT's ability in effective compositional analysis of Android inter-app vulnerabilities in the order of minutes.

The study was conducted in the Department of Conservative Dentistry and Endodontics, Sinhgad Dental College and Hospital, Pune with technical aid from the Department of Microbiology, SKN Medical College and Hospital, Pune. All the participants were informed about the study and necessary informed consent was taken. Ethical clearance was obtained from the ethical committee of college. Total 20 aprons of dental healthcare professionals (interns, PG students, faculty members) were included in the study.

## PROPOSED SYSTEM

As android applications are open source and can be developed by anybody, testing is not mandatory and hence it is more vulnerable. Android application developed by users are directly uploaded to Google play store and no code level testing's are done. Since the developers upload only compiled, packed (.apk) files no further investigation is done on the application.

A basic call graph can only give the number of permission checks but not the actual names of the checked permissions because of the lack of string analysis to extract permission names from the byte code CHA-Android which leverages the service redirection, service identity inversion and entry point construction components.

Spark specific issues such as entry point initialization or Android specific issues such as service initialization. Spark to get a first understanding of the main problems that occur when analyzing the Android API. This gives us a key insight, Spark discards 96 percent of the API methods to be analyzed. The reason is that Spark does not work on receiver objects whose value is null.

Android application, (ex: appwrong), which is able to communicate with external servers since it is granted the INTERNET permission. Moreover, appwrong has declared permission CAMERA while it does not use any code related to the camera. The CAMERA permission allows the application to take pictures without user intervention, i.e., the permission gap consists of a single permission: CAMERA. In this particular example the attacker would be able to write code to use the camera, take a picture and send the picture to a remote host on the Internet.

The problematic consequences of having more permissions than necessary and showed that the problem can be mitigated using compositional analysis. The approach has been fully implemented for Android, a permission-based platform for mobile devices. Android application stores indeed suffer from permission gaps.

We propose a High level Permission Checking Framework on Android Applications that were previously uploaded by breaking the .apk files to analyze in code level by decompiling it in a efficient way. We also innovate to recompile the vulnerable free code for secure use with the end users. We further make a proposal to Google Play Services to implement this kind of Frameworks so as to avoid Fake Applications that steals user's Private data and make some vulnerability.

Android 2.2 defines 134 permissions in the android. Manifest permission system class, whereas Android 4.0.1 defines 166 permissions. This gives us an upper-bound on the number of permissions which can be checked in the Android framework. Android has two kinds of permissions: "high-level" and "low-level" permissions. High-level permissions are only checked at the framework level (that is, in the Java code of the Android SDK). We focus on the high-level permissions that are only checked in the Android Java framework Compositional analyses for extracting permission checks. In essence, each analysis constructs a call graph from the byte code, finds permission check methods and extracts permission names.

We have presented a generic approach to reduce the attack surface of permission-based software in order to automatically add or remove permission enforcement points at the level of application or the framework.

## IMPLEMENTATION

Applications for Android are written in Java and compiled into Dalvik byte code.

Dalvik byte code is optimized to run on devices where memory and processing power are scarce. An Android application is packaged into an Android package file which contains the Dalvik byte code, data (pictures, sounds. . .) and a metadata file called the "manifest".

For installing an application, the user has to approve all the permissions the application's developer has declared in the application manifest. If all permissions are approved, the application is installed and receives group memberships. The group memberships are used to check the permissions at runtime.

Missing permission causes the application to crash. Adding too many of them is not secure. In the latter case, injected malware can use those declared, yet unused permissions, to achieve malicious goals. We call those unused permissions, "permission gap". Any permission gap results in insecure, suspicious or unreliable applications.

### Login / Registration and Upload

User enters the personal information for registration and the user input fields are validated and records are stored in Database. After registration the User can Login with his credentials and can upload source code. The uploaded source is securely stored in server side. If you are uploading a source code it should in a zip format which can be done by any zip until tools. The uploaded zip contents are automatically unzipped in code level in server side.

## Reverse Engineering the .apk file

In this Module, user can upload both source and apk files. The apk file is broken by using APK Tool and the generated (.dex) files are converted to (.jar) files by de2jar. The layout and resource files are retained. The jar files are extracted to get the .class files. Now we use the jad API to convert the .class files to .java files. Then these files are written to the src folder of android code base retaining the package name. Thus the Server automatically Decompile the .apk file by reverse engineering.

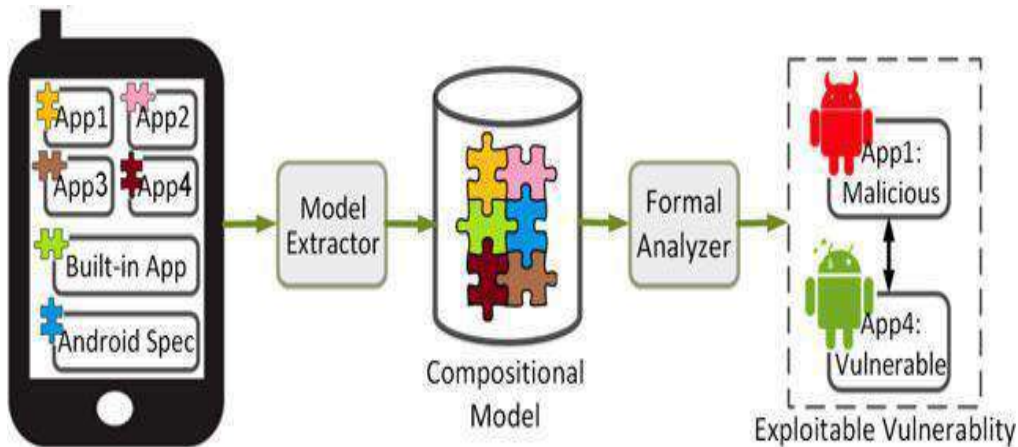


Fig. 1: Architecture

- Login / Registration and Upload.
- Reverse Engineering the apk file.
- Permission Check's in Source Code.
- Remove Unused Permissions.

### Permission Check's in Source Code

Android applications contain much permission to use the services. Developer must declare the permission in manifest to use that service. Once the permission is declared, the android application packager in the mobile phone will ask the users for accepting the permission usage while application installation.

For installing an application, the user has to approve all the permissions that application's developer has declared in the application manifest. If all permissions are approved, the application is installed and receives group memberships. The group memberships are used to check the permissions at runtime. Now the decompiled apk files are validated for permissions in the manifest.xml file. Now our high level permission checking framework examines the code written for each permission in java files and validates it. If the any of the permissions fails the validation process, it is tagged as Unused/Redundant permissions.

### Removing Unused Permissions

In this module, if unused permissions are declared, their respective service is also running in mobile. Missing permission causes the application to crash. Adding too many of them is not secure. Injected malware can use those declared, yet unused permissions, to achieve malicious goals. So the unused permissions found by our framework are removed in the Manifest.xml file. The modified/Permission checked source code is recompiled and harmless apk's are generated which can be downloaded using Qrcode. Only the uploaded source codes are recompiled.

## CONCLUSION

This paper presents a new approach for compositional learning of Android inter-app vulnerabilities. Our move toward employs static analysis to automatically recover models that reflect Android apps and interactions among them. It is able to leverage these models to identify vulnerabilities due to interaction of multiple apps that cannot

be detected with prior techniques relying on a single app analysis. We formalized the basic elements of our analysis in an analyzable specification language based on relational logic, and developed a prototype implementation, COVERT, on top of our formal analysis framework. The experimental results of evaluating COVERT against privilege escalation—one of the most prominent inter-app vulnerabilities—in the context of hundreds of real-world Android apps corroborates its ability to find vulnerabilities in bundles of some of the most popular apps on the market.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

### REFERENCES

- [1] *World Health Population* 11(3): 44–54. A. Shabtai, Y. Fledel, U. Kanonov, Y. Elovici, S. Dolev, and C. Glezer, “Google android: A comprehensive security assessment,” *IEEE Security Privacy*, vol. 8, no. 2, pp. 35–44, Mar./Apr. 2010.
- [2] E. Chin, A. P. Felt, K. Greenwood, and D. Wagner, “Analyzing inter-application communication in android,” in *Proc. 9th Int. Conf. Mobile Syst., Appl. Services*, 2011, pp. 239–252.
- [3] W. Enck, P. Gilbert, B. G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, “Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones,” in *Proc. 9th USENIX Conf. Operating Syst. Des. Implementation*, 2011.
- [4] P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall, “These aren’t the droids you’re looking for: Retrofitting android to protect data from imperious applications,” in *Proc. ACM Conf. Comput. Commun. Security*, 2011, pp. 639–652.
- [5] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, “Android permissions demystified,” in *Proc. ACM Conf. Comput. Commun. Security*, 2011, pp. 627–638.
- [6] W. Enck, M. Ongtang, and P. McDaniel, “On lightweight mobile phone application certification,” in *Proc. ACM Conf. Comput. Commun. Security*, 2009, pp. 235–245.
- [7] E. Fragkaki, L. Bauer, L. Jia, and D. Swasey. (2012). Modeling and enhancing android’s permission system, *Proc. ESORICS* [Online]. Available: [http://link.springer.com/chapter/10.1007/978-3-642-33167-1\\_1](http://link.springer.com/chapter/10.1007/978-3-642-33167-1_1)
- [8] S. Bugiel, L. David, Dmitrienko, T. A. Fischer, A. Sadeghi, and B. Shastri, “Towards taming privilege-escalation attacks on android,” presented at the NDSS Symp., San Diego, CA, USA, 2012.
- [9] M. Dietz, S. Shekhar, Y. Pisetsky, A. Shu, and D. S. Wallach, “Quire: Lightweight provenance for smart phone operating systems,” in *Proc. 20th USENIX Security Symp.*, 2011.
- [10] D. Jackson. (2002). Alloy: A lightweight object modelling notation. *TOSEM* [Online]. 11(2), pp. 256–290. Available: <http://portal.acm.org/citation.cfm?doid=505145.505149>
- [11] R. Valle\_e-Rai, P. Co, E. Gagnon, L. Hendren, and V. Lam, and P. Sundaresan, “Soot—a Java bytecode optimization framework,” in *Proc. Conf. Centre Adv. Stud. Collaborative Res.*, 1999, p. 13.
- [12] A. Bartel, J. Klein, Y. LeTraon, and M. Monperrus, “Dexpler: Con-verting android dalvik bytecode to jimple for static analysis with soot,” in *Proc. ACM SIGPLAN Int. Workshop State of the Art Java Program Anal.*, 2012, pp. 27–38.
- [13] J. Woodcock, P. G. Larsen, J. Bicarregui, and J. Fitzgerald, “Formal methods: Practice and experience,” *ACM Comput. Surv.*, vol. 41, no. 4, pp. 19:1–19:36, Oct. 2009.
- [14] P. Zave, “A practical comparison of alloy and spin,” *Formal Asp. Comput.* vol. 27, no. 2, pp. 239–253, 2015.
- [15] Android API reference document [Online]. Available: <http://developer.android.com/reference>, Oct. 2014.
- [16] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman, *Compilers: Principles, Techniques, and Tools*, 2nd ed. Boston, MA, USA: Addison-Wesley, 2006.
- [17] Android developers guide [Online]. Available: <http://developer.android.com/guide/topics/fundamentals.html>, Oct. 2014.
- [18] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie, “PScout: Analyzing the android permission specification,” in *Proc. ACM Conf. Comput. Commun. Security*, 2012, pp. 217–228.

\*\*DISCLAIMER: This published version is uncorrected proof. plagiarisms and references are not checked by IIOABJ. the article is published as provided by author and checked/reviewed by guest editor.



# SUPPORT VECTOR MACHINE BASED FRAMEWORK FOR DEMENTIA CLASSIFICATION

S. K. Aruna<sup>1\*</sup>, S. Chitra<sup>2</sup>, B. Madhusudhanan<sup>3</sup>

<sup>1</sup>Department of Computer Science and Engineering, Paavai Engineering College, Namakkal, TN- 637001, INDIA

<sup>2,3</sup>Department of Computer Science and Engineering, Er Perumal Manimekalai College of Engineering, TN, INDIA

## ABSTRACT

*Aims: Dementia is fast rising as a huge public health problem in recent times due to its extreme prevalence rate, huge burdens to patients in terms of health care costs and so on. Identifying alterable risk elements is significant for delaying or even preventing the onset of dementia. Magnetic resonance Imaging (MRI) is an affordable as well as non-radioactive imaging technique which does not have ionizing radiations. It possesses excellent spatial resolution and is commonly accessible within clinical environments. In the current work, image extraction is carried out through usage of Gabor as well as Grey-Level Co-Occurrence Matrix (GLCM). Classification is carried out through classifiers like K-Nearest Neighbor (KNN), Classification and Regression Tree (CART) as well as Support Vector Machines (SVM).*

Published on: 18<sup>th</sup>– August-2016

### KEY WORDS

Dementia, Magnetic Resonance Imaging (MRI), K-Nearest Neighbor (KNN), Classification and Regression Tree (CART)

\*Corresponding author: Email: [arunask.bir@gmail.com](mailto:arunask.bir@gmail.com)

## INTRODUCTION

Alzheimer's disease (AD) is the topmost common illness causing dementia amongst the older population. Through at present, there is no cure for the illness research is being carried out for developing novel treatment methods. Because most of these are greatly beneficial to pre-symptomatic patients, earlier detection of the disease is significant and so individuals suffering from Mild Cognitive Impairment (MCI), who are at heightened risk of falling prey to Alzheimer's are of interest here.

In recent times, it is possible to detect traces/biomarkers of AD in individuals who are suffering from Mild Cognitive Impairment through the usage of Magnetic Resonance Imaging (MRI) volumetric study, neurochemical analyses of cerebrospinal fluid, as well as Positron Emission Tomography (PET) scans [1]. This kind of research is costly, requires great technical expertise, invasive as well as available only in select locales. Longitudinal studies measuring the predictive values of neuropsychological tests in the advancement of individuals suffering from mild cognitive impairment to dementia reveal a region under the receiver operating characteristic (ROC) curve of 61-94% but with lesser accuracy as well as sensitivity values. It is significant to enhance values of neuropsychological tests for the prediction of advancement of MCI to dementia amongst patients. This is possible at the clinical level through raising the quantity of patients with longer clinical follow-up. Predictive capability of the tests can be improved by innovative statistical classification as well as data mining methods. Conventional statistical classification techniques for instance, Linear Discriminant Analysis (LDA) or Logistic Regression (LR) are wisely utilized in medical classification issues wherein criteria parameters are dichotomous [2]. Research is also being carried out into improving the accuracy as well as efficacy of data mining with classifiers such as Neural Network (NN), Support Vector Machine (SVM), Classification Tree (CT) as well as Random Forests (RF) utilized for medical predictions as well as classification tasks [3,4].

Dementia is a neurodegenerative illness whose causes are not yet known. AD is a typical kind of dementia wherein there is loss of neurons as well as synapses in the cerebral cortex as well as other sub-cortical areas. Although most individuals affected by dementia are elderly, not all elderly people are suffering from dementia which implies that it is not a natural effect of normal ageing. Dementia may impact anybody but typically it manifests in people above the age of 65. It is also noted that dementia is more prevalent amongst males [5].

In Alzheimer's, automated image classification was utilized in functional imaging as well as cortical thickness metrics for differentiating scans of patients with dementia and healthy patients. Of late, patterns recognition techniques employed to structural MRIs are utilized for distinguishing individuals suffering from mild cognitive impairment from healthy individuals.

MRI is an imaging method which has undergone evolution to become a clinical modality over a period of thirty years. Medical imaging tools assist healthcare professionals in reaching particular diagnoses. Medical image analyses as well as processing possess great importance in healthcare because of their non-invasive treatments as well as clinical studies. It also has a part to play in the identification as well as diagnosis of several illnesses. Imaging assists healthcare professionals in the visualization as well as analysis of images for understanding anomalies in internal structures. Medical images utilize tools such as CTs, MRIs, mammograms for the identification of lesions in the patient [6].

MRIs refer to scanning devices that utilize magnetic fields as well as computers for capturing images of brains on film. It does not utilize x-rays and yields images from several planes which permit healthcare professionals to view 3D images of the tumor. MRIs identify signals from normal as well as anomalous tissues and ensure clearer image scans of the tumors. It is vastly employed in high quality medical imaging situations, particularly when the brain is involved wherein soft tissue contrasts as well as non-invasive nature of the technique are extremely beneficial. MRIs are typically studied by radiologists and are interpreted visually for identifying anomalous tissues. Brain images are chosen for image references in the current work because brain injuries impact huge areas of the organ. The brain handles the control as well as coordination of movements, behavior as well as homeostatic bodily functions such as heart beat, blood pressure, body temperature and so on. The brain also handles cognition, memories, emotions as well as learning of all kinds.

The classification of brain MRIs as either normal or anomalous is significant in the pruning of healthy individuals and taking into consideration solely those with possible anomalies or tumors. There are only few radiologists and the quantities of MRIs to be examined are numerous, which makes it a very expensive process in labour as well as cost. Hence, there is a need for automatic systems for the analysis as well as classification of the images. Outcomes of human analyses regarding false negatives ought to be very low when handling human life. Double medical image readings result in improved tumors identification.

Classification refers to the assignment of physical entities or events to pre-specified categories. Medical image datasets for image classification or teaching possesses several modality images, obtained from various conditions with differing accuracy of annotations. This holds true for several online resource image scans, like those that access journals online content. Methods that combine visual as well as textual methods for classification show excellent promise in the classification of medical images.

## RELATED WORKS

Yasue et al [7] examined 783 patients as well as 2139 healthy controls who took part in a population-based study carried out in Japan. Sinusitis was tested for through usage of MRIs as per the Lund–Mackay scoring system. Sinusitis scores  $\geq 4$  were sorted as positive while scores  $\leq 3$  were sorted as negative. The presence of positive sinusitis was 6.3% in individuals with MMSE scores  $< 24$  ( $n = 507$ ), as well as 5.7% in individuals with AD ( $n = 280$ ). The presence of sinusitis was not considerably distinct between healthy individuals as well as those with dementia/AD after modifications for age/sex. The rate of positive sinusitis was greater for males than females in both groups.

Zheng et al [8] presented a summary of existing automatic dementia detection protocols in literature from the perspective of patterns classification. Because mostly these protocols comprise features extraction as well as classification, they offer a review on the three groups of features extraction techniques which are voxel-, vertex- as well as ROI-based ones as well as four groups of classifiers which are the LDA, Bayes classifier, SVM as well as ANN. The performance of the classifiers are contrasted and the comparison reveals that several protocols are capable of distinguishing AD from healthy controls with excellent accuracies although differentiating healthy controls from those suffering from MCI is still a difficult task.

Aruna & Chitra [6] suggested a model for the classification of MRIs for dementia. Dementia is an age-related disorder characterized by deterioration in cognition which is made manifest by the deterioration of cortical as well as sub-cortical structures. The characterization of these morphological alterations assists in the comprehension of development of diseases as well as earlier estimation as well as prevention of the illness. Modeling which is the capturing of the brain's structural variability and which still holds true in the classification as well as interpretation of diseases is a difficult task. Feature extraction is carried out through Gabor filter with 0, 30, 60, 90 orientations as well as GLCM. It is suggested for normalization as well as fusion of features. Independent Component Analysis (ICA) chooses attributes. SVM with various kernels is tested for efficacy in the classification of dementia. The work tests the suggested model through usage of MRIs from the OASIS database for the identification of dementia. Outcomes reveal that the suggested features fusion classifier attains excellent classification accuracy.

## METHODOLOGY

MRIs were collected from OASIS and utilized for evaluation of the suggested techniques in the current work. Feature extraction is carried out through usage of Gabor filter with 0, 30, 60, 90 orientations as well as GLCM. Features undergo normalization as well as fusion for obtaining fused features vector. mRMR is used for features selection. Naïve Bayes, Neural Network, Ensemble Neural Network classifiers are utilized for classifying images as dementia or non-dementia on the basis of the chosen attributes.

### OASIS data set

OASIS dataset comprises 416 subjects ranging between 18 and 96 years of age which also includes people with early-stage Alzheimer's [9]. 98 right-handed women (aged between 65 and 96) were chosen from the OASIS dataset. 200 subjects with incomplete records were discarded. For the current work, images of 49 normal subjects as well as 49 subjects with very mild to mild Alzheimer's are utilized. OASIS dataset was put together after strict imaging protocols for curbing imaging protocol variants that posed big image standardization problems. Several high-determination auxiliary T1-weighted Magnetization-Prepared Rapid Gradient Echo (MP-RAGE) images were obtained in one imaging session.

### Feature Extraction

Statistical Parameter Mapping (SPM12) is utilized for segmentation of brain images into Grey as well as White Matter. Textural features are chosen through Gabor as well as GLCM. Gabor filters are transform functions associated with Fourier transforms that may be utilized for conveying spatial data additional to frequency characteristics of signals. It is generally employed as band-pass filters in signal processing wherein it is utilized for the determination of sinusoidal frequencies as well as phase content of local sections of time-varying input signals and has been proven to be helpful in the case of image compressions. Amongst other helpful characteristics, Gabor filter has been discovered to perform minimization of conjoint time-frequency information resolutions of signals in a better manner. 1D Gabor filter is featured as a collection of cosine/sine (even/odd) waves with Gaussian windows,

$$g_e(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \text{Cos}(2\pi w_o x) \quad (1)$$

$$g_o(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \text{Sin}(2\pi w_o x) \quad (2)$$

Wherein refers to focus (frequency wherein filters yield most significant reactions) as well as  $\sigma$  refers to distribution of Gaussian windows.

The strength of Gabor filters responses depend on the filters' congruence with local signals; wherein the filters' sensitivities is defined through tuning of variables which are orientation, phase as well as frequency [10]. Because Gabor filters are not orthogonal [11], optimum set of variables for the design of Gabor filter jets which will identify favored range of object features with minimal redundancy is determined. Assume  $g(x, y, \theta, \phi)$  is a function that characterizes Gabor filters run at root with  $\theta$  as spatial frequency while  $\phi$  refers to orientation. Gabor filters are represented as (3):

$$g(x, y, \theta, \phi) = \exp\left(-\frac{x^2 + y^2}{\sigma^2}\right) \exp(2\pi\theta i(x \cos \phi + y \sin \phi)) \quad (3)$$

It was revealed that  $\sigma$ , standard deviations of Gaussian kernels rely on  $\theta$  measured.

Two-dimensional Gabor functions comprise sinusoidal plane waves of frequency as well as orientation, altered by two dimensional Gaussian envelopes. Canonical Gabor filters in space are given by (4):

$$h(x, y) = \exp \left\{ -\frac{1}{2} \left[ \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right] \right\} \cos(2\pi\mu_0 x + \phi) \quad (4)$$

Wherein  $\mu_0$  as well as  $\phi$  refer to frequency as well as period of sinusoidal plane waves along the z-axis (that is, the 0° orientation), refer to space constants of Gaussian envelopes along x-axis as well as y-axis, separately. Gabor filters with self-assertive introductions may be obtained through unbending revolutions of x-y coordinates model [12].

When investigating statistical textures, texture attributes are computed on statistical conveyance of pixel intensities at locations relative to others in pixels associating with image matrices. Contingent upon pixels or spots in combinations, there is first-order, second-order or higher-request statistics. GLCM based features extraction is second-order statistics which studies images as textures. A basic method for description of intensities, but not regarding the relative position of pixels with regard to one another in that particular texture is suggested. Utilizing a statistical method like co-occurrence matrix will assist in the provision of valuable data regarding about the relative position of neighboring pixels in a particular image. GLCM (similarly grey tone spatial reliance matrix) refers to frequency tabulation of how often a combination of pixel luminance quality in images. Apart from horizontal heading (0°), GLCM may be shaped for bearing of 45°, 90° and 135°. The feature vectors of Gabor and GLCM are normalized and fused to obtaining a combined features vector.

### Feature Selection

A filter-based features extraction framework known as minimum-Redundancy Maximum-Relevance (mRMR) that attempts to choose the important attributes with target class labels as well as decrease redundancy among chosen attributes all the while, the protocol utilizes Mutual Information  $I(X, Y)$  that assesses the degree of similitude between two discrete arbitrary parameters X as well as Y [14]:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p_1(x)p_2(y)} \right) \quad (5)$$

Wherein  $p(x, y)$  refers to the joint probability distribution capability of X as well as Y, while  $p_1(x)$  as well as  $p_2(y)$  refer to the minor probability distribution elements of X as well as Y separately.

Information theoretic positioning criterion mulls over nonlinear links between features as well as targets. Evaluation of features is carried out in an autonomous manner and features redundancy is not capable of being managed. For investigating mRMR systems as well as or handling the issue of redundancy, choosing of perfect features for classification is carried out. For features set S comprising  $n_0$  attributes  $\{x_i\}$ , ( $i = 1 \dots n_0$ ). The topmost priority is the identification of attributes so that common data characteristics between individual attributes as well as target class are to be amplified. Assume  $D(S, y)$  is the mean of common data between individual attribute as well as target  $y$ . It is mathematically given by:

$$\max D(S, y) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, y) \quad (6)$$

Although two attributes might possess solid separability on the target class, it is not desirable to include them just in case they have exceptional correlation. The notion of minimal redundancy is the choosing of attributes which have common maximum divergence. Assume  $R(S, y)$  is the mean of the common data between sets of features in S. It mathematically given by:

$$\min R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \quad (7)$$

The foundation bringing together the two equations given above is known as mRMR. mRMR features set is obtained through the boosting of  $D(S, y)$  as well as minimization of  $R(S)$  simultaneously that needs combining the two metrics into one model capacity (one criterion function).

### Classifiers

#### *K Nearest Neighbor (KNN)*

KNN refers to a supervised patterns recognition method that carries out segmentations through comparison of novel data to set of labeled samples in training sets. KNN classifiers are easier to utilize relative to others and this ensures that the procedure is more rapid. Hence, the primary benefit is that training models may be created more rapidly.

#### *Classification and Regression Tree (CART)*

CART [15] refers to an alternate method wherein data space is split into small sections wherein parameter interactions are clear. CART is a non-parametric, machine-learning technique which splits values of all predictor parameters in a recursive manner into two sets such that values of outcome parameters are homogeneous in all sets. All predictor parameters are regarded as potential splits, including all grouped values of numerical predictors. Optimal splits are those with greatest decrease in impurity indices that measure level of misclassification at a particular node.

### Support Vector Machine (SVM) Classification

Support Vector Machines refer to supervised, multivariate classification techniques, which imply that they possess training sets for learning about the differences between groups to be sorted. The technique was earlier employed to neuroimaging data. The data required for this technique is not required to fulfill presumptions of Random Field Theory, ensuring further smoothening not necessary [16]. Within the context of machine learning, individual MRIs are considered as points situated in high-dimensional space.

The space utilized for classification of image data is of greater dimensions, the total quantity of dimensions is defined by the number of voxels in all MRIs. Practically, linear kernel matrices are generated from normalized grey matter segmented scans. For this purpose, all MRIs undergo pair-wise multiplication with the others. All elements in the kernel matrices are thereby dot products of two images. Kernel matrices are also understood as similitude metric amongst subjects on a characterized set. Voxels are efficiently treated as coordinates of higher dimensional spaces and the position is defined by the intensity values at all voxels. The images do not span the entire higher-dimensional space and instead, they cluster in sub-spaces comprising images which are alike. This is the reason why image normalization into standard spaces is a significant pre-processing stage. Excellent normalization tightens clustering as well as reduces dimensionality.

The usage of SVMs for image classification is an instance of linear discrimination. In basic models, they are binary classifiers which imply that space is divided into which MRIs are sorted into two classes through identification of separating hyper planes. In simple two dimensional spaces, boundaries are denoted by lines but are known as hyper planes in higher dimensional spaces. Fisher's LDA or linear perceptron is capable of identifying linear discriminant hyper planes. But the reason or usage of SVMs is the fact that they utilize the principle of 'structural risk minimization' that focuses on the discovery of hyper planes which make maximum the distance between training classes.

Support Vector Machine (SVM) refers to a machine learning derived classifier that map vectors of predictors into high dimensional planes through linear or nonlinear kernel functions. In binary classification issues, the two groups, for instance  $\{-1\}$  as well as  $\{+1\}$ , are made separate in higher-dimension hyperplanes according to the structural risk minimization principle. The aim is the discovery of linear separating hyper planes

$$w'\phi(x) + b = 0$$

Generated from vector  $x$  of predictors mapped into high dimensional features space by non-linear features function  $\phi$ , vector  $w$  of weights as well as a bias offset  $b$ , which sorts all observations  $y_i$  into one of the two  $\{-1; +1\}$ . Classification function is given by:

$$f(x) = \text{Sign}(w'\phi(x) + b)$$

Because in binary classification issues there are infinite separation hyper planes, the aim is the discovery of optimal linear planes that group best. For finding optimal planes farthest from  $\{-1\}$  as well as  $\{+1\}$  groups, one method is the maximization of distances or margins of separation from the supporting planes, correspondingly  $w'\phi(x) + b \geq +1$  for  $\{+1\}$  as well as  $w'\phi(x) + b \leq -1$  for  $\{-1\}$ . The support planes are pushed apart till they turn into a small set of observations or training patterns which respect the limits mentioned and therefore are known as support vectors. Figure 2 shows this notion. Classification goals may be attained through the maximization of distances or margins of separation  $r$  between the two planes  $w'\phi(x) + b = +1$  and  $w'x + b = -1$  specified by  $r = 2 / \|w\|$ . This is the same as the minimization of the cost function

$$C(w) = \frac{\|w\|^2}{2} + c \sum_{i=1}^n \xi_i = \frac{1}{2} w'w + c \sum_{i=1}^n \xi_i$$

Subject to linear inequality limits

$$\gamma_i(w'\phi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

wherein  $c > 0$  is penalty variable which balances classification errors versus the complexity of the framework that is

monitored by margins of separation while  $\xi_i$  is known as the slack-parameter. The parameter is the penalty of misclassified observations which control how far on the wrong side of hyper planes points may lie when training data is not capable of being classified without errors, i.e. when objects are not capable of linear separation and soft separating non-linear margins are needed.

As features space may be infinite, non-linear mapping by features function  $\phi$  is calculated through special non-linear semi-positive definite K functions known as kernels. Hence, the above minimization is typically resolved by dual formulation issue:

$$\min \frac{1}{2} \sum \gamma_i \gamma_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^n \alpha_i$$

Subject to linear limits

$$\sum_{i=1}^n \gamma_i \alpha_i = 0 \text{ and } 0 \leq \alpha_i \leq C$$

Wherein  $\alpha_i$  ( $i = 1, \dots, n$ ) represent non-negative Lagrange multipliers while  $K(\cdot)$  represents kernel functions. In classification issues (c-SVM) the typical kernel functions are linear kernel  $K(x_i, x_j) = x_i x_j$  or Gaussian  $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$  wherein  $\gamma$  refers to kernel variable. The usage of kernel functions possesses the benefit of operating in the initial input parameters wherein solutions of classification problems are weighted sum of kernels tested at support vectors [18].

## RESULTS

In proposed method for experiments, 280 normal MRI image and 140 images with dementia are used. In this section, the classification accuracy, Sensitivity for Normal, Sensitivity for abnormal, Specificity for Normal and Specificity for abnormal are evaluated in table 1. Figure 1 to 5 shows the same.

Table 1 Summary of Results

	KNN	CART	SVM
Classification Accuracy	84.69	85.71	89.8
Sensitivity for Normal	0.8571	0.9184	0.9184
Sensitivity for abnormal	0.8367	0.7959	0.8776
Specificity for Normal	0.8367	0.7959	0.8776
Specificity for abnormal	0.8571	0.9184	0.9184

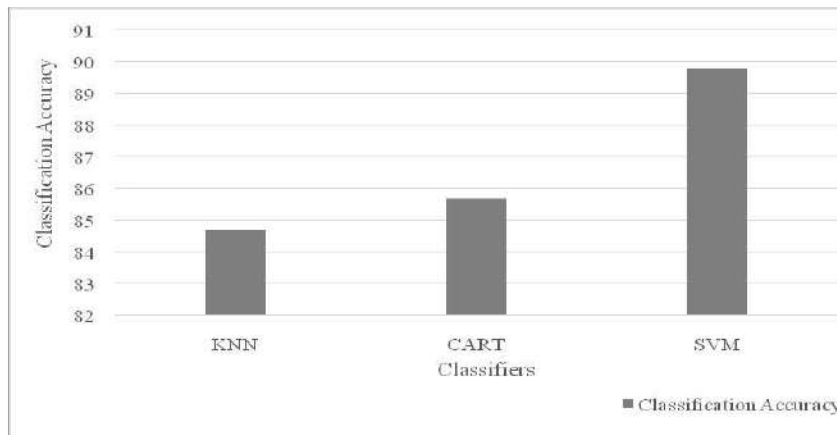


Fig. 1: Classification Accuracy

From the table 1 and figure 1, it can be observed that the classification accuracy for normal image of SVM performs better by 5.81% than KNN and by 4.66% than CART.

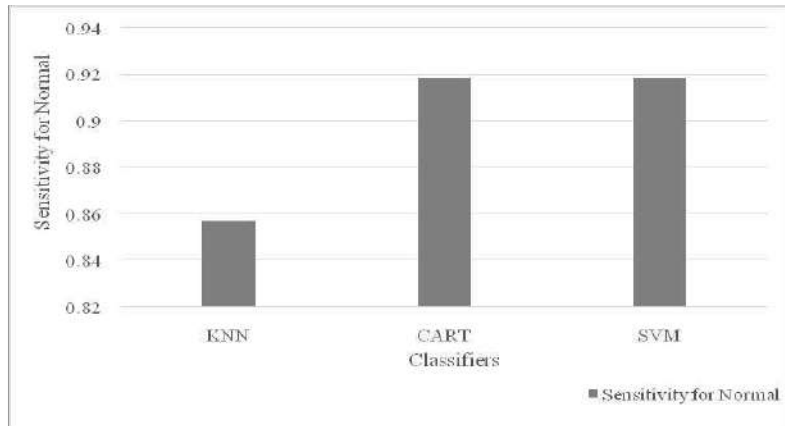


Fig. 2: Sensitivity for Normal

From the table 1 and figure 2, it can be observed that the sensitivity of SVM and CART performs in equal way. Both performs better by 6.91% than KNN.

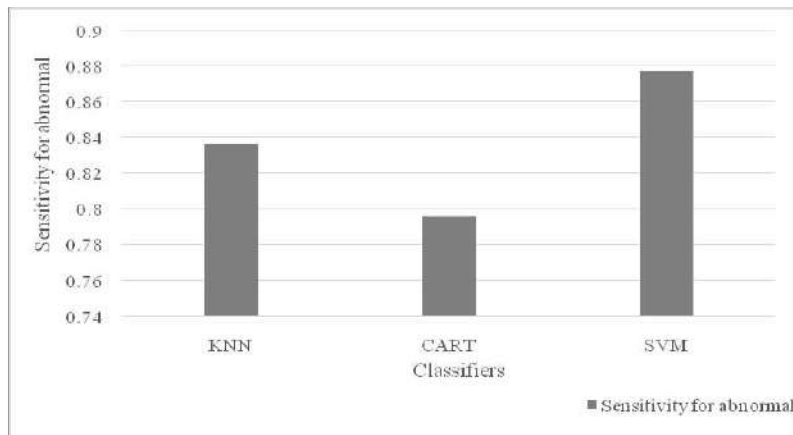


Fig. 3: Sensitivity for abnormal

From the table 1 and figure 3, it can be observed that the sensitivity for abnormal image of SVM performs better by 4.77% than KNN and by 9.8% than CART.

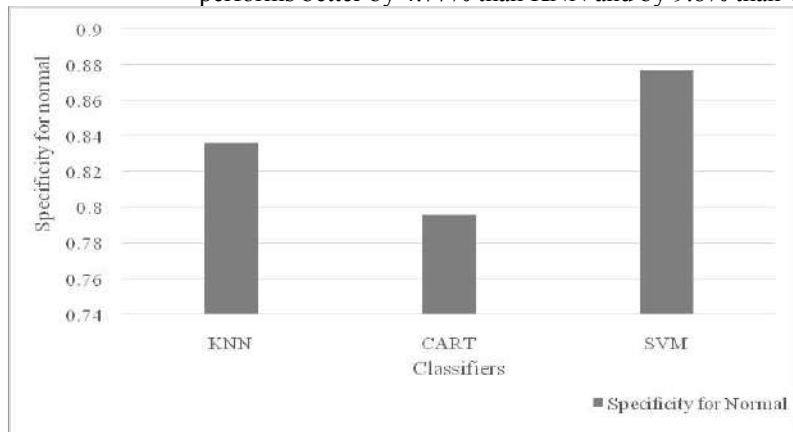


Fig. 4: Specificity for Normal

From the table 1 and figure 3, it can be observed that the specificity for normal image of SVM performs better by 4.77% than KNN and by 9.8% than CART.

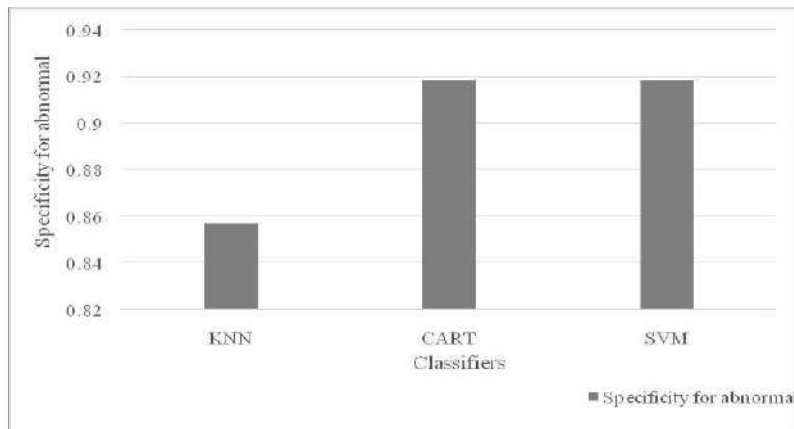


Fig. 5: Specificity for abnormal

From the table 1 and figure 3, it can be observed that the specificity for abnormal image of SVM and CART performs in equal way. Both performs better by 6.91% than KNN.

## DISCUSSION

AD is the most common cause of age-related dementia. Because of the rising proportion of older people in western societies, the presence of dementia is anticipated to double in the coming thirty years. MRI images collected from OASIS are used to identify dementia. Feature extraction through usage of Gabor filter with 0, 30, 60, 90 orientations as well as GLCM. For experiments, classifiers such as KNN, CART and SVM for obtaining performance measures such as Classification accuracy, sensitivity and specificity for normal and abnormal images. Results show that the classification accuracy for normal image of SVM performs better by 5.81% than KNN and by 4.66% than CART.

The lab coat can get contaminated by microorganisms due to improper handling practices. They get easily contaminated because patients continuously shed infectious microorganisms in the hospital environment, and the health care providers are in constant contact with these patients. *Staphylococci* are the pathogens belonging to the group of Enterobacter bacteria, which cause several infections to humans. They are facultative anaerobic gram-negative cocci mainly found in the skin and mucosa and are of three types *Staphylococcus aureus*, *Staphylococcus epidermidis* and *Staphylococcus haemolyticus* [7]. Health care professionals are most susceptible to colonization, and the main form of transmission is through temporarily colonized hands. Importantly, treatment of infections caused by *S. aureus* has become difficult because of their higher resistance to various drugs [8].

*S aureus* is part of the normal human microbial flora and it is found in the nasal passages, throat, gastrointestinal tract and skin. It is considered as one of the most important pathogenic bacteria, causing series of infections [9, 10] leading to the formation of abscesses. It causes infections such as furuncles, folliculitis, scalded skin syndrome, meningitis, and pneumonia. Coagulase-negative *Staphylococci* (CONS) which is a skin commensal has recently got attention as a potential pathogen, specifically for nosocomial infections [11-13]. CONS are a major cause of nosocomial infection and septicemia, especially in cases of immune-compromised patients [12].

This study evaluated the type of microbial flora present on the lab coats of the clinicians working in the Dept of Endodontics and their antibiotic sensitivity. Three sites were chosen i.e chest, pocket, cuff for determining the type of microbial flora. Microbial contamination was thought to be highest as these sites most commonly comes in contact with the patients [14,15]. This study showed that the numbers of gram positive cocci was the same as that of other studies and maximum of them were potentially pathogenic [15,16]. This is consistent with other studies that showed contamination of lab coats ranging from 23% to 95% [17]. They possess a risk of cross contamination if the host is immune compromised. *Micrococci* may act as an opportunistic pathogen in patients with compromised immune systems and they most commonly cause blood stream infection. Gram negative *Bacilli* were also isolated, but these were significantly lesser in number and they may be potentially infectious, as was reported by Zachary and Grabsch. They have shown that bacterial survival rate is of longer period of time on



hospital fabrics [18,19]. Chacko et al have shown that on lab coat fabrics made up of either cotton polyester or polyester material, bacteria can survive between 10-98 days [20]. Hence the lab coats should be washed daily or at least once in 3 days [20]. Of the two predetermined sites selected for examination on the lab coat, the mouth of the dominant pocket was more contaminated than the chest and cuffs of the sleeve. This is similar to the study of Nelly and contrary to that of Uneke and Ijeocoma which indicated that cuff has more bacterial load than the pocket [21, 22]. Pocket is the highly contaminated area because it frequently comes in contact with the hands of the health care professionals harboring bacterial contaminants.

Antibiotic sensitivity testing showed resistant species of microorganisms on the lab coats against Amoxicillin, Penicillin G, Gentamycin, and Cotrimoxazole. Antibiotic sensitivity results showed the organisms which were sensitive to most common antibiotics on 1st day got resistant on 3<sup>rd</sup> day [Figure-4]. Of the *S. aureus* isolated, 10% were MRSA. The MRSA has emerged as significant bacteria in hospital acquired infections. According to the Centre for Disease Control and Prevention, more than 60% of all hospital infections are caused by MRSA in United States. Because of frequent dermal contact, lab coats can harbor these resistant bacteria. In orders to prevent cross infection, guidelines should be followed for handling and washing procedures of lab coats.

This is a uni-centric study, done to create awareness among our dental colleagues. This study reflects center-specific microbial contamination in a dental operator. To reach to a more generalized conclusion, the study requires a multi-centric evaluation with a larger sample size.

## CONCLUSION

The present study highlights the fact that the lab coats may act as a vector for transmission of cross infection. In order to prevent transmission of cross infection, a strict protocol should be set in order to prevent cross contamination between doctor and patient. Efforts should be made to limit the use of coats outside the working area and they should be laundered every day. Wearing of plastic aprons or altering lab coat material to plastic-laminated clothing or closely woven waterproof cotton can reduce the bacterial transfer rate and cross-contamination

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delocourte A, Galasko D, Gauthier S, Jicha G, et al: Research criteria for the diagnosis of Alzheimer"s disease: revising the NINCDS- ADRDA criteria. *Lancet Neurology* 2007, 6:734-746.
- [2] Pohar M, Blas M, Turk S: Comparison of Logistic Regression and Linear. Discriminant Analysis: A Simulation Study. *Metodološki zvezki* 2004, 1:143-161.
- [3] Peter CA: A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine* 2007, 26:2937-2957.
- [4] Zollner FG, Emblem KE, Schad LR: Support vector machines in DSC-based glioma imaging: Suggestions for optimal characterization. *Magn Reson Med* 2010.
- [5] Chen, Y., & Pham, T. D. (2013). Development of a brain MRI-based hidden Markov model for dementia recognition. *Biomedical engineering online*, 12(Suppl 1), S2..
- [6] Aruna, S. K., & Chitra, S. (2016). Machine Learning Approach for Identifying Dementia from MRI Images. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 9(3), 881-888.
- [7] Yasue, M., Sugiura, S., Uchida, Y., Otake, H., Teranishi, M., Sakurai, T., ... & Nakashima, T. (2015). Prevalence of Sinusitis Detected by Magnetic Resonance Imaging in Subjects with Dementia or Alzheimer's Disease. *Current Alzheimer Research*, 12(10), 1006-1011.
- [8] Zheng, C., Xia, Y., Pan, Y., & Chen, J. Automated identification of dementia using medical imaging: a survey from a pattern classification perspective. *Brain Informatics*, 1-11.
- [9] <http://www.bio-edicine.org/Biology>
- [10] Li, C.T.; Wei, C.H; Wei, C.H.; Li, C. (2005); A Content-based Approach to Medical Image Database Retrieval, Database Modeling for Industrial Data Management: Emerging Technologies and Applications 10(6): 681-685.

- [11] Wei, C.H.; Chang-Tsun, L.; Roland, W. (2009); A Content-based Approach to Medical Image Database Retrieval, Database Technologies: Concepts, Methodologies, Tools, and Applications: 1062-1083.
- [12] Gulgezen, G., Cataltepe, Z., & Yu, L. (2009). Stable and Accurate Feature Selection. Springer
- [13] R. C. Gonzalez, R. E. Woods, Digital Image Processing, 3rd Ed. Prentice Hall, 2008.
- [14] Zhuo, Z. (2012). Automatic Glaucoma Diagnosis with mRMR-based Feature Selection. Journal of Biometrics & Biostatistics.
- [15] Koroukian, S. M., Schiltz, N., Warner, D. F., Sun, J., Bakaki, P. M., Smyth, K. A., ... & Given, C. W. (2016). Combinations of Chronic Conditions, Functional Limitations, and Geriatric Syndromes that Predict Health Outcomes. *Journal of general internal medicine*, 1-8.
- [16] Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., ... & Frackowiak, R. S. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3), 681-689.
- [17] Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests
- [18] Ivanciuc O: Applications of Support Vector Machines in Chemistry. In Reviews in Computational Chemistry. Volume 23. Edited by: Lipkowitz KB, Cundari TR. Weinheim: John Wiley 2007:291-400.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ, the article is published as provided by author and checked/reviewed by guest editor.

## AFSA DSR- ARTIFICIAL FISH SWARM ALGORITHM DYNAMIC SOURCE ROUTING PROTOCOL FOR MANET

Prasath. N<sup>1\*</sup>, P. Sengottuvelan<sup>2</sup>, B. Vinoth Kumar<sup>3</sup><sup>1</sup>Dept of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, TN, INDIA<sup>2</sup>Dept of Computer Science, Periyar University PG Extension Centre, Dharmapuri, TN, INDIA<sup>3</sup>Dept of EEE, Dr.Mahalingam College of Engineering & Technology, Pollachi, TN, INDIA

## ABSTRACT

**Aims:** Information technology is being extensively used during emergencies for efficient handling of data and communication. In emergencies due to natural disasters like earthquakes or tsunamis Wireless Ad Hoc Networks provide a reliable communication link when existing infrastructure has collapsed. Ad hoc networks do not require any fixed infrastructure and networks can be formed on demand and hence can be used in counter terrorism operations also. For efficient operation of Ad hoc network routing plays an important role and the routing technique should be capable of adapting itself to the scenario under which it is deployed. This work proposed an optimized Dynamic Source Routing (DSR) protocol for MANET. To locate optimal paths between communicating nodes, traditional DSR algorithm is modified based on multiple objectives and called as Link Quality Load Balanced Dynamic Source Routing (LQLB-DSR). Solutions obtained using LQLB-DSR is suboptimal due to dynamic variation of network parameters. To overcome this an Artificial Fish Swarm Algorithm (AFSA) to achieve deserved Quality of Service (QoS) is proposed. Extensive simulations show the improved performance of the proposed routing protocol

Published on: 18<sup>th</sup>– August-2016

## KEY WORDS

Mobile Ad-Hoc Networks (MANETs), Dynamic Source Routing (DSR), Link quality, mobility, end to end delay, Artificial Fish Swarm Algorithm

\*Corresponding author: Email: [prasadn.dsr@gmail.com](mailto:prasadn.dsr@gmail.com)

## INTRODUCTION

MANETs use the dynamic distributed system concept with many highly mobile devices communicating through a wireless medium. Every device has limited energy and communicates without fixed infrastructure. As each device/node moves arbitrarily, topology changes frequently and unpredictably. These characteristics make location prediction and reliable routing a challenge in MANETs. Each MANET node acts as host and router dynamically based on joining of new nodes and exit of current nodes. MANET can also work with existing wired local area networks. Key features of MANETs can be listed into [1]:

- All nodes are self-configurable dynamically.
- Every node may move in different directions and differing speeds.
- They work on constrained bandwidth.
- A node provides essential network functionalities like routing and maintenance, as there is no fixed infrastructure.
- Distributed algorithms are used for organizing, routing and scheduling.
- Security is limited as wireless medium is used for communication.

With availability of low cost devices and rugged operating systems for hand held devices MANET can become the first line of communication during natural emergencies like earthquakes, tsunamis and manmade emergencies such as war and terrorism [2]. Emergency applications of MANET include Search and rescue operations, Replacement of fixed infrastructure in case of environmental disasters, Fire fighting. Figure-1 shows a disaster communication setup using wireless network.

MANET efficiency depends on the cooperation of wireless nodes [3] and efficient routing. Conventional routing algorithms come under three categories including table driven or proactive algorithms, on demand or reactive algorithms and hybrid routing algorithms. Table driven routing algorithms are based on recently collected information on active paths stored in each node as a routing table and updated regularly. Periodic updates identify

when a path is invalid or establishes a new path. Updating routing table consumes node's power and wireless medium bandwidth due to additional overheads [4]. If updating interval is long, then routing table does not have most recent topology change leading to stale routes. Wireless Routing Protocol (WRP), Destination Sequence Vector Routing (DSDV) and Fisheye State Routing (FSR) are proactive routing protocols. Demand driven or On-demand routing algorithms begin route discovery to send data to destination node. When a source node wants to send a packet to a wireless node, source node starts the discovery process. A path is found and registered with agreement between source and destination nodes. In this technique, route discovery introduces delays [5]. Dynamic Source Routing (DSR) and Adhoc On demand Distance Vector Routing (AODV) are popular reactive routing protocols. Hybrid routing improves routing by combining table driven and on demand routing algorithms best features [6,7]. Zone Routing Protocol (ZRP), Zone based Hierarchical Link State Routing (ZHLS), and Hybrid Adhoc Routing Protocol (HARP) are examples of hybrid routing protocols.

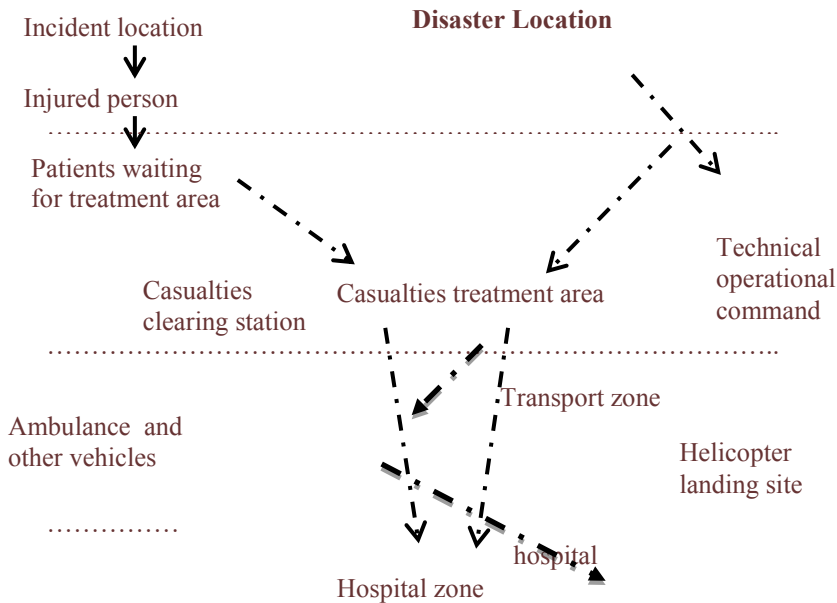


Fig. 1: A MANET based command setup during disaster

Conventional routing algorithms are not very efficient for MANETs distributed environment as network topology changes dynamically and Quality of Service (QoS) parameters differ for different applications. Traditional routing algorithms try to locate shortest path between source and the destination nodes. However, factors like latency, propagation path loss, and variable wireless link quality, protection against intentional jamming, interference, fading, power consumed, reliability, security and recovery from failure can be used for optimization the routes for specific applications. Failing to consider these parameters can degrade network performance and dependability [8].

Current MANET routing protocols do not consider parameters like network load along the path, processing capability of the intermediate node or the available bandwidth during the route selection process. To overcome the disadvantages this work proposed a Link Quality & Network Load Balanced Dynamic Source Routing is proposed (LQLB-DSR). Since multiple parameters have to be optimized, the routes discovered by LQLB-DSR are suboptimal and NP-Complete. To overcome this a Swarm Intelligence technique based on the behaviour of Fish Swarm is proposed. The rest of the work is divided into the following subsections. Section 2 discusses existing work, section 3 describes the proposed algorithm with section 4 discussing the experimental set up and results obtained. Section 5 concludes the paper.

## RELATED WORKS

The performance of various routing protocols like DSDV, AODV, and DSR for MANETs was investigated and compared by Manickam et al., [9], Shrestha and Tekiner [10] & Fan et al., [11]. After investigation, authors conclude that DSDV was suitable when number of nodes are limited and mobility is low due to storing routing

details in nodes. DSR suited moderate size networks with moderate traffic, as DSR used source routing and route cache at nodes. AODV provided best packet delivery ratio but had high end to end delay.

An Enhanced Multi-Path DSR algorithm (EMP-DSR) was proposed by Asl et al., [12]. The proposed algorithm used Ant Colony Optimization (ACO) to get global information about nodes and individual paths reliability. Simulation showed that EMP-DSR's overall reliability and end to end delay was higher than simple Multipath DSR but had additional overheads. Rajesh et al., [13] proposed an Artificial Neural Network (ANN) to secure MANET resources thereby enhancing DSRs performance. When routing, if a legitimate node is erroneously mistaken for a rogue node then QoS deteriorates. A new method to optimize route discovery in DSR routing algorithms was proposed by Hussein [14] using fuzzy logic. The objective of optimization was to minimize number of broadcast of RREQ packet, as this influenced overhead significantly in DSR routing. The author suggested a decision algorithm through a fuzzy logic system where each link was assigned a weight when destination path was constructed. Weights were based on the individual link's load and quality. When a node receives a RREQ, only good quality nodes reply. Simulations showed decrease in RREQ overhead by 30% compared to traditional DSR. An optimized routing algorithm was proposed by Arafat et al [15] using ACO. The multi objective optimization problem considered estimated link quality, delay and energy parameter. An algorithm to detect many disjoint routes from a specific source to a destination node was proposed by PrabhakaraRao et al., [16]. Route life was expressed by factors like Link Expiration Time (LET), Drain Rate (DR) and link stability. Link stability is proportional to LET and indirectly proportional to DR. An optimal path was selected to use bandwidth efficiently based on these factors. DSR routing found multiple disjoint paths by modifying RREQ and RREP packets at source node. An efficient algorithm to select multiple paths for source destination pairs based on location and bandwidth information and energy parameters was proposed by Shuchita et al., [17] to ensure constant, stable, healthy node disjoint paths satisfying many QoS parameters. A Multi-population Firefly Algorithm (MFA) for correlated data routing in Underwater Wireless Sensor Networks (UWSNs) was presented by Xu and Lin [18]. Three kinds of fireflies were used with coordination rules to improve adaptability of building, selecting, and optimizing routing path considering data correlation and sampling rate in sensor nodes. Different groups of fireflies conducted optimization independently to improve convergence speed and the proposed algorithm's solution precision.

Routing with TCP protocol was extensively studied by Swain et al., [19]. Kim et al., [20] proposed a disaster communication network and simulated the performance and showed the usefulness of MANET for telemedicine services. Kulla et al., [21] proposed data replication to support critical applications in disaster management. The proposed technique used fuzzy logic and optimization technique to improve the QoS. Lee & Yang [22] proposed an emergency escape system from disaster area using their own mobile devices. The proposed system utilized users' information such as weighted distances, weighted directions and revisit counts. The solution was implemented on a double-layered MANET to reduce traffic and to collaborate with each other proving that the proposed system could be an effective life-saving rescue system. Bai et al., [23] proposed an integrated communication system by composing heterogeneous wireless networks to facilitate the rescue teams and victim people to communicate inside and outside the disaster site.

## MATERIALS AND METHODS

Based on literature survey it can be concluded that QoS requirement changes based on the scenario where the MANET solution is deployed. To achieve this multiple objectives need to be optimized as solution outcome under different scenario will be sub optimal. This work proposed two new techniques Link Quality & Load Balanced –DSR (LQLB-DSR) and Artificial Fish Swarm Algorithm-DSR (AFSA-DSR) which is optimization of the LQLB-DSR.

In DSR routes are located when a source node needs to transmit data to a specific destination. This is a unicast routing protocol where a node uses cache memory to store route information of a node's recently taken routes. The two DSR routing steps are Route discovery and Route maintenance. When a source node attempts to send a packet to destination node, it checks its route cache [24, 25]. If route is available to destination node, then it transmits packet in the available path. Or else a route discovery process is initiated by broadcasting Route Request packets (RREQ). Nodes on receipt of RREQ check its route cache. If routing entry is not in it, the node appends its own address in RREQ packet and broadcasts it to neighbouring nodes. If RREQ packet reaches destination or a node with routing information to destination, it generates a Route Reply (RREP). Reply packet has nodes addresses traversed by the request packet. To maintain routing information, when data link layer finds a link failure or disconnection, a ROUTE ERROR packet is generated and transmitted from failed point to source node backward of sent data. So, all intermediate nodes delete route information via failed link. Then source node initiates another route location process. DSR benefits include route discovery control overheads reduction as it uses route cache. DSR's limitation is large size of packet header as per length of route due to source routing.

## PROPOSED LQLB MECHANISM

In the proposed technique Link Quality (LQ) and Network Load (NL) is measured and the path with high LQ and low NL is selected. Link quality (LQ) varies in a wireless medium compared to wired medium due to noise generated by outside source, distance between nodes being higher leading to higher packet loss. LQ can be evaluated based on the type of antenna in the node, the power capability of the antenna and can be computed by

$$P_r = T_{x_p} \times R_g \times T_g \times \frac{\lambda^2}{(4 \times \pi \times d)^2}$$

where

$P_r$  = received power,

$T_{x_p}$  = Power used for transmission,

$R_g$  = Gain of antenna at receiver,

$T_g$  = Gain of antenna at transmitter

$\lambda$  = wavelength,

$d$  = distance between two nodes.

The equation evaluates link quality based on received signal strength descriptive which is based on distance and mobility for a given time period [26]. Two more wireless link quality metrics, Hop Count metric for the entire route and Per-hop Round Trip Time (RTT) is also measured.

In many scenarios the traffic in the MANET is high in certain regions of the network leading to load balancing problems [27] leading to congestion, buffer overflow and packet drop. Effectively the channel is affected due to the low bandwidth typically available in MANETs [28]. In such scenarios the routing algorithms may not stabilize. Figure 2 shows how nodes can be affected by excessive traffic flow. It can be noted that node 4 faces congestion as it handles traffic from node 2, 3, 7, 6 and 1.

The overall data carried by each node for duration are computed. Route is selected based on the three parameters stated.

During route discovery process, the source node broadcasts Route Request (RREQ) packet with a LQLB query in packet header. Each node updates the LQLB once it receive the RREQ based on the sequence number concept of DSR. The LBLQ collects the total packet generated by each node and the number of packets it has received for a given duration and this represents the load balancing parameter identifying the weakest node. LQ identifies the weakest link in the route. LQLB field is added in each routing table entry and this gets updated when [29]:

- The received sequence number is higher than sequence number stored in the routing table.
- The sequence numbers are equal, but the hop count in the new RREQ is lower than the hop count stored in the routing table.
- The sequence numbers are equal, the LQLB in the routing table is smaller than the pre-defined
- LQLB value in the routing table is higher than current value.

The various routes are ranked and the best route selected. The ranking is based on the QOS parameter selected.

$$R_c = (w/H) + (1-w) \sum_{i=1}^n (P_{r,i} + (D_{g,i} + D_{f,i}))$$

where

$R_c$  is the route quality

$P_{r,i}$  is the normalized Link quality of node  $i$

$H$  is the hop count

$D_g$  is the normalized packet generated by node  $i$

$D_f$  is the normalized packet forwarded by node  $i$

$w$  is a constant between 0 and 1. Here  $w=0.5$

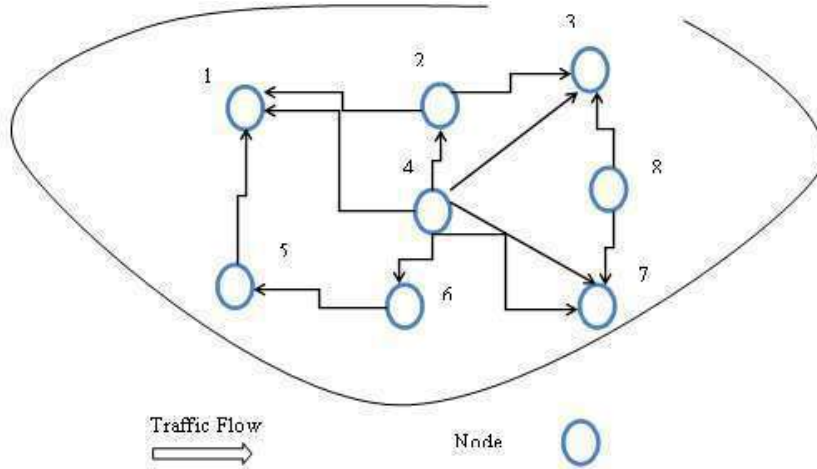


Fig. 2: Network Load

### Route Selection Mechanism

The protocol decides whether particular route should be eliminated at the routing layer instead of making the decision at the link layer and filtering out all routing traffic over the bad links and reducing the chance of selecting a bad route.

In DSR link failure is detected by regularly using hello packets. If the neighbour does not respond to the hello packet sent by a node then the Route Error (RERR) is initiated so that source reinitiates the route discovery process. This leads to high routing overhead and to overcome this disadvantage, the proposed system monitors the link quality and the source node starts the new route detection process before the current link is broke.

### Proposed AFSA - DSR

The proposed LQLB algorithm measures multiple parameters leading to suboptimal solutions. Also the ideal value of the weight cannot be found. To find an optimal solution various swarm intelligence techniques have been proposed in literature. Artificial Fish Swarm Algorithm (AFSA ) mimics the behaviour of fish swarms preying and has been found to converge very fast [30]. The search starts with random solutions mapped to each fish. Fitness is evaluated and the fish follow process is initiated and fish follows another fish with better solution. If the solution is suboptimal, the swarming process is initiated using steps. If the desired solution is not met, preying process is initiated. This process is continued till the desired threshold or termination criteria is met. The pseudocode of AFSA is given in figure 3.

```

Start AFSA
:: Initialize();
while termination criteria not met do
    switch (:: evaluate_AF ())
        case 1:
            :: AF_follow ();
        case 2:
            :: AF_swarm ();
        default :
            :: AF_prey ();
    end switch
    :: move_AF();
    Obtain_best_solution ();
end while
    
```

Fig.3: Pseudo code of Artificial Fish Swarm Algorithm

Artificial Fish (AF) model is represented by preying nature of fish, free move of individual fish, swarming of fishes to find better food and follow behaviour. AF searches the problem space by those behaviours. AFSA uses random search algorithm using variable current AF position, step, visual (visibility domain), try-number (maximum attempts for finding better positions in visual), and crowd factor  $\delta$  ( $0 < \delta < 1$ ) [31]. For our solution space

$AF = \text{Artificial fish} = (n, s_i), i=1,2,3,\dots,n$  the number of objectives

$(x_i, y_i)$  is the node  $i$  position at time  $t$

The distance between two fishes is computed using Euclidean distance

Visual <sub>$i$</sub> : Maximum transmission distance between of node

Step : The maximum step taken is twice the transmission distance

During preying mode the behaviour of fish is given by

$$\text{prey}(X_i) = \begin{cases} x_i + \text{step} \frac{x_j - x_i}{\|x_j - x_i\|} & \text{if } y_j - y_i \\ x_i + (2\text{rand} - 1) \cdot \text{step} & \text{else} \end{cases}$$

Here  $\text{rand}$  is random function with range  $[0,1]$ .

In the swarming phase the behavior of the swarm is given by

$$\text{swarm}(X_i) = \begin{cases} X_i + \text{step} \frac{x_j - x_i}{\|x_j - x_i\|} & \text{if } \frac{y_c}{nf} > \delta y_i \\ \text{prey}(X_i) & \text{else} \end{cases}$$

In the follow phase the behavior is given by

$$\text{follow}(X_i) = \begin{cases} X_i + \text{step} \frac{x_{\max} - x_i}{\|x_{\max} - x_i\|} & \text{if } \frac{y_{\max}}{nf} > \delta y_i \\ \text{prey}(X_i) & \text{else} \end{cases}$$

The three steps mentioned ensures both global and local search and the search direction following the best food source.

## RESULTS

Fifteen runs were carried out for DSR, proposed LQLB-DSR and AFSA-DSR. The communication and network parameters used in the experiment is shown in Table 1.

**Table 1 Summary of Results**



Parameters	Value
Data Rate	Fixed at 2 Mbps
Transmit Power	50 mW
Packet Reception-Power Threshold	-95 dBm
RTS Threshold	None
Path Loss Exponent	3.8
Route time out	4 second
Allowed hello loss	2
Hello interval	Uniform (1,1.15) second

Discrete Event Simulations were carried out for 1200 seconds and the average values computed. Table 2 shows the mean fitness and the best fitness obtained by AFSA-DSR.

Table 2 Fitness across 15 runs with standard deviation

	Mean fitness AFSA-DSR	Best fitness AFSA-DSR
1000 sq m	0.1349± 0.0231	0.1269
2000 sqm	0.1378± 0.0208	0.1157
3000 sqm	0.2073± 0.0316	0.1766

It can be observed from table 1 that the proposed algorithm performs consistently under multiple simulation scenarios proving the stability of the proposed technique. Figure 4 shows the Packet Delivery Ratio (PDR).

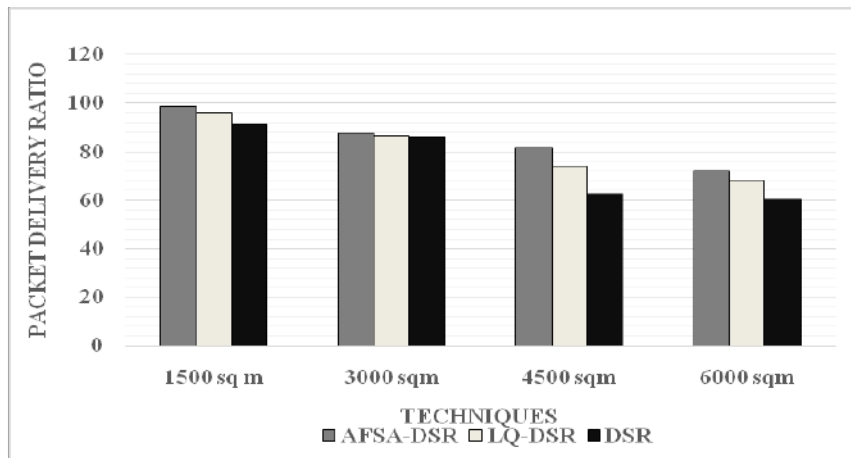


Fig. 4: Packet delivery ratio

From figure 4 it is observed that the average PDR of the proposed AFSA-DSR improves by 4.496 % and 12.453 %, when compared to the LQLB-DSR and traditional DSR routing. Figure 5 shows the end to end delay.

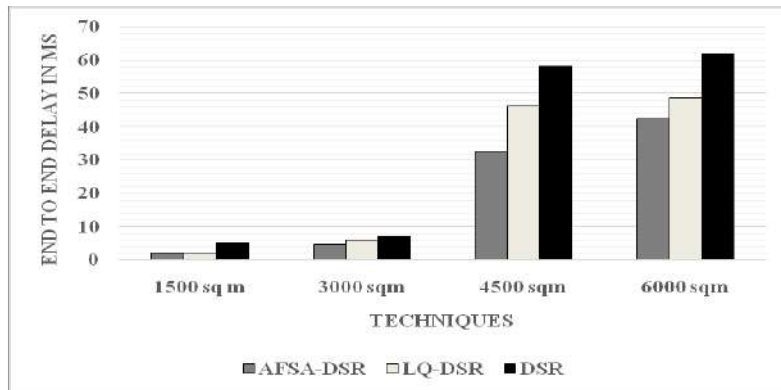


Fig. 5: End to end delay

From figure 5 it is observed that the average end to end delay of the AFSA-DSR reduces by 47.24 % and 23.04 % respectively, when compared to the Traditional DSR and LQLB- DSR routing for the simulated MANET respectively. It can also be seen that LQLB-DSR performance improves substantially over traditional DSR. From figure 6 it is observed that the average number of hops to the destination in the LQLB-DSR increases by 16 % when compared to DSR.

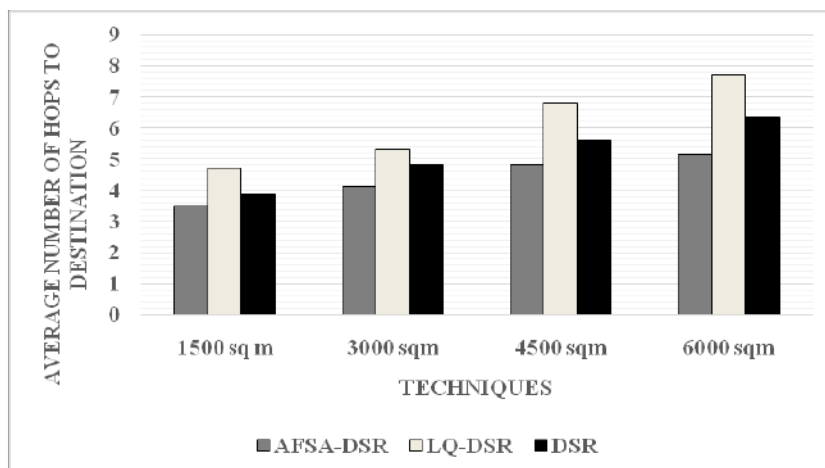


Fig. 6: Number of hops to destination

This is due to the suboptimal solutions generated based on pure computation. This is overcome by AFSA-DSR as it balances not only the network load but also avoids the sub optimal solution. However the number of hops is higher than DSR without affecting the network QoS.

### CONCLUSION

This work investigated performance of DSR and proposed two algorithm LQLB- DSR and AFSA-DSR to improve the Quality of Service adaptively for different application scenarios. Fish Swarm was chosen due to its faster convergence and easier computation. Simulations were conducted by varying the network size and 15 runs were conducted for each scenario. Numerical results show that throughput of AFSA DSR increases PDR by an average of 31% than traditional DSR. End to end delay and retransmission attempts decreased considerably which is statistically significant. The number of hops to destinations increased in both versions of modified DSR compared to traditional DSR which may require further investigation

**CONFLICT OF INTEREST**

The authors declare no conflict of interests.

**ACKNOWLEDGEMENT**

None

**FINANCIAL DISCLOSURE**

None.

**REFERENCES**

- [1] <http://www.comp.brad.ac.uk/~sburuha1/wirelessadhoc.html>
- [2] Bhimarao, L., & Uma, S. Disaster Area Communication Using Neighbor coverage in MANET.
- [3] Du, X., & Zhao, Z. (2010). A Group Key Agree Management Scheme for MANET. *International Journal of Nonlinear Science*, 10(1), 77-81.
- [4] Mario, G. and H. Xiaoyan, 2004. Fisheye State Routing Protocol draft-ietf-manet-fsr-03.txt 55<sup>th</sup> IETF Meeting in Atlanta, GA.2004.
- [5] Samir, R., D. Perkins, C.E Elizabeth and M. Royer,2000. Performance comparison of two on demand routing protocols for ad hoc network. In Proceedings INFOCOM, Tel Aviv, Israel.
- [6] Ashwini K. Pandey and Hiroshi Fujinoki, Study of MANET routing protocols by GlomoSim simulator, *International Journal of Network Management*, Volume 15, Issue 6, Pages: 393 -410,November, 2005.
- [7] David B. Johnson David A. and Maltz Josh Broch. DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks.
- [8] Saleh Ali K.Al-Omari, and Putra Sumari, “ An Overview of Mobile Adhoc Networks for the Existing Protocols and the Applications “, *International Journal of Applications of graph theory in wireless ad hoc networks and sensor networks*, Vol 2 , No 1, March 2010.
- [9] P. Manickam, T. Guru Baskar , M.Girija , and Dr.D.Manimegalai, “ PERFORMANCE COMPARISONS OF ROUTING PROTOCOLS IN MOBILE AD HOC NETWORKS “, *International Journal of Wireless & Mobile Networks (IJWMN)* Vol. 3, No. 1, February 2011.
- [10] Shrestha A and Tekiner F, “ On MANET routing protocols for Mobility and Scalability”, *IEEE International Conference on Parallel and Distributed Computing, Application and Technologies*, pages 451- 456, December 2009.
- [11] Fan Ya Qin, Fan Wen Yong and Wang Lin Zhu, “ OPNET based Network of MANET Routing Protocols- DSR routing Simulation “, *IEEE International Conference on Information Engineering*, Volume 4, August 2010, pp 46-49.
- [12] Asl E. K, Damanafskan. M, Abbaspour and Noorhosseini. M, “ MANETs based Ant Colony Optimization”, *Third International Conference on Modeling and Simulation*, 2009.
- [13] Rajesh Gargi, YogeshChaba, and R.B.Patel, “Improving the Performance of Dynamic Source Routing Protocol by Optimization of Neural Networks “, *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 4, No 3, July 2012.
- [14] Mamoun Hussein ,” A Proposed Route Selection Technique in DSR Routing Protocol for MANET “, *International Journal of Engineering & Technology IJET-IJENS* Vol: 11 No: 02, 2011.
- [15] Arafat S.M. Qaed& T. Devi, “Link Quality, Delay And Energy Aware Routing Protocol (Lqdearp) For Mobile Ad Hoc Networks “, *International Journal of Computer & Communication Technology* ISSN (PRINT): 0975 - 7449, Volume-4, Issue-1, 2013.
- [16] S.PrabhakaraRao, Dr.E.Nagabhooshanam, and S.RameshBabu, “Quality of Service Routing in Mobile Ad hoc Networks Using Node Mobility and Energy Depletion Parameters “, *International Journal of Network Security & Its Applications (IJNSA)*, Vol.5, No.3, May 2013.
- [17] Dr. ShuchitaUpadhayaya And Charu Gandhi “Quality Of Service Routing Inmobile Ad Hoc Networks Using Location And Energy Parameters” *International Journal Of Wireless & Mobile Networks (IJWMN)*, Vol 1, No 2, November 2009,Pp 138 -147.
- [18] Ming Xu, and Guangzhong Liu, “A Multipopulation Firefly Algorithm for Correlated Data Routing in Underwater Wireless Sensor Networks”, *Hindawi Publishing Corporation International Journal of Distributed Sensor Networks* Volume 2013,
- [19] Tanmaya Swain, and Prasant Kumar Pattnaik, “ Effect of Routing Protocols over renovated Congestion Control Mechanisms In Single-Hop Wireless “, *Journal of Theoretical and Applied Information Technology*, Volume 59, Number 1, January 2014.
- [20] Kim, J. C., Kim, D. Y., Jung, S. M., Lee, M. H., Kim, K. S., Lee, C. K., ... &Yoo, S. K. (2009, September). Implementation and performance evaluation of mobile ad hoc network for emergency telemedicine system in disaster areas. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE* (pp. 1663-1666). IEEE.
- [21] Kulla, E., Spaho, E., Xhafa, F., Barolli, L., & Takizawa, M. (2012, November). Using data replication for improving QoS in MANETS. In *Proceedings of the 2012 Seventh International Conference on Broadband, Wireless Computing, Communication and Applications* (pp. 529-533). IEEE Computer Society.
- [22] Lee, J. S., & Yang, S. B. (2010, December). An effective emergency escape system with service-oriented architecture in a double-layered MANET. In *Service-Oriented Computing and Applications (SOCA), 2010 IEEE International Conference on* (pp. 1-8). IEEE.
- [23] Bai, Y., Du, W., Ma, Z., Shen, C., Zhou, Y., & Chen, B. (2010, June). Emergency communication system by heterogeneous wireless networking. In *Wireless Communications, Networking and Information Security*

- (WCNIS), 2010 IEEE International Conference on (pp. 488-492). IEEE.
- [24] Johnson, "The Dynamic Source Routing Protocol (DSR)", RFC4728, Feb 2007.
- [25] Yih-Chun Hu and David B. Johnson "Caching Strategies in On-Demand Routing Protocols for Wireless Ad hoc Networks", ACM 2000.
- [26] Balaji, V., & Duraisamy, V. (2012). Improved Aodv Based On Link Quality Metrics. *International Journal of Advances in Engineering & Technology*, Vol. 5, Issue 1, pp. 269-275.
- [27] Bilandi, N., & Verma, H. K. (2012). Comparative Analysis of Reactive and Proactive Routing Protocols in MANETs Using Throughput, Delay and Network Load. *International Journal of Advanced and Innovative Research*, 1(1).
- [28] Ali, S., & Ali, A. (2009). Performance analysis of AODV, DSR and OLSR in MANET. In *Proceedings on Seventh International Conference on Wireless Systems* (p. 34).
- [29] Tsai, H. M., Wisitpongphan, N., & Tonguz, O. K. (2006, January). Link-quality aware ad hoc on-demand distance vector routing protocol. In *Wireless Pervasive Computing, 2006 1st International Symposium on* (pp. 6-pp). IEEE.
- [30] Romoozi, M., Vahidipour, S. M., & Babaei, H. (2009, December). Improvement of Connectivity in Mobile Ad Hoc Networks by Adding Static Nodes Based on a Realistic Mobility Model. In *Machine Vision, 2009. ICMV'09. Second International Conference on* (pp. 138-142). IEEE.
- [31] Singh, A., & Deep, K. (2015, January). How Improvements in Glowworm Swarm Optimization Can Solve Real-Life Problems. In *Proceedings of Fourth International Conference on Soft Computing for Problem Solving* (pp. 275-287). Springer India.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and checked/reviewed by guest editor.

# A NOVAL METHOD FOR OPTIMIZING MAXIMAL LIFETIME COVERAGE (OMLC) SCHEDULING OF NODES IN WIRELESS SENSOR NETWORKS

P. Kalyanasundaram<sup>1\*</sup> and T. Gnanasekaran<sup>2</sup>

<sup>1</sup>Dept of Electronics and Communication Engineering, Nandha Engineering College, Erode, INDIA

<sup>2</sup>Dept of Information Technology, RMK Engineering College, Chennai, Tamilnadu, INDIA

## ABSTRACT

**Aims:** Wireless Sensor Network (WSN) lifetime depends on nodes lifespan. In remote target coverage applications, random deployment provides high density. Many scheduling algorithms are proposed to improve the performance of network lifetime. **Materials and methods:** Among all, the Greedy algorithm addresses the optimization problems. In Greedy Activity Selection Algorithm the items are sorted in the decreasing order of values based on the finishing time, then scan the sorted list is scanned and the data is collected. Maximal Lifetime Coverage Scheduling (MLCS) trying to cover the target with maximum number of nodes in a schedule based approach. Both the approaches tried to maximize the lifetime but failed in reduction of unwanted data transmission. The proposed work Optimized Maximal Lifetime Coverage Scheduling (OMLCS) addresses the above mentioned problem by two methods. First one, Improved Sleep Scheduling algorithm is employed in which very limited number of nodes covers the target, while others in sleep state. **Results:** This idea reduces the energy consumption of nodes and reduces the redundant information about the target. Second idea depicts the periodic exchange of locally sensed information with neighboring sensors. This idea is utilized mainly for the purpose of sending the information only when the change occurs. The network lifetime is increased significantly. **Conclusion:** The simulation results show that 15% improvement in packet delivery ratio and throughput and 30% of reduction in end-to-end

Published on: 18<sup>th</sup>– August-2016

### KEY WORDS

Wireless Sensor Networks, Lifespan, Maximal Lifetime Coverage Scheduling, Sleep Scheduling, Packet delivery ratio, throughput, End-to-end delay.

\*Corresponding author: Email: [kalyanasundaram.pp@gmail.com](mailto:kalyanasundaram.pp@gmail.com); Tel.: +91 9942128326

## INTRODUCTION

The Wireless Sensor Network (WSN) comprises of a collection of active sensor nodes which are deployed in a well-defined area to collect the information about the physical or environmental weather parameters such as pressure, temperature, humidity, etc., and to transmit the sensed data to a centralized node or server cooperatively. In many cases, a group of targets need to be monitored in the defined geographical area. To achieve the assigned Quality of Service, every target should be covered by at least by one sensor node.

The Coverage plays vital role in a WSN, which determines how fine an area of interest can be monitored or followed by sensors [1]. The coverage is classified in to three types based on what is to be sheltered, namely discrete point coverage, area coverage and barrier coverage. The *Connectivity* is another parameter in WSNs which deals with delivering the sensed data from a source sensor to the sink node (destination node) through radio link. In the transceiver part of a sensor is equipped with different transmission power levels to attain different communication ranges. The maximum permissible power level ensures the maximum communication range. Two sensors are said to be connected whenever both of them are within each other's maximum allowable communication range. Multi-hop communications support a sensor to connect another if they cannot reach the sink node directly. In this way sensors in a WSN acts as repeaters to increase their coverage by relaying the data to other sensors to the remote destination. Therefore, both transmission and reception of data swallow a certain amount of energy. The time stamp between the periods from the time when the sensor network was set up to the time when the WSN cannot confirm the *coverage/ connectivity* requirements is defined by the term *network Lifetime* of WSN. i.e It specifies the time period of WSN which function well without any connectivity or coverage issues. It can be prolonged by scheduling merely a subset of sensors necessary to be active and scheduling remaining subset of sensors to be inactive. Hence the improved lifetime is guaranteed due to condensed idle listening, traffic load and collisions of *Media Access Control* (MAC).

## EXISTING METHOD

### Greedy Activity Selection Algorithm

Greedy is most suitable on optimization problems with the following uniqueness [2]:

1. Greedy-choice property: A global optimum can be achieved by picking a local optimum.
2. Optimal substructure: An optimal elucidation to the problem includes an optimal solution to sub problems.  
The property 2 may make greedy algorithms look like dynamic programming. However, the two methods are quite different.

### An Activity-Selection Problem

Let  $S = \{1, 2, \dots, n\}$  be the set of actions that compete for a resource  $S$ .

Every action  $k$  has its starting time  $S_k$  and ending time  $F_k$  with  $S_k \leq F_k$ , if selected,  $k$  takes place during time  $(S_k)$ .

The resource cannot be shared by two actions simultaneously at any period of time.

The actions  $k$  and  $l$  are compatible if their time periods are disjoint. The activity-selection problem is the setback of selecting the largest set of mutually compatible activities [Figure- 1].

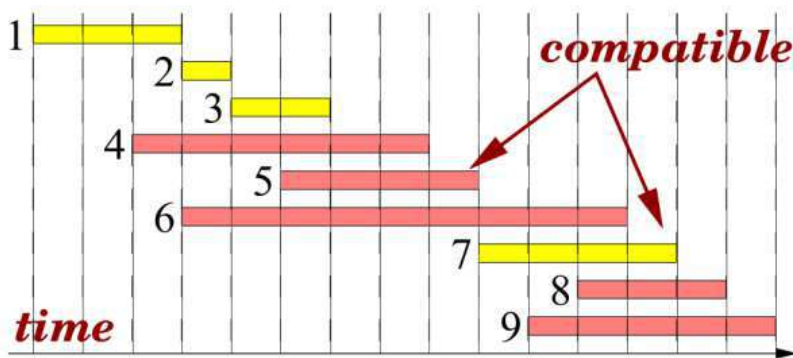


Fig.1 Activity-Selection

### Greedy Activity Selection Algorithm

In this algorithm, based on the finishing time, the activities are first sorted, from the most primitive to the most modern, where a tie can be broken arbitrarily. Then the activities are greedily preferred by referring the list and by selecting.[3]  
The running time of this method depends on the type of sorting algorithm used. The sorting part can be as small as  $O(n \log n)$  and the other part is  $O(n)$ , so the total is  $O(n \log n)$ .

Greedy-Activity-Selector get to the bottom of the activity-selection issues.

### Proof

The proof is by initiation on  $n$ . Initially, let  $n = 1$ . The statement trivially holds.

For the induction step, let  $n \geq 2$ , It is assumed that the claim holds for all values of  $n$  less than the current one.

Let us assume that the action are previously sorted based on their finishing time.

Let  $p$  be the number of activities in each optimal solution for  $[1, \dots, n - 1]$  and let  $q$  be the number for  $[1, \dots, n]$ .

Here  $p \leq q$  holds. It's because every optimal result for  $[1, \dots, n - 1]$  is a elucidation for  $[1, \dots, n]$ .

## PROPOSED METHOD

The above said methods are trying to prolong the lifetime of WSNs. During this process they failed in reduction of redundant data transmission. The lifetime of sensors are wasted in transmitting the repeated information which are collected from the neighboring sensor nodes.

### Maximal Lifetime Coverage Scheduling (MLCS)

All the sensors are powered by built-in batteries. The sensing of a target leads in the reduction of battery lifetime. In this scenario, there is a necessity of considering the reduction of power consumption. In this scenario, there is a necessity of considering the reduction of power consumption by turning OFF the power of the sensors, when they are inactive [4, 5]. Due to the critical issue of

power limitation, a novel method should be devised to prolong the life time of WSN to assume the Quality of Service. Thus the detailed study has been explored in the literature survey.

In target coverage problem in WSNs, the network lifetime is described as the time duration that each and every target point is examined. As depicted out in [1], the lifetime of the network can be extended by alternatively switch ON and OFF the different group of sensors. Actually the entire sensor nodes are organized into various sub set of groups. The scheduling is initiated in such a way that alternatively switch ON and OFF at given span of time. This scheduling is repeated with number of turns to cover all the targets. [Figure- 2] depicts an example. Four target points which are covered with four sensors are taken into consideration. The sensors SN1, SN2, SN3 and SN4 can monitor target points (T1, T4), (T1,T2), (T2,T3) and (T3,T4) respectively. It is indicated in Table 1.

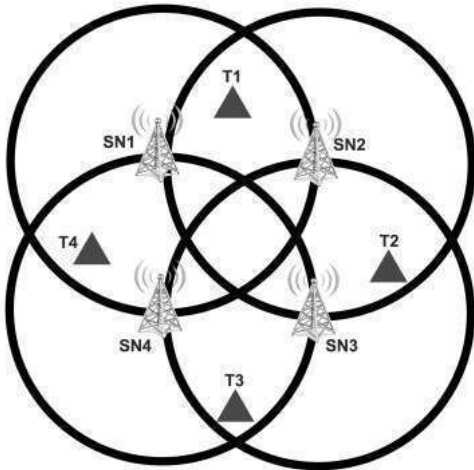


Fig.2 Target coverage in WSNs

Table.1 Coverage of sensors and targets

Sensor	Target Points covered
SN1	T1,T4
SN2	T1,T2
SN3	T2,T3
SN4	T3,T4

If all the sensors are working three time units, then by alternatively switching “on” and “off”, all the target points will be covered in four time units. The suitable schedule would be:

Table.2. Scheduling sensors by incorporating Sleep scheduling

Time units	Sensor Nodes		Target Points Covered
	ON	OFF	
1st Time unit	SN1,SN2,SN3	SN4	T1,T2,T3,T4
2nd Time unit	SN1,SN2,SN4	SN3	T1,T2,T3,T4
3rd Time unit	SN1,SN3,SN4	SN2	T1,T2,T3,T4
4th Time unit	SN2,SN3,SN4	SN1	T1,T2,T3,T4

During the first time unit, the nodes (SN1,SN2,SN3) are turned on and the node SN4 is turned off; in the second time unit, the node (SN1,SN2,SN4) are turned on and the node SN3 is turned off; in the third time unit, the node (SN1,SN3,SN4) are turned on and the node SN2 is turned off; in the fourth time unit, the node (SN2,SN3,SN4) are turned on and the node SN1 is turned off; As per this schedule, all the target points are covered in four time units by running all targets only in three time units.. If the sensor

nodes are not switched, then three time units are sufficient to monitor any target point. (Table.2) Therefore, to prolong the network lifetime, it is mandatory to build up efficient algorithms to schedule the sensors to perform the monitoring tasks [6, 7].

The proposed work Optimized Maximal Lifetime Coverage Scheduling (OMLCS) addresses the above mentioned problem by two methods. First one, Improved Sleep Scheduling algorithm is employed in which very limited number of nodes covers the target, while others in sleep state [1, 3, 8].

The area covered by a sensor is decided by  
Area of coverage

$$\pi r^2 (N) = \frac{(2\pi r^2 + \pi r^2)}{N} \tag{1}$$

Where r – Radius of the sensing field  
N – Number of Nodes deployed in the network.

In the deployment of new nodes some analogy should be followed to ensure the unceasing detection of all nodes in that networks. Random deployment of sensors may desecrate the performance of the entire network.

The distance between two nodes (i & j) are confirmed by  
$$d_{i,j} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{2}$$

where, x & y are coordinates.

With this equation we can find second nodes coordinates  $(x_j, y_j)$ , if we know the distance (may be fixed by us, as critical distance) and first node location. This idea reduces the energy consumption of nodes and reduces the redundant information about the target. Second idea depicts the periodic exchange of locally sensed information with neighboring sensors. This idea is utilized mainly for the purpose of sending the information only when the change occurs.

### Simulation

The Simulation is carried out with a network scenario of a fixed number of targets and sensors randomly deployed around the targets. In this simulation the sensor are considered as equal initial energy without any loss.

Various iterations are carried out on the simulation of Greedy Algorithm, Maximal Lifetime Coverage Scheduling and the proposed Optimized Maximal Lifetime Coverage Scheduling. The observations are recorded and analysis have been carried out. The network parameters average delay [Figure- 3], Packet loss [Figure- 4], Packet delivery ratio [Figure- 5], Control overhead [Figure- 6] and throughput [Figure- 7] are taken into account for analysis.

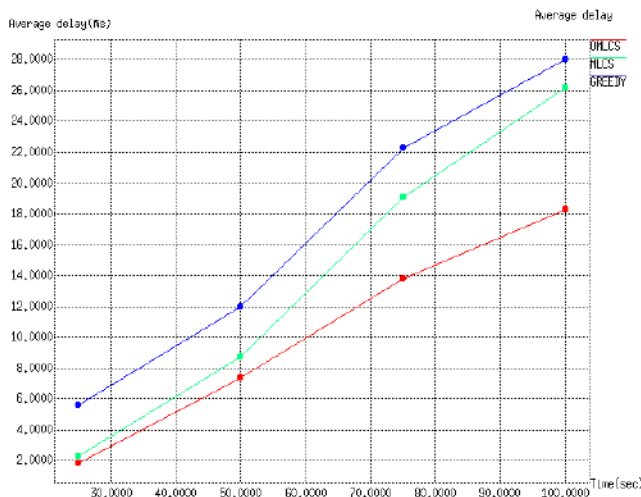


Fig. 3: Analysis of average delay

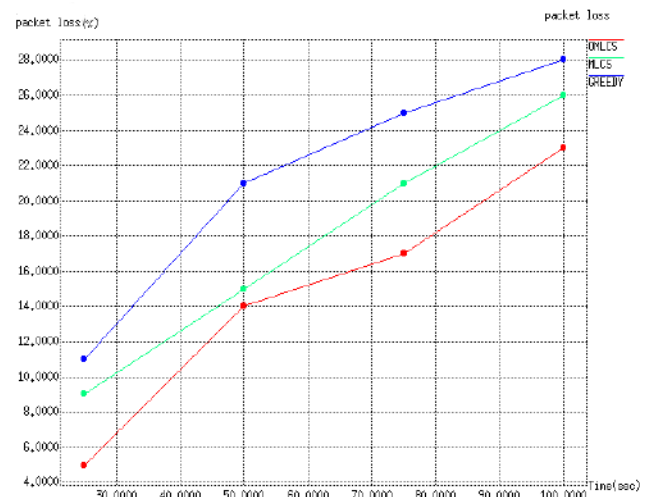


Fig.4: Analysis of Packet loss

## RESULTS

The major objective is to witness the increase in network lifetime as the Maximal Lifetime Coverage Scheduling is done along with optimization. In the optimization process, the periodic exchange of locally sensed information with neighboring sensors. This idea is utilized mainly for the purpose of sending the information only when the change occurs. It reduces redundant data transmission.

From Table-3, It is observed that the average delay encountered in the transmission of data under Optimized Maximal Lifetime Coverage Scheduling has been reduced considerably. The delay plays a vital role in pulling down the life time of sensors.



Table.3 Average Delay in GA, MCLS, OMCLS

Time (Sec)	15	50	75	100
Average Delay (ms)				
Greedy Algorithm	5.8	12	22.2	28
MCLS	2.2	9	19	26.2
OMCLS	1.8	7.5	13.8	18.2

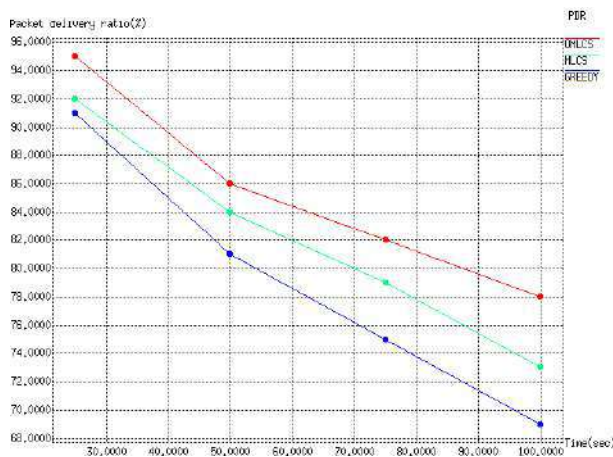


Fig.5: Analysis of Packet Delivery ratio (PDR)

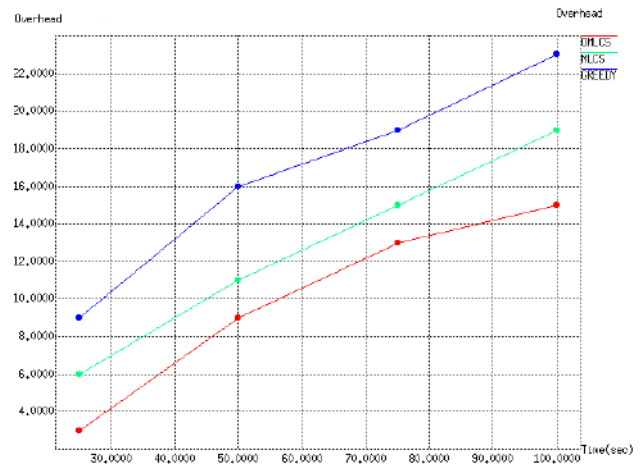


Fig.6: Analysis of Overhead

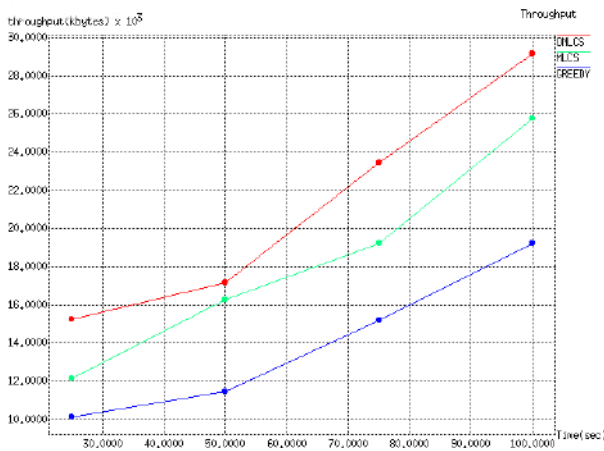


Fig.7: Analysis of Throughput

Table.4 Packet Delivery ratio in GA, MCLS, OMCLS

Time (Sec)	15	50	75	100
Packet Delivery Ratio (%)				
Greedy Algorithm	91	81	75	69
MCLS	92	84	79	73
OMCLS	95	86	82	78

The **Table-4** shows the progress the packet delivery ratio in OMCLS compare to the other schemes like Greedy Algorithm and MCLS.

**Table.5 Packet loss in GA, MCLS, OMCLS**

Time (Sec)	15	50	75	100
Packet Loss (%)				
Greedy Algorithm	11	19	25	28
MCLS	9	15	21	26
OMCLS	5	14	17	21

The **Table-5** is the evidence for the reduction in packet loss drastically in contrast to the other schemes. It ensures the effective packet transmission with minimum loss which in turn indirectly increases the life time of sensors.

**Table.6 Packet loss in GA, MCLS, OMCLS**

Time (Sec)	15	50	75	100
Overhead				
Greedy Algorithm	9	16	19	23
MCLS	6	11	15	19
OMCLS	2	9	13	13

The observation from **Table-6** indicates the diminution of overhead in OMCLS scheme.

**Table.7 Packet loss in GA, MCLS, OMCLS**

Time (Sec)	15	50	75	100
Throughput (Kbytes) X 10 <sup>3</sup>				
Greedy Algorithm	10.2	11.5	15	19
MCLS	12.2	16.5	19	25.8
OMCLS	15.2	17	23.5	29.2

Table.7 witnesses the augmentation in the throughput of Optimized Maximal Lifetime Coverage Scheduling which will escalate the connectivity of I/O devices to the sensors. Within given time span, more amount of data can be shared among the neighboring sensors which will increase the life time of sensors noticeably.

## CONCLUSION

The simulation results show that 15% improvement in packet delivery ratio and throughput and 30% of reduction in end-to-end delay.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] Zaixin Lu, Wei Wayne Li, Miao Pan [2015], Maximum Lifetime Scheduling for Target Coverage and Data Collection in Wireless Sensor Networks, *IEEE Transactions on Vehicular Technology*, 64:714-727.
- [2] Tong Zhao, Qing Zhao [2009], Lifetime Maximization Based on Coverage and Connectivity in Wireless Sensor Networks” *Springer, J Sign Process Syst*, 57:385-400.

- [3] Babacar Diop, Dame Diongue, Ousmane Thiare [2015] Greedy Algorithms for Target Coverage Lifetime Management Problem in Wireless Sensor Networks, *International Journal of Control and Automation* 8: 232-250.
- [4] Mihaela Cardei, My.T.Thai, Yingshu Li, Weili Wu [2005] Energy-efficient Target Coverage in Wireless Sensor Networks, *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies*.4:1976-1984.
- [5] Antonina Tretyakova, Franciszek Seredynski [2013] A Novel Genetic Algorithm with Asexual Reproduction for the Maximum Lifetime Coverage Problem in Wireless Sensor Networks *The Third International Conference on Advanced Communications and Computation* :87-93.
- [6] Qun Zhao, Mohan Gurusamy [2008] Lifetime Maximization for Connected Target Coverage in Wireless Sensor Networks, *IEEE Transactions on Networking* 16:1378-1391.
- [7] Vinay Kumar, Sanjeev Jain, Sudarshan Tiwari [2011] Energy Efficient Clustering Algorithms in Wireless Sensor Networks: A Survey *International Journal of Computer Science Issues* 8:259-268.
- [8] R.Saranya, V.Subathra, S.Mangai, [2014] Energy Efficient Optimized Sleep Scheduling and Target Prediction in Wireless Sensor Network Target Prediction, *International Journal of Modern Trends in Engineering and Research*: 207-214.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ, the article is published as provided by author and approved by guest editor.

# OPTIMIZED LEHE: A MODIFIED DATA GATHERING MODEL FOR WIRELESS SENSOR NETWORK

S. Senthil Arasu<sup>1\*</sup> and N. K. Karthikeyan<sup>2</sup>

<sup>1</sup>Department of Electronics and Communication, Sakthi Polytechnic College, Erode, INDIA

<sup>2</sup>Department of Information Technology, Karpagam College of Engineering, Coimbatore, INDIA

## ABSTRACT

**Aims:** Energy source of Sensor node in a large scale Wireless Sensor Network (WSN) is a non-replaceable and non-rechargeable small power source. Efficient use of the node's energy is the crucial issue in WSN. Lifetime of the network depends upon number of nodes alive to collect and transfer data to the remote base station. Different research proposals are tried to reduce the energy consumed by each node. Those methods are categorized as cluster based, chain based, mobile data gathering based ideas. Cluster based methods are used to subdivide the network in to groups called clusters and elect a node as cluster head whose responsibility is to collect and forward the data from members to base station. **Materials and methods:** Due to heavy usage of nodes in specific area its energy is drained and unable to participate in data forwarding. Those areas look like a hole in the network and that breaks the communication between two ends. **Results:** In proposed work, mobile data gathering method is used to prevent the network break from energy-hole issue and also tried to reduce the probability of formation of energy hole by implementing optimized sleep scheduling algorithm. **Conclusion:** The simulation result shows the improvement in average delay by 10% and packet delivery ratio by 20% due to the dynamic cluster size and multi pattern mobile data gathering.

Published on: 18<sup>th</sup>– August-2016

### KEY WORDS

Wireless Sensor Network,  
Network Lifetime, Mobile data  
Collection, Dynamic pattern.

\*Corresponding author: Email: [senthilarasuece@gmail.com](mailto:senthilarasuece@gmail.com); Tel.: +91 98947 23968

## INTRODUCTION

Smart network applications like home automation, remote area monitoring and agriculture development are implemented based on the concept of Wireless Sensor Network (WSN). WSN collects the sensed data, process it based on application requirement and transmit the aggregated data towards base station in single or multi-hop communication method. WSN consists of large number of small sized sensor elements also called nodes, which may sense more than one physical parameter. The nodes are deployed in unattended environment, so the power source is a non-replaceable and non-rechargeable small unit which is placed inside the node [1-3]. The energy source has to be utilized effectively to extent the lifespan of node as well as the network. Many research proposals are tried to reduce the energy consumption level of each node by different methods. Those methods mostly under the following three categories, namely cluster based, chain based and mobile data gathering. In Cluster based method, the network is subdivided in to groups called clusters. The cluster size may be equal [4] or unequal [5] based on hop count for forwarding the data. The cluster elects a node as cluster head (CH) whose responsibilities are collecting the data from member node and forwarding it towards a base station or sink node. These responsibilities drain the energy of CH node and lead to a quick dead, if the node continuous as CH for long period. Hence the CH role is rotated among all members to avoid early dead issues of selective nodes [6]. Cluster formation and CH elections methods are proposed in many algorithm forms [4-8].

Chain-Cluster based Mixed routing (CCM) protocol [9] divides the network into a number of chains, which means that continuous connection towards base, and it work with two phases. First phase collects data from sensor nodes through their chain head node with the help of improved chain routing protocol. Chain head nodes are grouped in a cluster form with self-organizing capability and elect a cluster head node to transfer the collected data to base in the second phase. It combines both the concept of chain process and cluster method to provide better performance.

Data gathering methods using WSN are applied in object tracing and monitoring applications. Some applications are time sensitive and some applications are data sensitive. Efficiency of the network is based on both these parameters. In Remote monitoring applications the node energy is the important parameter to enhance network lifetime. Most of the energy wasted in transmission of data in the form of communication cost. Mobile data gathering is the idea to overcome the energy wastage in the form of communication cost. A mobile device with more amount of resources (processing, memory, energy...) is used with dynamic path selection to collect data from different cluster head nodes which has the collected information from their member nodes [10].

Energy hole is the hidden area inside the network due to nodes in that area unable to collect or forward the data [11]. It is a critical issue for network lifetime calculation. The energy hole divides the network from base and creates unavailability of the network. Network lifetime is the time duration for three different points based on the application requirements. First Node Died (FND), Half Node Died (HND) and Last Node Died (LND) are those points for calculation of network lifetime. Time duration from initialization of network to death of first node is FND, time duration from initialization of network to death of 50 % of the nodes in the network is HND and time duration from initialization of network to death of last node in the network is LND.

The proposed method tried to overcome the energy-hole issue by considering optimized sleep scheduling and mobile data gathering methods in effective manner to provide better FND, HND and LND. Further the paper organized as literature review in section II, proposed work in section III, and performance analysis in section IV followed by conclusion in section V.

## RELATED WORKS

Clustering method introduced in LEACH (Low-Energy Adaptive Clustering Hierarchy) [4] protocol is the motivation for most of the energy based routing protocol researches. The overall network is subdivided in to equal portions, named it as clusters. Based on nodes density of that area, number of member nodes may vary from cluster to cluster. Cluster members share their availability through messages and use a clustering algorithm to select cluster head (CH) node as their leader. CH node collects the information from member nodes through TDMA slots and forwards the aggregated information to base station. CH role drains the energy of that node and leads to early dead. To overcome this issue, LEACH-C [6] uses the concept of rotating the CH role among all the members based on the threshold values which is shown in equation (1). Network nodes density measured by  $\rho$ , current round CH election is intimated by  $r$  and  $n$  is the node which is winning in the set  $G$  for calculation.

$$T(n) = \begin{cases} \frac{r}{1-r(\rho \bmod \frac{1}{G})} & , n \in G \\ 0 & , \text{Others} \end{cases} \quad (1)$$

Different cluster formation and CH election methods are proposed in [5-8] to improve the performance of the network by considering the sleep scheduling algorithm and node's residual energy. A hybrid model with chain and clustering concepts introduced in [9] utilizes the positive points in both the methods.

A modified Mobile Data Gathering protocol proposed in [10] with load balanced clustering technique and dual data collection methods in a mobile device. The path used in this method is fixed and this method leads quick energy drain of nodes along the fixed path.

Energy-hole inside the network is the challenging issue for lifetime analysis of WSN. It creates unavailability of the network by early death of the nodes [11-14]. The load balancing method is used to handle the energy-hole problem in a large-scale WSNs, and proposed a distributed solution to balance the consumed energy of nodes by tuning their transmission level. A mathematical model given in Lifetime and Energy Hole Evolution (LEHE) [15] analyze the entire network lifetime from network initialization to complete disable of network, and calculates the boundary of energy-hole inside the network. Based on boundary values the transmission power increased to overlap the data from the energy-hole. But this method not tried to prevent the formation of energy hole.

A Hybrid Energy Consumption model proposed in [16] give a way for multi pattern mobile data collection to avoid early death of nodes nearby fixed path used in other methods.

## MATERIALS AND METHODS

### Optimized LEHE

Network lifetime enhancement based on prevention of energy-hole formation consists of different stages. Network initialized with large cluster formation and the cluster uses modified clustering algorithm to elect the CH nodes and Data collection using TDMA slots. High node density leads to energy wastage in the form ideal listening, this issue is addressed by implementing optimized scheduling algorithm in the field. A Multi-pattern dynamic path selection method of mobile data collection idea reduces the energy wastage in routing and also prevents the formation of energy-hole inside the network.

### Cluster Formation

After initialization of network, the clusters are configured with sensor nodes self-configurable property. In Each round of the cluster it elects a CH with modified method [Figure 1]. The role of CH is to collect the data from member node and aggregate the information received. Aggregated information is kept for relay to mobile data collection.

Modified Threshold value calculation is given by equation (2), were  $E_{res(x)}$  is the residual energy of the node calculated in current round and  $E_{max}$  is the initial energy of node and it should be equal to all nodes.

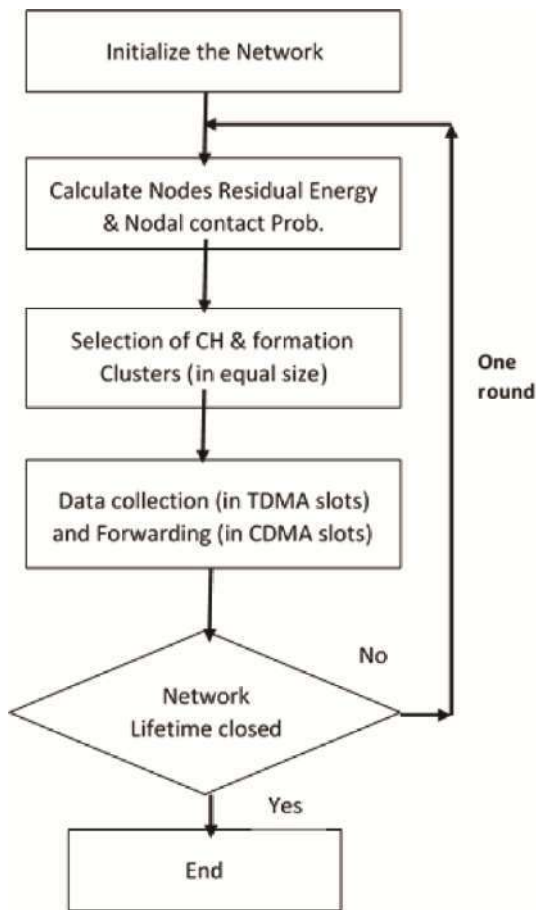


Fig. 1. Modified Cluster Head Election

$$T(x) = \begin{cases} \frac{rE_{res(x)}}{(1-r(p \bmod (\frac{1}{n})))E_{max}} , & x \in G \\ 0 , & \text{Others} \end{cases} \quad (2)$$

After the Calculation of threshold value for node 'x', it send for CH election. Elected CH node send TDMA slot to all member node and the member nodes send data to their respective slot to CH. CH aggregate the data to remove redundant information from received data. It also reduces the transmission cost of unwanted data. After the Collection of data from CH, it checks for the

availability of network and also residual energy of the node. If residual energy of the node is above the maximum reachability power level then it goes for next round of cluster otherwise it is considered as a dead node. The lifetime calculation using three different methods is shown in [Figure -2]. R1,R2,R3...are the cluster rounds and each of the round has clustering and data collection phases. The Network continuously monitor the death of each node by its energy level calculation after each round and it records the time duration for each death from network initialization.

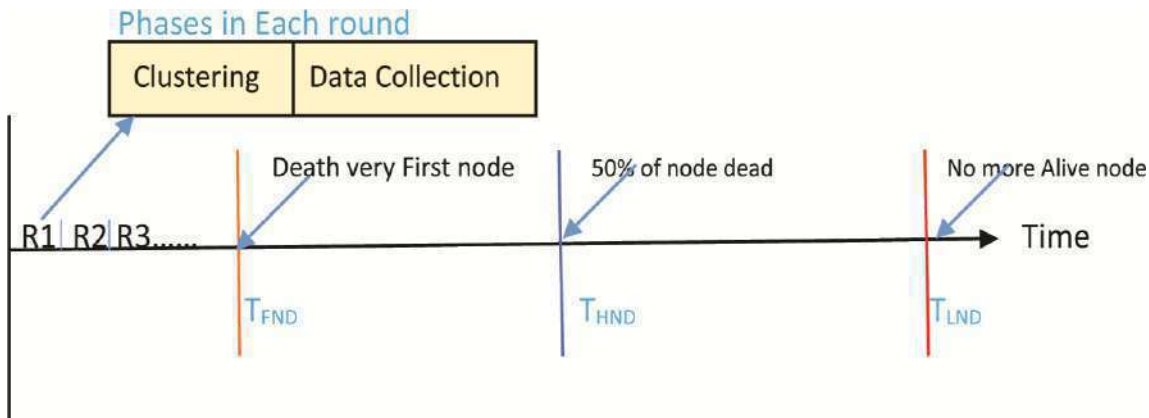


Fig: 2. Network Lifetime Calculation

### Network Model

Network model for the proposed work is shown in [Figure -3]. Each node acted as transmitter and receiver. Number of bits used for transmission is 'n' and the distance between two end points is 'd'. Energy loss due to transmission of n bit data is denoted by  $E_{Tx}$  and same for receiver is  $E_{Rx}$ . The distance threshold  $d_0$ , denotes the minimum distance of air interface with free space channels. If the distance  $d > d_0$ , then multipath fading model channels are used for transmission. These information's are highlighted in the equations (4) & (5), where  $\epsilon_{fs}$  is the energy for amplification in free space channel and  $\epsilon_{mp}$  is the energy for amplification in multipath channel.

$$E_{Tx} = \begin{cases} nE_{Elec} + n\epsilon_{fs}d^2 & ,d \leq d_0 \\ nE_{Elec} + n\epsilon_{mp}d^4 & ,d > d_0 \end{cases} \quad (4)$$

$$E_{Rx} = nE_{Elec} \quad (5)$$

### Mobile data Collection

Routing overheads in cluster based algorithms are overcome by a mobile device called RelayCar [10]. It is a mobile device with maximum sensing, processing and transmitting capabilities with rechargeable energy source. The device does not covered by lifetime issue, because of rechargeable power source. The device passes through the application field in air, collect data from CH node in each round and process it for removal of redundant information and forward it directly to the base with large transmitting range. The path, which the mobile device passes through the field is not a fixed one. Different pattern of paths are stored in the device and based on the requirement or environmental condition, the paths are selected dynamically. The following algorithm describes more in deep about the process of the proposed algorithm.

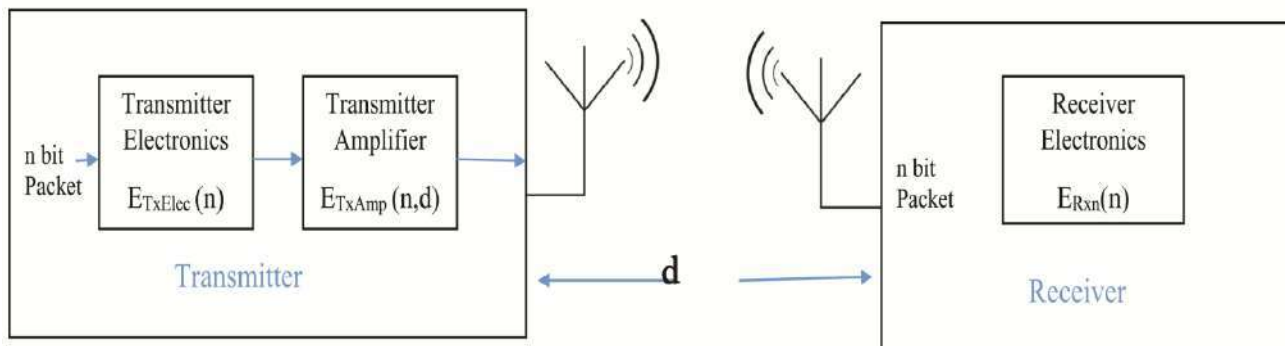


Fig: 3. Network Model

### Algorithm for Optimized LEHE

- 
- Step 1: Identify the group nodes for sleep schedule  
 Step 2: Set 10% duty cycle  
 Step 3: Cluster Head election with updated Residual energy values  
 Step 4: Data Collection through TDMA Slots  
 Step 5: RelayCar applied with Dynamic Mobile pattern  
 Step 6: Check for death nodes and record the time duration  
 Step 7: Repeat the steps 3 to 6, if network available  
 Step 8: End the process, if network unavailable after current round
- 

**Table 1. Simulation parameters**

Parameter	Value
Number of nodes	100
Network Grid	500×500m <sup>2</sup>
Channel BW	1 Mbps
Size of data packet	500 bytes
Initial energy of nodes	1J
Duty Cycle	10%
$E_{fs}$	10pJ/bit/m <sup>2</sup>
$E_{mp}$	0.0013pJ/bit/m <sup>4</sup>
$E_{elec}$	50nJ/bit
$E_{idle}$	0.88 mJ/s

## RESULTS

### Performance Analysis

Performance of the Optimized LEHE protocol is analyzed with the help of simulation tools. The parameters used for analysis is listed in table 1. 50 nodes are considered for analysis with 1J as initial energy for all nodes. 10 % duty cycle is considered for sleep scheduling. Single round time duration for clustering is set as 10s. Optimized LEHE protocol is compared with LEHE protocol and First Node Died Time (FNDDT) for better analysis. Analysis parameters Average Delay, Packet Delivery Ratio (PDR), Control Overhead and Average remaining energy are discussed one by one as below.

### Average Delay

End to end average delay is the time duration in which the collected information reaches the base. Compared with other data gathering methods mobile data gathering reduces the time delay. With increased number of nodes the delay reduces 10% compared with LEHE protocol due to implementation of mobile data collection method [Figure -4].

### Packet Delivery Ratio (PDR)

PDR is the ratio between number packets received at the base to the total number of packets send. After the variations PDR value improved with 20% difference [Figure -5]. The difference is due to considering the proper scheduling algorithm and implementation of mobile data collection method.

### Control Overhead

The ratio between control packets send to the total packets send in the network for specific duration. Control overhead value has to be minimum for a better protocol. The performance is improved by 5-10% when compared with LEHE protocol [Figure -6]. The variation is due to more number of control messages are required when number of nodes are increased.



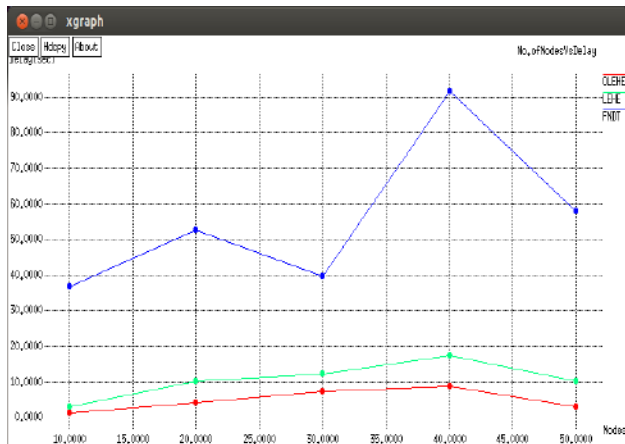


Fig. 4. Average Delay

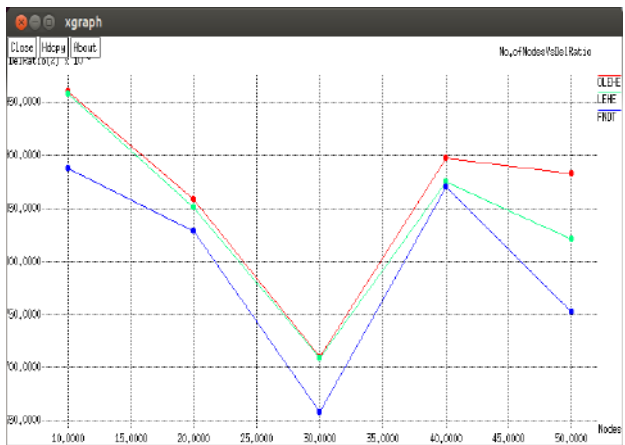


Fig. 5. Packet delivery ratio

### Nodes Remaining Energy Level

The sum of all alive nodes remaining energy is considered for analysis. Up to 40 nodes combination the remaining energy level is comparatively high in LEHE protocol and after that 50 nodes combination, need of more number of control messages and more number of active node conversions the remaining energy level is reduced considerably [Figure -7].

## CONCLUSION

The proposed work focused on the network lifetime enhancement by preventing the formation of energy-hole inside the network. Sleep scheduling and dynamic multipath mobile data gathering methods are used to reduce the energy consumption in each node. Nodes residual energy level consideration for CH election provides better improvement. First node died (FND), Half node died (HND) and Last node died (LND) are considered for analysis of the proposed work. Network lifetime is improved when compared with the other existing methods because of considering average residual energy level and Dynamic path selection for mobile data gathering. Packet delivery ratio and average delay are improved much better than the compared methods.

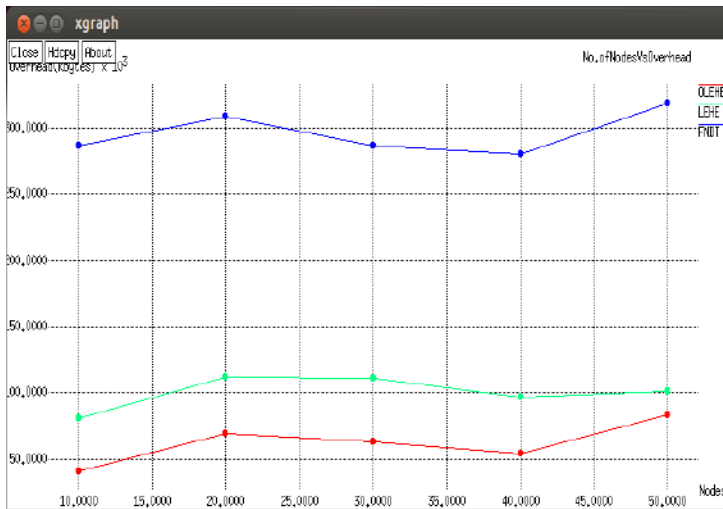


Fig. 6. Control overhead

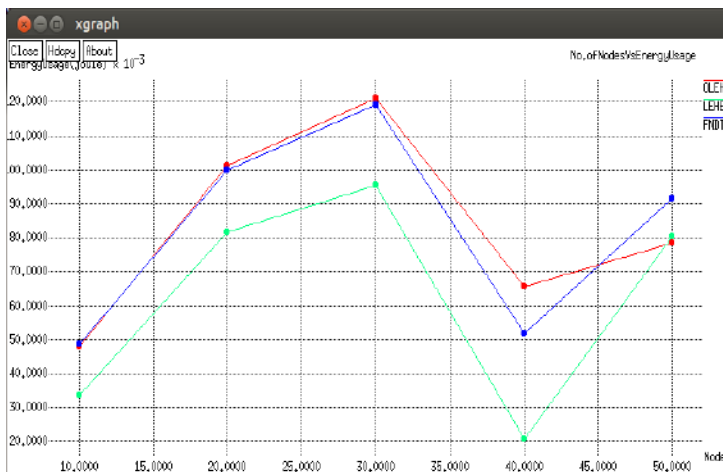


Fig. 7. Nodes Remaining Energy Level

**CONFLICT OF INTEREST**

The authors declare no conflict of interests.

**ACKNOWLEDGEMENT**

None

**FINANCIAL DISCLOSURE**

None.

**REFERENCES**

[1] Y. Tung et al., [2014]. The generic design of a high-traffic advanced metering infrastructure using Zigbee. *IEEE Trans. Ind. Informat.* 10(1): 836–844.

[2] C. Tung et al. [2014]. A mobility enabled inpatient monitoring system using a zigbee medical sensor network. *Sensors*14(2): 2397–2416.  
 [3] C. Caione, D. Brunelli, and L. Benini. [2012]. Distributed compressive sampling for lifetime optimization in dense wireless sensor networks. *IEEE Trans. Ind. Informat.*8(1): 30–40.

- [4] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan [200]. Energy- Efficient Communication Protocol for Wireless Sensor Networks. *Proceedings of the 33th Hawaii International Conference on System Sciences*.
- [5] Ramesh, K., K. Dr.Somasundaram, [2012]. Improved Fair-Zone technique using Mobility Prediction in WSN. *International Journal of Advanced Smart Sensor Network Systems* 2(2): 1-10.
- [6] W.B. Heinzelman, A.P. Chandrakasan and H. Balakrishnan, [2002]. An Application-Specific Protocol Architecture for Wireless Microsensor Networks. *IEEE Transactions on Wireless Communications* 1(4): 660-670.
- [7] K.Ramesh and K.Somasundaram, [2014]. Enhanced Energy Efficient method for WSN to prevent Far-zone. *International Journal on Communications Antenna and Propagation* 4(4):137-142.
- [8] K.Ramesh, S.Saritha and K.Somasundaram, [2016]. Enhancement of Network Lifetime by Improving the LEACH Protocol for Large Scale WSN. *Indian Journal of Science and Technology* 9(16): 1-6.
- [9] F.Tang, I.You, S. Guo, M. Guo and Y. Ma , [2010]A chain-cluster based routing algorithm for wireless sensor networks. *Journal of intelligent manufacturing*: 1-9.
- [10] Miao Zhao, Yuanyuan Yang, [2015]. Mobile Data Gathering with Load Balanced Clustering and Dual Data Uploading in Wireless Sensor Networks. *IEEE Transactions on Mobile Computing* 14(4):770- 785.
- [11] J. Li and P. Mohapatra, [2007]. Analytical modeling and mitigation techniques for the energy hole problem in sensor networks. *Pervasive Mobile Comput.* 3(3): 233–254.
- [12] S. Olariu and I. Stojmenovic,[2006]. Design guidelines for maximizing lifetime and avoiding energy holes in sensor networks with uniform distribution and uniform reporting. *Proc. IEEE Int. Conf. Comput. Commun.*: 1–12.
- [13] M. Perillo, Z. Cheng, and W. Heinzelman [2004]. On the problem of unbalanced load distribution in wireless sensor networks. *Proc. IEEE GlobeCom Workshops*: 74–79.
- [14] R. Kacimi, R. Dhaou, and A. Beylot, [2013]. Load balancing techniques for lifetime maximizing in wireless sensor networks. *Ad Hoc Netw* 11(8): 2172–2186.
- [15] Ju Ren, Yaoxue Zhang, Kuan Zhang, Anfeng Liu, Jianer Chen, and Xuemin,[2016]. Lifetime and Energy Hole Evolution Analysis in Data-Gathering Wireless Sensor Networks. *IEEE Transactions on Industrial Informatics* 12(2).
- [16] S. Senthilarasu and N.K. Karthikeyan., [2016]. Hybrid Energy Consumption model to improve the Wireless Sensor Network Lifetime. *Advances in Natural and Applied Sciences*. 10(3):104-110.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as provided by author and approved by guest editor.

## FLIERMEET: AN EXTENSION TO ONLINE SOCIAL NETWORKING SITE (OSNs)

T. Praveen\*, K. Karthick, M. Thapasya, S. Sai Preethika

Dept of Computer Science and Engineering, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai, Tamilnadu, INDIA

## ABSTRACT

**Aim:** FlierMeet, a mobile crowd sourcing platform for cross space public information reposting, intelligent tagging and pervasive sharing has served an important function for public information in modern society. So the main aim of the proposal is to focuses on distributed public flier information collection, enhance the cross-space transferring, intelligent tagging and pervasive sharing of distributed fliers. **Materials and methods:** First we extract the text from flier images & for that we use optical character recognition (OCR). Then detecting fuzzy words & non-word characters in the extracted text, pre-processing is conducted using regular expressions & dictionary (Medinet) search & then we employ Natural Language Processing(NLP) for category tagging. We also enable pervasive sharing of distributed fliers. **Results:** To obtain manifold views and comments about our system, both data contributors (a total of 10) and new users (20) were invited to participate this study. They were asked to evaluate FlierMeet on its attractiveness (overall impression), prospect (the perspective and acceptability), and simplicity (is it easy to use) on the scale of 1 to 5. Overall, the results indicate that most people were excited about FlierMeet and felt that it was easy to use. **Conclusion:** All the drawbacks are overcome in our proposal by using the popularity of online social networking (OSN) sites such as Facebook, Google Plus, Twitter, etc. OSN provide an important platform for the dissemination of news, ideas, opinions, etc.

Published on: 18<sup>th</sup>– August-2016

## KEY WORDS

Crowdsourcing, Cross space transferring, OSNs, FlierMeet, OCR, NLP, STA algorithm, intelligent tagging

\*Corresponding author: Email: [praveent738@gmail.com](mailto:praveent738@gmail.com); Tel.: +91 9566330222

## INTRODUCTION

Bulletin boards serve an important communication function within communities. They are usually placed in public settings, taking advantage of the movement of people through social spaces (on the streets, near to sport centers, in college campuses, at workplaces/cafes. The paper fliers posted on bulletin boards usually provide a means for people to seek and advertise local businesses, events (e.g., local art shows, gatherings), and services (e.g., bicycle wanted, lost object sought) the use of bulletin boards have proved its significance on community information sharing, socializing, viewpoint advertising, and marketing the usage of Bulletin boards have been part of the fabric of the social space and it presents an informal, nonintrusive, and inexpensive medium for mass communication[1-4]. The popularity of online social networks (OSNs) such as Facebook, Google Plus and Twitter has greatly increased in recent years. OSNs have become important platforms for the dissemination of news, ideas, opinions, etc.

Though the bulletin board has proved useful and significant in our daily lives, it suffers from several drawbacks. It has limited spatial-temporal coverage. For example, fliers on bulletin boards might quickly be covered by new fliers and are mainly visible to passers-by. Fliers on a board are often cluttered and lack order, making it a laboring task for people to identify the information needed. Therefore, there would be benefits to augment paper-based boards with digital techniques to facilitate information sharing and retrieval. There have been several techniques devoted to address this. For example, digital displays lead to a transformation from paper fliers to digital contents, but its deployment and publishing cost is high, thus creating barriers to average content providers. Though barcodes or RFID tags connect paper fliers with the Internet. The link is often directed to proprietary websites and thus making it difficult to have an open and universal platform for public information sharing. Due to their complementary features and merits, in the near future, we envision a co-existence of varied public information exchange mediums (paper fliers, digital contents, barcode-tagged fliers, etc.) and an effective and practical way is, however, to build an overlay above them for public information gathering and sharing. The consensus is that we should bridge the gap between these physical-space objects and their cyberspace counterparts to facilitate public information sharing, i.e., enabling cross-space content transferring and sharing [5].

With the recent surge of sensor-rich (e.g., accelerometer, GPS, camera, etc.) mobile phones and the prevalence of GPS-equipped cars, taxis, and buses, mobile crowd sensing (MCS) However, no existing method focuses on distributed public flier information collection and cross-space sharing. With the help online social networking sites we help the user to find a common place to search, share and retrieve their needs without moving from place to place and without navigating to different websites.

## SCOPE

The main aim of this project is to focuses on distributed public flier information collection, enhance the cross-space transferring, intelligent tagging and pervasive sharing of distributed fliers.

Besides the public information sharing service developed, FlierMeet can nurture numerous novel applications, some of which are listed below.

1. **Community Memory:** The usage of FlierMeet provides a community repository or memory of postings that may be browsed when their public shows are expired. For instance, we can send a reminiscence message to local residents about the popular events in the past year
2. **Peer-Enhanced Marketing:** For commercial-related fliers, we can build an economic model that motivates reposting activities and facilitates ad dissemination. User preferences can be learned from their reposting behaviors, enabling target advertising
3. **Cross-Space Lifelong Activity Management:** Activities in our daily life often follow a generic lifecycle that contains several stages: activity initiation, activity running, and activity completion. The current system focuses on the activity initiation process, which addresses the challenges on activity information advertising. Following the suggestions obtained during the user study, we can also leverage MCS to record the running activity information and share crowdsourced activity summaries after activity completion. Therefore, a doubled offline-online interaction (activity flier reposting, running activity recording and sharing) application will be developed, which allows lifelong crowdsourced activity management
4. **Pervasive Usage:** Since we are using online social networking sites to improve the efficiency, we develop it as a mobile application having its own timeline. Thus all the information needed are obtained in a single pervasive environment, making it easy for the users

## EXISTING SYSTEM

1. **Bulletin Board:** The bulletin board has proved useful and significant in our daily lives. It is a traditional way of sharing information through a board on any public space. All it needs is a pen and sticky note which makes it cheap and easy to use. It doesn't have any structured rules to be followed. It is the most informal of sharing the information [6].  
It suffers from several drawbacks like limited spatial-temporal coverage. For example, fliers on bulletin boards might quickly be covered by new fliers and are mainly visible to passers-by. Fliers on a board are often cluttered and lack order, making it a laboring task for people to identify the information needed.
2. **Digital Displays:** Digital display leads to benefit of augment paper-based boards with digital techniques to facilitate information sharing and retrieval. Digital display leads to a transformation from paper fliers to digital contents. Here the information appears one after the other making it systematic and readable. But this limits the information shown at a point of time. It suffers from the drawback of hiding information from the viewers. There have been several techniques devoted to address this. For example, digital displays lead to a transformation from paper fliers to digital contents, but its deployment and publishing cost is high, thus creating barriers to average content providers [7].
3. **Bar codes/ RFID tags:** The barcodes or RFID tags used to connect paper fliers with the internet connecting to proprietary websites but with a backdrop. Trying to give a more specific & detailed view of the flier we get disconnected from the universal platform for public information sharing. Due to their complementary features & merits, in the near future, we envision a co-existence of gathering & sharing. The consensus is that we should bridge the gap between these physical space objects & cyberspace counterparts [8-10].

## PROBLEM DEFINITION

**Limited spatial-temporal coverage fliers:** The flier boards are fixed to a particular point of place, thus making it reachable only to the passers-by. This takes advantage of the movement of people. Thus people are look out information, a cumbersome process within, instead of the information reaching the user directly.

**Cost is high:** Setting digital displays & using bar codes & RFID tags makes it construction complex which directly influences the price of the system. The alternative used must be efficient enough & at the same time economical in nature.

**Existing Gap between physical and cyberspace:** These systems proposed earlier have a gap that exists between physical space objects & their cyberspace. This decreases efficiency in information sharing to public which directly affects the cross space content transferring and sharing.

**Incentives:** To male FlierMeet a success, we should have numerous participants who are willing to share data. Since, it costs computational & communication resources for data capture & publicizing, how to motivate users to participate is challenging. Interest & reputation can be motivations but further mechanisms are needed. There are three stakeholders in the system: flier publishers, re-posters & publishers want to disseminate their information among a wide range of community & they are willing to pay to be well-informed re-posters.

**User Privacy:** Privacy is a major concern of crowd sensing system, since user locations, references & activity patterns can be revealed. This issue is alleviated in FlierMeet since data is captured in participatory manner & user can control their data.

**Outdated:** The temporal scope of flier vary with some of them are only relevant for a short period while others offer content that has ongoing relevance. Because of the presence of outdated materials, it is sometimes hard to tell what is still relevant and useful.

## PROPOSED SYSTEM

We propose FlierMeet- An extension to online social networking (OSNs) site to provide efficient public information sharing. The FlierMeet system enables cross-space transferring which is used to build the connection between mobile clients & the back end server. Here the user can capture interested fliers & transmit them to back end server. This component is also to support the detection & location of bulletin boards. This also enables intelligent tagging to facilitate public information sharing retrieval which has two implementation phase. First we need to extract the text from flier images & for that we use optical character recognition (OCR). The second steps involve detecting fuzzy words & non-word characters in the extracted text, pre-processing is conducted using regular expressions & dictionary(Medinet) search & then we employ Natural Language Processing(NLP) for category tagging. We also enable pervasive sharing of distributed fliers. We leverage crowd sourced i.e. Taking a picture of the flier, to re-post fliers from physical space to cyberspace. Context sensitive approaches are proposed to group the distributed crowd sourced re-posts & evaluate their quality for data selection. The clustered re-posts groups are further processed to predict their category e.g. sales or recruitment activity, and semantic e.g., is it widely noted, does it meet my preference tags based on the features extracted from crowd-flier interaction. The system can be applied to a variety of application areas, such as public information collection and sharing, targeted advertising, mobile socializing, and so on. Specifically, our work makes the following research contributions: Develops a mobile platform for participatory public information reposting, intelligent tagging and sharing. This system is now extended to online social networking sites where it has its own timeline, posts, shares, viewers, posters, etc. Here we bringing everything in a single environment which makes viewing, sharing & retrieving easier. The posts are grouped in timeline based on the flier grouping done using STA algorithm. This enables to have a structured view.

## SYSTEM ARCHITECTURE

The system architecture of FlierMeet has its major components such as explained below sections.

### Cross Space Transferring

In this module, Cross-space reposting builds the connection between mobile clients and the backend server. Using the application running at the mobile clients, users can capture interesting fliers from bulletin boards and transmit them to the backend server. Since prior knowledge about boards is often not available, this component also supports the detection and location of bulletin boards based on crowd reposting behaviors.

### Flier Grouping

This module clusters fliers with duplicate reposts from different re posters into flier groups. A context-sensitive approach is proposed to improve group performance by repost grouping and high-quality repost selection, using a set of contexts such as spatio-temporal constraints, flier publishing behaviors, reposting behavior associated contexts (e.g., GPS Position) and so on.

### Data Selection

FlierMeet chooses the best view of a flier in a flier group, which can be used for flier grouping and result display. The challenge is to design heuristics that can achieve reliable elimination and leave good candidates.

### Intelligent Tagging

Category tags are categorized based on content analysis, and semantic tags are predicted using the features extracted from crowd-flier interaction.

### User Interface

It displays the extracted information to users in a Multi view manner, e.g., browsing by categories or semantic tags.

### Integration with Social Networks

The application gives the user the option to share the flier detail on social networking sites for wide range of recipient for the information. When the user selects the option of social networks, he/she either selects to share among friends on the site to the public as a whole. Along with the scope of sharing of flier, the user can add user-defined user tags along with the semantic tags and post it on the networking site. The flier posted can be liked, disliked and even commented upon according to the user needs on their own networking site, making the use ease. These updates over the flier can be viewed in the “flier map view” option in the app, where the fliers can be searched location and tag accordingly as per information obtained with details of it.

In our implementation, we have hosted our own social networking site, similar to the widely used ones, from the local server. With suitable number of accounts on the social networking site, we chose all or few among it in friend list for sharing our fliers and getting response through the information shared.

### Pervasive Environment

We are extending the flier sharing and reposting system to online social networking (OSN) site by developing it as a mobile application having its own flier timeline to view the user’s timeline in chronological order. This way the system is used for information sharing in a pervasive environment to access from anywhere and anytime.

## ARCHITECTURE DESCRIPTION

The architecture of flier meet system has three major components:

**CLIENT:** Which consists of a mobile phone and allow operations like user registration, login, capture flier, flier timeline, map view, sharing and reposting. All these involves the client side operations.

**SERVER:** The server side consists of Fliermeet server, grouping and intelligent tagging and a database. Updating, storing and processing requests are carried out in the server side.

**FLIERMEET SERVER:** This is the processing unit of the Fliermeet system which is connected to database and algorithms. This processes the algorithms according to the user needs by retrieving data from the database which is connected two ways. After processing the results are again updated to the database. The database holds updated value from the server.

**DATABASE:** The database of the FlierMeet consists of all data such as the user information, credentials of the user, flier images, locations, groups, dictionary, posts, etc. All these are regularly updated by the operation updated by the operation carried out in the server which is the only input for database.

**ALGORITHM:** It contains the algorithms or steps used to carry out the process demanded. This is done by FlierMeet server, which retrieves necessary input data for the algorithm to lead the desired results. The appropriate algorithms are selected like grouping & intelligent tagging according to the need.

**WEB SERVICES:** Web services acts as the interface for the client & server communication which integrates the client request & server response. This decides where to send which process thus obtaining accurate results as needed.

**WORKING:** The client should first register in the applications which will be evaluated by the web service. If already existing user, then they can simply login using username and password. After a successful login, the user can capture flier which is against sent to web services, where the location, time and grouping is done and gets updated in the database.

The updated data is now posted in the flier timeline for public view. The flier timeline gives the map view of the post by retrieving location information from the database.

Then sharing the post or reposting is done accordingly to the user needs. This is again given to web services which updates the database.

## MODULES

Our project, Fliermeet – an extension to online social networking sites (OSN) is made up of four modules. They are:

- Cross Space Transferring
- Flier Grouping
- Intelligent Tagging
- Flier Sharing. Reposting and Map view

## MODULE DESCRIPTION

### Cross Space Transferring

In this module, Cross-space reposting builds the connection between mobile clients and the backend server. Using the application running at the mobile clients, users can capture interesting fliers from bulletin boards and transmit them to the backend server. Since prior knowledge about boards is often not available, this component also supports the detection and location of bulletin boards based on crowd reposting behaviors.

Crowd-powered cross-space reposting is a novel method of public information collection. In this section we first describe the reposting infrastructure and then present the flier grouping and selection methods.

*Bulletin Boards detection:* In FlierMeet, each repost is associated with a GPS point, which is captured during reposting at the mobile client side. Assuming that at a certain time  $t$ , there are  $n$  reposts in the system and the associated GPS points are

$P = \{p_1; p_2; \dots; p_n\}$ , and the detected board set is

$B = \{b_1; b_2; \dots; b_m\}$ . When a new repost from GPS point

$p_{n+1}$  arrives

- If the distance between  $p_{n+1}$  and a board  $b_j$  ( $0 < j \leq m$ ) is within a given distance threshold (DisThres), the new coordinate of board  $b_j$  will be the midpoint of  $p_{n+1}$  and  $b_j$ ;

- If the distances between  $p_{n+1}$  and every board are all above DisThres, a new board  $b_{m+1}$  is discovered and  $p_{n+1}$  is set to the initial coordinate point of  $b_{m+1}$ .



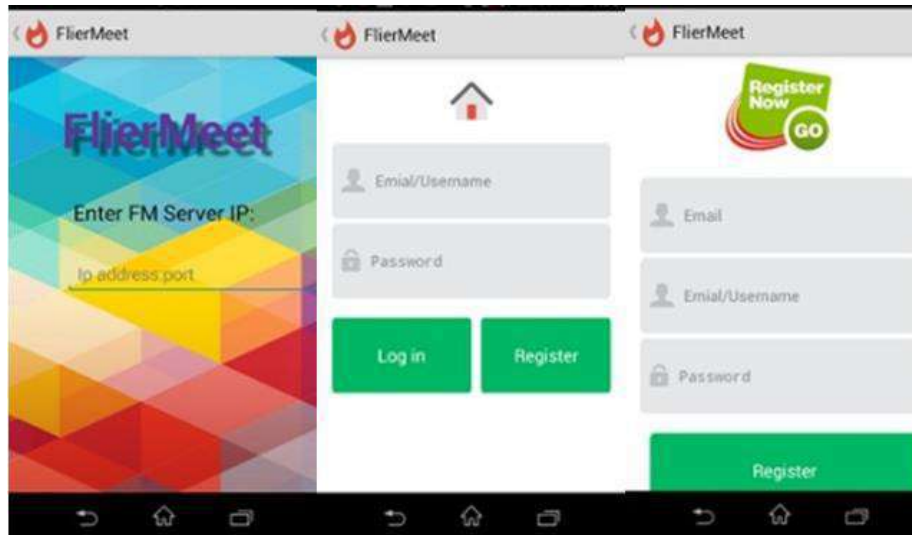


Fig. 1: The FlierMeet application interface for connecting to server from client side

**Flier Grouping**

This module clusters fliers with duplicate reposts from different re posters into flier groups. A context-sensitive approach is proposed to improve group performance by repost grouping and high-quality repost selection, using a set of contexts such as spatiotemporal constraints, flier publishing behaviors, reposting behavior associated contexts (e.g., GPS Position) and so on.

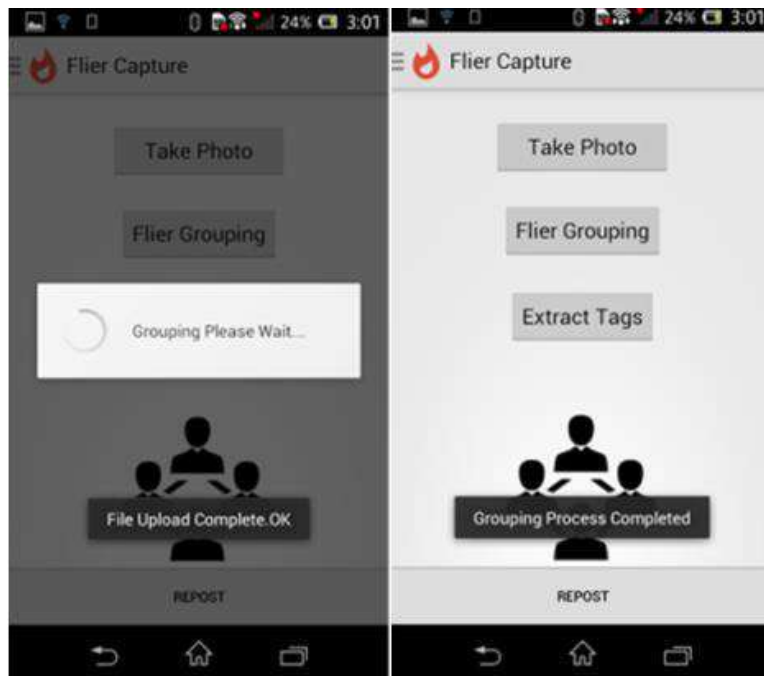


Fig 2.Flier Grouping process shown in the application

*Similarity-Based Flier Grouping:* The aim of flier grouping is to cluster similar reposts from different users. It can be framed as a near-duplicate (similar regions exist in two reposts) image detection problem. Grouping is challenging subject to two issues: (1) the reposts are transmitted to the server individually along the timeline, and thus dynamic clustering or matching should be conducted based on the new coming repost and existing records (in traditional clustering, the dataset is already there and thus it is not executed in a dynamic manner); (2) the

high computational cost on duplicate repost matching (if the new repost and an existing repost in the server is duplicate), especially when the number of reposts is large. To achieve dynamic, resource- conserved flier grouping, we should limit the number of duplicate matches when a new repost comes (i.e., matching with a subset of existing reposts under certain constraints). The spatiotemporal-association (STA) based grouping algorithm is thus proposed.

### Algorithm: The STA Grouping Algorithm

```

Input: New flier f
Output: group id gId for f
1:   bi ¼ GetBoard (f.gps)
2:   F ¼ GetFliers(bi, AgingThres)
    //F is the inner-board flier group set, ordered by flier
    groups' UpdateTime
3:   if MatchIn(F, f ,gId) then
4:     return gId;
5:   else
6:     B ¼ CRC(bi)
    //B is the board set found and ordered by CRC
7:     for each bi in B
8:       F ¼ GetFliers(bi, AgingThres)
        //F is the inter-board flier group set, ordered by
        flier groups' UpdateTime
9:       if MatchIn(F, f ,gId) then
10:        return gId;
11:      end if
12:    end for
13:  end if
14:  gId¼NewGroup(f)
15:  return
  
```

As illustrated in the above algorithm, STA will select the appropriate flier groups to match with. This is a filtering and ordering process based on spatiotemporal constraints and board associations, and  $f_1$  to  $f_3$  are selected accordingly. The selected flier groups are matched in order:  $f_1$  has fliers coming from the same board ( $b_1$ ) as  $N$ , so the match happens first with  $f_1$ ;  $f_2$  and  $f_3$  come from the same associated board ( $b_2$ ), but because the Update Time of  $f_2$  is more recent,  $f_2$  is placed ahead  $f_3$  in matching. Finally,  $f_2$  is matched with  $N$ .

### Spatial Constraint

Spatial constraint is based on the fact that flier information is often related to local contexts (e.g., a school campus, a street block). We partition the city into 150 m\_150 m region cells. When logging into the system, a user selects an interesting region, and only the reposts related to that region are displayed. It can significantly reduce the number of reposts to match with, considering that only the reposts within the same region of a new repost are matched for grouping.

### Temporal Constraint

It is not needed to match a new repost with outdated reposts, but it is usually hard to recognize which of the validation time (e.g., the date of a workshop), most others do not. Nevertheless, it is known that information without attention for a period of time might be outdated or uninteresting. Therefore, in the current study, we introduce an aging factor called AgingThres (e.g., six hours or two days). The last update time of a flier group is defined as Update Time. A match process is triggered only when Eq. meets

$$\text{CurrentTime} - \text{UpdateTime} < \text{AgingThres}$$

### Match by Association

We leverage the implicit association between fliers and boards to optimize the match order and the grouping performance. It is motivated by the fact that flier publishing behaviours often follow certain patterns. For example, due to the social preferences of boards, a certain type of flier is often posted in a similar set of boards within a social community. This implicitly reveals the logical connections among boards. We characterize the board-flier association at two levels

- \_ Inner-board association: Match within the flier group from current board.
- \_ Inter-board association: Match similar fliers from different boards.

### Intelligent Tagging

This module chooses the best view of a flier in a flier group, which can be used for flier grouping and result display. Flier category tagging is implemented in two steps. First, we need to extract the text from flier images. In this study, we use a commercial-grade optical character recognition (OCR) tool to recognize text.

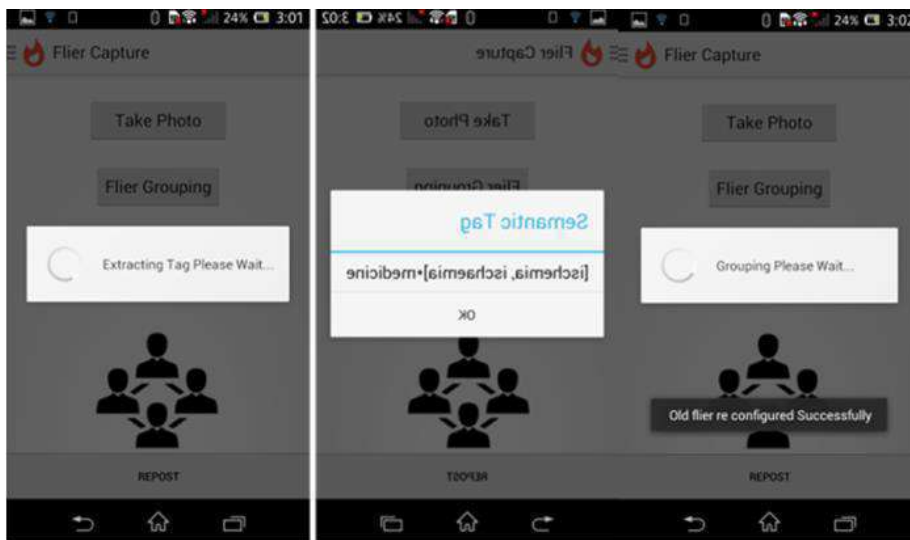


Fig.3 Extracting tags from the flier posted and re-posted

Since there will be fuzzy words and non-word characters in the extracted text, pre-processing is conducted using regular expressions and dictionary(Medinet) search and then we employ Natural Language Processing for category tagging.

### Flier Sharing, Reposting and Map view

In this Module, if the 'Flier timeline' button is pressed, a list of recently published fliers (e.g., the last six hours) will be displayed in a time order. 'Flier map' allows users to view the fliers on the map. Users can choose among different category or semantic tags and browse by 'tag' on the map. If the user clicks a repost on the map, detailed information about that repost, including its re-posters and user comments, will be listed.

## LITERATURE SURVEY

### Mobile Crowd sensing: Current State and Future Challenges

An emerging category of devices at the edge of the Internet are consumer centric mobile sensing and computing devices, such as smartphones, music players, and in-vehicle sensors. These devices will fuel the evolution of the Internet of Things as they feed sensor data to the Internet at a societal scale. In this paper, we will examine a category of applications that we term mobile crowd sensing, where individuals with sensing and computing devices collectively share data and extract information to measure and map phenomena of common interest. We will present a brief overview of existing mobile crowd sensing applications, explain their unique characteristics, illustrate various research challenges and discuss possible solutions. Finally we argue the need for a unified architecture and envision the requirements it must satisfy.

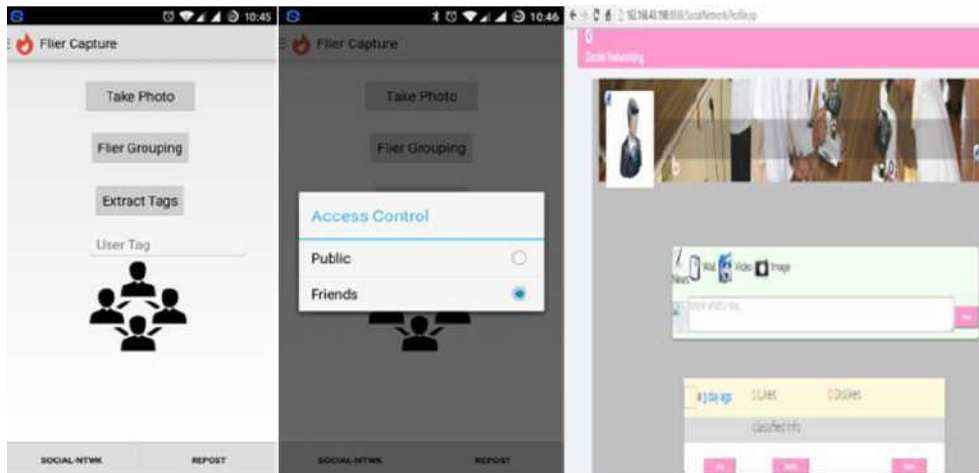


Fig.4. Extension to an OSN with post' update in the Application and OSN

### Supporting Mobile Service Usage through Physical Mobile Interaction

Although mobile services can be used ubiquitously, their employment and the interaction with them are still restricted by the constraints of mobile devices. In order to facilitate and leverage mobile interaction with services, we present a generic framework that combines Semantic Web Service technology and Physical Mobile Interaction. This interaction paradigm uses mobile devices to extract information from augmented physical objects and use it for a more intuitive and convenient invocation of associated services. For that purpose, the presented framework exploits Web Service descriptions for the automatic and dynamic generation of customizable user interfaces that support and facilitate Physical Mobile Interaction. This generic approach to mobile interaction with services through the interaction with physical objects promises to meet the complementary development of the Internet of Things. A user study with a prototype application for mobile ticketing confirms our concept and shows its limits

### Free Market of Crowd sourcing: Incentive Mechanism Design for Mobile Sensing

Off-the-shelf smartphones have boosted large scale participatory sensing applications as they are equipped with various functional sensors, possess powerful computation and communication capabilities, and proliferate at a breathtaking pace. Yet the low participation level of smartphone users due to various resource consumptions, such as time and power, remains a hurdle that prevents the enjoyment brought by sensing applications. Recently, some researchers have done pioneer works in motivating users to contribute their resources by designing incentive mechanisms, which are able to provide certain rewards for participation. However, none of these works considered smartphone users' nature of opportunistically occurring in the area of interest. Specifically, for a general smartphone sensing application, the platform would distribute tasks to each user on her arrival and has to make an immediate decision according to the user's reply. To accommodate this general setting, we design three online incentive mechanisms, named TBA, TOIM and TOIMAD, based on online reverse auction. TBA is designed to pursue platform utility maximization, while TOIM and TOIM-AD achieve the crucial property of truthfulness. All mechanisms possess the desired properties of computational efficiency, individual rationality, and profitability. Besides, they are highly competitive compared to the optimal offline solution. The extensive simulation results reveal the impact of the key parameters and show good approximation to the state-of-the-art offline mechanism.

### Image Quality Assessment: From Error Visibility to Structural Similarity

Objective methods for assessing perceptual image quality have traditionally attempted to quantify the visibility of errors between a distorted image and a reference image using a variety of known properties of the human visual system. Under the assumption that human visual perception is highly adapted for extracting structural information from a scene, we introduce an alternative framework for quality assessment based on the degradation of structural information. As a specific example of this concept, we develop a Structural Similarity Index and demonstrate its promise through a set of intuitive examples, as well as comparison to both subjective ratings and state-of-the-art objective methods on a database of images compressed with JPEG and JPEG2000.1

### The WEKA Data Mining Software: An Update

More than twelve years have elapsed since the first public release of WEKA. In that time, the software has been rewritten entirely from scratch, evolved substantially and now accompanies a text on data mining [35]. These days, WEKA enjoys widespread acceptance in both academia and business, has an active community, and has been downloaded more than 1.4 million times since being placed on Source-Forge in April 2000. This paper provides an introduction to the WEKA workbench, reviews the history of the project, and, in light of the recent 3.6 stable release, briefly discusses what has been added since the last stable version (Weka 3.4) released in 2003.

## EXPERIMENTATION RESULTS

We also made a user study to test the usability of our system. To obtain manifold views and comments about our system, both data contributors (a total of 10) and new users (20) were invited to participate this study. They were asked to evaluate FlierMeet on its attractiveness (overall impression), prospect (the perspective and acceptability), and simplicity (is it easy to use) on the scale of 1 to 5. As shown in Fig. 13, the average scores to the three system properties were 3.8, 3.6, and 3.9, respectively. Overall, the results indicate that most people were excited about FlierMeet and felt that it was easy to use. Only slight differences can be found to the evaluation results from data contributors and new users, where the data contributors felt the simplicity of the system could be improved. We also asked for their comments about the improvement of the system and presented the representative ones below. As a cross-space application, it is important to link online reposts with offline activity participation, such as allowing people to express their willingness to attend the activity and allowing people to crowdsource the important and interesting moments of running activities associated with the repost. By having crowdsourced activity information, a comprehensive characterization of the activity can be obtained and shared. It is useful to develop a ‘view by activity place’ function, because many people are interested about the activity place, while not the flier posting place. This can be achieved by analyzing the text extracted using named-entity recognition techniques.

## CONCLUSION

Thus our project flier meet – an extension to online social networking site has been successfully developed to focus on distributed public flier information collection, enhance the cross-space transferring, intelligent tagging and pervasive sharing of distributed fliers with the popularity of online social networks (OSNs) such as Facebook, Google Plus and Twitter has greatly increased in recent years, OSNs have become important platforms for the dissemination of news, ideas, opinions, etc.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] S. Brand, "What happens after they are built," in *How Buildings Learn*, London, U.K.: Penguin, 1994.
- [2] E. F. Churchill, L. Nelson, and L. Denoue, "Multimedia fliers: Information sharing with digital community bulletin boards," in *Communities and Technologies*, Dordrecht, Netherlands: Springer, 2003, pp. 97–117.
- [3] E. F. Churchill, L. Nelson, L. Denoue, and A. Girgensohn, "The plasma poster network: Posting multimedia content in public places," in *Proc. IFIP Int. Conf. Human-Computer Interaction*, 2003, pp. 599–606.
- [4] A. B. Jacobs, *Great Streets*. Cambridge, MA, USA: MIT Press, 1999.
- [5] J. D. Startt and W. D. Sloan, *The Significance of the Media in American History*. Northport, AL, USA: Vision Press, 1993.
- [6] F. Alt, T. Kubitzka, D. Bial, F. Zaidan, M. Ortel, B. Zurmaar, T. Lewen, A. S. Shirazi, and Albrecht Schmidt., "Digifieds: Insights into deploying digital public notice areas in the wild," in *Proc. 10th Int. Conf. Mobile Ubiquitous Multimedia*, 2011, pp. 165–174.
- [7] F. Alt, N. Memarovic, I. Elhart, D. Bial, A. Schmidt, M. Langheinrich, G. Harboe, E. Huang, and M. P. Scipioni, "Designing shared public display networks—implications from today's paper-based notice areas," in *Proc. 9th Int. Conf. Pervasive Comput.*, 2011, pp. 258–275.
- [8] N. Taylor and K. Cheverst, "Exploring the use of non-digital situated displays in a rural community," in *Proc. OZCHI Workshop Public Situated Displays Support Communities*, 2008, pp. 1–3.
- [9] E. M. Huang, K. Anna, and B. Jan, "Overcoming assumptions and uncovering practices: When does the public really look at public displays?" in *Proc. 6th Int. Conf. Pervasive Comput.*, 2008, pp. 228–243.
- [10] G. Broll, S. Siorpaes, E. Rukzio, M. Paolucci, J. Hamard, M. Wagner, and A. Schmidt, "Supporting mobile service usage through physical mobile interaction," in *Proc. IEEE 5th Annu. Int. Conf. Pervasive Comput. Commun.*, 2007, pp. 262–271.

\*\*DISCLAIMER: This published version is uncorrected proof; plagiarisms and references are not checked by IIOABJ; the article is published as it is provided by author and approved by guest editor.

# IMPLEMENTATION OF KERBEROS BASED AUTHENTICATED KEY EXCHANGE PROTOCOL FOR PARALLEL NETWORK FILE SYSTEMS IN CLOUD

C. Chandravathi<sup>1</sup>, K. Somasundaram<sup>2\*</sup>, Ramesh Kandasamy<sup>3</sup>, J. Velmurugan<sup>1</sup>

<sup>1</sup> Department of Information Technology, Vel Tech High Tech Dr.RR Dr.SK Engineering College, Chennai, TN, INDIA

<sup>2</sup> Department of Computer Science and Engg, Aarupadai Veedu Institute of Technology, Chennai, TN, INDIA

<sup>3</sup> Department of Information Technology, Nandha Engineering College, Erode, TN, INDIA

## ABSTRACT

**Aims:** In today's world, cloud computing is the developing technology. There exists some security risks and difficulties in accessing parallel network files in this type of virtual technology. To overcome this, concurrent access and user authentication is used for the defense purpose. This paper is based on kerberos protocol using visual cryptographic in cloud. Kerberos is one of the most popular authentication protocol used in networks **Materials and methods:** This protocol uses a trusted third party for authentication. Our work also focuses onto parallel Network File System(pNFS) and using kerberos to provide parallel session keys between client web service and cloud web service **Results:** Using visual cryptographic, the image is uploaded as an authentication services which is verified by the server and session key along with username and password is encrypted in the form of an image to reduce the impact of security risks. **Conclusion:** Thus, our proposed scheme will provide the kerberos protocol robust, secure, and escrow-free and provides full forward secrecy.

Published on: 18<sup>th</sup>– August-2016

### KEY WORDS

Cloud Computing, Kerberos,  
Parallel Network File System,  
Visual Cryptography

\*Corresponding author: Email: [soms72@yahoo.com](mailto:soms72@yahoo.com); Tel.: +91 9443467264 Fax: +91-44-2836 0198

## INTRODUCTION

Cloud computing is the transfer of computing services which is done over the internet. Cloud service allows using software and hardware by individuals and business that are managed by third parties at secluded locations. Examples of cloud services include online storage, social networking sites, webmail, and online business applications. When the network connection is available, the cloud computing model allows access to data and computer resources from anywhere. Cloud computing offers a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications. The features of cloud computing include on-demand self-service, network access broadcasting, pooling of resource, elasticity and measured service. Cloud services are typically made available via a private cloud, community cloud, public cloud or hybrid cloud. Cloud services are popular because they can reduce the cost and complexity of operating computers and networks [1]. While there are benefits, there are privacy and security concerns too. Data travelling over the internet and is stored in remote locations. In addition, cloud provides services to multiple customers simultaneously. This may give rise to several possible attacks [2].

Kerberos is an authentication protocol in this key Distribution Center (KDC) issues ticket encrypted with user's password. There are three main components of kerberos protocol:

- 1) Client: Clients are the users which request the service provider for service from the specific application servers.
- 2) Key Distribution Centre (KDC): The KDC provides authentication services and key distribution functionality. It contains user's and service's secret key. It consists of two components:

- a) Authentication server (AS): The AS authenticates the users. If a new user registers with the AS, it provides the user ID and secret password to the user. The database contains the username and corresponding passwords. The AS verifies the user, issues a session key and sends a ticket to the client.
- b) Ticket Granting Service (TGS): The TGS issues a ticket to the user for establishing session with the application server. It provides session key between user and application server. User verifies its ID just once with AS and can contact TGS multiple times to get tickets for different application servers.
- 3) Application server: The application server provides services for the requested user.
- Kerberos authentication process takes place as follows:

Step1: Client requests service by sending its user's ID together with the ID of Ticket Granting Service (TGS) to the Authentication Server (AS).

Step2: AS responds with the ticket that is encrypted with a key derived from user's password. Client decrypts the incoming message and if the password is correct, the ticket is successfully recovered.

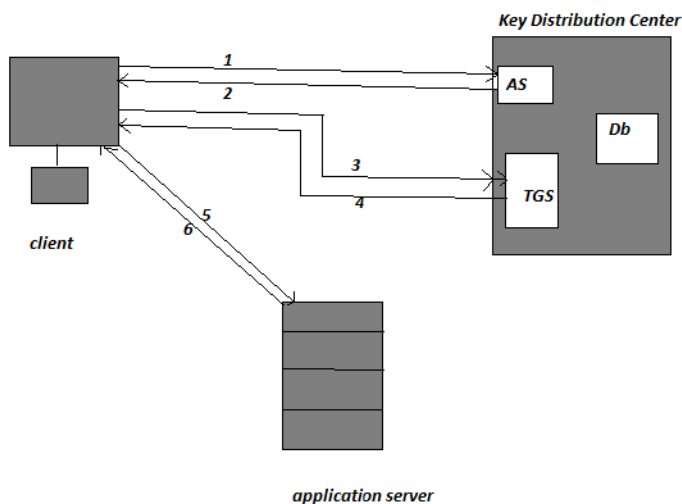


Fig.1. Kerberos architecture

Step3: Client requests Service Granting Ticket (SGT) by transmitting a message to the TGS containing the user's ID, service ID and TGT.

Step4: TGS decrypts incoming ticket and verifies the ID. It checks to make sure that the lifetime has not expired. Then it compares user ID and network address with the incoming information to authenticate the user. If successful, TGS issues a Service Granting Ticket (SGT) by using which user is allowed to access the server.

Step5: After getting SGT from TGS, client sends this ticket along with user ID to the server in order to access a service. The server verifies the ticket and authenticates the user.

Step6: Finally, server opens the conversation with client and perform reverse authentication after verifying user information successfully.

In this work, the problem of many-to-many communications in large scale network file systems (NFSS) that support parallel access to multiple storage devices is investigated. Parallel Network File System (pNFS) is a communication model where there is large number of clients accessing multiple remote and distributed storage devices in parallel. Here, key materials are exchanged and parallel secure sessions between clients and the storage devices in the pNFS are established [3].

### Security issues in cloud

Attacks are the most important problem in cloud services.

- 1) Replay Attack: Getting the data in advance and then replaying it after some time produces unauthorized effect. To avoid replay attack, Kerberos uses timestamp mechanism. This requires synchronization of clocks. When the



user's request is authenticated within stipulated time, the attackers monitor it and replay the information within that time and timestamp mechanism would become waste.

2) Dictionary Attack: The secret key generated from user's password may be vulnerable to dictionary attack, if the password is not strong.

3) Key Storage Problem: As symmetric key algorithm is used for encryption and decryption in kerberos, a secret key need to be shared between clients and KDC, between AS and KDC, between KDC and distant KDC. This makes key management and maintenance, a tedious problem.

4) Malware Attack: The system which is designed to act as KDC may be modelled by the attackers in such a way that it contains built-in listeners. Then, the attackers can fool the users by installing malware. This listens to the users operations including password. Thus, the attackers can directly attack KDC, and masquerade as KDC to complete the man in the middle attack [4].

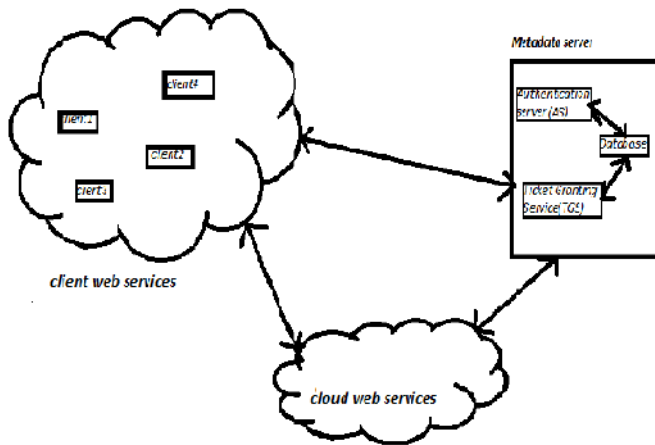


Fig. 2. Concept of parallel Network File System (pNFS)

5) Authentication Forward Problem: In Kerberos 5 a new feature called authentication forwarding is added. When a client is granted access to a server, it can let this server act as a client to apply for a other server. This leads to springboard attack. There is no such problem in Kerberos 4, as it does not support authentication forwarding.

6) Unauthorized Database Access: Kerberos database contains all the user credentials. It must be secured very carefully, otherwise the database may be compromised by the attacker and he/she can easily access the credentials such as usernames and passwords of the clients.

7) Single Point of Failure: KDC must be available continuously for Kerberos protocol. If the KDC is down, the system may undergo single point of failure. This may be solved by using multiple KDC in Kerberos protocol.

8) Clock Synchronization: Use of timestamps generates the problem of clock synchronization. For communication, the system clock of client must be synchronized with the server clock. If the client clocks are not synchronized with the Kerberos server clock, the authentication becomes unsuccessful.

9) Digital Signature: Kerberos verifies only the identities of the user and performs the exchange of keys, but it cannot fulfil the purpose of digital signature. Thus, it cannot provide the undisputable mechanism.

## EXISTING SYSTEM

The existing system will describes that authenticated key exchange protocol for concurrent access network file system this is achieved by three way authentication firstly reducing the workload of metadata server. and secondly providing forward secrecy at last thirdly providing escrow freeness [4]. It approaches to enhance the

performance and scalability of the system and parallel secure session between client and service provider. It provides escrow freeness and overcomes the forward secrecy issue.

Likewise the methodology of preserving the confidential information by image share security with the help visual cryptography whereas it provides high degree of correlation. This paper prevents from phishing attack and also identify whether it is an authentic user. And data security is provided by cloud service provider by implementation of kerberos authentication service. This is done with DES (data encryption standard) algorithm. It ensure the authenticated user to gain access. Basically this system implement the Kerberos authentication service in cloud service provider .therefore existing another service two factor authentication for secure communication one factor as secret share and another one for client private key [5]. It approaches to enhance the security and security attack by combination of visual cryptography and digital envelope in the Kerberos authentication protocol by this mutual authentication achieved therefore it solves the key distribution and clock synchronization issues and improves the efficiency. This is done with AES (advanced encryption standard) algorithm and ECC algorithm. and for parallel network file system implementing the key management in large scale distributed system by establishing the lightweight key management technique. This system introduce file system security architecture (FSSA) for key management problem and for improving the security [6].

## PROPOSED SYSTEM

In our proposed scheme, the main aim is to reduce the workload of metadata server and to provide strong authentication Here, multiple clients web service can access the application server simultaneously. In general, metadata server is used to generate all the service tickets and session keys between client web service and cloud server by placing heavy workload on it. In our solutions, client web service first pre-computes some key materials and forwards them to metadata server and issues the corresponding authentication tokens. It is not necessary that client web service must compute the key materials before each access request. Instead, this is at the done at the beginning of the pre-defined validity period. For each request to access one or more application servers at a specific time, client web service computes a session key from the pre-computed material. Thus, the workload of generating session key by metadata server is reduced.

The modified version of kerberos allows the clients to generate its own session keys. The key material is used to generate session keys. To address key escrow while achieving forward secrecy, visual cryptographic technique is incorporated into kerberos-based pNFS. In visual cryptography, the session key along with username and password and visual cryptographic image for enhancement to security layer. Two shares of images is maintained in both client and server , visual cryptography images to be kept with the client web service and the another one and the original image is to be kept with the KDC. The improved kerberos-based pNFS is as follows:

Step 1: In the first step, the client sends its user ID, sequence number (SN) and its secret shares of image to the AS.

Step 2: At KDC, the AS contains the secret share of image. The secret share sent by the client is stacked onto the secret share by the AS. This generates a computed image. The computed image is compared with the original image present in the database of KDC. If it is equal, the AS generates a Ticket Granting Ticket (TGT). The TGT along with session key, username, password and timestamp forms packet 1. The packet 1 is encrypted using One Time Password (OTP), which is symmetric encryption. This One Time Password is encrypted using the public key of the client. This forms the packet 2. Then, both these packets are sent to the client.

Step 3: After receiving the packets, client decrypts the packet 2 by using its private key. Thus, the One Time Password (OTP) is extracted. Using OTP, the session key and TGT is recovered. The client keeps the session key with itself and sends TGT to the TGS.

Step 4: TGS, present in KDC, verifies the TGT with the help of database. Then, it sends Service Granting Ticket (SGT) to the client, which contains the secret session key used for communication with the cloud service provider.

Step 5: Then, the client sends secret session key (which has been shared between the client and metadata server) to the cloud server.

Step 6: Finally, the cloud server responds the client by sending the acknowledgement for the requested service.

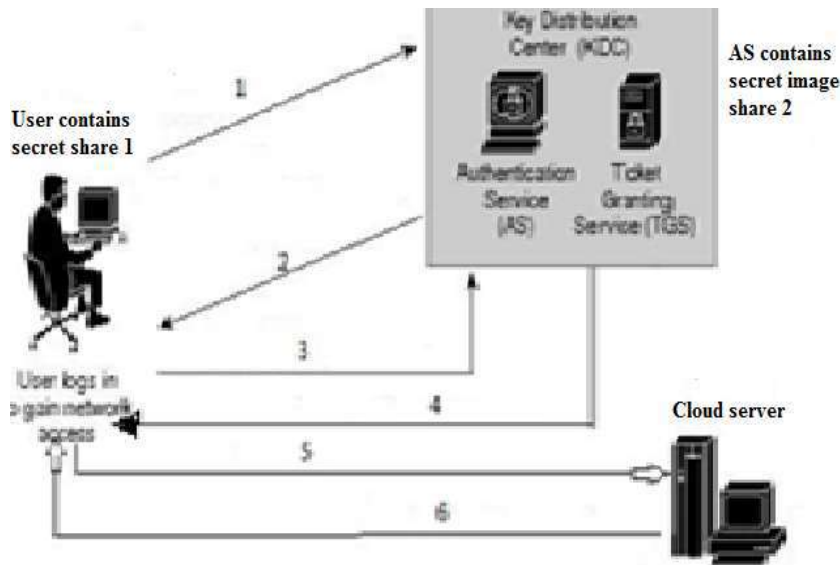


Fig. 3. Kerberos-based pNFS using visual cryptography

The following graph represents the security be achieved by Kerberos in cloud services.

Algorithm:

- In first step client send user details to the server for session establishment and access.
- In second step request will reach to KDC
- Thirdly, TGS in KDC will generate the ticket to ticket will be generate ticket to clients.
- Fourth step client will decrypt and send session key to server for session establishment and access to it.
- Final step cloud server will respond to requested client for access.

## RESULTS

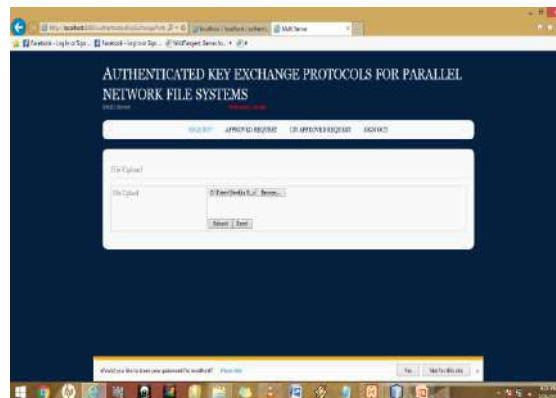
### Key generation and encryption

a) Secret key Authentication: The secret key submitted by the sender to the trusted center (TC), then the TC will verify the secret key and authenticate to the respective sender and gets the session key from TC, else TC doesn't allow the user transmission.

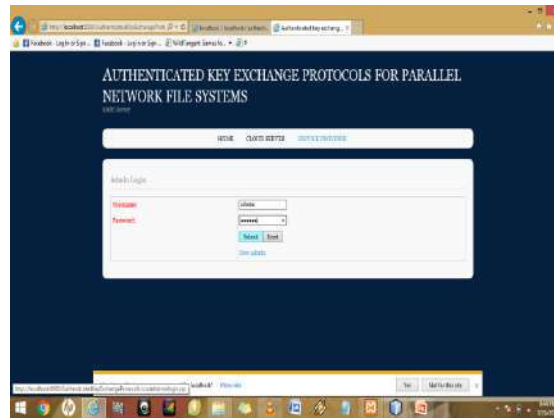
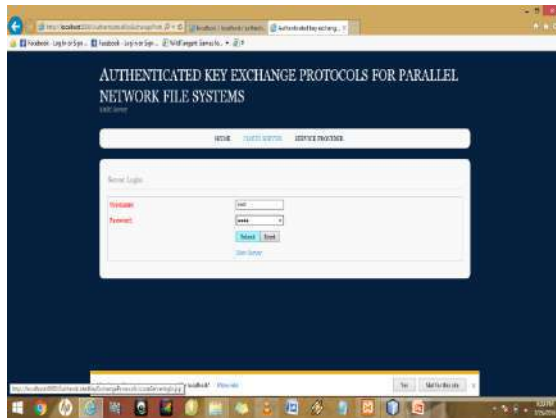
b) Encryption The message is encrypted using received session key and appends the qubit with that encrypted message, then transmits the whole information to the corresponding receiver.



a) Authentication service:



b) Session access:



c) Cloud server access:

d) Cloud server maintenance:

Fig. 4. Key generation and encryption

**Verification and decryption**

- a) Secret key Authentication: It receives the encrypted message with hashed session key and qubit, then verifies the qubit with TC and generates the master key and reverses the hash, the session key and also reverse hash the session key from sender then compare the session key which improve the key authentication.
- b) Decryption then finally decrypt the message using session key and show it to the user.

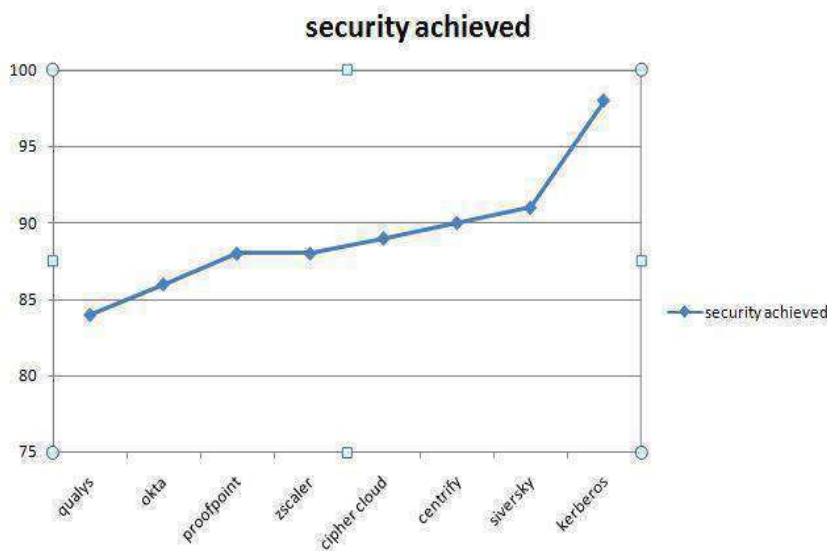


Fig. 5. Comparison – Security Levels

**CONCLUSION**

In this paper, we are incorporating the technique of Visual cryptography and digital envelope into Kerberos-based pNFS protocol. Visual cryptography adds an extra layer of security in Kerberos, which acts as a pre-authentication. Less computation intricacy, high security, decryption process requires no technical knowledge are some of the projecting features of Visual cryptography. For the secure exchange of the session key between the client web service and KDC, we use the idea of Digital Envelope. By using this technique, the high speed advantage of private key algorithm is combined with the key management advantage of public key algorithm.

This solves the problem of key distribution and password guessing attack. Here, we use parallel Network File System concept, where multiple clients can access the cloud server simultaneously. We ruminates this work as an innovative step towards the further improvement of Kerberos authentication protocol.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None.

### REFERENCES

- [1] Hoon Wei Lim and GuominYang,[2015]. Authenticated Key Exchange Protocols for Parallel Network File Systems. *IEEE Transactions on Parallel and Distributed Systems* 27(1): 92-105.
- [2] Mehdi Hojabri and K. venkat rao, [2013]. Innovation in cloud computing: Implementation of Kerberos version5in cloud computing in order to enhance the security issues. *International Conference on Information Communication and Embedded Systems*:452-456.
- [3] Abhishek Thorat, MaheshMore, Ganesh, Thombare, Vikram Takalkar, Manisha N.Galphade, [2015]. An Anti-phishing Framework using Visual Cryptography. *International Journal of Advanced Research in Computer and Communication Engineering* 4(2): 332-334.
- [4] S. Khandelwal, Pariza Kambo, [2015]. Two Factor Authentication Using Visual cryptography and Digital Envelope in Kerberos. *International Conference on Electrical, Electronics, Signals, Communication and Optimization*: 1-6.
- [5] Hoon wei lim . Key management for large scale storage distributed Storage Systems. *SPA Sophia anti polis research, France*.
- [6] Dr.G.Ananda Rao et al., [2011]. Three Party Authentication Key Distributed Protocols Using Implicit and Explicit Quantum Cryptography. *Indian Journal of Computer Science and Engineering* 2(2): 143-145.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# FEATURE SELECTION USING SWARMS IN PARALLEL FOR CLASSIFYING AFFECTIVE AND INFORMATIVE CONTENT

Neeba E. A and S. Koteeswaran

Department of Computer Science Engineering, VelTech Dr. RR & Dr. SR Technical University, INDIA

## ABSTRACT

**Aims:** Electronic media are being utilized in recent time for obtaining medical information and even advice. There is a variety of healthcare information present in the web. For instance, there are blogs on personal experiences of particular illnesses or even discussion forums for patients and even peer-reviewed journals and so on. In the current work, content analyses of healthcare information present in the internet is carried out for obtaining overview on medical content present that makes use of higher-level features delineating medical as well as affective content in blogs. Features selection is the process of selecting relevant attributes on the basis of particular measurements. Optimization of this process is carried out through Particle Swarm Optimization (PSO) as well as Bacterial Foraging Optimization (BFO). The former is an evolutionary computational as well as intelligent swarm method which owes its inspiration to the group activity of flocks of birds or schools of fish. The latter owes its inspiration to the foraging strategies of the bacteria for achieving required variable settings in a successful manner. The bacterium mimicked is the *E. coli* which exhibits chemo taxis, swarming, tumbling, reproduction, elimination as well as dispersal behavior. BFO is generally complex and so much research has been performed for making it simpler as well as obtaining more rapid convergence. The accuracy of classifications relies on the system being created through the usage of historical information which estimates labels of unlabeled instances in an accurate fashion. In the current work, Bacterial Foraging Particle Swarm Optimization (BFPSO) learning protocol is suggested.

Published on: 18<sup>th</sup>– August-2016

### KEY WORDS

Opinion Mining (OM), Particle swarm optimization (PSO), Bacterial Foraging Optimization (BFO) Algorithm and Bacterial Foraging Particle Swarm Optimization (BFPSO)

## INTRODUCTION

Blogs as well as other social media information are increasingly influencing large numbers of people and therefore sophisticated access to the information is required to be given. Because various user groups possess various requisites on the kinds of data queries, search engines are to be capable of enabling patients as well healthcare professionals to discover the appropriate results. The results should also be capable of being filtered with regard to authors (doctors or patients), information kind (affective or informative) or even polarity (negative or positive sentiments) [1].

A vast range of healthcare related data is present in the web. There are several sites with information regarding all kinds of diseases, treatments and even healthcare in general. These kinds of information are provided by various user groups such as doctors, patients, insurance companies and even hospitals [2]. Biomedical research is present in websites like PubMed. Vast quantities of social media technologies also possess healthcare related information. These may be query and answer type, wikis, reviews, encyclopedias and even blogs, which is the primary focus in the current study.

The distinction of affective and informative posts is identical to the issue of subjectivity analysis. The primary variation in the current method and existing techniques for subjectivity analysis is that the proportion of affective to informative content is inefficiently made use of for the purpose of classification and particularly, with the target of medical blogs. Ni et al suggested in [3] a machine learning protocol for the classification of informative as well as affective posts in blogs which is related to the method employed in the current work. Their method varies from the current one in the features made use of: they utilize words as features whereas in the current method, medical concepts as well as polarity are utilized. The focus in the current work comprises medical texts in English which is distinct from blogs.

At the time of classification of documents [4], the quantity of words that are utilized as features are regarded, although merely a few terms in the text denote sentiments in actuality. The additional attributes are to be discarded because they bog down the process of classifying the documents because there are too many words greater than what is required which in turn leads to loss in accuracy because the classifier employed has to take these words into consideration as well. Utilization of lesser number of features is beneficial and so, features selection is employed for removal of non-required features. Features selection refers to the procedure of the archive being run through before classifiers are trained for removing non-required attributes. This permits classifiers to fit models to the problem sets in an expeditious manner because there is lesser data for consideration, resulting in improved accuracy.

The objective of features selection is the simplification as well as reduction in time of the training procedure. Few classifiers like k-Nearest Neighbor perform poorly when the quantity of features is high. Hence, it is of great importance to choose features selection methods that reduce quantity of features with no reduction in the performance of OM. Features selection chooses subsets of the initial features set. Optimality of features subset is assessed through evaluatory criteria. When dimensionality of domains expands, the quantity of features (N) also rises. Discovering optimum features subset is an intractable process and several issues with regard to features selection are proven to be NP-hard [5].

Bacterial Foraging Optimization Algorithm (BFOA) is becoming increasingly popular across disciplines because of its biological motivations as well as elegant architecture. Hybridization of BFOA with several protocols is being explored for examining its local as well as global search characteristics in a separate manner. It has already been employed in several real world issues and has proven its efficacy over several variations of Genetic Algorithm (GA) as well as Particle Swarm Optimization (PSO). Mathematical models, adaptations as well as alterations of the protocol form a significant chunk of the studies into BFOA in the future.

The new Bacterial Foraging Particle Swarm Optimization (BFPSO) learning protocol performs integration of the advantages of the BFO global search capacity as well as the PSO rapid convergence learning machine for minimization of their shortcomings. Population-based BFPSO learning strategy resolves poorly-defined, non-linear, complicated, multi-dimensional optimization issues [6]. Evolutionary BFPSO learning protocols directed by particular fitness functions are an effective technique for acquiring approximate code books in a huge, complicated images space.

In this paper, feature selection based evolutionary BFO-PSO is evaluated. Section 2 shows the literature surveys, section 3 explains the methodologies used in research, section 4 discussed the obtained results and section 5 concludes the work.

## RELATED WORKS

Basari et al [7] focused on binary classifications that classifies into two classes which are positive as well as negative. The former displays good opinion messages while the latter expresses bad opinions messages regarding particular movies. Justification had its basis in the accuracy levels of SVM with the validation procedure utilizing ten-fold cross validation as well as confusion matrices. Hybrid PSO was utilized for improving selection of optimal variable for solving dual optimization issue. Results revealed an enhancement in accuracy levels from 71.87% to 77%. Li et al [8] improved the performance capacity of ABC protocol, with a hybridized ABC (HAB) protocol wherein swarming activity of BFO is brought into ABC for performing local searches. The suggested techniques' performances were studied with the usage of six numerical benchmark functions and the acquired outcomes were contrasted with that of ABC as well as BFO. The outcomes from experiments revealed that the suggested technique was very efficient in resolving numerical benchmark functions apart from providing excellent solution quality as well as convergence to the global optima, specifically on multi-modal functions.

Computational performances are enhanced through usage of basic features selection in almost all studies. Opinion mining involves the identification of the polarity of opinions conveyed on an entity in a particular test. However several OM applications are not viable due to the huge amounts of attributes that occur in the archive. Isabella& Suresh [4] tested a set of features selectors in a systematic manner with regard to their efficacy in the improvement of the performance of classifiers for opinion mining. Reviews of movies are utilized for opinion mining in the particular work. Gupta et al [9] suggested a technique for automated features selection for aspect term extractions as well as sentiments classifications. The suggested method has its basis in the PSO principle and carries out features selection within the learning model of Conditional Random Field (CRF). Experimental

evaluation was carried out on the benchmark setup of SemEval-2014. Aspect-based Opinion Mining Shared Task display F-measure values of 81.91 % as well as 72.42 % for aspect term extractions in laptop as well as restaurant fields, correspondingly. The technique provides classification accuracy of 78.48 % for the latter and 71.25 % for the former.

BFOA is vastly acknowledged as an excellent global optimization protocol which is popular because of its distributed optimization as well as control. BFOA owes its inspiration to the group foraging activity of the *Escherichia coli* bacteria. BFOA is already attracting several experts due to its efficacy in the resolution of real-world optimization issues occurring in various application fields. The biological base underlying the foraging scheme of the *E. coli* bacteria is simulated and is employed as a simplistic optimization protocol. Das et al [10] details the traditional BFOA and then presents an analysis of the dynamics of the simulated chemotaxis stage with the assistance of basic mathematical models. Picking up from the study, it offers a novel adaptive variation of BFOA, wherein chemotactic step size is modified on the fly as per current fitness of virtual bacteria. Analyses on the dynamics of reproductive operators are also detailed apart from hybridization of BFOA with other optimization methods.

## METHODS

Various source of healthcare related information may be found on the web. The current work has its focus on social media tools, specifically, answer portals, wikis, reviews as well as blogs that are popular or are published by huge institutions like the Mayo clinic or the National Library of Medicine.

### Dataset

Mayo Clinic provides excellent care to all patients each day via integrated clinical practices, education as well as research. The Mayo Clinic Model of Care is characterized by excellent quality, medical care provided with compassion in a multi-speciality integrated academic institution. The main objective is the fulfilling of requirements of the patients and this is achieved through the embracing of several core attributes.

Mayo Clinic possesses twelve general blog sites for individuals looking for information or support regarding particular health or medical topics, ranging from Alzheimer's to sexual health. Bloggers publish comments and interact with Mayo Clinic professionals and other users. Mayo has the following blogs: Mayo Clinic Health Policy Center blog ([healthpolicyblog.mayoclinic.org](http://healthpolicyblog.mayoclinic.org)) for news as well as conversations regarding health care reform efforts. A blog companion ([sharing.mayoclinic.org](http://sharing.mayoclinic.org)) to the Sharing Mayo Clinic newsletter for patients as well as the entire Mayo Clinic community for connecting as well as sharing stories and experiences.

### Bacteria Foraging Algorithm (BFO)

Bacteria foraging optimization (BFO) protocol is a novel division of meta-heuristics algorithms. It is a population-based optimization method formulated by the simulation of the foraging activity of *E. coli* bacteria [11]. In real life, locomotion at the time of foraging is attained through sets of tensile flagella. These assist *E. coli* bacteria to perform tumbling or swimming, the two fundamental operations carried out by bacteria for foraging [12]. When the flagella are rotated clockwise, all flagella pull on the cell which leads to movement of flagella in an independent manner and the bacteria move for finding nutrient gradients. Rotation of flagella counter-clockwise enables the bacteria to perform swimming at a rapid rate. In this protocol, bacteria undergo chemotaxis, wherein they favor movement toward nutrient gradients and avoidance of toxic environments. Typically, bacteria travel further distances in friendly environments. BFO imitates the four basic operations present in actual bacterial systems. These are chemotaxis, swarming, reproduction as well as elimination/dispersal for solving the non-gradient optimization issue. The fundamental operations of BFOA are detailed here:

**Chemotaxis:** At the time of foraging, wherein bacteria are to trace, handle as well as ingest nutrients, *E. coli* bacteria travel towards nutrients through the assistance of flagella by either swimming or tumbling. In the former, they travel in a particular direction and in the latter, they alter the direction of searches. The above mentioned two ways of movement are constantly performed during the entire lifetime of the bacteria for moving in arbitrary routes and discovering appropriate amounts of positive nutrients.

**Swarming:** Here, after successfully discovering the direction of optimal food position, bacteria that possess knowledge regarding the best route toward the nutrients try to transmit this information to the others by means of an attraction signal. This signal communications between cells in *E. coli* bacteria is denoted by (1):



$$\begin{aligned}
 J(\theta, D(j, k, l)) &= \sum_{i=1}^N J_{CC}(\theta, D(j, k, l)) = A + B \\
 A &= \sum_{i=1}^N \left[ -d_{attract} \exp \left( -W_{attract} \sum_{m=1}^D (\theta_m - \theta_m^i)^2 \right) \right] \\
 B &= \sum_{i=1}^N \left[ h_{repell} \exp \left( -W_{repell} \sum_{m=1}^D (\theta_m - \theta_m^i)^2 \right) \right]
 \end{aligned} \tag{1}$$

wherein  $\theta$  refers to the location of the global optimal bacterium until  $j$ th chemotactic,  $k$ th reproduction, as well as  $l$ th elimination stage while " $\theta_m$ " refers to the  $m$ th variable of the global optimal bacterium.  $J(\theta, D(j, k, l))$  denotes objective function assessments, "N" refers to the total quantity of bacteria while "D" refers to the total variables to be optimized. The other variables like " $d_{attract}$ " refers to the depth of attracting signals transmitted by a bacterium while " $W_{attract}$ " refers to the width of attracting signals. The signals " $h_{repell}$ " as well as " $W_{repell}$ " refer to the height as well as width of repellent signals between bacteria (wherein attractants refer to signals for nutrients whereas repellents refer to signals for toxic environments).

**Reproduction:** During the procedure of swarming, bacteria form groups in the positive nutrients gradients that lead to increases in bacteria concentrations. Once the groups of bacteria are ranked as per their health value, bacteria with the worst health value die whereas bacteria with greatest health value reproduce and divide into two so as to maintain constant population.

**Elimination-Dispersal:** On the basis of environmental conditions like temperature changes, toxic environments or even presence of food, the population of bacteria might either alter in a steady or abrupt manner. At this phase, set of bacteria in restricted regions (local optimum) will be discarded or the group might be dispersed into novel food locations in the 'D' dimensional search space. Dispersal potentially flattens chemotaxis advancements. Once dispersal is done, bacteria might be situated near excellent food sources and chemotaxis is supported for identification of presence of nutrients. The processes mentioned above are iterated till optimal solutions are attained.

### Feature Selection based Bacteria Foraging Optimization

The extricated features are decreased more through usage of BFO for removal of redundant as well as non-relevant attributes. Resultant features subset is the most representative one. In all dimensions of search spaces, bacteria positions are their 0 or 1, wherein they indicate whether the feature is chosen or not correspondingly as needed features for the subsequent generation. In every iteration of the chemotaxis stage, all bacteria tumble to novel arbitrary positions. Position of  $i$ th bacteria in  $j$ th chemotaxis as well as  $k$ th reproduction stage is given by (2):

$$\Theta^i(j, k) = F_1, F_2, \dots, F_m \tag{2}$$

Wherein  $m$  refers to the length of features vector extricated. Every  $F_z = 1$  or 0 ( $z=1,2,\dots,m$ ) on the basis of whether  $z$ th feature is chosen or not for the subsequent round.

### Particle Swarm Optimization (PSO)

Particle Swarm Optimization [13] is begun with a set of arbitrarily distributed particles designated with certain random velocities. The particles travel in the  $d$ -dimensional problem space, cluster and result in convergence at global optima. The motion of particles in search space is as per the flying experiences of all individuals as well as their neighbors in the swarm population (swarm intelligence (SI)). Assume the  $i$ th particle in the swarm is at positioned at  $x_{id}(t)$  travelling with velocity  $v_{id}(t)$ . Then, position as well as velocity of the particle at subsequent iteration is  $x_{id}(t+1)$  as well as  $v_{id}(t+1)$ , correspondingly, which is represented as(3):

$$\begin{aligned}
 V_{id}(t+1) &= w.V_{id}(t) + c_1.r_1[p_{id}(t) - x_{id}(t)] + c_2.r_2[g_d(t) - x_{id}(t)], \\
 x_{id}(t+1) &= x_{id}(t) + V_{id}(t+1)
 \end{aligned} \tag{3}$$

In the equation above, variable  $w$  refers to inertia constant which maintains a balance between local as well as global search.  $c_1$  as well as  $c_2$  refer to acceleration constants.  $r_1$  as well as  $r_2$  refer to two independently created arbitrary numbers that are uniformly distributed in the interval  $[-1, 1]$ .  $p_{id}(t)$  denotes coordinates of the optimal position found so far by the  $i$ th particle (local optimum), while the coordinates of optimal position found as of yet by the complete swarm (global optimum) is denoted by  $g_d(t)$ .

### Feature Selection based Particle Swarm Optimization

A novel features selection method is suggested by investigation of how PSO [14] may be employed for finding optimum features subset or rough set decreases. Particle Swarm Optimization is certainly beneficial for features selection because particle swarms will find optimal features combination when they travel through the problem space. Particle Swarm Optimization frequently discovers optimum solutions rapidly with such limits. Fitness functions are denoted by (4):

$$Fitness = \alpha * \gamma_R(D) + \beta * \frac{|C| - |R|}{|C|} \quad (4)$$

Wherein  $\gamma_R(D)$  refers to the classification quality of condition feature set  $R$  related to decision  $D$ ,  $|R|$  refers to the '1' number of a position or length of chosen features subset.  $|C|$  refers to the total quantity of features.  $\alpha$  as well as  $\beta$  refer to two variables relating to the importance of classification quality as well as subset length,  $\alpha \in [0, 1]$  as well as  $\beta = 1 - \alpha$ .

### Bacterial Foraging-Particle Swarm Optimization (BFPSO) in parallel

In this kind of hybrid combination, PSO carried out global searches and yields almost perfectly optimum solutions in a rapid manner after which follows a local search through BFO that fine-tunes solutions and provides optimal solutions of excellent accuracy. PSO possesses a basic shortcoming of being forced into local optima however it has excellent convergence speeds while BFOA possesses the shortcoming of low convergence speeds but the advantage of not being forced into local optima.

After a certain set of complete swims, resultant solutions are stored in descending order. In the current method, after chemotactic steps are completed, all bacteria further get mutated by a Particle Swarm Optimization [15] operator. In this phase, all bacteria are stochastically attracted toward gbest positions and local searches in various regions are handled by BFOA.

The primary aim of BFPSO features selection phase is the reduction of features of the issue prior to supervised NN classification. In all the wrapper protocols utilized, BFPSO resolves optimization issues through usage of evolution techniques and has proven to be an excellent one.

The stages for PSO-BFOA comprise:

1. Population is initialized and this is common to both PSO as well as BFOA.
2. The protocols of PSO as well as BFOA are run in parallel.
3. Optimal solution is acquired amongst PSO as well as BFOA.

### Classification Algorithm

#### Naïve Bayes (NB)

Naïve Bayes [16] is a popular probabilistic classifier and was built for incorporating unlabeled data. The job of learning of generative models is the estimation of variables through usage of labeled training data solely. The predicted variables are utilized by the protocol for classifying novel documents through the calculation of which class the specified document is a part of. Naïve Bayesian classifier functions thus:

Let there be a training set of instances with class label  $T$ .  $k$  classes  $C_1, C_2, \dots, C_k$  are present. All samples comprise  $n$ -dimensional vectors  $X = \{x_1, x_2, \dots, x_n\}$ , denoting  $n$  assessed values of  $n$  features,  $A_1, A_2, \dots, A_n$  correspondingly.

Classifiers sort the provided sample  $X$  so that it is part of the class possessing the greatest posterior probability. This means that  $X$  is estimated to be a part of the class  $C_i$  if and only if

$$P(C_i | X) > P(C_j | X) \text{ for } 1 \leq j \leq m, j \neq i \quad (5)$$

Hence class that makes maximum  $P(C_i | X)$  is discovered. Maximal value of  $P(C_j | X)$  for class  $C_i$  is known as the maximal posterior hypothesis. Bayes' theorem states:

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)} \quad (6)$$

- Solely the value of  $P(X | c_i)P(c_i)$  is to be made maximum because for every class, value of  $P(X)$  is equal. If priori probabilities,  $P(C_i)$  of the class are unknown, then it is presumed that all classes are probably equal, i.e.  $P(C_1) = P(C_2) = \dots = P(C_k)$ , and will hence maximize  $P(X | C_i)$ . Else, value of  $P(X | c_i)P(c_i)$  is made maximum. Priors probabilities of a class are predicted by (7):

$$P(C_i) = \text{freq}(C_i, T) / |T| \quad (7)$$

For computing  $P(X | C_i)$ , a great deal of computational cost is required because the provided datasets comprise of various features. For reducing computations when evaluating  $P(X | c_i)P(c_i)$ , conditional class independence of naïve assumptions is made. Values of class label features of the provided instance are assumed to not be conditionally dependent on each other. This is represented by (8):

$$P(X | C) \approx \prod_{k=1}^n P(x_k | C_i) \quad (8)$$

### K-Nearest Neighbor (KNN)

K-Nearest Neighbor [17] classifier for patterns recognition as well as classification wherein particular test tuples are contrasted with set of training tuples which are almost identical. kNN protocol is a very simple technique for resolving classification issues. It frequently provides competitive outcomes and possesses several considerable benefits over many other data mining techniques. Offering more rapid as well as accurate recommendations to the user with favored qualities as an outcome of direct application of similitude or distances for the purposes of classification, kNN is considered extremely effective as well as dependable for understanding customer behavior as well as trends regarding a specific event or entity.

## RESULTS

Table 1 to 3 shows the classification accuracy, precision and recall respectively. Figure 2 to 4 shows the result graph for classification accuracy, precision and recall respectively.

**Table 1 Classification Accuracy**

	EBFO	PSO	PSO-EBFO
Naïve Bayes Classifier	0.8547	0.8705	0.9032
KNN	0.8495	0.8558	0.8989

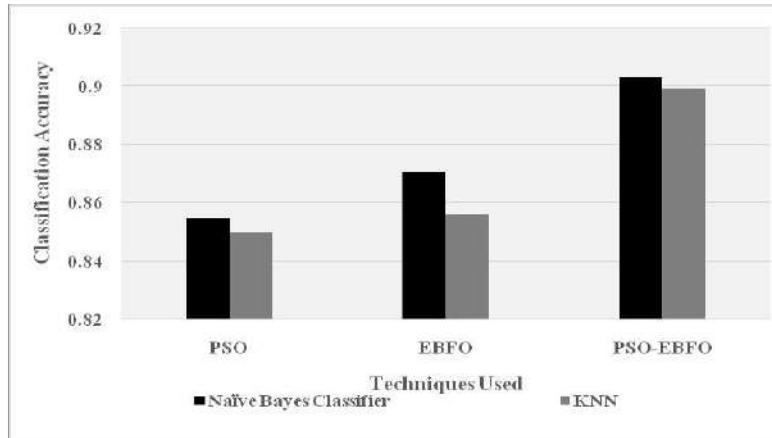


Fig. 2. Classification Accuracy

Table 1 and figure 2 shows the classification accuracy of Naïve Bayes performs better than KNN. Results shows that the accuracy of Naïve Bayes with PSO-EBFO performs better by 10.79% than Naïve Bayes with PSO and by 3.69% than Naïve Bayes with EBFO. Similarly the accuracy of KNN with PSO-EBFO performs better by 6.02% than KNN with PSO and by 5.46% than KNN with EBFO.

Table 2 Precision

	EBFO	PSO	PSO-EBFO
Naïve Bayes Classifier	0.815833	0.859933	0.8943
KNN	0.8366	0.845533	0.890533

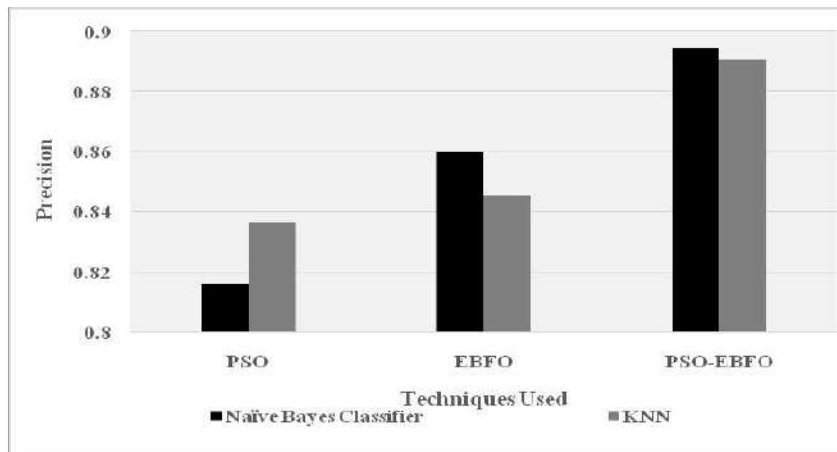


Fig. 3. Precision

Table 2 and figure 3 shows the classification accuracy of Naïve Bayes performs better than KNN. Results shows that the accuracy of Naïve Bayes with PSO-EBFO performs better by 9.18% than Naïve Bayes with PSO and by 3.92% than Naïve Bayes with EBFO. Similarly the accuracy of KNN with PSO-EBFO performs better by 6.25% than KNN with PSO and by 5.18% than KNN with EBFO.

Table 3 Recall

	EBFO	PSO	PSO-EBFO
Naïve Bayes Classifier	0.805	0.8644	0.8969
KNN	0.840467	0.845233	0.892633

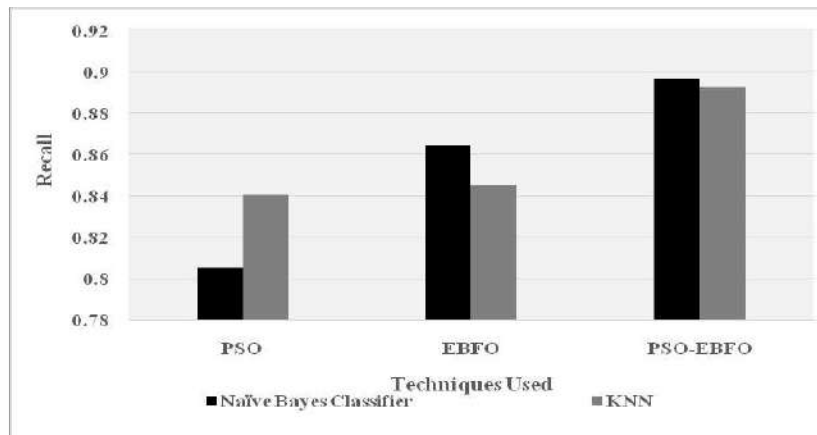


Fig. 4. Recall

Table 3 and figure 4 shows the classification accuracy of Naïve Bayes performs better than KNN. Results shows that the accuracy of Naïve Bayes with PSO-EBFO performs better by 5.52% than Naïve Bayes with PSO and by 3.69% than Naïve Bayes with EBFO. Similarly the accuracy of KNN with PSO-EBFO performs better by 5.65% than KNN with PSO and by 4.91% than KNN with EBFO.

## CONCLUSION

Automatic tracking of attitudes, sentiments as well as opinions on online forums, blogs as well as news sites is a favored tool for supporting statistical analyses by organizations and even private users. In the current work, a new method for classification of affective as well as informative posts in medical datasets is suggested. A novel BF oriented by PSO optimization protocol is suggested. The protocol joins PSO as well as BFO for exploiting PSO's capacity for exchanging social information as well as BF's capacity for discovering novel solutions through eliminations/dispersals. For experiments, classifiers such as Naïve Bayes and k nearest neighbor is used. PSO-EBFO performs better than PSO and EBFO. Experimental result shows that the classification accuracy of Naïve Bayes performs better than KNN. Results shows that the accuracy of Naïve Bayes with PSO-EBFO performs better by 10.79% than Naïve Bayes with PSO and by 3.69% than Naïve Bayes with EBFO. Similarly the accuracy of KNN with PSO-EBFO performs better by 6.02% than KNN with PSO and by 5.46% than KNN with EBFO. Also the precision and recall for proposed PSO-EBFO performs in a better way than PSO and EBFO techniques.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] Denecke, K., & Nejd, W. (2009). How valuable is medical social media data? Content analysis of the medical web. *Information Sciences*, 179(12), 1870-1880.
- [2] Goeriot, L., Na, J. C., Min Kyaing, W. Y., Khoo, C., Chang, Y. K., Theng, Y. L., & Kim, J. J. (2012, January). Sentiment lexicons for health-related opinion mining. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 219-226). ACM.
- [3] X. Ni, G.R. Xue, Y. Yu, Q. Yang, Exploring in the Weblog Space by Detecting Informative and Affective Articles, in: *WWW'07: Proceedings of the 16<sup>th</sup> International Conference on World Wide Web*, 2007, pp. 181-190.
- [4] Isabella, J., & Suresh, R. (2012). Analysis and evaluation of Feature selectors in opinion mining. *Indian Journal of Computer Science and Engineering (IJCSE)*, 3(6).

- [5] A.L. Blum and R.L. Rivest. Training a 3-node neural networks is NP-complete. *Neural Networks*, 5:117 – 127, 1992.
- [6] Korani, W. M., Dorrah, H. T., & Emara, H. M. (2009, December). Bacterial foraging oriented by particle swarm optimization strategy for PID tuning. In *Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on* (pp. 445-450). IEEE.
- [7] Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53, 453-462.
- [8] Li, L., Zhang, F. F., Liu, C., & Niu, B. (2015, June). A hybrid Artificial Bee Colony algorithm with bacterial foraging optimization. In *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on* (pp. 127-132). IEEE.
- [9] Gupta, D. K., Reddy, K. S., & Ekbal, A. (2015). PSO-ASent: Feature Selection Using Particle Swarm Optimization for Aspect Based Sentiment Analysis. In *Natural Language Processing and Information Systems* (pp. 220-233). Springer International Publishing.
- [10] Das, S., Biswas, A., Dasgupta, S., & Abraham, A. (2009). Bacterial foraging optimization algorithm: theoretical foundations, analysis, and applications. In *Foundations of Computational Intelligence Volume 3* (pp. 23-55). Springer Berlin Heidelberg.
- [11] Passino, K. M. (2002). Biomimicry of bacterial foraging for distributed optimization and control. *Control Systems, IEEE*, 22(3), 52-67.
- [12] Rajinikanth, V., & Latha, K. (2012). Controller parameter optimization for nonlinear systems using enhanced bacteria foraging algorithm. *Applied Computational Intelligence and Soft Computing*, 2012, 22.
- [13] Patnaik, S. S., & Panda, A. K. (2012). Particle swarm optimization and bacterial foraging optimization techniques for optimal current harmonic mitigation by employing active power filter. *Applied Computational Intelligence and Soft Computing*, 2012, 1.
- [14] Wang, X., Yang, J., Teng, X., Xia, W., & Jensen, R. (2007). Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters*, 28(4), 459-471.
- [15] Kora, P., & Kalva, S. R. (2015). Hybrid bacterial foraging and particle swarm optimization for detecting Bundle Branch Block. *SpringerPlus*, 4(1), 1-19.
- [16] Buche, A., Chandak, D., & Zadgaonkar, A. (2013). Opinion mining and analysis: a survey. *arXiv preprint arXiv:1307.3336*.
- [17] Sharma, V., & Gonnade, M. S. (2015). A Survey on Recommendation System Based on K-Nearest Neighbor Algorithm and Sentiment Analysis.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as it is provided by author and approved by guest editor.

# AN IMPROVED CLUSTER HEAD SELECTION TECHNIQUE FOR WIRELESS SENSOR NETWORK USING MODIFIED GENETIC ALGORITHM

E. Akila<sup>1</sup> and Baby Deepa<sup>2</sup>

<sup>1</sup>Bharathiar University, Coimbatore and Valluvar College of Science and Management, Karur, INDIA

<sup>2</sup>Government Arts College, Karur, INDIA

## ABSTRACT

**Aims:** Wireless Sensor Networks (WSN), a popular medium of low cost infrastructure communication is slowly emerging as an emergent form of wireless technology among the various classes of communication networks such as Cellular Networks, Adhoc Networks and Mesh Networks. Clustering, a classification process in which nodes are divided into categories via a set of partitioned subset of data, which are commonly known as clusters. A set of predefined categories become a part of the clusterhead and through wireless clustering algorithms, Low-Energy Adaptive Hierarchy (hereafter abbreviated as LEACH) is a popular. It involves a set of cluster heads, which are selected as predefined criteria. In the clustering routing algorithms for wireless networks, Low-Energy Adaptive Clustering Hierarchy (LEACH) a well-known hierarchical routing protocol applied in clustered WSN. The above protocol segregates wireless sensory networks into numerous clusters, and sensory nodes within the same cluster where nodes are capable of direct communication. Cluster head selection based on QoS is NP hard. In this work a novel clustering technique using a modified genetic algorithm is proposed. Extensive simulation show the performance improvement of the proposed technique

Published on: 18<sup>th</sup>– August-2016

### KEY WORDS

Wireless Sensor Networks (WSN), Clustering, Low-Energy Adaptive Clustering Hierarchy (LEACH) and genetic algorithm.

\*Corresponding author: Email: [akilasivam10@gmail.com](mailto:akilasivam10@gmail.com)

## INTRODUCTION

A network containing numerous sensory nodes and every sensory node involve the potential to process, transmit and sense information gathered from the environment, and such networks are otherwise known as Wireless Sensor Networks or WSNs [1]. Sensory nodes which are employed into the environment have limited capacity in energy, memory and other resources as they are often assisted with batteries. Sensory nodes transfer knowledge collected from base stations or gateways and pass them on to other nodes. Clustering is an efficient and scalable energy management technique which is commonly used for large amounts of wireless sensor nodes and it involves sensory organizations which are divided into groups or clusters. Usually among such clusters, work is divided among all the present nodes and each cluster has a central node which is also known as the Cluster head. The main duty of Cluster heads is to ensure the maintenance of information affiliated to each cluster and node. Also, these heads filter and compress data proposed to be transmitted, apart from the mere collection of data. The freshly compressed data is transferred to other nodes and cluster base stations through gateways or any associated Cluster Head.

Election of cluster heads occurs at the nodal level wherein all nodes of a cluster are involved in this overhead process and during which more energy is consumed by sensory nodes. After Election processes, it is difficult to revitalize sensory nodes and researchers have proposed a variety of schemes to evaluate the limitations of such nodes in terms of battery life, energy levels and memory. But researchers also have to consider multiple parameters to select a sensor node as a cluster head [2]. Parameters like residual energy, location, battery and localized distance are considered to be important.

Hierarchical Routing Protocols are the focus of research among wireless sensor networks. For wireless sensor network protocols, the focus of the research is the hierarchical routing protocol. One of the common hierarchical routing protocols theorized by researchers, namely the LEACH, can prolong network lifetime by 15% compared with the ability of flat routing. Routing Protocol Researchers have mainly focused on improving LEACH protocols either at home or the workspace. This consists of improving selecting cluster heads, creating clusters and transmitting data. Computational difficulties in the CH marks a stark problem in the routing process, however these can be handled through efficient heuristic algorithms which are popularly employed and quickly cover local optimum area. Recent studies have seen an increase in genetic computations and algorithms with LEACH protocols. Crossover mutations are commonly performed where the chromosomes are mutated and used for election purposes.

## RELATED WORKS

Numerous techniques to improve WSN lifetimes were introduced by Desai & Rana [3]. They found clustering to be a strong enough approach to be used for linking hierarchies in a network. The efficiency in gathering aggregate data to enhance lifelines of networks is the main goal of clustering algorithms and in the proposed CH algorithm choices are made using node and nodal energy distances. The process is carried out in a way to ensure using approximate distancing between nodes and nodal energy points. Data is then transferred from the CH of every nodal point and all the data is sent to the CH located closest to the NS and the aggregated data is transmitted to the Base Station (BS).

Reviews on WSN protocols were conducted by Khan et al [4] especially in the fields of "Low Energy Adaptive Clustering Hierarchy" (LEACH), "Power-Efficient Gathering in Sensor Information System" (PEGASIS) and "Threshold Sensitive Energy Efficient Sensor Network" (TEEN). By accessing "Hop Counts" and other performance metrics protocols like the Expected Transmission cost and time and "Energy Consumption" levels were consecutively analyzed to protocols of general analysis methods. After the above processes, Khan et al gives us a comparative study among three Wireless Sensor Network's Protocols, namely, LEACH, PEGASIS and TEEN.

WSN has been a common focus among actual users and researchers alike and it is an important task. The energy utilization in these wireless sensor networks is very important task that increases the lifetime of the sensor network. In WSNs the researchers had explored numerous new protocols by considering the energy utilization as crucial task. There is a prime importance to give preference to hierarchical routing protocols based on scalability, even though there might be multiple WSN protocols available. Battery-powered sensory nodes have to assist in reducing energy consumptions in order to increase lifelines of networks. LEACH is the most commonly used sensory network protocol and is also used as a reference for other protocols. Various LEACH based protocols were accessed by Singh [5].

GP-Leach and HS-Leach algorithms proposed by Karimi et al [6] helped improve energy consumption levels and optimized cluster head selection systems with WSN nodes positioning and residual energy of partitioned systems. The results gained from simulations show how the proposed algorithm has an efficient and increased lifetime network.

A modern combination method proposed by Barekatin et al [7] with K-means and improved GA based energy consumption patterns helped to improve Gas and extend the lifetime of networks. Under the above method energy consumption is reduced by finding optimum number of CH nodes via Genetic Algorithms (GA). K-means-based algorithms dynamically cluster networks to balance energy distributions. Simulations in NS-2 portray how the proposed algorithm has a longer lifetime network than popular formulas like GAEEP, GABEEC AND LEACH protocols.

A combinatory EA clustering process was suggested by Martínez-Estudillo et al [8] which assisted in evolutionary design based local-search procedures especially in product-units neural networks. Only a few individuals are subjected to local optimization methods in the methodology presented. It should also be noted that local optimization algorithm can only be applied to specific evolutionary stage processes. The proposed results witnessed a favorable performance under the regression method as compared to other standardized methods.



An efficient clustering method was postulated by Gupta et al [9] which helped in the formation of CHs and assists in sending data to BS and the role of CH modified in every rotation. The final CH is then chosen on the basis of energy distribution and optimum selection procedure via GA. This approach ensures stable operating periods through stable results when it is compared with the probabilistic EC algorithm.

Traditional protocols were reviewed by Dongare & Mangrulkar [10] where energy efficient methods fostered the improvement of appropriate cluster head approaches. Selected formulas chose residual sensory energy clusters via the understanding of optimized cluster heads for proceeding rounds of cluster head operations. Following this equation ensures the survival of the whole network and improves the holistic performance of wireless sensory networks, especially in reducing latent WSN and bandwidth consumption and lifelines of sensory nodes. The distribution of energy balance among all the nodes increases the round number by which point the first node becomes extinct after the reduction of energy holes within WSNs.

A modern scheme is provided by Maraiya et al [11] which relates to data aggregation clustering and is also known as "Efficient cluster head selection scheme for data aggregation in wireless sensor network" (ECHSSDA). It is comparable to the proposed LEACH clustering formula and differences can be seen in terms of energy consumption especially in cluster formations and heads. This suggests that the above scheme is predicted to be better than LEACH especially in the case of consuming less energy among cluster node and head sending data to the base station which consume less energy than LEACH programming.

## METHODS

### Low-Energy Adaptive Clustering hierarchy (LEACH)

LEACH (Low-Energy Adaptive Clustering hierarchy) [12], a self-organized and adaptive clustering protocol adopts randomization which is chosen on the basis of the probability to distribute loads of energy equally among network sensor nodes. The nodes have ability to organize themselves into clusters in LEACH systems especially with one of the nodes acting as a router or data aggregator for other node. This initiates a process of randomized rotation of the high energy nodes as cluster heads to ordinary node and vice versa, and helps in preventing the faster draining of battery life among sensor nodes and enhances the lifelines of network. LEACH also performs the data aggregation and data fusion (data compression) [13] at cluster head level before transmitting data to base station, further reducing the energy consumption and enhancing the network lifetime.

The selection method of cluster head in LEACH protocol [14] is that the sensor node generates a random number between [0,1], if the random number is less than or equal to the node's threshold  $T(n)$ , the node is elected as the head node of the cluster.

$$T(n) = \begin{cases} \frac{p}{1 - p * [r \bmod (1/p)]}, & n \in G \\ 0, & n \notin G \end{cases} \quad (1)$$

In which the letter  $p$  stands for the probability of cluster head nodes each round that is the ratio of the total number of cluster head nodes and sensor nodes, the current number of rounds is represented through  $r$  and the letter  $G$  is the set of the nodes never become cluster head in recently  $1/p$  round.

A few drawbacks of LEACH systems are:

The non-deterministic nature of the setup phase due to randomness, can elongate the entire setup period. Instability during setup phase depends on the density of sensor nodes and this is not applicable on larger networks due to its usage of single hop communication methods. The consumption of energy depends on the location of the CH from the BS. It does not guarantee the good cluster head distribution and it involves assumption of uniform energy consumption of cluster heads during setup phase.

Other problems of LEACH cluster mechanism are the complete dependence on randomized nodally generated numbers for other attributes of the nodes, such as the current residual energy, location are not considered, which has the following problems:

- 1) The selection of low energy nodes as cluster heads without considering any residual energy of nodes when select cluster head and this causes the quick exhaustion of energy in nodes.

2) Nodal location is not considered during the distribution of cluster heads and this cannot guarantee uniform distribution of cluster head. This may also cause some cluster heads to be distributed densely, or cluster heads are too sparse, even no cluster head in certain areas.

### Improvement of Cluster Mechanism

Taking note of the position of energy and nodal positions into account, especially in view of problem in LEACH protocol, in order to optimize the selection mechanism the improved algorithm introduces three parameters include the number id neighbor nodes, energy, and the distance between node and base station to correct threshold.

1. Considering the current residual energy of the node when select the cluster head, and the energy adjustment parameter is introduced.

$$T_1(n) = \begin{cases} S(i).E / E_{ave}, & S(i).E > E_{ave} \\ 0 & , S(i).E < E_{ave} \end{cases} \quad (2)$$

Where  $S(i).E$  is the current residual energy of the node  $i$ ,  $E_{ave}$  is the average energy of all nodes.

2. The distance between the node and the base station is considered when select the cluster head, and the distance adjustment parameter is introduced.

$$T_2(n) = \begin{cases} S(i).dis / Dis_{ave}, & S(i).Dis > Dis_{ave} \\ 0 & , S(i).Dis < Dis_{ave} \end{cases} \quad (3)$$

Where  $S(i).dis$  is the distance between node  $i$  and base station,  $Dis_{ave}$  is the average distance of all nodes.

3. The density of nodes is considered when select the cluster head, and the number of neighbor nodes adjustment parameter is introduced.

$$T_3(n) = \begin{cases} S(i).Node / Node_{ave}, & S(i).Node > Node_{ave} \\ 0 & , S(i).Node < Node_{ave} \end{cases} \quad (4)$$

Where  $S(i).Node$  the number of neighbor nodes is,  $Node_{ave}$  is the average number of neighbor nodes of all nodes. The improvement of threshold for LEACH-H is expressed as follows:

$$T(n) = [w_1 T_1(n) + w_2 T_2(n) + w_3 T_3(n)] * p \quad (5)$$

Where  $w$  is the weight of the factors, its range is  $[0,1]$ ,  $w_1$  is the weight value of the residual energy of the node,  $w_2$  is the weight value of the distance between node and the base station,  $w_3$  is the weight value of the number of neighbor nodes, and  $\sum_{i=1}^3 w_i = 1$

### Genetic Algorithm (GA)

An adaptive Genetic algorithm (GA) was introduced by J.Holland for usage as search algorithm [15, 16]. GAs successfully handled many areas of applications and was able to solve a wide variety of difficult numerical optimization problems. GAs requires no gradient information and is much less likely to get trapped in local minima on multi-modal search spaces. GAs found to be quite insensitive to the presence of noise. The pseudo code of the GAs method is shown in figure

```

begin GAs
  g = 0 generation counter
  Initialize population
  Compute fitness for population P (g)
  While (Terminating condition is not reached) do
    g = g + 1
    Select P (g) from P (g - 1)
    Crossover P (g)
    Mutate P (g)
    Evaluate P (g)
  end while
end GA
  
```

Fig. 1: Pseudo code for Genetic Algorithm

The above problem is encoded via Gas within chromosomes which represent every possible solution. Fitness Functions investigate individual quality of each population members and these members undergo mutations and crossovers to recreate the next generation. Crossover functions create concatenated new solutions which are a part of two chosen chromosomes. Whereas a mutation is beneficial in overcoming local-minima entrapments and this continuous and repetitive process leads to an eventual solution.

### Local Search

This is the basis of multiple combinatorial optimization methods especially in terms of Local search [17, 18]. This is a simple iterative method for searching good approximate solution and it is based on the trial and error method. For instance combinatorial optimization problem is described through  $(S, g)$  in which  $S$  signifies the set of every feasible solutions and  $g$  is defined as the objective function which can maps every element  $s$  in  $S$  to a given real value. The end result is finding a solution  $s$  in  $S$  which will minimize the objective function  $g$ .

The problem is visualized through the following equation:

$$\min g(s), s \in S \quad (6)$$

Where  $N$  represents the function of the neighborhood or problem format  $(S, g)$  where it is represented from  $S$  to its powerset by the given mapping format:

$$N: S \rightarrow \mathcal{P}(S) \quad (7)$$

$N(s)$  is also symbolic of the value of the neighborhoods and it contains each possible solution which is reached via a single move from  $s$ . The move represents operators who convert multiple solutions with minute changes.  $x$  then represents the solutions which is otherwise known as the local minimum of  $g$  with respect to the neighborhood  $N$  iff:

$$g(x) \leq g(y), \forall y \in N(x) \quad (8)$$

The process of minimizing cost functions  $g$  or the Local search function is the successive steps in each of which the current solution  $x$  is being replaced by a solution  $y$  such that:

$$g(y) < g(x), \forall y \in N(x) \quad (9)$$

Most local search begins with arbitrary solution and end with the selection of local minimums. There are multiple ways to conduct local searches and the complexities in local search computations are dependent on neighborhood set sizes and its approximate time required to evaluate moves. It is thus noted that neighborhood size grows in size and this effects the time required to search for it, in order to determine a better local minima.

Local Search uses notion of state space, neighborhood and objective function.

i. State space  $S$ : the set of possible states that can be reached during the search.

ii. Neighborhood  $N(s)$ : the set of states, neighbors that during which can be reached from the state,  $s$  in one step.

iii. Objective function  $f(s)$ : A value that represents the quality of the state,  $s$ . The optimal value of the function is achieved when  $s$  is a solution.

Pseudo code for Local Search is as follows:

```

Select an initial state  $s_0 \in S$ 
While  $s_0$  is not a solution DO
    Select by some heuristic,  $s \in N(s_0)$  such that  $f(s) < f(s_0)$ 
    Replace  $s_0$  by  $s$ 
  
```

### Modified GA using Local Search

In genetic algorithm [19, 20], four parameters are presented. The size of the population, cross probability, mutation probability and weight accuracy of influence factors. The figure 2 shows the flowchart for proposed method.

- Coding the chromosome according to the required accuracy.
- Initial population of weight values: By the size of the population and the length of the individual obtained and the initial population of weights can be obtained.
- Calculating the fitness value of chromosome combined by weight values.

- Perform selection, crossover and mutation operators
- Perform local search operation
- If the new solution value generated by GA operators still can't satisfy the optimization condition, then go to 3). Else draw the optimal solution value

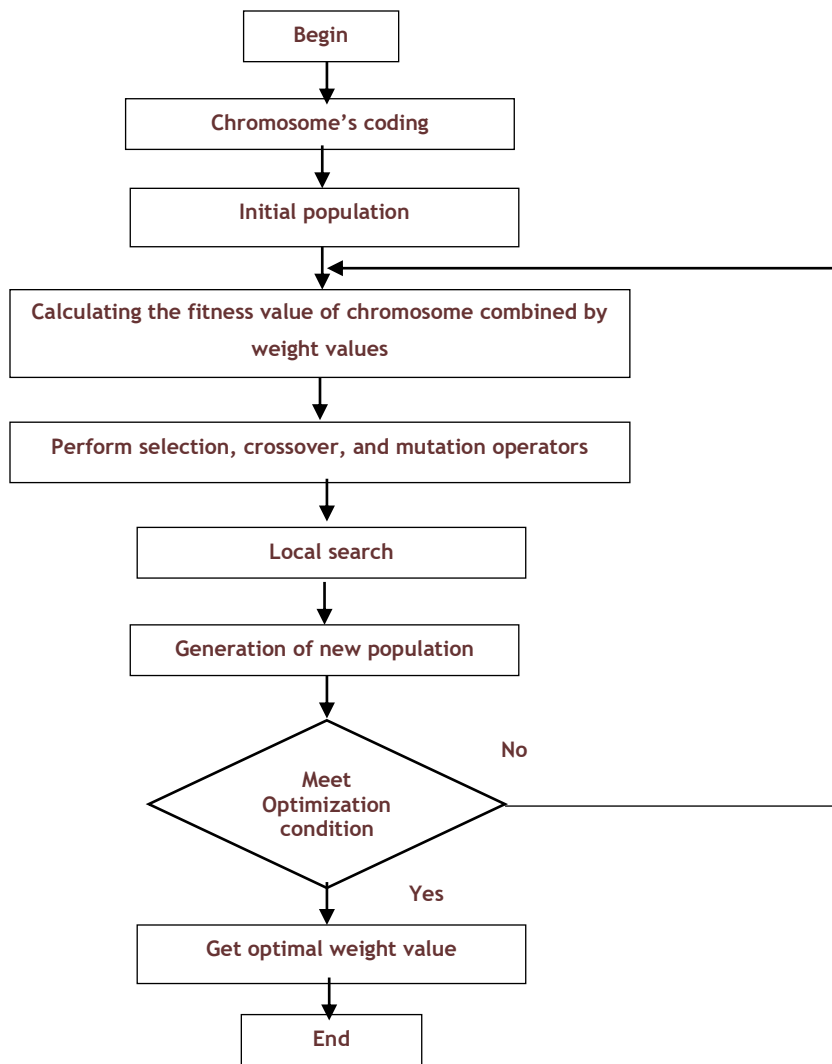


Fig. 2: Flowchart for Proposed Method

## RESULTS AND DISCUSSION

Table 1 to 4 and Figure 3 to 6 shows the results of number of clusters formed, average end to end delay (sec), average packet loss rate (%) and lifetime computation respectively. For experiments number of nodes considered is 100 to 600.

Observations from Table 1 and Figure 3 suggest that the number of clusters formed for GA and modified GA performs better than LEACH. When the number of nodes increases, the number of cluster formation also increases. The average of modified GA performs better by 3.15% than LEACH but reduces by 1.23% than GA.

Table 1: Number of Clusters Formed

Number of nodes	LEACH	GA	Modified GA
100	10	11	11
200	15	16	17
300	26	29	26
400	34	33	34
500	32	35	34
600	39	39	39

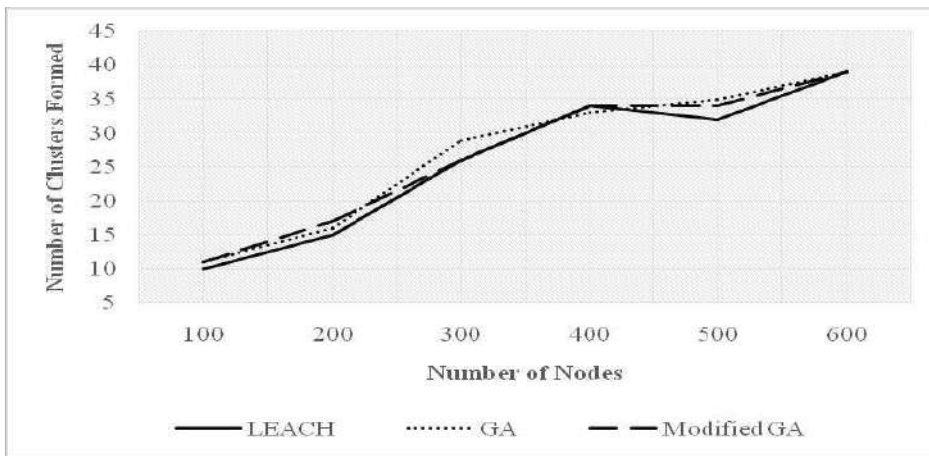


Fig. 3: Number of Clusters Formed

Table 2: Average End to End Delay (sec)

Number of nodes	LEACH	GA	Modified GA
100	0.00157	0.00159	0.00141
200	0.00161	0.00157	0.00153
300	0.01604	0.0165	0.01455
400	0.02632	0.02551	0.0244
500	0.05805	0.05988	0.05246
600	0.06473	0.06066	0.05305

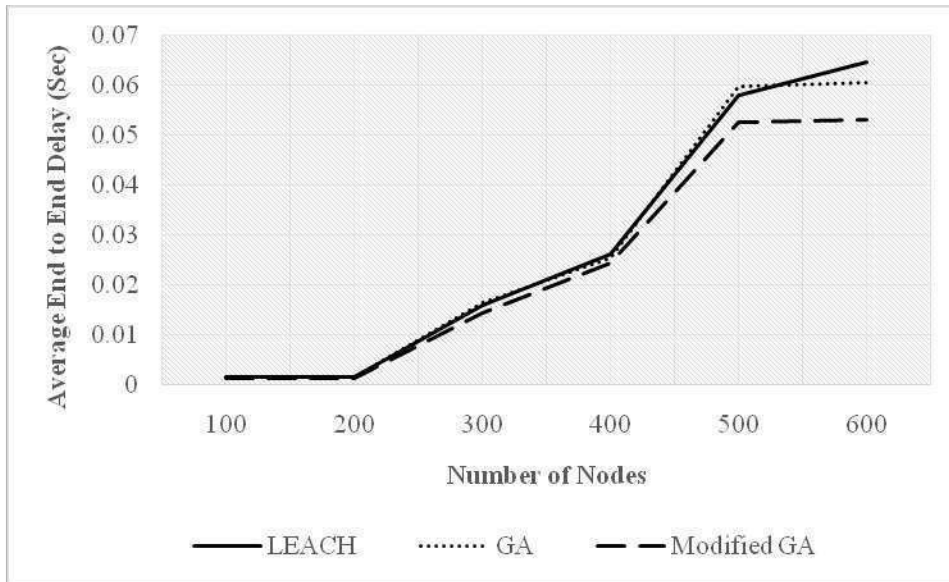


Fig. 4: Average End to End Delay (sec)

Observations from Table 2 and Figure 4 suggest that the average end to end delay for modified GA performs better by reducing the delay than LEACH and GA. When the number of nodes increases, the delay also increases. The average of modified GA performs better by 13.25% than LEACH and reduces by 11.69% than GA.

Table 3 Average Packet Loss Rate (%)

Number of nodes	LEACH	GA	Modified GA
100	11.07	8.77	8.35
200	17.64	13.75	13.47
300	18.19	13.26	12.63
400	23.01	20.88	18.25
500	31.99	26.9	24.57
600	43.89	30.73	28.89

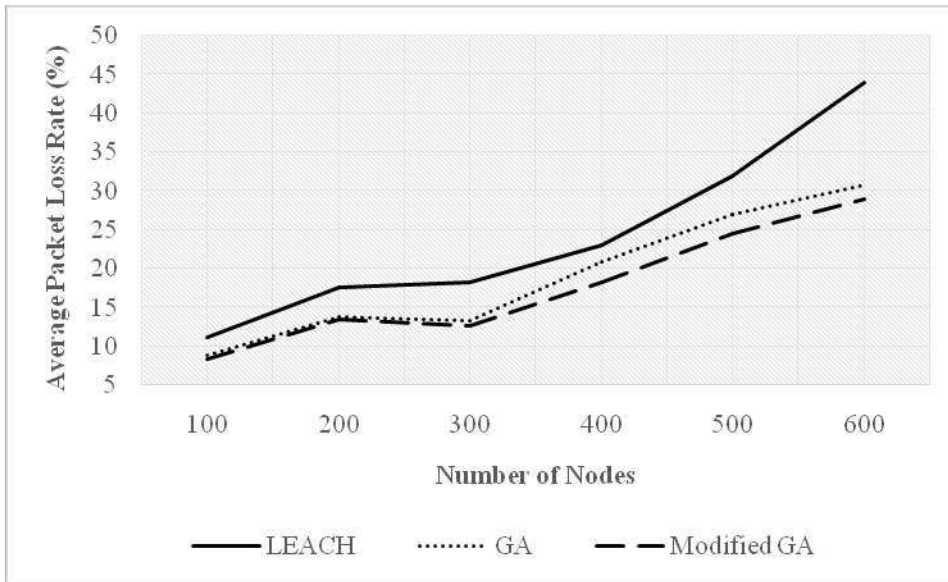
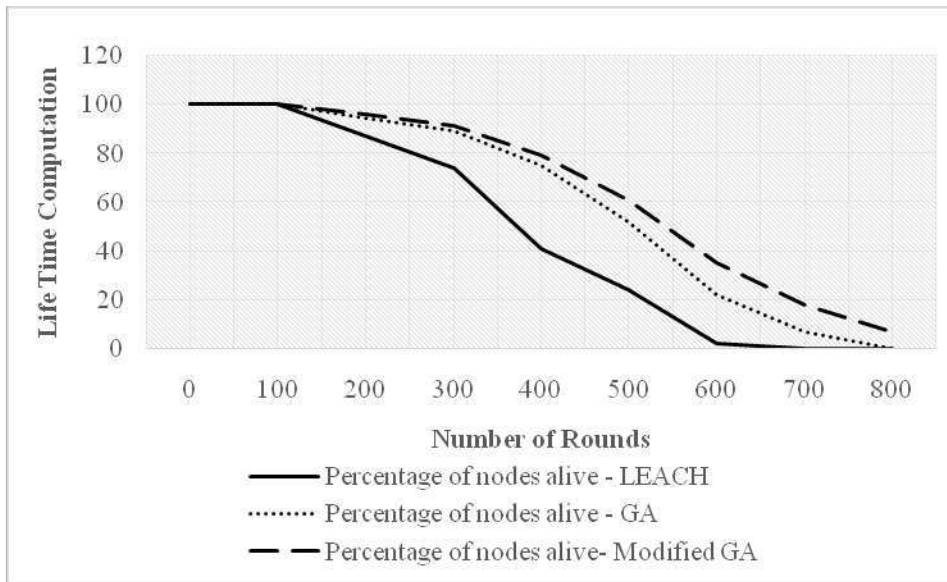


Figure 5 Average Packet Loss Rate (%)

Observations from Table 3 and Figure 5 suggest that the average packet loss rate for modified GA performs better by reducing the packet loss than LEACH and GA. When the number of nodes increases, the packet loss also increases. The average of modified GA performs better by 31.46% than LEACH and reduces by 7.37% than GA.

Table 4” Lifetime Computation

Number of rounds	Percentage of nodes alive - LEACH	Percentage of nodes alive - GA	Percentage of nodes alive- Modified GA
0	100	100	100
100	100	100	100
200	87	94	96
300	74	89	91
400	41	75	79
500	24	52	61
600	2	22	35
700	0	7	18
800	0	0	7



**Fig. 6: Lifetime Computation**

Observations from Table 4 and Figure 6 suggests that the lifetime computation for percentage of nodes alive - modified GA performs better by increasing lifetime than percentage of nodes alive -LEACH and percentage of nodes alive - GA. When the number of rounds increases, the lifetime computation decreases. The average of modified GA performs better by 31.33% than percentage of nodes alive - LEACH and by 8.53% than percentage of nodes alive - GA.

## CONCLUSION

A generic procedure, clustering is most commonly used to reduce distance in communication and help preserve nodes energies. Since genetic algorithms(GA) is superior to traditional optimization methods for its simplicity to operate and high stability in solving combinatorial optimization problems, GA is applied to obtain the optimal solution of weights of every impact factors, enabling the network to use the node energy more efficiently and balance the overall energy loss of the network. Results show that the average end to end delay for modified GA performs better by reducing the delay than LEACH and GA. When the number of nodes increases, the delay also increases. The average of modified GA performs better by 13.25% than LEACH and reduces by 11.69% than GA. Similarly performs better cluster formation, lifetime computation and reduces packet loss rate in a better way.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] D. Nguyen, P. Minet, T. Kunz and L. Lamont, " New Findings on the Complexity of Cluster Head Selection Algorithms", in Proceedings of the 2011IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks.
- [2] Yadav, J., & Dubey, S. K. (2014, July). Analytical Study of Cluster Head Selection Schemes in Wireless Sensor Networks. In *Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on* (pp. 81-85). IEEE.
- [3] Desai, K., & Rana, K. (2015, September). Clustering technique for Wireless Sensor Network. In *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on* (pp. 223-227). IEEE.
- [4] Khan, A. R., Rakesh, N., Bansal, A., & Chaudhary, D. K. (2015, December). Comparative study of WSN Protocols (LEACH,



- PEGASIS and TEEN). In *2015 Third International Conference on Image Information Processing (ICIIP)* (pp. 422-427). IEEE.
- [5] Singh, K. (2015, October). WSN LEACH based protocols: A structural analysis. In *Computing and Communication (IEMCON), 2015 International Conference and Workshop on* (pp. 1-7). IEEE.
  - [6] Karimi, M., Naji, H. R., & Golestani, S. (2012, May). Optimizing cluster-head selection in wireless sensor networks using genetic algorithm and harmony search algorithm. In *Electrical Engineering (ICEE), 2012 20th Iranian Conference on* (pp. 706-710). IEEE.
  - [7] Barekatin, B., Dehghani, S., & Pourzaferani, M. (2015). An Energy-Aware Routing Protocol for Wireless Sensor Networks Based on New Combination of Genetic Algorithm & k-means. *Procedia Computer Science*, 72, 552-560.
  - [8] Martínez-Estudillo, A. C., Hervás-Martínez, C., Martínez-Estudillo, F. J., & García-Pedrajas, N. (2005). Hybridization of evolutionary algorithms and local search by means of a clustering method. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(3), 534-545.
  - [9] Gupta, S. R., Bawane, N. G., & Akojwar, S. (2013, December). A Clustering Solution for Wireless Sensor Networks Based on Energy Distribution & Genetic Algorithm. In *Emerging Trends in Engineering and Technology (ICETET), 2013 6th International Conference on* (pp. 94-95). IEEE.
  - [10] Dongare, S. P., & Mangrulkar, R. S. (2015, January). An improved cluster head selection approach for energy efficiency in wireless sensor networks: A review. In *Pervasive Computing (ICPC), 2015 International Conference on* (pp. 1-6). IEEE.
  - [11] Efficient Cluster Head Selection Scheme for Data Aggregation in Wireless Sensor Network
  - [12] D. Singh, C. K. Panda, Performance analysis of modified stable election protocol in heterogeneous wsn, in: *Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on*, IEEE, 2015, pp. 1–5.
  - [13] W. Xinhua, W. Sheng, Performance comparison of leach and leach- c protocols by ns2, in: *Distributed Computing and Applications to Business Engineering and Science (DCABES), 2010 Ninth International Symposium on*, IEEE, 2010, pp. 254–258.
  - [14] Improvement and Application of LEACH Protocol based on Genetic Algorithm for WSN
  - [15] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA: University of Michigan Press, 1975.
  - [16] Sheta, A., & Solaiman, B. (2015). Evolving a Hybrid K-Means Clustering Algorithm for Wireless Sensor Network Using PSO and GAs. *International Journal of Computer Science Issues (IJCSI)*, 12(1), 23.
  - [17] Ekelin, C., & Olovsson, M. (1996). *Local search and configuration problems* (Doctoral dissertation, Master's thesis, Uppsala University).
  - [18] Voudouris, C., Tsang, E. P., & Alsheddy, A. (2010). Guided local search. In *Handbook of metaheuristics* (pp. 321-361). Springer US.
  - [19] Miao, H., Xiao, X., Qi, B., & Wang, K. (2015, September). Improvement and application of LEACH Protocol based on Genetic Algorithm for WSN. In *Computer Aided Modelling and Design of Communication Links and Networks (CAMAD), 2015 IEEE 20th International Workshop on* (pp. 242-245). IEEE.
  - [20] Yu Z, Bo J, Zhou Y. An Improved LEACH Routing Protocol Based on Genetic Algorithms for Wireless Sensor Network[J]. *Journal of Computer Research and Development*, 2010: S2.

\*\*DISCLAIMER: This published version is uncorrected proof, plagiarisms and references are not checked by IIOABJ; the article is published as it is provided by author and approved by guest editor.

# CERTAIN INVESTIGATIONS ON BIG DATA APPROACHES IN EDUCATION AND LEARNING ANALYTICS

Velmurugan, Kannaiya Raja, Saravanan Marimuthu Raja

Dept. of Computer Science, Ambo University, ETHIOPIA

## ABSTRACT

*The applications of computerized tools for learning management systems in education have been increasing in the last few years. Online contents by means of mobile phones, primarily smart phones enable the students to keep the learning processes around the clock. The enormous amount of data captured from student's online activities is wasted as traditional learning analytics are not capable of processing them. So in this paper it is proposed to use Big Data technologies and tools into education for capturing learning process of students, strategic operational decisions at institutions. The computing platform and teaching methods to accommodate emerging technologies for related courses are also investigated in this paper. In addition, visualization-based data discovery tools are focused on the front end of big data on helping institutions explore the data more easily and understand it more fully.*

Published on: 24<sup>th</sup>– August-2016

### KEY WORDS

*Big Data, Business intelligence (BI), Data warehouse, Enterprise resource planning (ERP), Visualization-based data discovery*

## INTRODUCTION

Higher education institutions are collecting and have access to more data than ever before and data-based decisions drive institutional effectiveness. The age of big data has come for higher education as IT becomes increasingly embedded in the process that comprises 'going to college,' such as course enrolment, classroom tutoring, and student services. Also, data about student successes and failures can be had to improve both individual and cooperative outcomes across all higher education in useful ways. Data warehouses and the cloud make it possible to collect and maintain massive records on students, alumni, operations, pedagogical impact, competition, marketplace and more. Today's sophisticated analytics technology makes it easier than ever to sift through and find meaningful patterns in all that data. As a result, demand for evidence to guide and support decision making is on the rise. Visualization-based data discovery tools allow institutions to mix up disparate data sources to get custom analytical views without rigidity and easy usage that was not available earlier. Advanced analytics are used to create interactive, animated graphics on desktops, and on mobile devices and laptops. End users can see the graphics on either the same devices, or on mobile devices such as tablets, smartphones etc.

## LITERATURE REVIEW

Big Data enables to mine learning information for insights regarding student performance and learning approaches from the institution database. Rather than rely on periodic test performance, [1] instructors can evaluate what students know and what methods are most effective for each student. By focusing on data analytics, professors can study learning in far better ways. Online tools enable evaluation [2] of a much wider range of student actions, such as how long they devote to readings, where they get electronic resources, and how quickly they master key concepts.

Computerized learning modules help to assess students in systematic, real-time ways. Data analytic software gives feedback about academic performance to students and professors. It also helps to predict students who need extra help or needing more hard tests. It also identifies scholastic approaches that are effective with specific students [3].

Employing software helps teachers to find out how the students learn statistics, chemistry, mathematical principles and experimental designs. This is done through pre-test and post-test evaluation by embedded assessment [4].

Beck and Mostow [5] used intelligent tutor software to comprehend whether a pupil learnt words better when re-reading an old story or a new story. Results showed on basis of reading time, mistakes, help requests and word knowledge that re-reading a story leads to half as much learning as to reading a new story.

School systems give more importance to feedback got to increase the learning process. Measurement of time spent on a specific test, skills realised and concepts mastered are used to get the feedback which is embedded in the process, so that real-time results is utilized to find what is learnt and to observe performance over time. Also, computer can alter test questions based on how the students answer earlier questions.

Kellen et al., [6] describes SAP's HANA, a Big Data analysis tool in University of Kentucky. By monitoring and evaluating the student's background data, the system calculates a "K-Score" for each student. This score shows the involvement of students in learning activities. A low score signifies an underperforming student whose needs should be taken care of.

## USING ANALYTICS TO DRIVE BETTER INSTITUTIONAL PRACTICES

An educational office relies on financial, student and other institutional data to report on and manage processes. Therefore, it's a natural shift to extend that kind of data-based decision making to all areas of college or university, including the most critical area for success in today's globally competitive environment enhancing the student experience. By utilizing information from the whole campus, linking it with outside data, and using technology to determine the "right data" to analyze and utilize, a school can easily begin asking (and answering) important questions like:

- Based upon projections from degree plans, how prepared is our institution for future classes, faculty, facilities, and other needs?
- What are the programs we should offer or drop to keep up with demand and expand our value to our constituents?
- How can financial statements from previous years help us predict future budgets?
- What is the actual return on investment for our advancement events and appeals?
- What is being said about our school on social media and what are we adding to that conversation?
- How do our continuing education, workforce development, online education, and/or schools of extension stack up against the competition?
- Based on our institution-level mission and our students' personal attainment goals, what are the best metrics to predict student success at our school?

The good news is that institutions already have a big pool of data to work with. The not so great news is, "although considerable amounts of data are being collected and stored (by higher education institutions), the data is not being used effectively to make predictions or trigger proactive responses". So, let's start small with big data. What do you have to work with that you can easily access? Most institutions find that grouping data into two main categories student data and institutional data helps clarify the kinds of big, strategic questions it can address. In the first bucket, if there are like most schools, have ample data on prospects, students, and alumni it can be posted with today's interconnected and social-media driven world, institutions have a lot of information about what they are doing, saying, thinking, and buying. The second large bucket of data that have revolves around institution. Whether institutions use an enterprise resource planning (ERP) system to connect systems across campus or are still operating in the "silo" world of disparate systems that have ample information on what institution is up to and how effective it is. That includes information and data collected from and reported to the local and federal government. Examples of data that have access to right now include:

- Data on your prospects, students, and alumni
  - + Demographics
  - + SATs, GPAs, transcripts
  - + Course selection, registration, add/drop
  - + Purchased/returned text books, library activity
  - + Financial aid applications, employment to support education
  - + Financials, fees, expenses (e.g., cafeteria)
  - + Online courses (data on how your students learn)

- + Social media
- + Internships
- + Alumni connections and post-CE job placements
- + Donations
- Data on institution
  - + Retention rates, graduation rates, transfer rates
  - + Marketing statistics
  - + Enrollment, yield rate
  - + Advancement effort progress
  - + Tenured faculty and adjunct
  - + Classroom allocation
  - + Instructional design effectiveness
  - + Financials
  - + Competitive information
  - + National Center for Education Statistics, IPEDS

Now that after recognizing the abundance of information the next step is collecting and organizing data with a business intelligence (BI) analytic tool that provides a clear view of what institutions are working with. Analytic tools offer common metrics for all datasets and give the needed dashboards, reports, visualization options, and real-time monitoring that brings data to life by making it understandable. With this new knowledge and ability institution can empower every area of an institution to make better business decisions and achieve optimal performance. Some example areas where applying the basic principles of BI analytics to big data, even minimally, can have profound impact. They are:

**Improve institutional operations.** In an age when all of higher education is being asked to do more with less, analytics can help reduce costs by providing information needed for streamlining and refining business processes. “Many colleges and universities have confirmed that analytics can help considerably advance an institution in areas as resource allocation, student success and finance.” Powerful analytic tools allow you to study patterns of performance over time, from one semester to another or from one year to another.

**Enhance pedagogy and learning.** Computerized testing, tablets and other mobile devices, online learning, course management/learning management systems, and other educational technologies are giving rise to a new era of learning analytics. Using real- or near-real-time monitoring of student activity such as postings on discussion boards, class material downloads assessment results, wiki activity, and the many other transactions per student per course—faculty can more easily create optimal learning environments and continually refine pedagogies along the way. In addition, data on faculty productivity can help drive positive learning outcomes. “Measuring faculty productivity is understandably a sensitive and controversial topic, but there is increasing acceptance that it is essential to sustainability. The issue here is largely cultural, rather than technology. However, the most successful institutions will employ technology to track, manage, measure, and improve faculty productivity.

**Increase student success.** Many colleges and universities now employ predictive analytics to improve their student success and retention rates. Best practices call for analyzing three years’ worth of historical data to find the risk factors and positive factors that inhibit or promote student success at an individual institution. Integrating data from multiple sources also improves at-risk student intervention efforts. These risk factors are different for every institution and will only be revealed by analyzing the existing data.

In addition, Big Data in online learning space will help institutions to improve learning outcomes for distinct students. By planning a curriculum that gathers data at every step of student learning process, universities can help student needs with customized modules, learning trees in the curriculum, assignments, and feedback, which will promote improved and richer learning.

Applying analytics to big data is quickly becoming imperative for successful higher education institutions. In higher education institutions Big Data may not solve all the issues and decisions, but they play an integral role in administrative and instructional functions. Even still, change is sometimes slow at many institutions. Using big data for impactful analysis on the campus can help with that problem, too. A recent EDUCAUSE study reported that “many study participants provided examples of how analytics programs can improve processes such as communication and decision making while increasing morale. Analytics programs can foster communication

between executive leadership, IR, and IT. Institutions should not wait for a cultural shift to be fully in place before beginning an analytics program. Initiating an analytics program may help establish that culture. Facing unprecedented demands for accountability, efficiency, and effectiveness, modern colleges and universities need to use big data to identify and evaluate strategies for improving the student experience, ensuring institutional success in every area from recruitment to alumni fundraising and everything in between. Higher education institutions that leverage the power of the large quantities of data at their disposal are better equipped to make impactful data-based decisions and thrive in today's fast-moving and competitive higher education world.

There are, however, no established best practices in higher education for what to measure or which measurement methodologies produce the most meaningful results. Schools are still learning how to harvest the power of the large quantities of data sets they collect daily. This paper provides guidelines [Figure -1] and best practices for making impactful data-based decisions and identifies mistakes to avoid when handling big data on the campus.

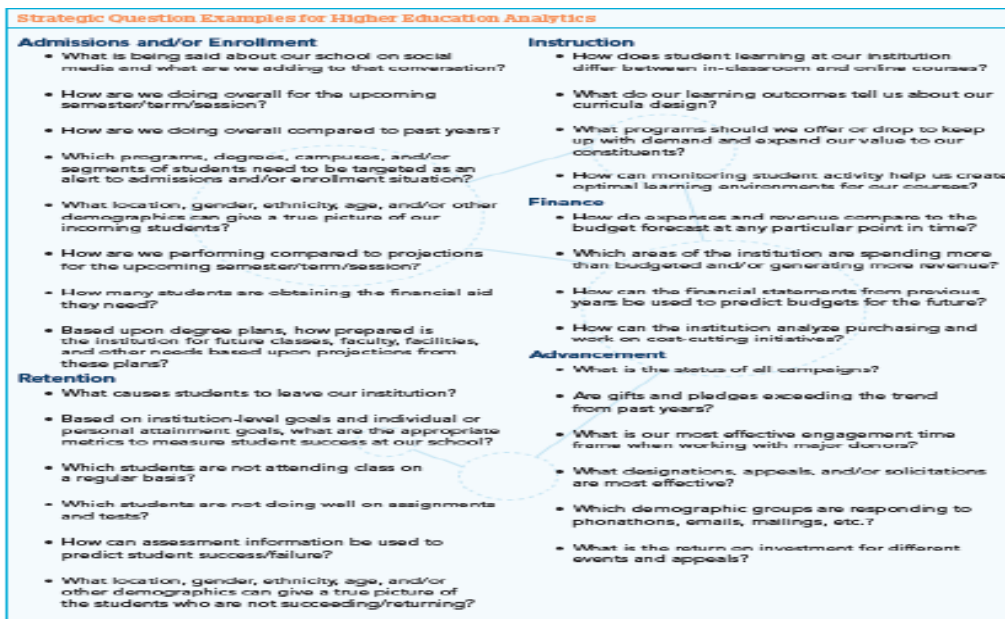


Fig. 1: Guidelines and best practices for making impactful data-based decisions

## TOOLS AND TECHNIQUES

**Techniques:** Various techniques are utilized for the problems faces in Big Data processing. Some of the techniques utilized in educational data mining are:

- Regression – It is used in predicting values of a dependant variable by appraising the relationship among variables using statistical analysis
- Nearest Neighbor – Here, the values are predicted based on the predicted values of the records that are nearest to the record that needs to be predicted.
- Clustering – This involves grouping of records that are alike by identifying the distance between them in an n-dimensional space where n is the number of variables.
- Classification – This is the identification of the category or class to which a value belongs to, on the basis of previously categorized values.

**Open Source Tools:** Several Open source tools exist which help in taming Big Data. Some of top tools are:

- MongoDB is a cross platform document oriented database management system. It uses JSON like documents instead of table based architecture.
- Hadoop is a structure that allows distributed processing of big datasets across clusters of networked computers using simple programming models.

- MapReduce is a programming model and framework used by hadoop. It enables processing huge amount of data in parallel on large clusters of compute nodes.
- Orange is a python based tool for processing and mining big data. It has an easy to use interface with drag and drop functionalities with variety of add-ons.
- Weka is a java based tool for processing large amount of data. It has a vast selection of algorithms that can be used in mining data.

## APPLICATIONS IN LEARNING

Big Data techniques can be utilized in different ways in learning analytics as listed:

- Performance Prediction: Student's performance can be predicted by analyzing student's interaction in a learning environment with other students and teachers.
- Attrition Risk Detection: By analyzing the student's behavior, risk of students dropping out from courses can be detected and measures can be implemented in the beginning of the course to retain students.
- Data Visualization: Reports on educational data become more and more complex as educational data grow in size. Data can be visualized using data visualization techniques to easily identify the trends and relations in the data just by looking on the visual reports.
- Intelligent feedback: Learning systems can provide intelligent and immediate feedback to students in reponse to their inputs which will improve student interaction and performance.
- Course Recommendation: This helps the student to find new courses based on their interest. This is done by analysing their activities.
- Student skill estimation: Estimation of the skills acquired by the student.

Behavior Detection: Detection of student behaviors in community based activities or games which help in developing a student model.

## BIG DATA VISUALIZATION

Apache Hadoop and other technologies are used to support storage and processing, and visualization-based data discovery tools help educational institutions explore the data more easily and understand it more fully [7].

Self-service BI also assists businesses to take advantage of mobile workforces. For instance, remote and on-site members of a product development team can easily view and share visualizations that explore potential product defects or customer preferences. This bring-your-own-device (BYOD) trend means that these users can use their own mobile devices to easily explore the data, discover trends and patterns, and communicate their findings to fellow team members and other audiences.

## PREDICTIVE ASSESSMENTS

Predictive and diagnostic assessments are some of the other ways of learning. In predictive assessment evaluate how students will achieve on standardized tests and diagnostic assessment highlights which techniques work for specific students and the better way to modify learning. Digital evaluation's main virtue is it gives students information needed for their learning and performance.

Performance is the key word for online predictive assessment. McGraw-Hill's Acuity Predictive Assessments [8] tool gives an initial indication of how students will perform on state NCLB assessments. It finds what the students know and what the students need to know on tests and recommends what the student should focus on to better their tests.

Similarly, the assessment tool helps "teachers probe student understanding of state standards, grade-level expectations, and specific skills, and quickly diagnose their strengths and instructional needs." By following how students solve problems and evaluate information, the tool provides guidance concerning preferred learning styles and works instruction to that preference.

Research has found that some pupils like to go through problem-solving step-by-step in a linear manner. Some others favour visual or graphical presentation in a non-linear fashion. So, assessment of learning styles is vital to personalization and tailoring instructional presentation. Digital tools that assist parents and teachers to comprehend student learning approaches are vital to educational attainment. Fifteen variables such as the number of discussion messages posted, time online, visits to course chat area, number of emails sent, number of assessments completed, and time spent on the assignments can be used to identify performance of students.

## CONCLUSION

Digital systems support real-time assessment for mining information. This increases learning, transparency, and accountability, and makes it easier to evaluate trends in educational institutions. Most schools have information systems like academic performance, student discipline, attendance etc. that do not link with one another. The fragmented nature of technology constrains the integration of school information and for mining useful trends. Also, educational institutions want to format data in related ways so that results can be linked. Too often there is inconsistent terminology or coding on issues related to graduation or school dropouts. Information entered into data systems should be easy to understand and coded in comparable ways. Teachers along with parents and students will benefit from advances in research and analysis.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Byers. Big Data.[2011] The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, May.
- [2] Felix Castro, Alfredo Vellido, Angela Nebot, and Francisco Mugica.[2007] Applying Data Mining Techniques to e-Learning Problems, *Studies in Computational Intelligence*, 62: 183-221.
- [3] US Department of Education Office of Educational Technology, Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics, 2012.
- [4] Sharona Levy and IriWilensky.[2011]Mining Students' Inquiry Actions for Understanding of Complex Systems, *Computers & Education*, 56:556-573.
- [5] Joseph Beck and Jack Mostow.[ 2008] How Who Should Practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students," Proceedings of the 9th International Conference on *Intelligent Tutoring Systems*, pp 353-362.
- [6] Vince Kellen, Cutter Consortium, Adam Recktenwald and Stephen Burr.[2013] Applying Big Data in Higher Education: A Case Study", *Cutter Consortium*, 13( 8): 1-39.
- [7] Dan Sommer, Rita L. Sallam, James Richardson.[ 2011]Emerging technology analysis: Visualization-based data discovery tools.
- [8] McGraw-Hill.[2009] Building the Best Student Assessment Solution," New York.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# AODV- QRS: A MULTIPATH QUALITY ROUTE SELECTION AODV FOR HIGH MOBILITY NETWORK

Indhumathi and Baby Deepa

Computer Science, Bharathiyar University, Coimbatore, INDIA  
Government Arts College- Autonomous, Karur, INDIA

## ABSTRACT

Mobile Ad-hoc Network (MANET), a collection of self-configured mobile nodal networks which can function without the need for any significant infrastructure. Routes have a tendency of switching very frequently and swiftly, especially because of the dynamic nature of network topology and due to this factor issues in route protocols have an important role and they are efficiently handled. Ad hoc On demand Distance Vector (AODV) routing systems are a popularly adopted MANET routing protocol system which is known for its adaptive capacities, especially to highly dynamic topologies even though it has issues in delay and scalability. The proposal of an On-demand Node-Disjointed Multipath Routing is suggested to overcome the shortcomings of on-demand AODV routing protocol. The proposed method is based on two concepts: multiple discoverable routes from the source to its destination and the mobility of node mobility (which is measured using RSSI signals) and there are significant changes in the route patterns and the number of packets dropped (if they increase over a set limit). Under the proposed technique, there have been significant improvements in QoS parameters, especially when comparing AODV and DSR simulations.

Published on: 28<sup>th</sup>– August-2016

### KEY WORDS

Mobile Ad-hoc Network (MANET), Routing, Ad hoc On Demand Distance Vector (AODV), Dynamic Source Routing (DSR), Received Signal Strength Indicator (RSSI).

\*Corresponding author: Email: [eindhumathi79@gmail.com](mailto:eindhumathi79@gmail.com)

## INTRODUCTION

A collection of self-configured nodular mobile networks, MANET systems perform well without any significant infrastructures. Nodes are connected to radio interfaces via wireless links in which every instrument within the MANET system is independent and free to randomly change its links frequently to other devices. A dynamic network topology, route networks shift quickly and frequently and this requires the efficient routing systems to handle important protocol roles. As a multihop process, limited transmission ranges constrain multiple mobile nodes, with each network topology acting as a router for itself. MANET networks have the ability to ensure the safe delivery of packages and handling any malfunctions within nodal systems, through reconfiguring the network [1].

Typical applications of MANET consist of:

- Application in military battlefields: Military bases have the advantage of maintaining proper network connections between soldiers, military information headquarters and vehicles through the use of Ad Hoc networking systems.
- Collaborative work applications: The need to create collaborated computational data for any form of business space, outside the office environment where it is difficult for people to have meetings and have proper exchanges of project information.
- Local level: The automatic linking of Ad-Hoc networks to temporary media can create instant connections using computer notebooks to share and spontaneously spread information among participants in a classroom or a conference. Home networks are an alternative locally used application system where devices and home networks directly communicate and exchange information.
- Personal area network and Bluetooth: A short ranged personal area network, where nodes are localized and commonly linked with a given individual. MANET based short range devices such as Bluetooth enables machines can help make inter-communication channels between portable devices like mobile phones and



laptops.

- Applications built for the Commercial Sector: In the case of an emergency rescue operation Ad hoc networks are popularly used for engaging in disaster relief efforts, for instance: in the case of an earthquake, flood or fire. It is essential for using uninterrupted communication systems for engaging in emergency rescue operations and especially for rapidly deploying networks wherever needed [2].

Thus, it is a big challenge to design routing patterns for MANET systems, especially the task of creating a dynamic topological network. Important reason for this change has been the multiple changes in the topology due to the higher degree of nodular mobility. Numerous protocols were developed in order to achieve this task and especially through selective path routing processes within a network, data packets were shifted from one node to another to transfer data in networks. A conventional MANET routing protocol is a standardized control of data flow in the network and also decide that which path should be followed by the packets to reach the particular destination [3].

Routing protocols in MANET are categorized based on implementing strategies in routing. The different kinds of Routing Protocols are as follows: (1) Table driven- Proactive Protocols, (2) On demand or Reactive Protocols and lastly, (3) Hybrid Protocols. Table driven Protocols are determined by conjoining nodes which connect a location to its destination and these are periodically maintained by updates in the routes. Whereas among on-demand routing protocols pathways are discovered when required and after a certain period of time expire. The last category, namely Hybrid routing protocols have the combined features of both reactive and proactive routing systems to scale network size and calculate the density of nodes in a network.

The AODV routing protocol algorithms are structured to cater to ad hoc modular networks. It has the capacity to both multicast and unicast routing systems and as an on-demand equation, it has created routes between nodes as a desired source route. A reactive protocol creates routes among nodes as required by sourced nodes. This protocol also maintains the above paths as per the requirement of the sources. But, apart from these, AODV systems help in creating new trees which help conjoin members together. These systems utilize systematic and sequenced numbers to promise freshness among routers. It is also free of loops, scalable at large numbers and self-starting mobile nodes [4].

Route REQuests messages are commonly used in AODV protocols to discover new paths needed by source nodes among flooded networks. Intermediate nodes which are present in this setting receive replies from RREQ and use it to route its correspondence with destination points when the sequence number is greater than or equal to what is contained into the RREQ. Most cases, the RREP sources its units back to its origin otherwise it will be rebroadcasted in the RREQ [5].

RREQ sources are tracked by nodes through IP address and ID of the source. Sources are already processed if they already possess an RREQ ID and do not forward it. RREP propagates source nodes back to its origin, or forward pointers to its destination and once source nodes receive RREP, it soon starts to forward pointers and data to the appropriate destination. Any RREP is containing a greater amount of sequenced numbers or the same amount with smaller hop counts, it updates routing information for destinations through better routes. The route will continue to be maintained as long as it is active.

Routes are active as long as there are periodic travels of data packets from the source to the path destination. Links are eventually deleted after the source stops sending data packets from intermediate node tables. If a link break occurs while the route is active, the node upstream of the break propagates a Route ERRor (RERR) message to the source node to inform it of the now unreachable destination(s).

AODV protocols benefits favor least congested routes and not short routes, where it supports multicast and unicasted packet transmissions for constant movements in the nodes. The quick response to any form of topological changes affects the functioning of active routes. And AODV systems do not add any additional overhead changes to data packets which are not a part of source routings [6].

AODV protocols have certain limitations in terms of requiring broadcasting medium nodes to help detect any signals from other broadcasts. It is possible for validated routes to expire and not be able to predict a reasonable expiry time because sending time widely differs from various nodes and can change dynamically. Additionally,

as the performance metrics begin to decrease, the size of a network increases. Due to this, AODV networks are vulnerable to multiple forms of attacks as it is based on the assumption that every node cooperates with one another and the failure of this can lead to the breaking of nodes.

A pure example of an On-Demand routed protocol is the DSR system which is based on the theory of source routings. Designed to handle multihop ad hoc networks mobile nodes. It provides complete self-organization, self-configuration and does not require existent network administration or infrastructure [7]. DSR uses no periodic routing messages like AODV, thereby reducing network bandwidth overhead, conserves battery power and avoids large routing updates. Instead DSR needs support from the Medium Access Control (MAC) layer to identify link failure

DSR is composed of the two mechanisms of Route Discovery and Route Maintenance, which work together to allow nodes to discover and maintain source routes to arbitrary destinations in the network. A prominent benefit of DSR based protocols is the lack of requirement for to keep track of routes through routing table systems, to ensure the entire packet is contained within the packet header. A unique feature of DSR exists in its source routing abilities and since packet routes itself in loops of either short or long lived patterns, they cannot be formed immediately as they will be eliminated on detection. The above property is a useful optimization protocol to open up features. But neither DSR systems or AODV can ensure a small path, as even if the route maybe the shortest, the destination always responds to route requests in which the source request is always the initiator. This study proposes a multipath quality route selection in AODV for high mobility network. Section 2 reviews related work in literature. Section 3 describes methods used and section 4 discusses experiments results. Section 5 concludes the paper.

## RELATED WORK

A multipath routing protocol was suggested by Ahn et al., [8] in MANET systems which are composed of highly mobilized nodal networks. New multi-path routing establishes mechanisms for the main route AODV based systems, and then the process data transmission immediately begins. Backup search route processes taking place when data is transmitted by a lower than minimum transmission delay. Unconnected node routes are selected among main node route by avoiding other main body nodes. When main or back up routes break, the data transmitted continuously with another route and the broken route recovered through route maintenance processes. The result of simulation based problems on Qualnet simulator shows how the proposed routing protocol has the backup route 62.5% of the time when the main route was broken, improves the packet transmission rate by 2-3% and reduces the end-to-end delay by 10% compared with AODV and AODV-Local Repair.

Venkataraman, et al., [9] proposed a generalised trust-model over routing protocols in MANETs. The novelty of the approach, that the notion of trust can be easily incorporated into any routing protocol in MANETs. The vector auto regression based trust model was introduced to identify malicious nodes that launch multiple attacks in the network. The proposed trust model was incorporated over AODV routing protocol and Optimised Link State Routing (OLSR) protocol in MANETs. The performance evaluations showed that by carefully setting the trust parameters, a substantial benefit in terms of throughput can be obtained with minimal overheads.

Kuppusamy et al., [10] described the characteristics of ad hoc routing protocols OLSR, AODV and TORA based on the performance metrics like Packet Delivery Ratio (PDR), end-to-end delay, routing overload by increasing a number of nodes in the network. This comparative study proves that AODV, TORA performs well in dense networks than OLSR in terms of PDR.

Bagwari et al., [11] analyzed the performance of reactive routing protocol via increasing number of nodes and observing its effect on Quality of Service (QoS) of MANET. The routing protocols make an important role in improving QoS in MANET. The QoS depends on upon several parameters like end-end delay, throughput, data drop and network load. The reactive routing protocol which was considered AODV for this scenario with Multiple Cluster Head Gateway (MCHG). The author observing the performance of Routing Protocol via enhancing the network size on the basis of following parameters: delay, throughput, traffic sent, traffic received, data dropped and network load. Network simulation tool used in the simulation was OPNET Modeler (Ver. 14.0). Finally, the author has conducted simulation experiments in the conditions can be improved QoS of MANET Network performance.

Sarma & Nandi [12] proposed a Route Stability based QoS Routing (RSQR) protocol in MANET through which QoS routed extensions are controlled with constraints in delay and throughputs. This can ensure that the path was chosen for data transfer is validated and can survive for longer periods. Due to its complex nature, MANET systems suffer this critical issue and the authors suggested a simple method to address link and route stability based on the strength of signals received. By including some extra fields in route request/reply packets, the route stability information can be utilized to be selected higher stability routes within all possible routes among situated route pairs to the given destination of the source. Moreover including the strength of signals on the basis of control at the time of admission can enhance performance factors in routing processes. Results of the experiments show performance improvements in terms of PDR, control overhead and average end-to-end delay in comparison with a QoS routing protocol.

Moussaoui et al., [13] proposed a new mechanism to be established stable and sustainable paths between all pairs of nodes in a MANET. In this mechanism, the author used a stability function as the main path selection criterion based on the calculation of the mobility degree of a node relative to its neighbor. The author applied this mechanism on the OLSR protocol to be elected stable and sustainable Multi-Point relays (MPR) nodes and topology. This mechanism can be significantly minimized the recalculation of MPR and the routing tables recalculation process. Moreover, it guarantees other QoS metrics such as the packet loss and the response time. The simulation results show the effectiveness of the mechanism and encouraged further investigations to be extended it in order to be guaranteed other QoS requirements.

Chatterjee & Das [14] proposed an enhanced version of the well-known Dynamic Source Routing (DSR) scheme based on the Ant Colony Optimization (ACO) algorithm, which can be produced a high data PDR in the low end to end delay with low routing overhead and low energy consumption. In this proposal, in the situation where a node needs to transfer the packages from one node to a different one, similar to DSR systems, the node needs to initially evaluate existing route caches. If there is a lack of availability of known nodes, the sender can find the route by locally broadcast Route Request control packages (also known as Req. Ant packets) to find out the routes. This was similar to the biological ants initially spreading out in all directions from their colony in search of food. The author also proposed a novel pheromone decay technique for route maintenance. The simulation results show that the ACO based Enhanced DSR (E-Ant-DSR) outperforms the original DSR and other ACO based routing algorithms.

Mohanapriya & Krishnamurthi [15] presented a Modified Dynamic Source Routing Protocol (MDSR) to be detected and prevented selective black hole attack. Selective black hole attack was a special kind of black hole attack where malicious nodes drop the data packets selectively. The author proposed an Intrusion Detection System (IDS) where the IDS nodes are set in promiscuous mode only when required, to be detected the abnormal difference in the number of data packets being forwarded by a node. When any anomaly was detected, the nearby IDS node broadcast the block message, informing all nodes on the network to be cooperatively isolated the malicious node from the network. The proposed technique employs Glomosim to be validated the effectiveness of proposed IDS.

Zhao et al., [16] proposed a novel Opportunistic routing (OR) protocol - Context-aware Adaptive Opportunistic Routing (CAOR) for MANETs. CAOR abandons the idea of candidate list and it allows all qualified nodes to be needed in packet transmitted participation. CAOR can transfer simultaneous packets by using multi-cross layered knowledge, like the progress in geography, energy, mobility and quality of linkages quality, geographic progress. Through the assistance of the Analytic Hierarchy Process theory, CAOR adjusts the weights of context information based on their instantaneous values to be adapted the protocol behavior at run-time. Moreover, CAOR uses an active suppression mechanism to be reduced packet duplication. Simulation results show that CAOR can be provided efficient routing in highly mobile environments. The adaptivity feature of CAOR was also validated.

Moussaoui & Boukeream [17] presented a survey of recent routing solutions. The author started by giving general definitions related to the mobility and the link stability. Then, the author proposed a classification for the routing protocols based on the link stability. For each proposed class, the author will list examples of routing protocols. Finally, a conclusion and future research directions are discussed.

Yadav et al., [18] theorized an alternate method for calculating the availability of signal strength predictions in AODV based routing systems. Estimate link breakages in nodes and the time is taken to precaution other nodes about breakages in the pathways and on the basis of this available information, local repair links or newly discovered paths are used in advance to breakages in the route path. This can reduce the impact of daily losses in data usage packages. By the above proposed method and knowledge gained locally route repair or new route discovery, are compared with AODV systems without the need for link prediction. The results show that there was a significant reduction in packet drops and average end-to-end delay. There was also an improvement in data PDR for AODV with link prediction. Proposed approach results in improvement in the QoS.

Amara Korba et al., [19] presented a comprehensive survey of security threats in MANET. In particular, the author examined all routing threats that can target the operation of routing protocol, whether they belong to selfish behaviors or malicious attacks, as well as countermeasures against such attacks. In order to be analyzed the existent countermeasures in a structured manner it has been classified them into three classes; solutions based on cryptography; IDSSs; and trust management and reputation-based solutions.

Su [20] focused on the wormhole attack, and proposed a secure routing protocol based on the AODV routing protocol, which was named Wormhole-Avoidance Routing Protocol (WARP). WARP systems use multipathed disjointed links especially at the time of multiple path discoveries, and can significantly give a greater choice in paths and which routes to be avoided due to the presence of malicious nodes, but eventually uses only one path to be transmitted information. It is based on the feature wherein wormhole nodes can access routes with ease, from source nodes to its appropriate destination nodes. Especially via the WARP neighbors are enabled to be discovered inside wormhole nodes which have abnormal attraction paths. After which point wormhole nodes consequently become alone and isolated from neighboring nodes, and after this point they will be separated from the entire network.

Yerneni & Sarje [21] proposed modifications to the AODV protocol and justified through the implementation of appropriate simulations via the NS-2.34, the solutions of the given problems. The proposed protocol makes use of the number of RREQ and RREPs forwarded by the nodes to be detected the attack. The analysis shows that modified protocol improves PDR even in the presence of attack.

Bhalaji & Shanmugam [22] proposed and analysed a new routing protocol based on the trust model. Here each node has been calculated trust value and association status for all its neighboring nodes through monitoring its behavior in the network. Then this trust model was integrated into the DSR protocol which was the most common on demand routing protocol used in

MANET. The above idea theorizes that selected routes are not allocated based on the premise of initial RREP arrivals and it waits till this factor receives data from neighboring nodes and critically decides a pathway to be chosen based on the relation between each other. Therefore, Greyhole nodes are identified based on the above factors and they are not given any selected preferences based on route decisions wherein the existent rules within routes are examined on the basis of comparing simulation results of it with the standard DSR in the presence of Greyhole nodes. Simulated results can demonstrate how proposed routing protocols can be effectively detected Greyhole Nodes and isolated them from routing.

## METHODOLOGY

RSSI portrays relations amongst transmitted and received powers by the following equation (1):

$$p_r = p_t \times \left( \frac{1}{d} \right)^n \quad [1]$$

Wherein  $p_r$  refers to the amount of power received and,  $p_t$  is the amount of power transmitted. Distance  $d$  is the space which exists among the sending and receiving nodes, whereas  $n$  is the amount of factors which is inputted into transmissions where the value depends on environments which are propagated. [23].

Now it needs to show the relation between RSSI and distance, for calculating the received power based on this model, it first calculates the received power at a reference distance using the Friis formula (given in equation (1)). Then, it incorporate the effect of path loss exponent and shadowing parameters.

$$RSSI = -(10 \times \log_{10}(d_{i,j}) - A) \quad [2]$$

Estimated hypothetical space amongs nodes is represented by the following equation given below:

$$d_{ij} = 10^{\frac{RSSI - A}{-10 \cdot n}} \quad [3]$$

In which the symbols represent the following factors:

$d_{ij}$  as the representative of the estimated distance between node  $i$  to node  $j$ .

RSSI as the abbreviation of Receiving Signal Strength Indicator.

"A" symbolizes the amount of power received from reference distance = 1 meter

$n$ : is the transmission factor whose value depends on the propagation environment.

Every node is aware of the distance from their neighbors and able to decide the choice of node as the next hop route. This is also visible in the given equation (3)

Power consumption is an important issue for transmitting data via Wi-Fi nodes and this is controlled by the mode of operation and data consumed; by which it can derive the amount of power consumed for transmitting an amount of data during a period of time ( $t$ ) is presented as follows:

$$E(t) = \sum_j E_j(t_j) + \sum_j \sum_k E_{j,k} \times C_{j,k}(t) \quad (4)$$

Where  $E(t)$  is the total energy consumed by the hardware component over the duration  $t$ ,  $t = \sum_j t_j$ ,  $t_j$  is the duration spent in power state  $j$  and  $E_j(t_j)$  is the energy spent during  $t_j$ . Assuming that  $P_j$ , the rate of energy consumption in power state  $j$ , is constant during  $t_j$ ,  $E_j(t_j)$  can be calculated as the product of  $t_j$  and  $P_j$ ,  $E_{j,k}$  is the overhead caused by the transition from power state  $j$  to  $k$ , while  $C_{j,k}(t)$  shows how many times this transition has occurred during  $t$ .

The remaining power of each node after transmission of desired data and costs are calculated through the following equation given below. Also, the leftover power is known as Remaining Battery Power (RBP):

$$RBP = \frac{AVLBP - E(t)}{MPB} \quad (5)$$

Where:

RBP : Remaining Battery Power.

AVLBP : Available Battery Power.

$E(t)$  : the total energy consumed by the hardware component over the duration  $t$ .

MPB : Maximum Battery Power.

The optimal node should be chosen as an intermediate routing node will be the one with higher RBP after calculating the amount of power will be consumed as described in equation (4).

The RSSI value [24] is calculated with the help of two ray ground model in equation (6):

$$P_r(d) = \frac{P_t * G_t * G_r * h_t^2 * h_r^2}{d^4 L} \tag{6}$$

- $P_r$  : Power received at distance d
- $P_t$  : Transmitted signal power
- $G_t$  : Transmitter gain (1.0 for all antennas)
- $G_r$  : Receiver gain (1.0 for all antennas)
- D : Distance from the transmitter
- L : Path loss (1.0 for all antennas)
- $h_t$  : Transmitter antenna height (1.5 m for all antennas)
- $h_r$  : Height of the receiver antenna (All antennas are estimated to be 1.5 m)

NS2 systems adopt RSSI standardized measurements where the strengths of signals are easured at one node at a time. Assuming that at the point of the simulation, two wireless nodes are at different coordinates. Transmissions are started by one of the nodes, especially Transmission Control Protocols (TCP) and User Datagram Protocol (UDP) packages transferred through a wireless interface, with the provided transmitted powers and gains by the antennas. It is propagated through Random way and the threshold for receiver and carrier sensitive models. The given thresholds help to define the probable success of receiving packages.

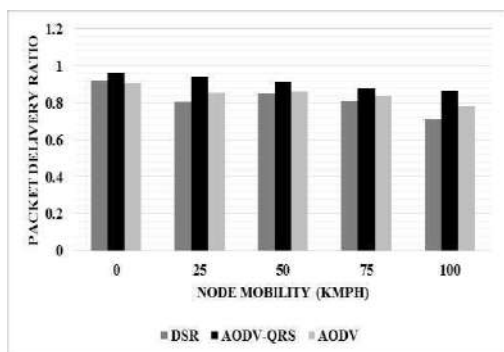
Request Signal Strength (NSS) nodes gain their value from the latest updates and the different between previous and new RSS values differ greatly. Each hop matches new RSS values at fixed intervals with the difference calculated from the Threshold Value (THRS) few link is usually established by this value. After calculations, if any of the parameters is found below the threshold value then the link is considered to be having a breakdown. At this point Possible Route Maintenance Algorithms (PRMA) is considered a solution to fix links between nodes in the midst of a breakdown.

## RESULTS

In this section, 80 nodes in 4 sq km. Each node has 250 m range are used. The DSR, AODV-QRS and AODV are evaluated. The Packet Delivery Ratio (PDR), end to end delay in second and number of hops to the destination as shown in [Table -1], [Table -2], [Table -3] and [Figure -1], [Figure -2], [Figure -3].

Table: 1. Packet delivery ratio

Node mobility kmph	DSR	AODV-QRS	AODV
0	0.9183	0.9628	0.9074
25	0.8058	0.9422	0.8543
50	0.8486	0.9127	0.8634
75	0.813	0.8783	0.8363
100	0.7127	0.8662	0.7823

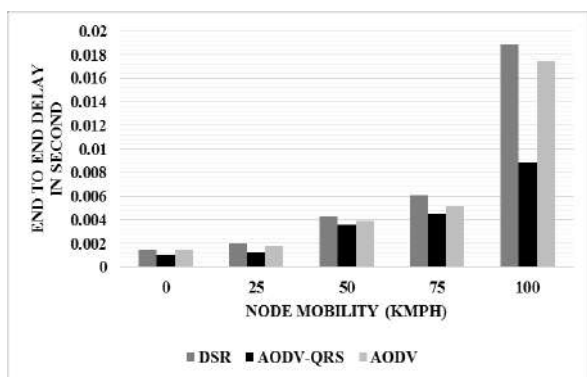


**Fig:1. Packet Delivery Ratio**

[Figure- 1] shows how AODV-QRS has higher PDR by 4.73% & 5.92% for 0 node mobility, by 15.6% & 9.7% for 25 node mobility, 7.27% & 5.55% for 50 node mobility, by 7.72% & 4.89% for 75 node mobility and by 19.44% & 10.17% for 100 node mobility when compared with DSR and AODV.

**Table: 2. End to end delay in second**

Node mobility kmph	DSR	AODV-QRS	AODV
0	0.0015	0.001	0.0015
25	0.002	0.0012	0.0018
50	0.0043	0.0036	0.0039
75	0.0061	0.0045	0.0052
100	0.0189	0.0089	0.0175

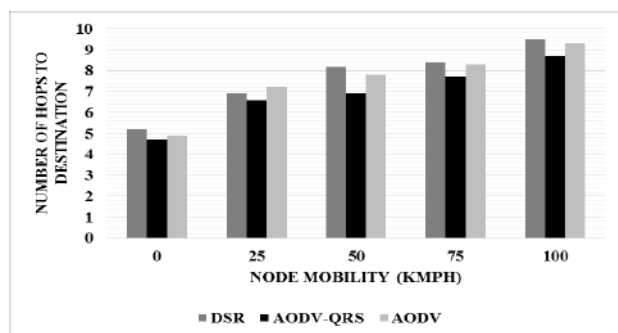


**Fig:1. End to End delay in Second**

[Figure -1] suggests that AODV-QRS has lower end to end delay in second by 40% & 40% for 0 node mobility, by 50% & 40% for 25 node mobility, 17.72% & 8% for 50 node mobility, by 30.18% & 14.43% for 75 node mobility and by 71.94% & 65.15% for 100 node mobility when compared with DSR and AODV.

**Table: 3. Number of Hops to destination**

Node mobility kmph	DSR	AODV-QRS	AODV
0	5.2	4.7	4.9
25	6.9	6.6	7.2
50	8.2	6.9	7.8
75	8.4	7.7	8.3
100	9.5	8.7	9.3



**Fig: 2. Number of Hops to Destination**

It is observed from the above [Figure -3] that the AODV-QRS has lower number of hops to destination by 10.1% & 4.16% for 0 node mobility, by 4.44% & 8.69% for 25 node mobility, 17.21% & 12.24% for 50 node mobility, by 8.69% & 7.5% for 75 node mobility and by 8.79% & 6.66% for 100 node mobility when compared with DSR and AODV.

## CONCLUSION

A MANET contains self-configuring, self-organizing and self-operating nodes, each of them communicates with other nodes directly, without any help of centralized administration or fixed infrastructure, within transmission range of nodes. Due to the quick installation behavior, dynamic configuration, various advantages and different application areas, the field of MANETs is rapidly growing and changing. Although there are still many challenges and issues that need to be faced by the MANET. In order to secure and effective communication within a MANET, an efficient routing protocol is required to discover routes between mobile nodes. The common objective of routing protocol is to provide better efficient energy aware and secure routing schemes to MANET. In this paper, proposed AODV routing protocol and measurement of node mobility using RSSI signal. Experimental results show that the AODV-QRS has higher PDR by 4.73% & 5.92% for 0 node mobility, by 15.6% & 9.7% for 25 node mobility, 7.27% & 5.55% for 50 node mobility, by 7.72% & 4.89% for 75 node mobility and by 19.44% & 10.17% for 100 node mobility when compared with DSR and AODV.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Nand P, Sharma DS. Performance Study of Broadcast Based Mobile Adhoc Routing Protocols AODV, DSR and DYMO. *International journal of security and its Applications*, 5(1):53-64.
- [2] Aarti DS. [2013] Tyagi, Study of MANET: Characteristics, Challenges, Application and Security Attacks. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(5):252-257.
- [3] Bhosle AA, Thosar TP, Mehatre S. [2012] Black-Hole And Wormhole Attack in Routing Protocol AODV in MANET". *International Journal of Computer Science Engineering and Applications*, 2(1):45.
- [4] Ade SA, Tijare PA. [2010] Performance Comparison of AODV, DSDV, OLSR And DSR Routing Protocols in Mobile Ad Hoc Networks. *International Journal of Information Technology and Knowledge Management*, 2(2):545-548.
- [5] Manikandan SP, Manimegalai R. [2012] Survey on Mobile Ad Hoc Network Attacks and Mitigation Using Routing Protocols. *American Journal of Applied Sciences*, 9(11):1796.
- [6] Taneja S, Kush A. [2010] A Survey of Routing Protocols in Mobile Ad Hoc Networks. *International Journal of Innovation, Management and Technology*, 1(3):279.

- [7] Gupta AK, Sadawarti H, Verma AK. [2010] Performance Analysis of AODV, DSR & TORA Routing Protocols. *International Journal of Engineering and Technology*, 2(2):226.
- [8] Ahn CW, Chung SH, Kim TH, Kang SY. [2010, ]. A Node-Disjoint Multipath Routing Protocol Based on AODV In Mobile Ad Hoc Networks. In Information Technology: New Generations (ITNG), 2010 *Seventh International Conference on IEEE* pp. 828-833.
- [9] Venkataraman R, Pushpalatha M, Rama Rao T. [2012] Regression-Based Trust Model for Mobile Ad Hoc Networks. *Information Security, IET*, 6(3):131-140.
- [10] Kuppusamy P, Thirunavukkarasu K, Kalaavathi B. [2011] A Study and Comparison of OLSR, AODV And TORA Routing Protocols in Ad Hoc Networks. In Electronics Computer Technology (ICECT), 2011 *3rd International Conference on IEEE* 5:143-147.
- [11] Bagwari A, Jee R, Joshi P, Bisht S. [2012] Performance of Aodv Routing Protocol with Increasing the MANET Nodes and its Effects on QoS Of Mobile Ad Hoc Networks. In Communication Systems and Network Technologies (CSNT), 2012 *International Conference on IEEE* , 320-324.
- [12] Sarma N, Nandi S. [2010] Route Stability Based QoS Routing in Mobile Ad Hoc networks. *Wireless Personal Communications*, 54(1):203-224.
- [13] Moussaoui A, Semchedine F, Boukerram A. [2014] A Link-State QoS Routing Protocol Based on Link Stability For Mobile Ad Hoc Networks. *Journal of Network and Computer Applications*, 39:117-125.
- [14] Chatterjee S, Das S. [2015] Ant Colony Optimization Based Enhanced Dynamic Source Routing Algorithm for Mobile Ad-Hoc Network. *Information Sciences*, 295:67-90.
- [15] Mohanapriya M, Krishnamurthi I. [2014] Modified DSR Protocol for Detection And Removal of Selective Black Hole Attack In MANET. *Computers & Electrical Engineering*, 40(2):530-538.
- [16] Zhao Z, Braun T, Rosario D, Cerqueira E. [2014] CAOR: Context-Aware Adaptive Opportunistic Routing in Mobile Ad-Hoc Networks. In *Wireless and Mobile Networking Conference (WMNC), 2014 7th IFIP (pp. 1-8). IEEE*.
- [17] Moussaoui A, Boukerram A. [2015] A Survey of Routing Protocols Based on Link-Stability in Mobile Ad Hoc Networks. *Journal of Network and Computer Applications*, 47:1-10.
- [18] Yadav A, Singh, YN Singh RR. [2015] Improving Routing Performance in Aodv with Link Prediction in Mobile ADHOC Networks, *Wireless Personal Communications*, 83(1):603-618.
- [19] Amara Korba A, Nafaa M, Salim G. [2013] Survey of Routing Attacks and Countermeasures in Mobile Ad Hoc Networks. In Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on *IEEE* pp. 693-698.
- [20] Su MY. [2010] WARP: A wormhole-avoidance routing protocol by anomaly detection in mobile ad hoc networks. *computers & security*, 29(2):208-224.
- [21] Yerneni R, Sarje AK. [2012] Secure Aodv Protocol to Mitigate Black Hole Attack in Mobile Ad Hoc. In Computing Communication & Networking Technologies (ICCCNT), 2012 *Third International Conference on IEEE* pp.1-5.
- [22] Bhalaji N, Shanmugam A. [2012] Dynamic Trust Based Method to Mitigate Greyhole Attack in Mobile Adhoc Networks. *Procedia Engineering*, 30:881-888.
- [23] Yasin A, Jabareen S, Al Suqi I. [2014] Enhancing the Connectivity of Mobile Ad-Hoc Networks by Considering the Power, Mobility and Activity of Nodes. *International Journal of Computer Science Issues (IJCSI)*, 11(2):140.
- [24] Gupta C, Sharma P. [2014] An Approach to Link Failure in MANET. *International Journal of Computer Science and Network Security (IJCSNS)*, 14(11):108.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.



# IMPACT OF BLACK HOLE ATTACK UNDER DIFFERENT SCENARIOS ON AD HOC ON-DEMAND DISTANCE VECTOR

Keerthika<sup>1</sup> and Malarvizhi<sup>2</sup>

<sup>1</sup>Veltech Dr.RR & Dr.SR Technical University, Avadi, Chennai, TN, INDIA

<sup>2</sup>Dept of CSE, Veltech Dr.RR & Dr.SR Technical University, Avadi, Chennai, TN, INDIA

## ABSTRACT

Mobile Ad hoc Networks (MANETs) are self-configuring networks with nodes being connected through wireless links forming a multi-hop radio network without infrastructure or administration. Security is a major issue in MANETs due to dynamically changing topologies, lack of centralized monitoring, open medium, and bandwidth constraints. It faces security issues not addressed by the security services of infrastructure based networks. Routing performance deteriorates in MANETs due attacks. Ad hoc On-demand Distance Vector (AODV) is a suitable MANET routing protocol that is highly vulnerable to black hole attack by malicious nodes. This study simulates/analyses the impact of black hole attack on an AODV protocol.

Published on: 28<sup>th</sup>– August-2016

### KEY WORDS

Mobile Ad hoc Networks (MANETs), Ad hoc On-demand Distance Vector (AODV), Black Hole Attack.

\*Corresponding author: Email: [keerthivenkatt@gmail.com](mailto:keerthivenkatt@gmail.com); Tel.: +91-9790428279

## INTRODUCTION

Mobile Ad hoc Networks (MANETs) are groups of independent mobile nodes communicating with each other through radio waves (wireless links). A node has a wireless interface to communicate and mobile nodes in radio range communicate directly whereas intermediate nodes are required to route packets to nodes beyond the radio range. Such networks are self-configuring, fully distributed, and work at any place without fixed infrastructure like access points/base stations [1].

MANETs have advantages over conventional networks including reduced infrastructure costs, easy establishment, and fault tolerance as routing is by individual nodes using intermediate network nodes for packet forwarding which reduces bottlenecks. The major MANET attraction is greater mobility compared to wired solutions. MANETs have dynamic topology with mobile nodes having limited resources like battery, processing power and onboard memory. Such infrastructure-less networks are used in situations where ordinary wired networks are infeasible including battlefields and natural calamities. Nodes within transmission range communicate directly or communication is via intermediate nodes which forward packets. Hence, these networks are called multi-hop networks.

Routing protocol design is a challenge due to ad hoc network's dynamism. Routing is an interesting MANET research area, which received tremendous attention from researchers. Guaranteeing delivery and ability to handle dynamic connectivity are important issues for wireless MANET routing protocols. When there is source to destination path for a specific time, routing protocols should deliver data through that path [2].

Routing protocols use metrics to calculate best path to route packets to destinations. Metrics are standard measurements that are number of hops used by a routing algorithm to determine the optimal path for a packet to reach a destination. Path determination process is that routing algorithms initialize/maintain routing tables with a packet's total route information [3].

Routing protocols define rules set, which govern the transmission of message packets from source to destination in networks. In MANETs, there are different routing protocols each used according to network circumstances. Routing protocols based on properties are classified as proactive, reactive and hybrid routing protocols [2].

MANET's dynamic environment's routing protocol security is a challenge. A self-organizing environment introduces security issues not addressed by security services meant for infrastructure based networks. Secure routing protocols deal with malicious nodes that disrupt routing protocol functioning by modifying routing information, fabricating routing information and impersonating nodes [4].

The first line of defence to reduce attacks are attack prevention measures like authentication and encryption. But, these do not suit MANET's resource constraints, i.e., limited bandwidth and battery power, as heavy traffic load emanates to exchange/verify keys.

MANET attacks are classified into passive and active attacks. A passive attack exchanges data in a network without disrupting communications while active attacks involve information interruption, modification/fabrication disrupting MANET functioning.

The attacks are divided into external and internal attacks, according to the attack domain. External attacks are mounted by nodes not belonging to a network whereas internal attacks are due to compromised nodes within the network. Internal attacks are more disruptive as an insider knows and can access confidential information. An attacker in external attacks causes congestion, propagates fake routing information or disturbs nodes from providing services. In internal attacks, an adversary participates in network activities, through impersonation as a new node, or by compromising a current node and using it as a conduit for its nefarious work.

AODV is a reactive routing protocol where routes are determined when required. It exchanges messages [5] and is used for unicast, broadcast, and multicast communication. It adopts basic Route Discovery and Route maintenance mechanism on demand from Dynamic Source Routing (DSR) and Destination Sequenced Distance Vector (DSDV) hop by hop routing sequence number and periodic beacons. This study discusses the impact of black hole attacks on MANETs with AODV routing protocol. Section 2 deals with literature related to this work, Section 3 reveals methods used in the work, Section 4 provides results and discussions of the obtained results, and Section 5 concludes the work.

## LITERATURE REVIEW

Highlighting attack scenarios in multicast routing protocols that exploited vulnerabilities was attempted by Singal et al., [10]. Attacks were from a multicast routing protocols perspective were implemented. They analyzed blackhole, jellyfish drop, and neighbourhood attacks impact on ODMRP routing protocol for MANETs. Some techniques forwarded by researchers to detect and prevent black hole attack in MANETs using AODV protocol were discussed by Sarma et al., [11] and a new methodology based on their flaws was proposed. The effects of the attacks on applied AODV protocol based on performance metrics like throughput, packet drop ratio, normalized routing load, and dropped packets number on parameters like varying speed, nodes, and pause time was exhibited by Chadha and Jain [7]. A measure to reduce black hole attack was also analysed on metrics. A method against black hole attacks in MANETs presented by Narayanan and Radhakrishnan [13] used destination MAC address to validate a node in its path thereby ensuring a direct secure route. Simulation was carried out on the new scheme to prove the effectiveness of the mechanism in attack mitigation while maintaining a reasonable throughput, packet delivery ratio, and end to end delay.

A mechanism like trust based routing, intrusion detection system, sequence number comparison, and data routing information table to overcome black hole attack was proposed by Venkanna and Velusamy [8]. Trust based on-demand routing mechanism identifies and decreases hazards by malicious node on a path. A survey to prevent and identify black hole attack using trust management mechanism in MANET was undertaken. Serrat-Olmos et al., [9] proposed a collaborative approach to detect black holes and selfish MANET nodes, using a Bayesian watchdogs set, which enhanced individual/collective performance. Results revealed that misbehaved nodes detection time was reduced, and false positives and false negatives impact was minimised while overall accuracy increased. The results were confirmed by simulation. The collaborative Bayesian watchdog performed better than standard Bayesian watchdog regarding accuracy and quick detection.

Chaubey et al., [6] proposed a network size, Trust based Secure on Demand Routing Protocol (TSDRP) and AODV routing protocol secured it against black hole attack. AODV is a MANET routing protocol, without in-built security measures and so is vulnerable to attacks. Black hole attack at network layer is a major attack in this routing protocol. It considers average end-to-end delay average throughput, packet delivery fraction and normalized routing load to evaluate performance. Modification of AODV routing protocol was suggested by Wahane and Lonare [12]. A mechanism was used to detect and defend against cooperative black hole attacks. The authors suggested two concepts including Maintenance of Routing Information Table and second being node Reliability checking. This decreased end to end delay and Routing overhead.

Lu et al., [17] proposed and implemented Bad AODV (BAODV) Routing by simulating black hole attack in MANET. SAODV protocol based on BAODV, addressed AODV protocol's security weakness and withstood black hole attack. Analysis showed that SAODV was more secure than basic AODV. DOA and AODV routing protocols used in large scale networks were analysed by

Jeni et al., [14] who used them in black hole attack and evaluated quality parameters like packet delivery ratio and average end to end delay. Sharma and Sharma [15] presented two solutions. The first was locating more than one destination route and the second was exploiting packet sequence numbers in packet headers. Simulation revealed that compared to the original AODV routing scheme, the second solution verified 75% to 98% destination routes depending on pause time with minimum network delay. The objective was to analyse black hole attack in a MANET and its solutions.

Thachiland Shet [16] presented a collaborative approach to mitigate black hole nodes in MANET's AODV protocol where a node monitors neighbouring nodes and calculates their trust value dynamically. When a monitored node's trust value is below a predefined threshold, then the monitoring node assumes it as malicious and avoids it. Experiments revealed that the new scheme mitigated black hole nodes and secured AODV routing protocol for MANETs. Conquering black and gray hole attacks was proposed by Yang et al., [18] whose watchdog mechanism based neighbour observed model detected one black hole attack by focusing on direct trust value. Historical evidence was considered against gray hole attacks. A neighbour recommendation model accompanied by indirect trust value figured out a cooperative black hole attack. Both revealed good results and proved the new method's advantages by punishing malicious actions to prevent attack camouflage/deception.

## METHOD

This section discusses black hole attack's impact on MANETs with AODV routing protocol.

### AD HOC ON-DEMAND DISTANCE VECTOR (AODV)

AODV routing protocol is a MANET routing protocol. AODV router is a state machine processing incoming requests from a network. When a network has to send a message to a node, it asks AODV to determine the next-hop [5]. Though AODV uses DSDV sequence numbers and routing beacons it performs route discovery with on-demand Route Requests (RREQ) similar to DSR protocol. AODV uses sequence numbers to handle node mobility to identify/discard out-dated routes. Route Error (RERR) messages for detecting broken links. RERR packets travel to source informing nodes to delete broken links triggering new route discovery when alternative routes are unavailable.

AODV, unlike DSR is an on-demand, single path, loop-free distance vector protocol using source routing through a hop-by-hop routing approach. AODV is better than DSR for high mobility but has high routing load problems compared to DSR as the latter resorts to aggressive route caching which AODV does not. So, there are researches which solve AODV's problems using cache memory. AODV has DSDV's properties which include DSR's loop free properties using cache memory. A route is created on demand by a network connection in AODV and information regarding a route is stored in nodes routing tables on the route path [19].

AODV shares DSR's on-demand characteristics as it also discovers routes on demand by flooding networks with RREQ packets. A node, on receipt of an RREQ rebroadcasts it unless it is a destination or has a route to it in its cache. The node replies to RREQ with an RREP packet reverted back to the original source. AODV maintains routing information through conventional tables, one entry per destination.

AODV, relies on routing table entries to propagate RREP back to source without source routing and to route data packets to a destination. AODV uses sequence numbers in destinations to determine routing information freshness and to prevent routing loops. Maintenance of timer-based states in nodes is an important AODV feature, on use of individual routing table entries [20].

In AODV, routing updates through RREQs/RREPs are 'route advertisements.' Update rules in [Figure -1] are invoked by nodes on receipt of route advertisements which help maintain loop freedom.

Consider the tuple  $(-seq\_num_i^d, hop\_count_i^d)$  where  $seq\_num_i^d$  represents the sequence number for destination  $d$  at node  $i$ , and  $hop\_count_i^d$  represents the hop count from node  $i$  to destination  $d$ .

Define  $(-seq\_num_i^d, hop\_count_i^d) > (-seq\_num_j^d, hop\_count_j^d)$  if and only if either  $seq\_num_i^d < seq\_num_j^d$ , or  $seq\_num_i^d = seq\_num_j^d$  and  $hop\_count_i^d > hop\_count_j^d$  (i.e., lexicographic ordering among  $(-seq\_num_i^d, hop\_count_i^d)$  tuples).

A node  $i$  applies these rules on receipt of a route advertisement for destination  $d$  from a neighbour  $j$ . Variables  $seq\_num_i^d$ ,  $hop\_count_i^d$ , and  $next\_hop_i^d$  denote destination sequence number, hop count and next hop respectively for destination  $d$  at node  $i$ .

```

1: if (seq_numid < seq_numjd) or ((seq_numid = seq_numjd) and (hop_countid > hop_countjd)) then
2:   seq_numid := seq_numjd;
3:   hop_countid := hop_countjd + 1;
4:   next_hopid := j;
5: end if
    
```

Fig:1. AODV route update rules

**BLACK HOLE ATTACK**

Black hole attack is a Denial of Service (DoS) attack in MANETs where a malicious node advertises about a best path it has to a destination node during route discovery. When it receives a RREQ message, it sends a fake RREP to source node immediately. The source node receives the RREP from the malicious node before other RREPs. But, when the source node starts sending data packet to the destination using the given route, the malicious node drops packets instead of forwarding [21] them.

A black hole problem is seen in [Figure -2], where node "A" asks node "D" to send data packets and starts discovering a path. So, if node "C" is a malicious node, it claims it has a positive route to a specific destination, till the road receiving a request (RREQ) packet is open. It responds to node "A" through any node. Thus, node "A", ie; on the path takes a positive discovery initiative. Node "A" ignores other responses and plants package node "C". So, all lost packets are consumed/lost.

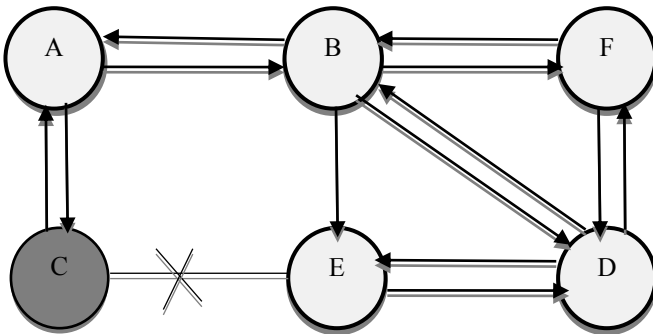


Fig: 2. Black hole attack in AODV

Detection is complicated in MANETs due to limited resources like bandwidth, battery life, and storage. Concerns about minimum possible rise in routing overhead and end-to-end delay affect implementing any detection process. Related research implemented detection method against AODV routing through Dynamic Learning Method[22].

An attacker sends fake RREQ messages to form black holes. In a RREQ black hole attack, an attacker pretends to rebroadcast a RREQ message with a non-existent node address. Other nodes update their route to bypass non-existent node to the destination node. The attacker forms a black hole between source and destination nodes through a fake RREQ message.

**RESULT AND DISCUSSION**

Network simulation was through NS2, and the performances of AODV were compared for non-malicious and malicious network. Simulations are conducted with number of nodes that 30, 60, 90, 120 and 150. The nodes are spread over an area of 4000 sqm and has a transmission range of 200 m. Traffic transmitted are in Constant Bit Rate (CBR). The impact of AODV without malicious nodes, with 15% malicious node, with 30% malicious node and with 45% malicious node. The simulation parameters are summarized in [Table- 1].The results achieved for packet delivery ratio, average end to end delay and number hops to sink are presented in this section.

Table: 1. Simulation Parameters

Parameter	Value
Number of Nodes	30, 60, 90, 120 and 150
Network area	4000 m <sup>2</sup>
Transmission range	200 m
Traffic	CBR
Routing	AODV
Maliciousness	15%, 30% and 45% maliciousness

Table: 2. Average End to End Delay in second

Number of nodes	AODV without malicious nodes	AODV with 15%malicious	AODV with 30% malicious node	AODV with 45% malicious node
30	0.00086	0.00109	0.00131	0.0013
60	0.00107	0.00129	0.00173	0.00152
90	0.0013	0.00281	0.00356	0.00173
120	0.00129	0.0042	0.00534	0.00182
150	0.00794	0.01287	0.01639	0.01093

It can be observed from [Table -2] that the maliciousness in network tends to increase the end to end delay. It is observed from the simulation results that the end to end delay increases by 28.38% to 40.74% for a 45% malicious network.

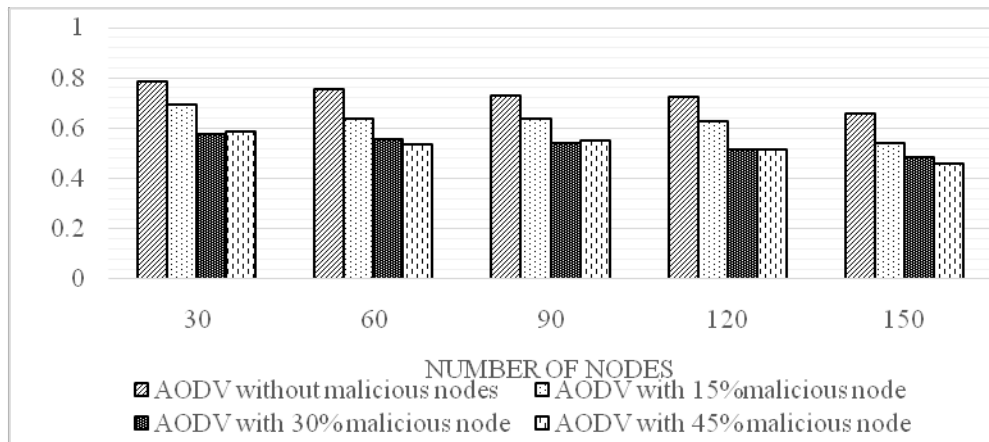


Fig: 3.Packet Delivery Ratio

It can be observed from [Figure -3] that the packet delivery ratio decreases with increase in number of nodes and increase in maliciousness in the network. The packet delivery ratio ranges from 0.66 to 0.79 for AODV without maliciousness in the network. As the maliciousness increases the delivery ratio decreases drastically. For a network with 15% maliciousness, the packet delivery ratio decreases in the range of 12.36% to 19.6%.

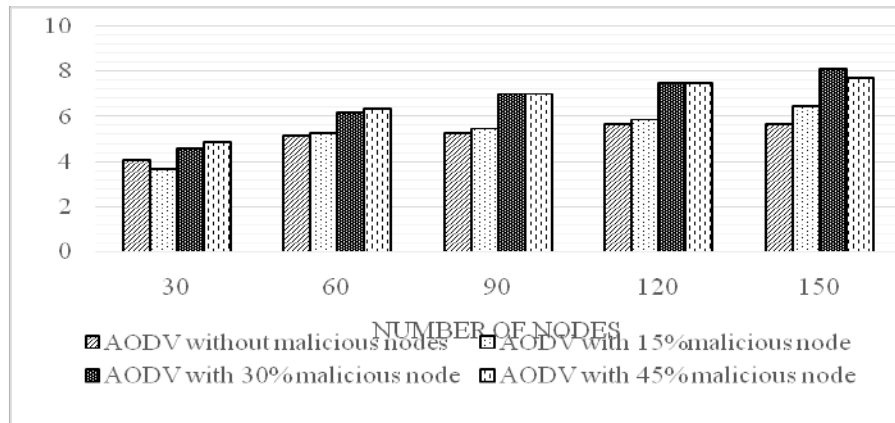


Fig: 4. Average Number of hops to sink

It can be observed from [Figure -4] that the number of hops increases as the maliciousness in the network increases.

## CONCLUSION

This study analysed the effect of black hole attack in AODV protocol performance. Metrics like packet delivery ratio, average end to end delay in second and average hops to sink were evaluated and analysed with variable node mobility, number of nodes. Simulation shows that when black hole node exists in a network, it can affect and decrease performance of AODV routing protocol. So black hole attack detection and prevention in a network is a challenge. Future direction of work to create a solution for black hole attack and to compare its performance with AODV

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Aarti DS. [2013] Tyagi, Study Of Manet: Characteristics, Challenges, Application And Security Attacks". *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(5):252-257.
- [2] Rhee I, Shin M, Hong S, Lee K, Kim SJ, Chong S. [2011] On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, 19(3): 630-643.
- [3] Gorantala K. [2006] Routing protocols in mobile ad-hoc networks. *A Master thesis in computer science*, -1-36.
- [4] Zhang D, Gogi SA, Broyles DS, Çetinkaya EK, Sterbenz JP. [2012] Modelling wireless challenges. In *Proceedings of the 18th annual international conference on Mobile computing and networking*, ACM, 423-426.
- [5] Chakeres ID, Belding-Royer EM. [2004] AODV routing protocol implementation design. In *Distributed Computing Systems Workshops, 2004. Proceedings. 24th International Conference on*, IEEE, (pp. 698-703)
- [6] Chaubey N, Aggarwal A, Gandhi S, Jani KA. [2015] Performance Analysis of TSDRP and AODV Routing Protocol under Black Hole Attacks in MANETs by Varying Network Size. In *Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on* (pp. 320-324). IEEE.
- [7] Chadha K, Jain S. [2014] Impact of black hole and grayhole attack in AODV protocol. In *Recent Advances and Innovations in Engineering (ICRAIE), 2014* (pp. 1-7). IEEE.
- [8] Venkanna U, Velusamy RL. [2011] Black hole attack and their counter measure based on trust management in manet: A survey. In *Advances in Recent Technologies in Communication and Computing (ARTCom 2011), 3rd International Conference on* (pp. 232-236). IET.
- [9] Serrat-Olmos MD, Hernández-Orallo E, Cano JC, Calafate CT, Manzoni P. [2012] Accurate detection of black holes in MANETs using collaborative bayesian watchdogs. In *Wireless Days (WD), 2012 IFIP* (pp. 1-6). IEEE.
- [10] Singal G, Garg H, Laxmi V, Gaur MS, Lai C. [2014] Impact analysis of attacks in multicast routing algorithms in MANETs.

- In Industrial and Information Systems (ICIIS), 2014 9th International Conference on (pp. 1-6). *IEEE*.
- [11] Sarma KJ, Sharma R, Das R. [2014] A survey of Black hole attack detection in Manet. In *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on* (pp. 202-205). *IEEE*.
- [12] Wahane G, Lonare S. [2013] Technique for detection of cooperative black hole attack in MANET. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1-8). *IEEE*.
- [13] Narayanan SS, Radhakrishnan S. [2013] Secure AODV to combat black hole attack in MANET. In *Recent Trends in Information Technology (ICRTIT), 2013 International Conference on* (pp. 447-452). *IEEE*.
- [14] Jeni PJ, Vimala Juliet A, Parthasarathy R, Messiah Bose A. [2013] Performance analysis of DOA and AODV routing protocols with black hole attack in MANET. In *Smart Structures and Systems (ICSSS), 2013 IEEE International Conference on* (pp. 178-182). *IEEE*.
- [19] Al-jubori MJ, Pawale SS, Shinde SR. [2011] Efficient Ad-Hoc On-demand Distance Vector Routing Protocol using Link State Algorithm. *International Journal of Computer Applications*, 26(2).
- [20] Barua G, Agarwal M. [2002] Caching of routes in ad hoc on-demand distance vector routing for mobile ad hoc networks. In *proceedings of the international conference on computer communication*, 15( 2): 768
- [21] Chaudhary A, Malhotra P. [2014] Impact of Black Hole Attack on AODV Routing Protocol. In *International Journal of Engineering Development and Research, IJEDR* 2( 3) .
- [22] Khandelwal V, Goyal D. [2013] BlackHole Attack and Detection Method for AODV Routing Protocol in MANETs. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 2(4):1555.
- [15] Sharma N, Sharma A. [2012] The black-hole node attack in MANET. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on* (pp. 546-550). *IEEE*.
- [16] Thachil F, Shet KC. [2012] A trust based approach for AODV protocol to mitigate black hole attack in MANET. In *Computing Sciences (ICCS), 2012 International Conference on* (pp. 281-285). *IEEE*.
- [17] Lu S, Li L, Lam K.Y, Jia L. [2009] SAODV: a MANET routing protocol that can withstand black hole attack. In *Computational Intelligence and Security, 2009. CIS'09. International Conference on* , *IEEE*, 2:421-425.
- [18] Yang B, Yamamoto R, Tanaka Y. [2014] Dempster-Shafer evidence theory based trust management strategy against cooperative black hole attacks and gray hole attacks in MANETs. In *Advanced Communication Technology (ICACT), 2014 16th International Conference on* (pp. 223-232). *IEEE*.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

## RESEARCH BASED LITERATURE SURVEY AND ANALYSIS ON VARIOUS SHARDING TECHNIQUES

Sasikala

School of Computer Science and Engineering, VIT University, Vellore, INDIA

## ABSTRACT

**Aims:** Bigdata applications with large data set and complex structure can challenge the capacity of single server. Moreover the Bigdata applications with high query rate overload the capacity of a single server. Hence to address the issues, database servers follows two techniques : **Materials and Method:** Vertical scaling and Horizontal scaling. Increasing the capacity of the server(vertical scaling) leads to more expensive and the problem of single node failure. Hence the Enterprises supports partition(sharding) and distribution of database into multiple servers called Horizontal scaling. Partitioning the data with multiple backups(replication) make the system scalable and fault tolerant. **Results:**The paper investigates various sharding and replication schemes and analyze the performance parameters and issues. The paper identified various research issues from the literature survey on partitioning and sharding techniques such as dynamic sharding, initial partitioning with an analysis of load balancing, identifying novel domain based sharding schemes and keys, hybrid solutions like MongoDB-Hadoop connector, sharding techniques on complex queries and In-memory computing techniques. **Conclusions:**The paper assists the researchers to move forward to find solutions on setting up scalable and reliable shards in data centres.

Published on: 28<sup>th</sup>– August-2016

## KEY WORDS

Bigdata, NoSQL, Partitioning, Sharding, Replication, Sharding Key

## INTRODUCTION

Infection is a dynamic process involving invasion of body tissues by pathogenic micro-organisms and their toxins. Nosocomial/ hospital/ acquired infections are those which are not present or incubated before admission of patient to the hospital but obtained during the patient's stay in hospital. Lab coats, nurses' uniforms and other hospital garments, materials and articles may play an important part in transmitting pathogenic bacteria in a hospital setting .The hands of healthcare personnel are most commonly implicated in transmitting the pathogens [1]. Various nosocomial pathogens, such as methicillin-sensitive *Staphylococcus aureus* (MSSA), methicillin-resistant *Staphylococcus aureus* (MRSA), vancomycin-resistant *Enterococci* (VRE) and gram negative organisms is well documented [2]. Specifically in the area of dentistry, health care professionals are routinely exposed to potentially pathogenic microorganisms which are present in the surrounding environment. Most of them originate from the mouths of patients [3]. Contamination may occur from instruments through contamination vectors. These contaminated object infections may be transferred from patient to patient or from patient to professionals [4]. Methicillin resistant *Staphylococcus aureus* which is the most pathogenic microorganism, comes in contact with health care professionals via direct hand contact with contaminated body fluids, devices, items or environmental surfaces [5].

There are very few studies regarding the wearing and laundering of lab coats in hospitals and medical practice. This study highlights the role of lab coats acting as vector for transmitting health care infections to the patients and the common areas where contamination occurs.

Internet of Things(IoT), Enterprise data, Bio medical, scientific and Web applications data are growing every day and also the structure of data are getting more complex. The way the applications storing Databases are varying by number of data in each record, data type of each field, complex format and hence structure of Database plays significant role. Generally the term Bigdata describes massive amount of data. In addition, the term Bigdata [1] defines unstructured nature of data as 3v model : Volume, Velocity and Variety [2]. Here Volume refers size of data, Velocity refers timeliness and Variety refers unstructured nature of database. The massive amount of data creates more opportunities in business, social network and research world through data analytics, data mining algorithms



and data storage. Various technologies are used for handling Bigdata. The four different phases of handling Bigdata are : Data generation, Data acquisition, Data storage and Data analysis.

**Data Generation** : Internet, Web applications, IoT, Business, Research and Scientific applications are various sources of generating massive amount data. Not only the data size is big, the data structure are semi structured and unstructured mostly.

**Data acquisition** : Data acquired from various sources such as sensors, social network and packet capturing technologies called acquisition. The data are stored locally in centralized server before transfer and data are applied to pre-processing technologies to remove unwanted and noisy data. These three steps are under Data acquisition.

**Data Storage** : Various storage architectures are used for storing the data permanently and future use. This storage technologies assist the system to improve performance of Bigdata management like scalability, reliability, availability, consistency, query performance and fault tolerance. The technologies used for handling database are NoSQL(No SQL) data store[3], Partitioning(Sharding) & Replication, distributed data nodes and MapReduce.

**NoSQL Datastore** : Relational Database management system is popular Database technology for decades. Relational databases supports ACID (Atomicity, Consistency, Isolation, Durability) properties for data management systems. Now Bigdata systems led to creation of new technologies. One such technology is NoSQL databases. NoSQL technologies have been popular from 2009 for non relational databases. Various NoSQL Databases evolved such as MongoDB, Cassandra, HBase, CouchDB etc. Also the data models key-value data store, Column, document oriented data store, Graph databases are used for database storage.

**Partitioning(Sharding) & Replication** : As database size increases, single server may not be sufficient to store entire data and rises an issue of centralized failure. Hence entire database is partition(shard) into chunks and distributed into multiple nodes. Failover mechanism is achieved by replication and replication factor is normally fixed from 3. The partitioning and replication achieves high availability, scalability, reliability and faster query execution time. Sharding key is used to assist data partitioning but sharding key cannot be fixed and common for all the applications. The objective of this research review paper is to discuss various database partitioning & replication strategies, sharding keys to partition the data and performance analysis of various schemes. Section 2 describes survey papers on Bigdata handling mechanisms and NoSQL DB stores. Section 3 describes various sharding schemes, sharding keys and analysis of performance parameters. Section 4 gives conclusion and future research directions on sharding.

## EXISTING SURVEY ON BIG DATA

From 2009, various NoSQL and Bigdata handling technologies are developed and issues of bigdata management are solved effectively. The existing survey papers describes general study on Bigdata, Bigdata management schemes and NoSQL DB technologies.

### *Survey of Large-Scale Data Management Systems for Big Data Applications*

Lengdong Wu, Liyan Yuan et al have done detailed survey on large scale data management system for Bigdat, describe Data management model, consistency model and architecture and analyse critical aspects and scalability limitations[4]. The data management systems are divided into two categories : Relational and Non Relational. The relational data model are traditional data model to support database of type structured. The non relational database model supports Big data with semi and unstructured databases. For this research review we have taken only non relational data part of this survey paper[4]. Non relational data model representatives are Key-value data store, Hadoop and Big Table. The Data model are classified as Conceptual and Physical data model. How the data is stored in database is called as Physical data model and how schema is used to represent the structure of database is called Conceptual data model. The data model is classified further as Structured, Semi structured and unstructured in conceptual model and Row oriented, Column oriented and Hybrid oriented in physical model. Consistency model is classified based on two properties: ACID and BASE. Moreover my research review required how big data is handled by contemporary method of Database storage architecture. Hence this survey paper concentrated the technologies related to storage system architecture of Big Data management. The System architecture of Big data management are classified as,

**Symmetric Multi-Processing (SMP) on Shared-Memory Architecture** : It consists of same type(homogenous) and tightly coupled pooled processors. The data will be stored on shared memory. When the database size increases, complex nature of this architecture is not suited to scale well. The paper concludes that SMP on shared memory has limited scalability for Big data management.

**Massively Parallel Processing (MPP) on Shared disk Architecture** :The program will be executed parallelly on SMP cluster of nodes with Shared disk is scalable. But the system is not fault tolerant due the centralized behaviour of shared disk.

**Sharding on Shared-Nothing Architecture** : This is the technique widely used by large scale big data management systems. The data will be distributed among number of nodes and process on shards will be parallel. The nodes can share the routing data using centralized server or all the nodes are treated equally called decentralized topology[5-11].

**Mapreduce/Staged Event-Driven Architecture(SEDAs)**:Some systems use MapReduce and SEDA as hybrid solution[12].

Finally this survey paper[4] concluded that MapReduce/SEDA architecture, unstructured data model with strong consistency and BASE properties are well suited for handling Bigdata management systems.

### ***Choosing the Right NoSQL Database for the Job: a Quality Attribute Evaluation***

Ricardo Lourenço<sup>1</sup>, Bruno Cabral et al have done survey on NoSQL Technologies and analyse quality attributes which helps to map NoSQL technology to particular usecase or job[13]. The paper summarized the following points :

- A survey of the literature on NoSQL Technologies
- Future research directions on NoSQL technologies with respect to software quality attributes
- Identifying the NoSQL technology to particular Job

The authors have done detailed study on NoSQL databases Aerospike, Cassandra, Couchbase, CouchDB, MongoDB, Voldemort and HBase and analysed quality attributes on these technologies. For this research survey, only Partitioning scheme and Native partitioning are taken into consideration

**Aerospike** : A key-value data store with Proprietary based partitioning scheme and it basically supports native partitioning. It performs auto partitioning and replication on several layers[14].

**Cassandra** : A Column oriented data store with Consistent Hashing partitioning scheme and it basically supports native partitioning. It supports different types of partitioning schemes and auto replication[15,16].

**Couchbase** : A document oriented data store with Consistent Hashing partitioning scheme and it basically supports native partitioning. It allows partitioning and performs inter and intra cluster sharding schemes[14].

**CouchDB** : A document oriented data store with Consistent Hashing partitioning scheme and it does not support native partitioning. It does not support sharding and allows master-master or master-slave replication[14].

**MongoDB** : A document oriented data store with consistent hashing partitioning scheme and it basically supports native partitioning. It supports both partitioning and replication[14,15].

**Voldemort** : A key-value data store with consistent hashing partitioning scheme and it basically supports native partitioning. It supports replication and partition. Also it can add and remove nodes dynamically for partitioning[14].

**HBase** : A Column oriented data store with Range based partitioning scheme and it basically supports native partitioning

Authors have taken software quality parameters : Availability, durability, maintenance, consistency, performance, reliability, robustness and scalability. Considering above parameters, the NoSQL data store MongoDB, Cassandra and Aerospike strongly supports 4 parameters. Next Couchbase supports 3 parameters and other data stores supports less than three. With the above parameters and theoretical study on different type of NoSQL data store Cassandra, Aerospike and MongoDB supports most of the important parameters. This is listed in Table 1. Also authors showed research direction as, this work can be extended with implementation of a particular usecase on all the above datastores directs the Software Architects to choose right kind of data store.

Table 1 Analysis of Bigdata Management Survey Papers

Survey Paper Title	Survey on	Nature of the paper	Parameters	Conclusion	Future scope
<b>Survey of Large-Scale Data Management Systems for Big Data Applications</b>	Large scale data management systems	Study	Storage architecture, data model, Consistency	MapReduce/SEDA architecture, unstructured data model with strong consistency and BASE properties are well suited for handling Big data management systems	Implementation for detailed analysis
<b>Choosing the Right NoSQL Database for the Job: a Quality Attribute Evaluation</b>	Aerospike, Cassandra, Couchbase, CouchDB, MongoDB, Voldemort and HBase	Study	Availability, durability, maintenance, consistency, performance, reliability, robustness and scalability	The NoSQL data store MongoDB, Cassandra and Aerospike strongly supports 4 parameters. Next Couchbase supports 3 parameters and other data stores supports less than three.	Implementation of usecase with all DBs to perform performance analysis

## SURVEY ON SHARING SCHEME

Sharding schemes performance is depends on sharding key, number of shards, load balance and distribution of related information. The sharding key plays an important role to fix number of shards, load of each shard and distribution behaviour. To partition the data, shard key need to be selected effectively. In addition, to increase the query execution performance and improve migration cost, shard key must be fixed and static. The existing schemes Range and Hash based sharding are not well suited for all the applications and node balancing. Hence various sharding schemes are developed and analyzed its performance with existing schemes.

### Knowledge Driven Query Sharding

Adam Krasuski and Marcin Szczuka proposed how the knowledge of data structure is used to perform sharding and proved that how sharding improves performance of data analytics on large database[17]. The proposed work provides the solution to search the repositories of scientific information using semantic content. The system is called SONCA(Search based ONtologies and Compound Analytics). The queries taken by the authors involve join and GROUP By operations on large databases and the resultant database size exceeds RAM allocation. Here Sharding is used for decomposing complex queries into small queries(Query Sharding). The smaller queries are executed concurrently and independently in a multicore processor or multiple machines in a networked environment. The Explicit Semantic Analysis(ESA) method assists to find semantic relationship between documents and the knowledge base MeSH. The MeSH is a knowledgebase consists of vocabulary for the purpose of indexing journal articles and book in the life sciences. The experiments are conducted with and without sharding. Three major database technologies used for experiments : Infobright(column Oriented), PostgreSQL(row oriented) and MongoDB(column oriented). The results showed that query sharding in the SONCA system utilize the computing resources optimally and execution time is considerably less than traditional(without sharding) technique.

### Clustering-based Fragmentation and Data Replication for Flexible Query Answering in Distributed Databases

Flexible query assist the system to find related information if query cannot be answered exactly. Lena Wiese proposed the clustering based fragmentation and replication for finding related information in a sharded or replicated databases if exact information is not retrieved[18]. The paper proposed clustering based fragmentation and derived fragmentation to distribute the database into cluster of nodes. Here fragmentation means divides the database into number of fragments and fragments are assigned into servers. In addition the paper suggested query rewriting and redirecting to decompose and direct query into multiple servers. The main focus of this paper is how sharding and replication improves performance. Hence we focus more on fragmentation and replication part of this paper.

**Bin Packing Problem** : Objects of different volumes must be stored in a finite numbers of bins and provides a way to minimize number of bin used.

**Replication** : To achieve availability, reliability and fault tolerance, fragmented data should be replicated into several servers. An extension of Bin Packing Problem(BPP), BPP with conflict proposed two constraints: Conflict objects(fragments) should not be placed in same server and fragmented part of database stored in one sever will be replicated in  $m-1$  other servers, where  $m$  is a replication factor.

**Clustering Fragmentation**: The fragments are clustered and number of cluster are formed to support flexible query system. Lena Wiese proposed the following rules as Definition[18] :

- Horizontal fragmentation
- Clustering
- Threshold
- Completeness

- Re-constructability
- Non-redundancy

**Derived Fragmentation:** In fragmentation, data accessing together should also be in same cluster or server will make the system faster. This is achieved by derived fragmentation technique. To achieve better locality, derived fragmentation also supports redundancy in fragments. That is few fragments might be stored in more than one servers to give better performance in query answering.

Implementation is based on PostgreSQL and the UMLS and the database is MeSH taxonomy. The results shows high execution time and number of fragments, when row count is more than 1000. The authors suggested to improve scalability by the implementation of parallel clusters

#### **Social BIMCloud: a Distributed Cloud-based BIM Platform for Object-based Lifecycle Information Exchange**

The AEC(Architecture, Engineering, Construction) industry adopted the BIM information exchanges through networks for sharing files, designs, models etc. The current file based information exchange leads problems like slow data transfer, lack of interoperability etc. Moumita Das, Jack CP Cheng et al proposed Social BIMCloud, a simple and less expensive distributed cloud based BIM platform for information exchange[19]. The existing BIM do not provide any platforms for social interaction between teams. Hence the author proposed a Social BIMCloud that integrates NoSQL database management systems, dynamic splitting and merging of BIM information, social interactions and partitioning(sharding) of BIM models. The Social BIMCloud framework divided into three layers: data capture and flow controller layer, data upload and extraction layer, and data storage layer.

- The data capture and flow controller layer obtains the input from users in four different ways : BIM files, web pages, BIM software and system commands
- The data upload and extraction layer receives data from input extracts the key information then convert it into NoSQL format
- The Data Storage layer uses partitioning data into number of shards. Partitioned data or shards stores in different nodes of cloud

The Data storage layer of BIMCloud proposed partitioning of Big BIM model using horizontal partitioning and fragments are stored in cloud nodes. This shards facilitate the parallel read and write and faster query execution rate. The horizontal partitioning makes the faster query performance. The BIMCloud also supports automatic replication and replication factor is decided by the end users. The Social BIMCloud supports

- large volume of data
- data can be added dynamically with different size
- dynamic schema

The BIMCloud supports column oriented data model. Each row consists of unique key with set of column values. The value of column can be added dynamically. The column family has two types : Regular column and Super column. The super column supports tree like data structure i.e., one column is nested with another column. Since Building elements consists of many properties and sub layers, BIM cloud supports tree like data structures. New instances can be created and deleted. For example, more number of cloud instances are needed during construction phase than maintenance phase. For demonstration, the BIMCloud is deployed and Tested using Tomcat webserver, Cassandra NoSQL database and PHP scripting language and hosted in Amazon Web Services(AWS).

#### **Research on the Improvement of MongoDB Auto-Sharding in Cloud Environment**

MongoDB Auto sharding creates shards in network environment without external intervention but most of the time data are not distributed evenly among shards. Yimeng Liu, Yizhi Wang et al proposed an algorithm which effectively balance the data among shards[20]. Data in shards are distributed evenly based on Frequency Of Data Operation(FODO). The paper focuses MongoDB Auto sharding which supports

- Auto Balancing
- Easy addition of new servers, if needed
- No single point of failure
- Automatic failover

MongoDB uses shard key for distributing the data among servers. The key assists to create the chunks and chunk will be divided again, if size is increasing beyond the limit. This is called auto sharding. The components of MongoDB sharding are

Shards : Servers which stores partitioned data

Config servers : Servers stores metadata which gives the details of which data stored in which server

Mongos : Servers receiving and directing requests from user to shards.

Auto-sharding moves the chunk among shards based on the size, but the data operation is not taken into consideration. Hence, Auto-sharding cannot achieve effective balance. The proposed algorithm takes FODO as value based on number of Insert, Delete and Update operations on particular chunk. The FODO algorithm works as follows:

Step 1: Calculates FODO<sub>i</sub> value for each chunk.

Step 2 : If number of chunks between shards is greater than 8, then balancer algorithm will start work till it reaches 2.

Step 3 : Migrate the chunk from one shard to another shard based on FODO<sub>i</sub> value. Move the chunk to front shard if FODO<sub>i</sub> value higher and keep lower FODO<sub>i</sub> value shard at the end.

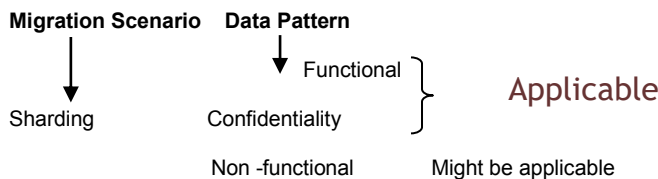
The implementation part chooses 10000000 documents(records) of random data and auto sharding cluster consists of 10 virtual machines. The results showed that performance of concurrent read and write operations is improved than traditional auto-sharding algorithm.

### Migrating Application Data to the Cloud using Cloud Data Patterns

The advantage of cloud computing is to design application for cloud or after implementation, we can migrate the application to cloud for utilizing resources economically. Steve Strauch, Vasilios Andrikopoulos et al analyzed various migration scenarios and characteristics to migrate the application data to the cloud[21]. This paper mentioned the following scenarios :

- Database layer outsourcing : Complete or partial database layer outsourcing without changing the datastore.
- Using Highly scalable data store : Migrating non-highly scalable database to NoSQL or BLOB data store
- Geographical Replication : Replicating and maintaining consistency of the database layer
- Sharding : Distributes the data into cluster of nodes and distribution can be done geographically
- Cloud bursting : outsourcing database temporarily to cloud for managing resources
- Working on Data copy - Keeping complete or partial database layer on cloud
- Data synchronization - Working with database offline and synchronize with the cloud copy to maintain consistent database

In the direction of supporting application data migration to cloud, authors have done literature research focusing on cloud data patterns. The research is focused reports from various industries who already moved to the cloud for application layer data. Strauch et al., says " A Cloud Data Pattern describes a reusable and implementation technology independent solution for a challenge related to the Data Layer of an application in the Cloud for a specific context" [21]. There data patterns have been identified : functional, non functional and confidentiality. The next step is mapping of migration scenarios to cloud data patterns. Various migration scenarios and its data patterns are discussed in the paper. For example,



In the case of cloud data store selected for migration does not support sharding, to scale the database layer read and write operations, Sharding-based Router is used. It decides which part of database layer should not move into cloud store, for example business confidential data. Sharding based router direct confidential data to local server instead of public cloud. Geographical replication can also be combined with sharding when we replicate frequently accessed data. For Health Insurance Company(HIC), the client health records are stored in local server and financial transaction are stored in both local and public cloud. Moreover part of financial transaction are necessary for auditing, hence critical data filter and Pseudonymizer are used for configuration before Router based sharding operation. The authors guiding researchers to analyze various mapping of scenarios to data patterns and choose one among to implement as a single framework. The authors also concluded that the query results should not contain any client medical records and queries from the auditors must be analyzed in advance to save confidential data. This review focused more on sharding part and other mappings also discussed in the literature review[21].

### Point Collection Partitioning in MongoDB Cluster

Spatial data has its own properties such as topological relationships, spatial locality and spatial resemblance. Partition and distribute the data into cluster of nodes and balance the data size on node is a critical issue. Shuai Zhang and Bolei Zhang et al proposed the following techniques for applying sharding on spatial data[22]:

**Random Partitioning Strategy (RPS)** - In order to shard the collections of data, shard key must be required. The shard key must be in every document so that data can be partitioned. RPS chooses the objectID(like Serial no) as the shard key. Hence partitioning is random

**Space filling Curve Partitioning Strategy(SPS)** - It is a mapping technique to map multi dimensional data to 1-D data. The data is divided into multiple blocks and thread is passing through all the blocks atleast one time to create a single shard. This will be repeated to create cluster of shards.

**K means Partitioning Strategy (KPS)** - It creates K centroids (shards) and the partitioning of data into number of shards is based on K Mean clustering formula. Distance measure between a data point and the cluster centre , is an indicator of the distance of  $n$  data points from their respective cluster centres. The above three algorithms are applied on 3 set of data and the average response time was measured. The results concluded that KPS provided better results than SPS and RPS techniques.

### Data Management in Cloud Environments: NoSQL and NewSQL Data Stores

Katarina Grolinger<sup>1</sup> and Wilson A Higashino et al have done detailed study on NoSQL and NewSQL data stores considering various features such as read and write requests, partitioning, replication, consistency and concurrency control[23]. In addition

this paper identified challenges, benchmarking criteria and diversity and inconsistency of terminologies. They have done detailed analysis of various usecases on different sets of data. This paper focusing on three scalability oriented parameters partitioning, Replication, consistency and concurrency.

**Partitioning** : Two techniques of horizontal partitioning techniques are discussed: range partitioning and consistent hashing. In range based partitioning, range of data on particular key will be grouped into one shard. Hence adjacent data will be stored in a same shard. The disadvantages are unbalance shards and processing node always be focused on one or cluster of shards. The data will be represented as ring and the ring will be divided into ranges based on number of shards. The range of data will be assigned to particular shard based on hashing algorithm. Partitioning Graph database is more challenging task than other NoSQL Databases. The NewSQL databases clustrix and VoltDB use consistent hashing algorithm for partitioning. Google spanner uses spanservers for partitioning.

**Replication** : Replication can be done in three ways : Master-Slave, Master-Master and Master less replications. In Master-Slave, one node is assigned for write operation(Master). slave nodes are informed by Master. In Master-Master, all the nodes are updated and communicated simultaneously. In Masterless is working same as Master-slave, but all the nodes are treated equally.

#### An Overview of Newly Open-Source Cloud Storage Platforms

Gu, Genqiang et al provided an overview of open source cloud stores in terms of architecture, implementation technology, architecture and cloud storage[24]. They have taken MongoDB, Cassandra and Hadoop open source technologies for comparison. This paper focusing more on storage architecture for analysis which gives insight to readers to choose technologies for their applications.

Storage mechanisms of Cassandra and MongoDB are taken for reviewing.

**Cassandra** : Cassandra supports Column Family based data model which is more effective than traditional key-value data store. Cassandra follows Write-back cache technique for writing and updating the data into database. Before updating, the commitlog file should be submitted. Then the data will be updated in Memtable. Memtable content will be flushed into disk batch wise after some conditions are satisfied. The data are stored as Sorted String Table (SSTable). Read operation is not allowed before the write. Bloom Filter algorithm is used for searching the data by key.

**MongoDB** : MongoDB supports document oriented database. The row of table is called document. The storage architecture of MongoDB comprises of Config server, Shard server and Router Process. Shard server contains actual database partition and its replicas. Config server assists to identify which data stored in which shard. Route process redirect the client request to exact shards with the help of Config server.

The storage mechanism of Hadoop is Hadoop Distributed File Systems(HDFS), Abicloud is appliances repository virtual storage systems, Cassandra is structured key value and MongoDB is document oriented data store. The technology used in storage mechanism for Hadoop is multiple replicas of data blocks, Abicloud is virtual storage factory, cassandra is column family data store and MongoDB is auto-sharding. Authors concluded that there is a perfect balance between performance and the scalability, MongoDB has been selected as storage mechanism for their own experiments.

#### High Dimensional Biological Data Retrieval Optimization with NoSQL Technology .

In medicine studies, microarray experiments data are stored and accessed frequently. The RDBMS technologies used for data warehouse transSMART for accessing patient gene expressions make the query performance slow. Wang, Shicai, et al. introduced a key value data model implemented in HBase to support faster queries on large scale microarray[25]. The result analysis shows that the new data model on HBase outperforms traditional RDBMS and NoSQL Data store MongoDB.

Microarray Data model Implementation using HBase:

**Determining Column Key and Row key** : Transcriptomic Database is taken for implementation. PATIENT\_ID and TRAIL\_NAME are used as row key for two reasons: Column Family can manage single data type easily and BigTable cache will help to retrieve the data related to single patient from StoreFiles quickly.

**Optimizing Row Key to Speedup Query Performance**: Order of placing keys in a row is also important. Here authors recommended to place TRAIL\_NAME before PATIENT\_ID. The composite key with proper placement keys will assist to retrieve the identical patient easily.

**Optimizing Column Key to Increase Cache Hit Ratio** : Design a column key which consists of different type of data to increase hit rate. For example, GENE\_SYMBOL+PROBESET\_ID as column key retrieve millions of record of patients who have same type of PROBESET\_ID. The theoretical performance of ideal key value data model performance is 83% higher than RDBMS model. Also the key value data model is implemented using HBase with dataset loaded in transSMART. Various test cases are tried with HBase, MongoDB and RDBMS and the results showed that HBase retrieval rate is higher than MongoDB and RDBMS. Authors also concluded that RDBMS consumes more memory than NoSQL data store

#### Sharding for Literature Search via Cutting Citation Graphs

Haozhen Zhao proposed a sharding policy for search document using cutting citation and co-citation graphs[26]. Scientific literature is growing rapidly every year, hence the shards are used to partition and distribute the documents into different nodes.

Hence the data are distributed to number of nodes. Distributed Information Retrieval(DIR) is required to enhance the performance. DIR has the following features:

- Parallel execution of the queries : Speedup the process
- Searching promising shards : instead of searching all the shards, search only promising shards gives faster response
- Small size shards : even it is unscalable, it makes effective search on specific shards

First step is to build a shards based on citation and co-citation graph partition. Secondly, effectiveness of sharding based on citation and co-citation graph cutting is experimented and results are demonstrated. Citation and Co-citation graphs are constructed based on documents are considered as nodes and edges are relationship of citation and co-citation. The subgraphs are constructed to build meaningful shards. Hence the shard contains the documents under same category makes searches are easy and fast. Here iSearch test collection and Graph clustering algorithm Graclus is used for experiments. The results showed that the cutting citation and co-citation graph effectively producing relevant information retrieval. The experimental results also showed that co-citation graph performance is better than citation graph.

### **Morphus: Supporting Online Reconfigurations in Sharded NoSQL Systems**

Mainak Ghosh and Wenting Wang et al proposed a system called Morphus proposed the features and incorporated into MongoDB[27]:

- Online reconfiguration
- allows read and write operations concurrently
- Flexible data placement
- Master-slave replication
- range partitioning

Here reconfiguration phase is significant which allows read and write operations concurrently. The reconfiguration allows the shard key changes, splitting the chunks and altering the chunk size dynamically. The proposed Morphus : Online reconfiguration can be done in five sequential steps:

- Create partition with shard key
- Isolate one secondary server from replica set
- Operations can be done on primary server
- Recovery phase - reconfiguration can be done on isolated secondary server
- Finally secondary server copies everything into primary server and same will be repeated in all secondary servers.

The purpose of reconfiguration is online resharding operation which achieves load balance the new shards to be placed to reduce network traffic. This paper proposed new techniques : Greedy Assignment and Bipartite matching.

**Greedy Assignment** : A centralized server that runs greedy algorithm collects all the information and inform its new decisions to servers. But it creates a bottleneck by allocating more chunk at few servers.

**Bipartite matching** : The algorithm has more advantages than Greedy Technique: reduces read and write bottlenecks and latency and prevent allocation of too many chunks to few servers. The techniques are implemented by using pymongo interface with Amazon review as datasets. ProductID as old key and userID as new shard key. The results showed that mild degradation in read and write latency during reconfiguration phase. Morphus scales well with increasing replica set.

### **Research on Improvement of Dynamic Load Balancing in MongoDB**

Data partitioning on distributed nodes and migration dynamically makes the system costly. Xiaolin Wang, Haopeng Chen proposed heat based dynamic load balancing algorithm which reduce the cost of the sharding process[28]. The following are the steps of Heat based dynamic load balancing algorithm :

**Exception Detection Algorithm** : To identify the load of each shard, upper and lower bound are fixed for three resources : CPU, Memory and bandwidth. The monitored utilization of each resource is compared with upper and lower bound value and decision is taken whether the shard is overloaded or underloaded.

**VM overloaded balancing** : This step identify the hot chunk(more requests) and that chunk will be divided and then autosharding will be executed to balance the load of shard.

**VM underloaded balancing** : Each overloaded node will identify one pair which has status of underload. Then overloaded node will migrate data to this underloaded node.

**Physical overloaded balancing** : once the physical node is identified as overloaded, that primary server status will be changed as secondary. One secondary node will also be changed as primary. Hence the write request will be reduced. Also the migration step is prevented. If no node is identified to convert from secondary to primary, migration step comes into action.

Simulation work has been done for client access and virtual nodes are created with Xenserver and MongoDB is used in virtual nodes. The results showed that resource utilization is controlled and reduced migration cost.

### **Scalable Transactions in Cloud Data Stores**

The hash and range based partitioning schemes are easy but not good at current online transactions scenario. Swati Ahirrao1 and Rajesh Ingle proposed workload driven data partitioning for Online Transaction processing(OLTP) web applications[29]. The advantages and disadvantages of various partitioning schemes are discussed

**Static Partitioning** : once designed, it will not be reconfigured. lower number of migrations and more number of distributed transaction

**Dynamic Partitioning** : Partitioning can be done online. lower number of distributed transactions and more migration steps

**Scalable workload driven partitioning** : not fully static or dynamic. initial transaction logs are analyzed and partition can be done. Periodically partitions can be reconfigured. Less number of migrations and distributed transactions

The objectives of the scalable workload driven data partitioning are reduce number of distributed transactions and even load distribution in shards. The algorithm steps are as follows :

- The algorithm takes partitions and complete transaction log file. The algorithm starts with static distribution of shards.
- Genetic algorithm step mutation is applied on partitions to find all the combination of shards to form the partition.
- Calculate load distribution for that partition
- The combination association is calculated based on executed transactions and number of distributed transactions.
- Rank value will be assigned and the values will be ordered based on the rank
- The lower rank value combination is taken as a partition which reduces distributed transaction and efficient load balancing among servers

Experimental setup is done on Amazon SimpleDB, EC2 and TPC-C standard benchmark and results are analyzed on parameters throughput, response time and distributed transactions. The results showed that it provides better results than Graph and Schema based partitioning

#### Sharding Social Networks

Quang Duong and Sharad Goel proposed a sharding technique for social website data. Since social network website user communalities are geographically closer, tightly knit cluster of users are stored in single shard[30]. This solution solves the problem of reduced distributed access for a single query. This paper showed that Random partitioning is not suitable for social network database and network aware sharding is an NP complete problem. This paper proposed two steps to apply sharding on social network databases:

- VLabelProp is a technique for identifying tightly group of knit nodes and place the nodes in a single shard.
- BlockShard is a greedy method for zero replication partitioning which minimizes sharding cost. The inputs of Blockshard is adjacency matrix, Maximum sharding capacity and VLabelProp output. The BlockShard assigns node to only a single shard. The technique used excess storage by replicating locally popular nodes in each shard.

The system is implemented on two social network websites LiveJournal and Twitter. The experiments done on with and without replication in terms of load balancing and average access on shards. The results are compared with Random partitioning, Geo, network aware sharding with METIS and VLabelProp. The results showed that network aware sharding with METIS and VLabelProp outperform Random and Geo based sharding techniques. The authors concluded that knowing network architecture in sharding provides better performance

#### Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis

E. Dede, M. Govindaraju et al have taken scientific data from Advanced Light Source, Joint Genome Institute and Materials Project for analysis[31]. This paper evaluated the performance, fault-tolerance and scalability of using MongoDB with Hadoop framework for scientific data analysis. The MongoDB and Hadoop connector is an open source framework supported by Hadoop. The sharding facility of MongoDB is used by Hadoop instead of HDFS. But the results showed that MongoDB exhibits poor performance for parallel write operation because of global write block.

Experimental setup configured Hadoop, HDFS, MongoDB and Hadoop-MongoDB connector. Hadoop framework selected number of Maps and set Reduce as single. Each mapper selects MongoDB shard and this leads to load balancing. Java and Python scripts are used for implementation and data set as US Census records. Authors concluded that HDFS performs better than MongoDB. The MongoDB with Hadoop framework supports isolation of data nodes from process nodes that increase fault tolerance percentage. The following are the key insights :

- MongoDB with shards will have better performance than single node
- Output of the analysis could be written into HDFS provides better performance
- Serving MongoDB for storage and Hadoop for computing nodes make the system more elastic
- MongoDB with Hadoop connector provides better performance than MongoDB map-reduce
- MongoDB with Hadoop connector makes worse performance in read and write operations. It is due to design difference of HDFS and MongoDB
- MongoDB with Hadoop connector provides fault tolerant system

#### Using Paxos to Build a Scalable, Consistent, and Highly Available Datastore



Jun Rao, Eugene J. Shekita, et al proposed Spinnaker's Paxos a replication protocol for spinnaker experimental data store[32]. Its performance on read and write operation is better than traditional method. Generally master-slave replication is fault tolerant for single node failure and not working well for two node failure with replication factor 3. The Paxos protocol is the proven solution for the replication factor 3 or more.

#### Spinnaker's Architecture:

- Spinnaker apply range partitioning and row of a table is spread over cluster.
- The data in a node is replicated into  $N-1$  node with  $N=3$ . This step is similar to chained declustering [33].
- The replicated cluster is called cohort. Cohorts can overlap. For example  $A-B-C$  is one cohort and  $B-D-E$  is another cohort. Each cohort will have one leader.
- Leader election phase selects leader in case of failure. In the quorum phase, leader propose write operation and followers will accept it. After write operation, the message will be informed to followers. The followers will send acknowledgement after write is over. This is the way it maintains consistency among nodes in cohort.
- If leader fails, the follower who committed all the write operation initiated by the old leader will be selected as new leader. Even the old leader comes back, it can join as follower.

The experiment results are compared with Cassandra. The results showed that read operation latency is significantly reduced compared to Cassandra. Write operation results are 5% to 10% worse than Cassandra. This is due to wait for acknowledgement from followers in Cohort. The authors also concluded that this is little pay to achieve strong consistency.

#### SWORD: Scalable Workload-Aware Data Placement for Transactional Workloads

The main objectives of any database sharding on nodes are how to partition the database effectively, high availability, reducing number of distributed transactions and fault tolerance. Abdul Quamar, K. Ashwin kumar et al proposed a scalable workload aware data partitioning(SWORD) scheme for online transactions[34]. The following are the different steps applied in this paper to improve the performance of system:

- hypergraph compression technique to reduce the overheads of partitioning
- Incremental data repartitioning technique for dynamic load changes
- SWORD manages load balancing and increases availability
- fine grained quorum improves throughput and reduce the cost of distributed access with different read and write patterns

The experiments uses TPC-C bench mark for evaluations and the results showed that improved query routing time, overall end-to-end time and throughput. The results are compared with Random technique and baseline approach.

#### The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data

Aaron McKenna and Matthew Hanna et al proposed a Genome analysis toolkit for analyzing Next Generation DNA Sequencing(NGS) data[35]. The Genome Analysis ToolKit(GATK) architecture uses the MapReduce functional programming to design the framework. The MapReduce framework allow the researchers to write efficient NGS toolkit. Here MapReduce divides the computation into two steps: Map and Reduce. Single Nucleotide Polymorphism (SNP) discovery and genotyping are fed into Map function and chromatin immunoprecipitation is fed into Reduce step. GATK partitioning Genome, SNP and SMP format of genome into shards for Map Reduce operation. In addition, the shards are divided further by traversal engine. The authors also suggested to implement sharding with data localizing systems such as Hadoop or Sun Grid Engine or LSF for performance improvement. The distributed clusters(shards) and shared memory machines facilitate to optimize GATK engine, CPU and memory efficiency and parallelize analysis tools.

#### A Novel Clustered MongoDB-based Storage System for Unstructured Data with High Availability

Wenbin Jiang, Lei Zhang et al proposed various strategies to support complex query functions and provide high availability of the storage system[36]. The various strategies proposed are

- MyStore a new highly available data storage system for unstructured data
- Consistent hashing algorithm and virtual node are used for partitioning and distribute data on multiple nodes
- NWR mode is used for automatic backup and data consistency
- Gossip protocol is used for failure information exchange among nodes
- Cache module and user friendly interface are used for improving the usability
- VeePalm virtual education experiment platform is used for implementation

This research review is focusing on how data partitioning and distribution are effectively done. Hence consistent hashing algorithm is taken for study and results are discussed. The nodes in cluster is assigned a random value. The data to be written is applied to hash function called Ketama hashing algorithm[37], the output of hash function is a key to map data to the nodes. When the number of nodes are limited in a cluster, the basic consistent hashing algorithm is not effective. Hence the virtual nodes are introduced, which assist to place the data on physical node effectively.

The system uses replication with replication factor  $N$ . The data firstly stored in a primary node and replicated in  $N-1$  nodes. replication done in clock wise direction of a ring. During *Put* operation, if number of replication is greater than  $W$  then *Put* operation is successful. The same is repeated for *Get* operation, if number of *Get* is greater than  $N$  then read operation is

successful. For high consistency, the algorithm suggest  $N=W$  and  $R=1$ . for high availability  $W=1$  and  $R+W<N$ . This step improves write latency and performance.

The system is implemented with various XML data set and requestes are generated using Microsoft web application stress tool and results are compared with MySQL and MongoDB for storage operations. The parameters Get and Put performance, load balancing, throughput, response time and number of hits have taken for evaluations. The results showed that MySTore is providing better performance than MySQL and MongoDB.

### Analysis of Sharding Schemes

Having set of research work done on various sharding schemes, implementation and it analysis, in this survey, we present critical aspects and key insights of sharding performance for Bigdata applications. The essential aspects of analysis are

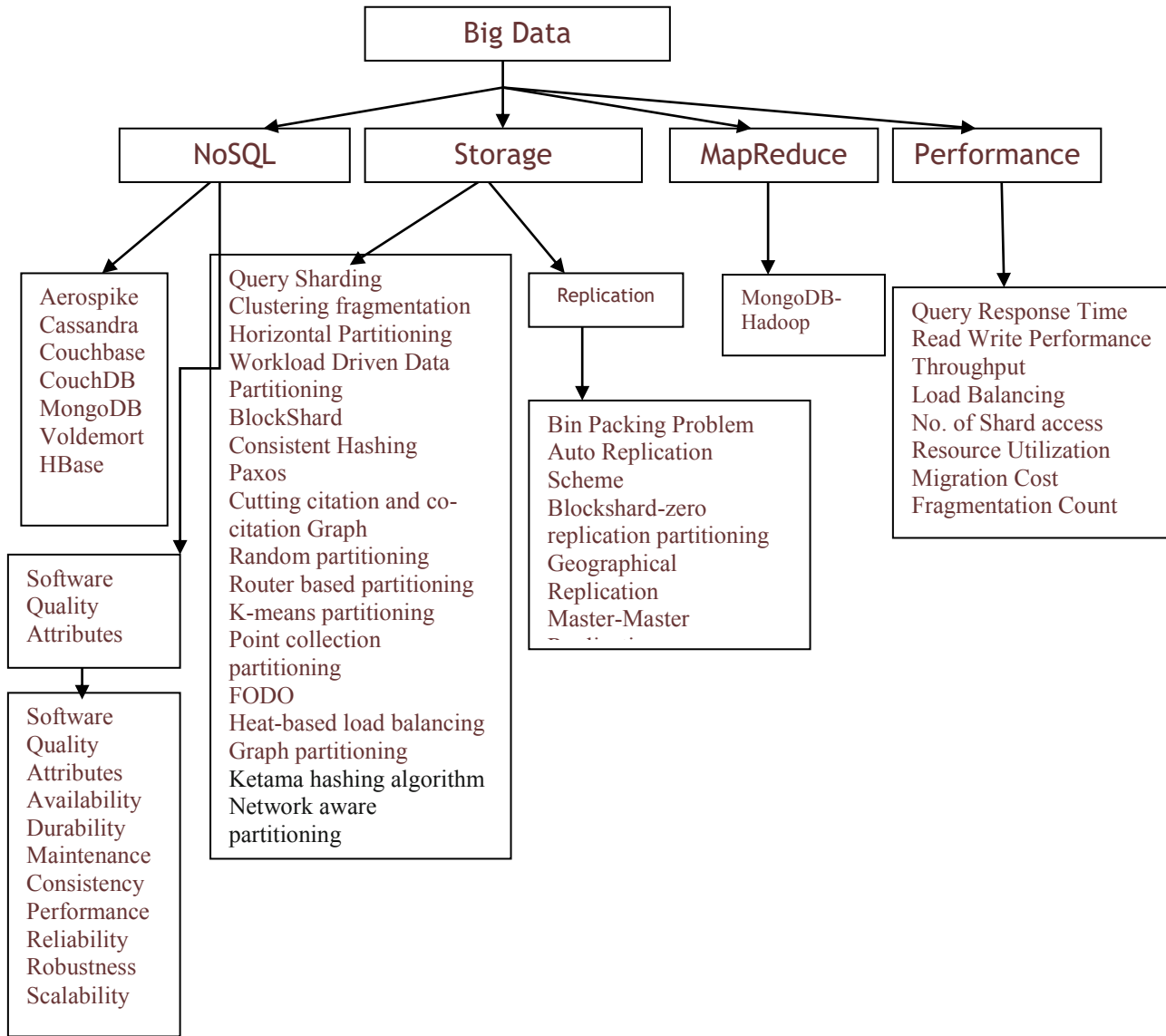
- The existing survey work covered only architectural and Bigdata management elements. This research survey work focus more on storage management specifically sharding and replication strategies and detailed analysis of sharding implementations
- In addition to partition the database, query can also be sharded to provide better results
- Instead of applying general sharding schemes, the knowledge of data structure is significant factor to find new sharding schemes
- The segments frequently accessed can be replicated in more than servers reduce execution time
- Implementation of sharding schemes on variety of database showed that Researchers previously started the research works on bigdata in the fields of Biological data(Genome analysis), US census Dataset, Construction industry, Health insurance, E-commerce, Social network and Scientific literature
- MongoDB autosharding, Repartitioning, FODO algorithms and heat based load balancing supports dynamic sharding techniques for load balancing online
- Mapreduce merging programming(Mapreduce) with storage(HDFS). But Sharding and Mapreduce connector assists to isolate storage from programming increases performance.
- The MongoDB with Hadoop framework supports isolation of data nodes from process nodes that increase fault tolerance percentage. MongoDB-Hadoop connector is an open source software tools supported by MongoDB
- Sharding on Cloud servers with router based sharding schemes assists the Business layer to choose partial data to move on cloud and allows confidential data to stay in local server
- The survey also identified that Random and Range based partitioning are not well suited for load balancing
- The implementation results concluded that MongoDB and Cassandra will have built-in auto sharding facilities
- Depends on the application, sharding the data with same category reduces query response time
- Greedy algorithm and Gossip protocols assist to setup fault tolerant systems
- Scalable work driven partitioning makes static(initial level) partitioning as strong and reduces dynamic migration steps
- Flexible query systems assist the system to retrieve relevant data even if exact data is not able find.
- Mostly replication factor is fixed as  $N = 3$ , replicated at  $N-1$  nodes
- Microsoft web application stress tool, iSearch test collection, VbLabelProp, Hadoop-MongoDB connector and Graclus are the tools used by sharding schemes
- Research work can be further moved on
  - Dynamic online sharding with minimum migration steps
  - Static partitioning with proper load balance
  - Algorithms to maintain consistency in replicas
  - Identifying new sharding schemes and keys
  - Hybrid solutions like MongoDB-Hadoop connector
  - In-memory computing
  - Query sharding

**Table 2 : Analysis of Sharding Schemes**

Sl. No	Title	Sharding Technique	Sharding on	Parameters	Conclusion
1	Knowledge Driven Query Sharding	Query Sharding Explicit Semantic Analysis	MeSH	Execution Time	Query sharding in the SONCA system utilize the computing resources optimally and execution time is considerably less than traditional(without sharding) technique.
Sl. No	Title	Sharding Technique	Sharding on	Parameters	Conclusion
2	Clustering-based fragmentation and data replication for flexible query	Clustering Fragmentation, Derived	Biological data MeSH	Execution Time Fragmentation Count	Few fragments stored in more than one servers gives better performance in

	answering in distributed databases	fragmentation query re-writing Bin Packing Problem			query answering
3	Social BIMCloud: a distributed cloud-based BIM platform for object-based lifecycle information exchange	Horizontal Partitioning, Automatic Replication, Dynamic instance creation	AEC Industry	Not implemented	Performance depends on Bandwidth of Internet and reliability of service providers
4	The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data	Shared memory parallelization and distributed clusters(shards)	Sub region of Genome, SNP and SAM format of genome	Processor count and Elapsed time	Reduces elapsed time for end user analysis Parallelize analysis tools
5	Research on The Improvement of MongoDB Auto-Sharding in Cloud Environment	FODO algorithm	Random data with four fields int, long, string and double	Concurrent write and query operation	Concurrent number of operations increased
6	Migrating Application Data to the Cloud using Cloud Data Patterns	Router based sharding, critical data filter and Pseudonymizer	Health Insurance Company and External Auditing company	Literature Research	The authors guiding researchers to analyze various mapping of scenarios to data patterns and choose one among to implement as a single framework.
7	Point collection partitioning in MongoDB Cluster	RPS SPS KPS	3 sets - random and uniform distribution range	Response time	KPS provided better results than SPS and RPS techniques.
8	Data management in cloud environments: NoSQL and NewSQL data stores	Study on partitioning, Replication, consistency and concurrency	Study paper	-	Range and Hash based partitioning schemes are compared
9	An Overview of Newly Open-Source Cloud Storage Platforms	Comparison	MongoDB Cassandra ABICloud	Selection of best storage mechanism	MongoDB has been selected as storage mechanism for their own experiments.
10	High dimensional biological data retrieval optimization with NoSQL technology	Microarray Key-value data model on HBase	transcriptomic data set from NCBI GEO	Data retrieval time	HBase retrieval rate is higher than MongoDB and RDBMS
11	Sharding for Literature Search via Cutting Citation Graphs	Cutting the citation and co-citation graphs, iSearch test collection, Graclus - Graph clustering software	Scientific Literature	Retrieval of relevant documents	co-citation graph performance is better than citation graph
<b>Sl. No</b>	<b>Title</b>	<b>Sharding Technique</b>	<b>Sharding on</b>	<b>Parameters</b>	<b>Conclusion</b>
12	Morphus: Supporting Online Reconfigurations in Sharded NoSQL Systems	Online reconfiguration, Greedy Algorithm,	Amazon reviews	Read and Write Latency Scalability Network	Morphus scales well with increasing replica set

		Bipartite matching		optimization	
13	Research on Improvement of Dynamic Load Balancing in MongoDB	Heat based load balancing system	Simulation Xenserver	Resource utilization Migration cost	Resource utilization is controlled and reduced migration cost.
14	Scalable transactions in cloud data stores	Workload driven data partitioning	E-commerce Transactions	throughput, response time, number of distributed transactions	Less number of migrations and distributed transactions
15	Sharding Social Networks	VbLabelProp BlockShard	LiveJournal and Twitter	Load balance No of shard access	Network aware sharding with METIS and VbLabelProp outperform Random and Geo based sharding techniques
16	Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis	Hadoop-MongoDB connector	US Census Data set	Read and write performance	The MongoDB with Hadoop framework supports isolation of data nodes from process nodes that increase fault tolerance percentage
17	Using Paxos to Build a Scalable, Consistent, and Highly Available Datastore	Spinnaker's Paxos Zookeeper	Spinnaker	Read and Write latency	Read operation latency is significantly reduced compared to Cassandra. Write operation results are 5% to 10% worse than Cassandra
18	SWORD: Scalable Workload-Aware Data Placement for Transactional Workloads	hypergraph compression technique, Incremental data repartitioning, SWORD, fine grained quorum	TCP-C Benchmark	Throughput Query response time End-to-end transaction time	Improved query routing time, overall end-to-end time and throughput
19	A novel clustered MongoDB-based storage system for unstructured data with high availability	Consistent Hashing algorithm with virtual nodes	XML data set	Get and Put performance Throughput Response time Load balancing	MyStore is providing better performance than MySQL and MongoDB



**Fig 1 Technologies and Parameters Listed in this Survey**

The Table 2 shows various sharding techniques and its concluding remarks. Fig 1 listed various sharding schemes, replication strategies, NoSQL DBs, Software quality parameters and performance parameters taken for survey in this research review paper. The above detailed analysis, technologies mentioned in Fig 1 and the concluding points in Table 2 help out the researchers to move further in sharding schemes.

**CONCLUSION**

In this survey, we studied, investigated, categorized, and analyzed critical aspects of sharding schemes for Bigdata management systems. The survey on sharding schemes implemented from the year 2010 have been done and the results are analyzed. The existing survey papers also direct the researchers to implement the usecase with different architecture and NoSQL DBs assists to compare and find the optimal solution in Bigdata management. This research review paper also identified that dynamic load balancing, sharding keys and algorithms to find minimum migration steps, in-memory computing, maintaining consistency in replicas and query sharding are the major research areas in this direction.

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the family member and colleagues for given support.

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] Min Chen, Shiwen Mao, Yunhao Liu. "Big Data: A Survey" *Journal of Mobile Network Applications* 19 (2014) :171–209.
- [2] Laney D 3-d data management: controlling data volume, velocity and variety. META Group Research Note, 6 February 2001
- [3] Cattell R Scalable sql and nosql data stores. *ACM SIGMOD Record* 39(4) (2011):12–27
- [4] Wu, Lengdong, Liyan Yuan, and Jiahuai You. "Survey of large-scale data management systems for big data applications." *Journal of computer science and technology* 30.1 (2015): 163-183.
- [5] Campbell D, Kakivaya G, Ellis N. Extreme scale with full SQL language support in Microsoft SQL Azure. In *Proc. the 2010 ACM SIGMOD International Conference on Management of Data*, June 2010, pp.1021-1024.
- [6] DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W. Dynamo: Amazon's highly available key-value store. In *Proc. the 21st ACM SIGOPS Symposium on Operating Systems Principles*, October 2007, pp.205-220.
- [7] Cooper BF, Ramakrishnan R, Srivastava U *et al.* PNUTS: Yahoo!'s hosted data serving platform. In *Proc. the 34th International Conference on Very Large Data Bases*, August 2008, pp.1277-1288. Lakshman A, Malik P. Cassandra: A decentralized structured storage system. *ACM SIGOPS Operating Systems Review*, 2010, 44(2): 35-40.
- [8] Joshi A, Sam H, Charles L. Oracle NoSQL database scalable, transactional key-value store. In *Proc. the 2nd International Conference on Advances in Information Mining and Management*, October 2012, pp.75-78.
- [9] Ghemawat S, Gbioff H, Leung S T. The Google file system In *Proc. the 19th ACM Symposium on Operating Systems Principles*, December 2003, pp.29-43.
- [10] Chang F, Dean J, Ghemawat S *et al.* Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems*, 2008, 26(2): Article No.4.
- [11] Baker J, Bond C, Corbett J *et al.* Megastore: Providing scalable, highly available storage for interactive services. In *Proc. the 5th Biennial Conference on Innovative Data Systems Research*, January 2011, pp.223-234
- [12] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 2008, 51(1): 107-113
- [13] Lourenço, João Ricardo, et al. "Choosing the right NoSQL database for the job: a quality attribute evaluation." *Journal of Big Data* 2.1 (2015): 1-26.
- [14] Kuznetsov S, Poskonin A (2014) Nosql data management systems. *Program Comput Softw* 40(6):323–332
- [15] Haughian G (2014) Benchmarking replication in nosql data stores. Dissertation, Imperial College London
- [16] Gudivada VN, Rao D, Raghavan VV (2014) Nosql systems for big data management. In: *Services (SERVICES), 2014 IEEE World Congress On*. IEEE, Anchorage, AK, USA. pp 190–197
- [17] Krasuski, Adam, and Marcin S. Szczuka. "Knowledge Driven Query Sharding." *CS&P*. Vol. 225. 2012
- [18] Wiese, Lena. "Clustering-based fragmentation and data replication for flexible query answering in distributed databases." *Journal of Cloud Computing* 3.1 (2014): 1-15.
- [19] Das, Moumita, Jack CP Cheng, and Srinath S. Kumar. "Social BIMCloud: a distributed cloud-based BIM platform for object-based lifecycle information exchange." *Visualization in Engineering* 3.1 (2015): 1-20
- [20] Liu, Yimeng, Yizhi Wang, and Yi Jin. "Research on the improvement of MongoDB Auto-Sharding in cloud environment." *Computer Science & Education (ICCSE), 2012 7th International Conference on*. IEEE, 2012.
- [21] Strauch, Steve, Vasilios Andrikopoulos, and Thomas Bachmann. "Migrating application data to the cloud using cloud data." *3rd International Conference on Cloud Computing and Service Science, (CLOSER)*. 2013.
- [22] Zhang, Shuai, et al. "Point collection partitioning in MongoDB Cluster."
- [23] Grolinger, Katarina, et al. "Data management in cloud environments: NoSQL and NewSQL data stores." *Journal of Cloud Computing: Advances, Systems and Applications* 2.1 (2013): 22

- [24] Gu, Genqiang, et al. "An overview of newly open-source cloud storage platforms." *Granular Computing (GrC), 2012 IEEE International Conference on*. IEEE, 2012.
- [25] Wang, Shicai, et al. "High dimensional biological data retrieval optimization with NoSQL technology." *BMC genomics* 15.Suppl 8 (2014):
- [26] Zhao, Haozhen. "Sharding for literature search via cutting citation graphs." *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014.
- [27] Ghosh, Mainak, et al. "Morphus: Supporting Online Reconfigurations in Sharded NoSQL Systems." (2014).
- [28] Wang, Xiaolin, Haopeng Chen, and Zhenhua Wang. "Research on Improvement of Dynamic Load Balancing in MongoDB." *Dependable, Autonomic and Secure Computing (DASC), 2013 IEEE 11th International Conference on*. IEEE, 2013
- [29] Ahirrao, Swati, and Rajesh Ingle. "Scalable transactions in Cloud Data Stores." *Advance Computing Conference (IACC), 2013 IEEE 3rd International*. IEEE, 2013.
- [30] Duong, Quang, et al. "Sharding social networks." *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013.
- [31] ede, Elif, et al. "Performance evaluation of a mongodb and hadoop platform for scientific data analysis." *Proceedings of the 4th ACM workshop on Scientific cloud computing*. ACM, 2013.
- [32] Rao, Jun, Eugene J. Shekita, and Sandeep Tata. "Using Paxos to build a scalable, consistent, and highly available datastore." *Proceedings of the VLDB Endowment* 4.4 (2011): 243-254
- [33] H. Hsiao and D. J. Dewitt. Chained Declustering: A New Availability Strategy for Multiprocessor Database Machines. In *ICDE*, pages 227–254, 1990
- [34] Quamar, Abdul, K. Ashwin Kumar, and Amol Deshpande. "SWORD: scalable workload-aware data placement for transactional workloads." *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 2013.
- [35] McKenna, Aaron, et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *Genome research* 20.9 (2010): 1297-1303.
- [36] Wenbin Jiang, Lei Zhang et al. " A novel clustered MongoDB-based storage system for unstructured data with high availability, *Journal of Computing* (2014) 96:455–478
- [37] Petrovic J (2008) Using memcached for data distribution in industrial environment. In: *Proceedings of 2008 3rd international conference on systems (ICONS'08)*. IEEE, Piscataway, pp 368–372

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

## NEW APPROACH FOR TRACING STOLEN LAPTOP AND DATA PROTECTION

Menaka, Subashree, Narendran

Dept. of Computer Science and Engineering, SRM University, Chennai, INDIA

## ABSTRACT

**Aims:** Around the world everywhere, everyday many laptops are kept on stolen by strangers. It's been a critical issue since the laptops may have the confidential data, files or some personal info as well. Letting such information's to strangers will lead unexpected events. Tracing such laptops and the strangers is the big issue in the computing world. Many solutions provided by the many vendors but none of them really resolve this issue. Many of them require external intervention; require expensive additional hardware's like GPS, Hardware Lock, point-point sensors key servers etc. **Results:** This paper provides a new approach for Tracing the Stolen Laptops, suggesting a solution which does not need any additional resource. **Conclusion:** It is useful for tracing the location of the laptop, identity of stranger (captures image and video), protecting the content by locking the Laptop etc.

Published on: 28<sup>th</sup>– August-2016

## KEY WORDS

Laptop, Trace, GPS, IP address,  
MAC address, Hardware Lock,  
Sensors\*Corresponding author: Email: [menanand7@gmail.com](mailto:menanand7@gmail.com)

## INTRODUCTION

As per FBI, in 2005 laptop theft causes \$3.5 loss. The Computer Security Institute (CSI)/FBI Computer Crime & Security Survey, they found that the average theft of a laptop to cost a company \$31,97[1]. The survey done by 325 private and public organizations which is published by Intel in 2010 says 10% of Employee laptops were before the laptops usefulness lifespan [2]. Average total negative economic impact of a due to laptop theft was \$49,256 due to data stealing. The cost of laptops lost for the organization is \$2.1 billion [3]. In US, 28% of \$48billion total lost in economy is due to data breaches in stolen laptops and other mobile devices[4]. The NSW Bureau of Crime Statistics and Research prepared a report which states that laptop theft has been increased in last 10 years. The laptops were stolen for the cost initially, but now a day it has been stolen for accessing the data inside them. Most of the laptops are stolen during the travelling at train, hotels/motels, airports, taxi cars, and public gathering events.

It is necessary to protect laptops being stolen since it leads vulnerable effects. For the business people lose of laptop is lose of their business, since the business information's which is confidential shouldn't be leaked to others. For example, the information regarding the tenders, project business logics are highly confidential, those are not sharable or exposed to others. The private personal photos, videos are another secret set of things not sharable or exposed outside. So what is more important, tracing Laptop? Or protecting data in it? Most of the people will say both for this question.

Many solutions exist for protecting the laptop being stolen, protecting the data in them, tracing the laptop even it has been stolen by any. These solutions use various techniques like GPS, hardware locks, point-point tokens, inside data protection, centralization of data and WiFi based etc. Next section in this paper give brief about the techniques listed here.

## RELATED WORKS

Many researches have been done for preventing laptop theft, protecting data inside and tracing the stolen laptop. Still many working on improvising the existing solutions and to develop a new solution. But very few among them really protects the laptop from thief, protects data and trace stolen laptop.

## Basic Security

The security can be provided simply by using any of the following ways



- Pre-boot authentication  
Authenticating users of Laptop by asking username and password before the boot up, (i.e) verifying user even before Operating System launched[5]. This will be useful for completely protecting the Laptop data.
- Disk Encryption  
Encrypting the entire disk from un-authorized access is another solution for data protection [6]. Using this one can access the data in a disk only after the authentication.
- Locks  
Varieties of Laptop locks are available to protect it physically from strangers. These locks are available in market to lock the Laptop in a physical location to make it immovable.

### Laptop Tracking Software's

The methods specified above are used for protecting the data but they are not sufficient since many advanced security attacks will breach the data security. We need a solution need to protect the data as well as prevent the Laptop from stolen and track them even if it has been stolen.

This section describes some of the solutions for tracking the stolen Laptops.

- Tracking Laptop using Adeona

Adeona [7] is one of the Laptop Tracking solution for tracing the stolen laptop, it works on two step process. First the Tracking software needs to be installed later the recovery software. The tracker keep on send some data which used by the recovery software to analyse those data and predict its geo graphic location.

If at one point your laptop gets stolen and is connected to the internet, the Adeona will send you criminal's IP address. The IP address can then be retrieved using the Adeona Recovery Wizard. Knowing the IP address is enough in most cases to track the geographical location of the device.

Unlike similar commercial tools, Adeona is decentralized and doesn't store your password on its servers. It means no one besides the owner can use Adeona to track your laptop.

- Computrace Lojack for Laptop

It is another Laptop recovery tool brings back the lost or stolen laptop. It keeps working in background. When the laptop lost or stolen has to be reported to the Computrace team, they will trace it with the help of the tool installed in Laptop. Before reporting the lost or theft to the computrace it is mandatory to give a police complaint and the complaint copy is needed.

- MyLaptopGPS

Similar to above mentioned products it is used for protecting data as well as tracking it. It has an exclusive feature which allows us to re-encrypt the data with new encryption key in a Laptop even after lost. It is useful for encrypting even the USB devices, thump devices etc.

### Other Related Works

Transient Authentication [8] force the user to use of wearable token that constantly attests to the user's presence. When the user departs, the token and device lose contact and the device secures itself. The small, lightweight nature of mobile devices, combined with their common usage environments (in public places, amid many un authenticated people) makes them an easy target for theft. More important than loss of hardware may be the cost of information exposure. To overcome problem using Wearable hardware token provides authentication and security using wireless communication.

A boot system that uses a U-Key [9] can help ensure the integrity of fairly static PC components. The U-Key approach is designed both to provide boot integrity and to enforce access control. The basic idea is that the host computer actually boots from a USB disk loaded with the operating system and the loader. Smart card is a secure way to store certificates and keys. Along with hardware tokens, smart cards deliver user benefits in four major areas: easy portability of user credentials, drastic simplification of platform, better protection of personal credentials and a higher level of personal privacy.

Role Based Access control in location based services (LBS) using the methodology GEO-RBAC [10]. a location based technique for with access control. These approaches are referred to as user-driven and event-driven respectively. Under the user-driven approach the position of the user is checked only upon an access request; conversely under the event-driven approach the position of the user is tracked so that the set of enabled roles can be changed dynamically and transparently with respect to the user. The conclusion of this project using this approach the access to certain features can be restricted based on current location. Computer vision applications for mobile phones are gaining increasing attention due to several practical needs resulting from the popularity of digital cameras in today's mobile phones. The face and eye detection [11]. modules are implemented and running quite well. For simplicity, they are trained using Haar-like features and AdaBoost. The main advantage of using Haar-like features is their low computational cost. However, the discriminative power of such simple features is limited.

2D-Barcode Processing solution [12]. to support mobile applications and 2D Barcode Applications in M-Commerce, Using barcodes for providing mobile security. It is clearly understood that barcodes can be used to hide security information's which can be read only through bar code reader. The future work of this research includes two areas: a) to enhance the existing 2D barcode solution to deal with non-perfect 2D barcode images, and b) building a 2D barcode enabled mobile payment system. From the conclusion of this paper it is clearly understands that barcodes can be used to hide security information's which can be read only through bar code reader.

PCA & LDA based teeth-image personal identification method [13]. in this, the teeth image failed from the matching in the PCA & LDA based system is reconsidered by feeding back the image to eliminate the reflection and the rotation problems. In the experiments, 500 teeth images are tested with 200-teeth database. The results revealed that of the 10% errors caused by the two problems, 5% are correctly identified because of the proposed method. A method for improving a teeth image personal identification was proposed.

Multimodal biometric authentication approach using teeth image and voice as biometric traits [14], this method is evaluated using 1000 teeth images and voices, in which these are collected by smart-phone, i.e., one mobile device for 50 subjects. In this paper, the Ada Boost algorithm is used based on Haar-like features for teeth region detection, and the EHMM algorithm with 2DDCT as feature vector is additionally accommodated in process of teeth authentication. The proposed multimodal biometric authentication system exhibits an EER of 2.13%. Thus confirm that the performance of the proposed system is better than the performance obtained using teeth or voice individually.

TrustVisor [15], a special-purpose hypervisor that provides code integrity as well as data integrity and secrecy for selected portions of an application. TrustVisor achieves a high level of security. The protection granularity in these systems is too coarse to provide strong security properties, because the entire application is in the TCB. Dynamic Root of Trust for Measurement (DRTM) mechanism provides memory protection from DMA accesses, and integrity measurement of the launched code before it executes. TrustVisor is a small hypervisor that enables isolated execution of Pieces of Application Logic (PAL) with a TCB containing only the TrustVisor runtime and the PAL itself. This system enforces code and execution integrity, data secrecy and integrity for PALs. TrustVisor supports unmodified legacy OSES and their applications.

"Mobile User Location-Specific Encryption (MULE): using your office as your password [16]. Data breaches due to stolen laptops are a major problem. The goal of this work is to provide encryption of sensitive data while requiring minimal administrative effort and zero user effort during common accesses and moderate effort otherwise. The proposed system allows access to restricted files only from specific location. This approach is to use information and services available only in a trusted location to assist in key derivation without user involvement and without authenticating the laptop to any outside service. In this paper, the Home Key Derivation and Corporate Key Derivation protocols are designed which allow a laptop to automatically derive the key needed to access sensitive files based on a location.

Keypad [17], an auditing file system for theft prone devices, such as laptops and USB sticks. This paper described Keypad, an auditing file system for loss- and theft-prone devices. Keypad provides users with evidence that sensitive data either was or was not accessed following the disappearance of a device. Keypad achieves its goals through the integration of encryption, remote key management, and auditing. It demonstrates the advantage of separating encryption and key management to enforce auditing for mobile device data. If data was accessed, Keypad gives the user an audit log showing which directories and files were touched, and also allow users to disable file access on lost devices, even if the device has been disconnected from the network or its disk has been removed.

Intelligent anti-theft and tracking system for automobiles [18]. a system for preventing the automobile theft since Vehicle theft is a serious issue around the entire earth. So an efficient solution is needed to track them. The vehicle owners use lot of approaches to safeguard their vehicles providing a lock, SMS alert when a vehicle start, alarm when a vehicle start apart from the owner etc., This paper provides a solution using GPS receiver, Google Earth and SMS. Using GPS device the current location of a vehicle can be obtained, which can be sent to the owner through the SMS and the using the co-ordinates the vehicle can be tracked using Google Earth.

## BACKGROUND

The works described in previous sections having several shortfalls. This section will give a brief background about the new approach proposed in this paper.

### Idea Behind

The solution proposed in this paper is to trace the stolen laptop using its current IP address and MAC address. It's known to every one that the IP address is unique and using IP address we can trace the person using that IP address. The Internet Service Providers (ISP) providing internet services to any System using Mobile or Telephone lines. So it is clearly understand that every IP address is mapped on any of the Mobile number or Telephone number. So if we know the IP address, we can know the telephone or mobile number through which the system is connected. Using the telephone or mobile number we can know the person who is using the connection. This is the idea behind this solution. Apart from this identity, the photo of the person will help to trace him/her. Even If the photo is not clear the video will be useful. At least in a single frame the identity of a person can be found easily.

### New Approach in Tracing & Protecting

The new approach described in this paper is aimed at tracing the stolen laptop and restricting access to the files so that preventing data expose to others. In this, Laptop tracing is done with the help of IP Address and file protection is done using OTP (One Time Password).

IP addresses are the unique ID assigned to every system connected in a network which is used for communication between systems. If a system or laptop is connected in an Internet the IP address is assigned by the ISP. Usually the IP addresses are static or dynamic. The static IP addresses are fixed permanent IP assigned to a specific connection. i.e. Whenever the connection established between from the system through a specific connection always the same IP address will be assigned.

The next kind of IP addresses are dynamic IP address i.e. the IP address of a system connected using the specific line will change or not same for all sessions.

Globally region wise the IP addresses are allocated and managed by IANA (Internet Assigned Numbers Authority) which allocates IP addresses among following five Regional Internet Registries (RIR):

- African Network Information Centre (AfriNIC) for Africa
- American Registry for Internet Numbers (ARIN) for the United States, Canada, several parts of the Caribbean region, and Antarctica.
- Asia-Pacific Network Information Centre (APNIC) for Asia, Australia, New Zealand, and neighboring countries
- Latin America and Caribbean Network Information Centre (LACNIC) for Latin America and parts of the Caribbean region
- Réseaux IP Européens Network Coordination Centre (RIPE NCC) for Europe, Russia, the Middle East, and Central Asia

Each of the above RIR responsible for assigning and managing IP addresses in the specific region of the world. The next level allocation is done by these RIR's, these allocations include allocating IP for the ISP's in that regions, educational, government organization and large level private organizations.

## PROPOSED WORK

The idea proposed in this paper is to trace the Laptop using its current IP address also protecting the data in it by making the laptop unusable once it is found missed or stolen. The proposed work having two main objectives, one is to find the IP address if it is stolen another is to protect the data.

### Find IP address of a stolen Laptop

As specified in the previous section the IP address of the System no matter whether it is a Desktop PC or a Laptop is always unique one. So if the Laptop is missed the Tracer Located inside will fetch current IP address and check whether it is matched with the exiting stored IP address. If it is matched with the existing IP address it concludes that the Laptop is still with the Owner and is not missed. Whereas if the IP addresses doesn't match it is clearly indicates that the Laptop is moved to another network or location.

It is possible that the Laptop user/owner can use his Laptop anywhere wherever he/she goes. He may use it in his home, or in travel, in an office, in public gathering like malls, airports, railway stations etc. Everywhere he connects his laptop in different network using wireless (WiFi) or wired connection with different IP address.

So it is necessary to differentiate whether the Laptop has been stolen or it is used indifferent place with different IP address by the owner. In both the cases the tracer will fetch the current IP address, compare it with stored IP address. If the doesn't match sends the new IP address as a mail to the registered mail ID of the owner. This IP address is useful to trace the current location.

Apart from sending the IP address the photo image and video of the current user taken using web cam will also be send along with the IP in a mail as an attachment. This photo will be useful to trace the person who is using the Laptop right now. Even if the image is not clear the video may give us more information, regarding the person, his voice also his surrounding environment, these are all useful for finding the person hence the laptop.

So the IP address and the photograph and video will reveal the person who has stolen the Laptop. With these identities a complaint can be given in police to recover the Laptop and to catch the person. Instead of reporting in police that my Laptop is missing, a complaint can be given on specific person along with his photo, IP address with the mail copy. These two evidences will be useful for them to trace the person quickly [Figure -1].

Examples of how video evidence has been used in the criminal justice process:

- Film evidence for using child soldiers by Lubanga: video clips were important to sparking an investigation, while not proving age of children
- Symptoms of Ghouta and Bhopal: video clip was used as lead evidence, but not linkage evidence
- Case in Darfur, Sudan: while the evidence did not link the attacks to the alleged attackers, the video and picture evidence was enough to move the case forward in court
- Footage of police torture of prisoners: despite authentication questions, footage was still used to call an investigation
- Footage of Al-Houlah Massacre in Syria: footage called attention to the massacre and led the UN to call for a special inquiry to conduct an investigation

- Video of General Ratko Mladic Speech: video speech of General Mladic's speech served as linkage evidence, as it linked Tolimir to the "inner command circle" of Mladic
- Case of Video Manipulation: Serbian journalist manipulated video evidence in order to protect himself from incrimination
- *Trust Alaska*: a story of how a young man named Nelson took a story he told in court with his Standing Declaration and made it into an influential video
- Videos on human rights abuses in Sri Lanka: videos used in the U.N High Commissioner for Human Rights in Geneva
- Solidary Uganda & oil company representatives: video was used to promote accountability in case representatives of oil companies went back on promises later
- Abuses of Uganda Wildlife Authority: video used to show abuses of the Wildlife Authority, but difficult to do with little artificial lighting
- Footage of the Mostar Bridge destruction: while there was footage, it was not established who did it, so the ICTY followed the bottom up strategy.

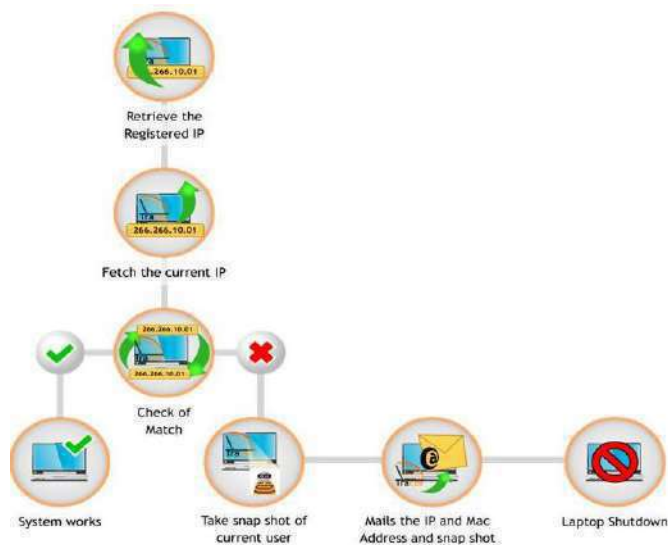


Fig: 1. Laptop Tracer

### Protecting Stolen Laptop data

Before the Laptop recovered from the robber it is highly important to protect the data in such Laptops. It is comparatively complex task. In Some earlier approach for protecting the data they used to provide a separate sever which have the key required to access the files in the Laptop. Normally this kind of servers is available in a concern need to protect their data. Every file access in such a environment need a encryption key without which the fie can't be accessed. Since the server is in office, if a person working in such office want to access any file from his home will be denied even the file exist in his laptop. So if such Laptop has been stolen by some strangers he is not able to access any of its content.

It seems to be a good idea but this approach having many drawbacks.

- It denies the file access even for a owner who is not in the office premises. In some emergency situation this approach gives a headache to the user.
- Need a dedicated server for maintaining file access keys.
- Slows down file access comparatively since every file access need a key from server and parallel request from others for the key causes heavy load at the key server.
- (iv) If the key server down or network down, none of the files from any laptop is accessible.

Keeping in mind the above limitations a simple solution for protecting the data is proposed in this paper.

When a Tracer finds that the laptop is in a suspicious location possibly with the stranger stolen it. Along with sending the IP address and photograph a One Time Password (OTP) will be send in a mail. Once the mail has been send the tracer will shutdown the Laptop.

If the Laptop restarted again the Tracer will be launched and this time it will prompt the user to enter the OTP to continue. If the Laptop is with the owner he knows the OTP since it has been delivered to his mail ID, so he could enter and continue to work in the Laptop. Also the current IP address will be stored in a Laptop.

Otherwise if the correct OTP is not entered the tracer will again send the new OTP with IP and photo of the user. This will continue for three time after third attempt the Tracer will not allow the System to boot to protect the data from the strangers. The following [Figure -2] above shows the working of data protection module.

The current IP address, mail id's, OTP code and stolen status every thing will be kept in a system registry in a encrypted format. The encryption key and password to access or modify the settings like changing IP, mail ids are also stored in registry.

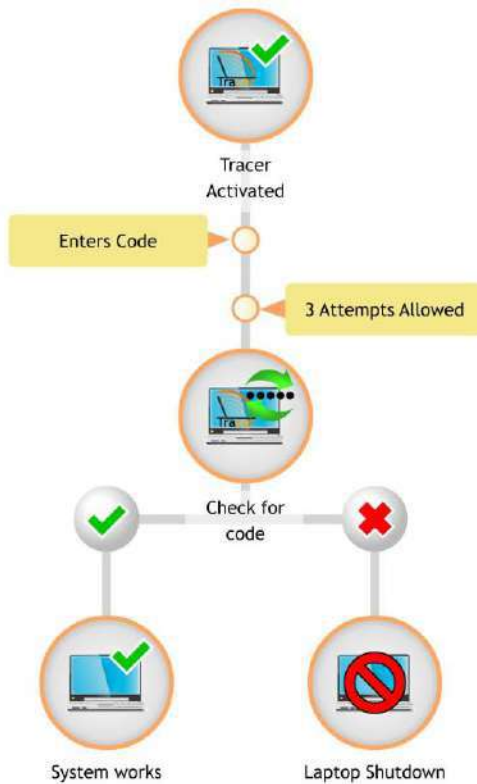


Fig:2. Laptop data protection

The following [Figure -3] is the sample mail from the Tracer.

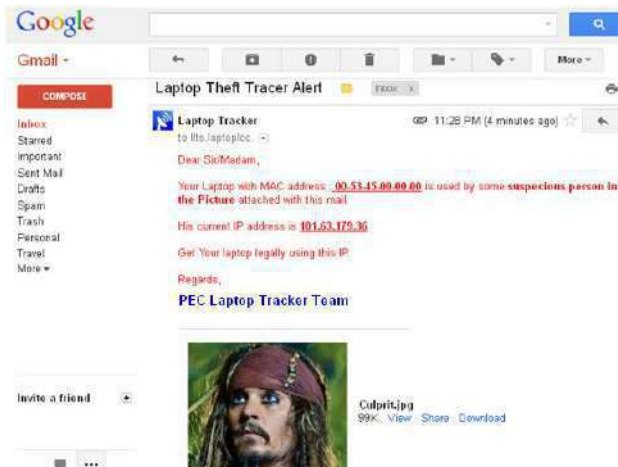


Fig: 3. Sample Alert Mail

The entire process of Laptop tracking shown in the following [Figure -4].

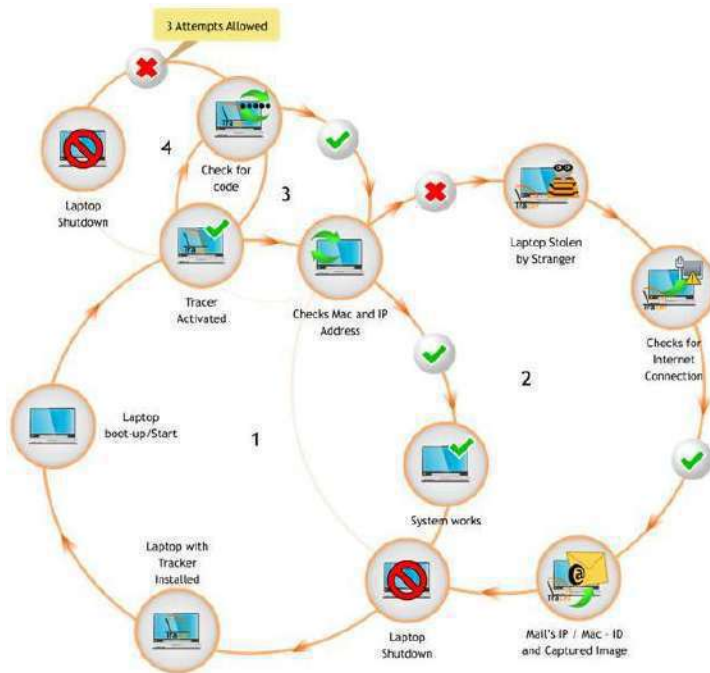


Fig: 4. The entire process of Laptop tracking

## CONCLUSION

The proposed system of Laptop Tracking and Protection of data is cost effective and doesn't require any third party intervention in tracking them. Every operation can be done in Laptop itself and all operations are completely hidden from the user. It can be enhanced by including data encryption module based on the request from the user.

## CONFLICT OF INTEREST

Authors declare no conflict of interest.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the technical support given by M.Narendran, Assistant Professor, Dept. of Computer Science and Engineering, SRM University, Chennai.

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] 2005 FBI Computer Crime & Security Survey
- [2] "The Billion Dollar Lost Laptop Problem." Page 2 Intel. Ponemon Institute, 2009. Web. 13 Feb. 2013.
- [3] "The Billion Dollar Lost Laptop Problem." Page 11. Intel. Ponemon Institute, 2009. Web. 13 Feb. 2013.
- [4] "Security Breaches Are On The Rise But Preventable." Druva, 2012. Web. 15 August 2012.
- [5] [http://www.softexinc.com/pre\\_boot\\_authentication.asp](http://www.softexinc.com/pre_boot_authentication.asp)
- [6] <https://www.lumension.com/data-protection/full-disk-encryption.aspx>
- [7] <http://adeona.cs.washington.edu/>
- [8] Anthony J Nicholson, Mark D Corner, and Brian D Noble, [2006] Mobile Device Security Using Transient Authentication, *IEEE*.
- [9] Peng Shaunghe, [2006] Enhancing PC Security with a U-Key, *IEEE* 4 (5)
- [10] Maria Luisa Damiani, "GEO-RBAC: A Spatially Aware RBAC", ACM 2006.
- [11] A Hadid, JY Heikkil, [2007] Faces and eye detection for person authentication in mobile phones, *IEEE International Conference* 25-28

- [12] Jerry Zeyu Gao, Lekshmi Prakash, and Rajini Jagatesan, "Understanding 2D-BarCode Technology and Applications in M-Commerce - Design and Implementation of A 2D Barcode Processing Solution, Published in Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International , 2
- [13] Nadee C, Kumhom.P, Chamnongthai. K.[2005] Improved PCA-Based Personal Identification Method Using Invariance Moment,*ICISIP*
- [14] Dong-Ju Kim, Kwang-Seok Hong.[ 2008] Multimodal biometric authentication using teeth image and voice in mobile environment, *IEEE Transactions* 54( 4).
- [15] McCune.JM, Yanlin Li, Ning Qu, Zongwei Zhou, Datta.A, Gligor.V, Perrig.A,[2010]TrustVisor: Efficient TCB Reduction and Attestation,*IEEE*.
- [16] Ahren Studer, Adrian Perrig.[2010] Mobile user location-specific encryption (MULE): using your office as your password,ACM.
- [17] Montaser N Ramadan, Mohammad A Al-Khedher, and Sharaf A Al-Kheder,[ 2012] Intelligent Anti-Theft and Tracking System for Automobiles, *Intl Journal of Machine Learning and Computing*, 2(1)
- [18] Dong-Ju Kim, Kwang-Seok Hong. [2008]Multimodal biometric authentication using teeth image and voice in mobile environment, *IEEE Transactions on*, 54(4).

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# CLUSTER BASED PUBLIC AUDITING FOR SHARED DATA WITH EFFICIENT GROUP USER REVOCATION IN THE CLOUD

Parimala Raghavan\*, Subasree, Sakthivel

Dept. of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Thrissur, Kerala, INDIA

## ABSTRACT

Cloud computing is a very familiar term used for the recent development of internet. It is computed in which very large group of remote servers is networked and provide centralized data storage and online access to computer services. Considering Cloud computing, Data security becomes more and more important. When users put their large size of data in the cloud, the data integrity protection is challenging. Public auditing of cloud data storage security is very essential. In the existing system users who share data as a group. In that group, one original user and number of group users. The original user creates data and other user's shares and accesses that data. The TPA (Third Party Auditor) verifies the data and after verification process cloud stores that verified data. TPA will help the data owner to make sure that his data are safe in the cloud and less burdening to the data owner. In the case of a large number of users single TPA can do the verification process it is very much time consuming process. To overcome this problem we modify the existing system. In that users can be grouped and each group has its own third party auditor. In the modified system the verification time is less as compared to the existing system. From the analysis we have identified that modified system is best for cloud environments.

Published on: 28<sup>th</sup>– August-2016

### KEY WORDS

computing; Public Auditing;  
Third Party Auditor; Verification  
time

\*Corresponding author: Email: [neelima.raghavan@gmail.com](mailto:neelima.raghavan@gmail.com)

## INTRODUCTION

Cloud computing, is a kind of Internet-based computing, where data, information and shared resources are provided with computers and other devices on-demand. It is the new technology that shares computer resources through internet instead of using the software. Cost saving is the main advantage of cloud computing and the prime disadvantage is data security. The data stored in the cloud are accessible to everyone so security is not guaranteed. To ensure data security effective third party auditor is introduced. Public verifier efficiently checks the correctness of data without downloading the entire data this is commonly referred to as a public auditing mechanism. In the existing system single TPA performs audits for multiple users simultaneously and efficiently [1],[11]. But sometimes users create a large number of data in that case single TPA can make the auditing process it is time consuming process. To overcome this problem we modified the system.

In the modified system users can be grouped and each group has its own TPA. Group users upload large number of data to the cloud. To ensure the integrity of data cloud saves these data only after the verification process. TPA collects these data and verifies without downloading the entire data. Each user group has its own specified TPA. . In the public auditing system single TPA can do the auditing process of all uploaded data, but in the cluster based public auditing system multiple number of TPA can auditing the uploaded data. From the analysis we have identified in the modified system the verification time is less as compared to the existing system.

## RELATED WORKS

Cloud service providers provide mainly three services including Software as a service (SaaS), Platform as a service (PaaS) and Infrastructure as a Service (IaaS). The cost for users to rent cloud service is cheaper than the cost for users to build cloud environment. Cloud storage service is the most common and popular service among many cloud services (e.g. Google Drive, Dropbox, Amazon S3 and Microsoft OneDrive) for general users.

To protect the integrity of data in the cloud, numbers of mechanisms have been proposed. All these mechanisms, each block of data



a signature is attached, and the integrity relies on the correctness of these signatures. Most of the previous work focus on auditing the integrity of personal data but some works [2],[3],[4],[9],[10] focus on how to preserve identity privacy when auditing the integrity of shared data. The public mechanism proposed by Wang *et al.* [7] is able to preserve confidential data from the TPA based on random masking. In that paper use the technique of providing more security by using the TPA. The TPA allows the user to know the information about the data stored in the cloud. When anyone tries to modify the data TPA informs the user by verifying the data. The TPA does not even allow CSP (cloud service provider) to read the data of the user. To operate multiple auditing tasks from different users efficiently this mechanism support batch auditing.

One recent work [2] proposed a mechanism for public auditing shared data in the cloud for a group of users. This is based on a ring signature scheme with homomorphism authenticators, the TPA can verify the integrity of shared data, but is not able to reveal the identity of the signer on each block. It supports an external auditor to audit user's outsourced data in the cloud. The main advantages of this mechanism are public auditability, storage correctness and privacy preserving but one main drawback is it is not supported user revocation when auditing the data [5],[8]. The auditing mechanism in [6] is designed to preserve identity privacy for a large number of users. However, it fails to support public auditing.

## MATERIALS AND METHODS

The below figure shows the Cluster based public auditing system model. In this users can be grouped in the cloud network. Each group has its own Third Party Verifier.

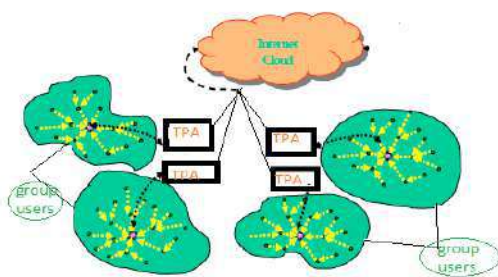


Fig. 1. Cluster based public auditing system model

System architecture consisting three entities: the cloud, TPA or public verifier and users who share data as a group.

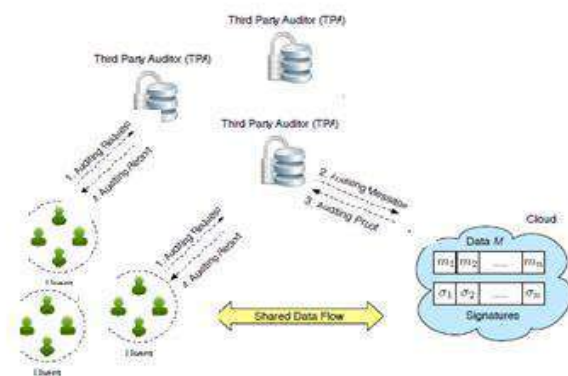
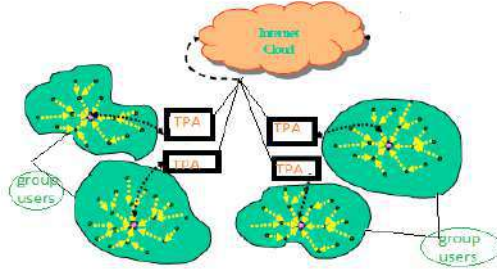


Fig. 2. System Architecture

The cloud provides data storage and sharing services. The public verifier or third party auditor utilizes cloud data for particular purposes such as searching, computation and data mining, etc. TPA provides verification services via challenge-and-response protocol. In a group, there is one original user who creates the data and share data with other users in the group through the cloud. In the modified system number of groups creates and each group consisting number of group members. Each group has its own TPA. Once a user is revoked in the group, the signatures computed by the revoked user become invalid. In this case the cloud is able to re-sign the blocks, which were already signed by the revoked user.

The important design objectives are correctness, efficient user revocation, public auditing, scalability and network security. The public verifier checks the correctness of data. The cloud data can be efficiently shared among group users. In the existing system

single TPA is able to handle large number of auditing tasks simultaneously this is time consuming. Considering this paper one of the important design goals is to decrease the auditing time using multiple number of TPA to increase the efficiency of the system



each block of data this mechanism commonly referred to as a public auditing mechanism. It checks the correctness of data without downloading the entire data. In the verification process simply downloading the data copy the data. TPA can do it and provide an audit report. In the auditing phase user uses keys and computes MAC (message authentication code) PA. In auditing phase TPA gives a key to cloud service provider and send the random number of blocks and code from the cloud service provider see the data so data in cloud keep being confidential.

In cluster based public auditing system the internal architecture is same as a public auditing mechanism. In public auditing scheme introducing third party verifier. Single TPA can do the auditing process and they provide audit report. In cluster based public auditing mechanism using multiple TPA'S for the auditing process. In this case multiple audit report provides simultaneously so the efficiency of the system will increase. In cluster based public auditing system has two phases setup phase and audit phase. In the setup phase KeyGen, SigGen algorithm using and the audit phase using GenProof and VerifyProof. TPA sends Challenge-response protocol to the CSP. Challenge-response protocol helps the verifier for verification process of blocks of data. Multiple TPA sends multiple numbers of challenge-response protocols so the verification process became fast as compared to using single TPA.

### RESULTS

In this paper, we are focusing public auditing in the cloud using multiple TPA's with efficient user revocation. Public auditing mechanism single TPA performs audits for multiple users simultaneously, but it is a time consuming process. To overcome this problem user in the network can be grouped and each group has its own individual TPA. In this scheme we have been using a different range of users and we analyze the auditing time of these range users using single TPA and multiple TPA'S.

**Table: 1. Verification time Comparison between Public auditing mechanism Vs Cluster based public auditing mechanism using less than 100 users.**

Number of existing users	Auditing Time in ms(Cluster Based Public Auditing Mechanism)	Auditing Time in ms(Public Auditing Mechanism)
20	190	280
40	220	440
60	300	720
80	350	800
100	400	980

**Table: 2. Verification time Comparison between Public auditing mechanism Vs Cluster based public auditing mechanism using 100 to 1000 users.**

Number of existing users	Auditing Time in seconds(Cluster Based Public Auditing Mechanism)	Auditing Time in seconds(Public Auditing Mechanism)
200	3	22
400	6	35
600	8	42
800	11	48
1000	13	58

**Table: 3. Verification time Comparison between Public auditing mechanism Vs Cluster based public auditing mechanism using 1000 to 5000 users.**

Number of existing users	Auditing Time in minutes(Cluster Based Auditing Mechanism)	Auditing Time in minutes(Public Auditing Mechanism)
2000	1	15
3000	2	24
4000	6	37
5000	9	45

The above table shows auditing time for different range of users using public auditing and cluster based public auditing mechanism. Public auditing mechanism using single TPA and the cluster based public auditing mechanism using multiple TPA'S. The auditing time can be taken in milliseconds, seconds and minutes depend upon the uploaded data. The above three tables showing three categories of users. The first table shows less than 100 users. In that auditing time can be taken in milliseconds. The second table shows range of users is in between 100 and 1000 and the auditing time taken in seconds. The last table the existing users are less than 5000 in that case auditing time taken in minutes. The auditing time for uploaded data files using different range of users is different for using single TPA and multiple TPA. The following figure shows the graphical representation of the table values.

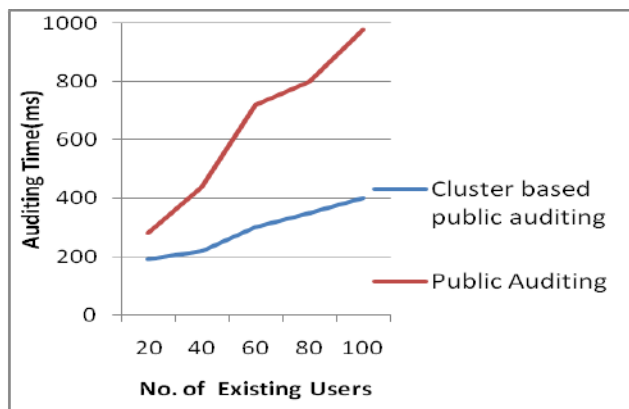


Fig: 3. Verification time between Public auditing mechanism Vs Cluster based public auditing mechanism(<100 users)

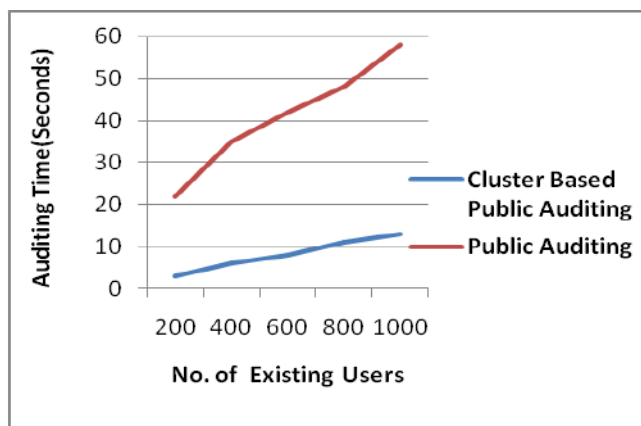


Fig: 4. Verification time between Public auditing mechanism Vs Cluster based public auditing mechanism (upto 1000 users)

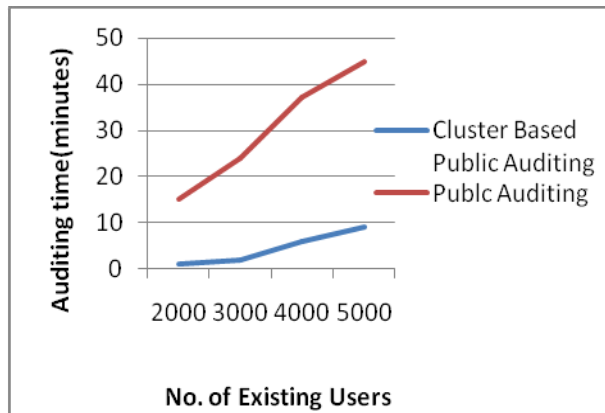


Fig. 5. Verification time between Public auditing mechanism Vs Cluster based public auditing mechanism (upto 5000 users)

## DISCUSSION

In this paper, we have compared existing and modified system in terms of verification time. We have implemented public auditing in the cloud network using different ranges of users. In the existing system, all users can upload data and single TPA can do the verification process. In modified system users can be grouped and each group has its own third party verifier. We identified the modified system the verification time is less as compared to the existing system. From the analysis we have identified that modified system is best for cloud environments.

## CONCLUSION

In cloud computing, data security is the biggest challenge. A number of research work carried out in this area. To ensure data security effective third party auditor is introduced. In this mechanism provides a number of advantages in cloud computing. The main advantage is TPA can save encrypted data file on cloud and perform the integrity verification without downloading the entire file. Once the user is revoked in the group, the cloud themselves re-sign the blocks so the efficiency of the user revocation is significantly improved in this scheme. TPA can perform multiple auditing tasks simultaneously this provides better efficiency. In the cluster based public auditing system each group consist number of group members, and they are uploaded large number of data. Sometimes some TPAs are very busy and the other one is idle depends on uploaded data. In this case we have plan to implement load balancing of TPA'S for the verification process. This is much more effective than the modified system. In this paper, we have compared existing and modified system in terms of verification time. Based on the comparison results we identified the modified system verification time is less as compared to the existing system. From the analysis we have identified that modified system is best for cloud environments.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Boyang Wang, Baochun Li, and Hui Li.[2015] "Public Auditing for Shared Data with Efficient User Revocation in the Cloud," *IEEE Trans. On Services Computing*, 8(1): 92-106.
- [2] B Wang, B Li, and H Li.[2014] Oruta: Privacy-Preserving Public Auditing for Shared Data in the Cloud, *IEEE Tns On Cloud Computing*,2(1): 43-56.
- [3] C Wang, Q Wang, K.Ren,W Lou.[2013] Privacy Preserving Public Auditing for Secure Cloud Storage, *IEEE Transactions on Computers*, 62.
- [4] B Wang, H Li, M Li.[2013] Privacy-Preserving Public Auditing for Shared Cloud Data Supporting Group Dynamics, *Proc IEEE Int'l Conf Comm. (ICC'13)*, 539-543.

- [5] B Wang, SS Chow, M Li, H Li.[ 2013] Storing Shared Data on the Cloud via Security-Mediator, *Proc. IEEE 33rd Int'l Conf. Distributed Computing Systems* (ICDCS'13), pp. 124-133.
- [6] B Wang, B Li, and H Li.[2012] Knox: Privacy Preserving Auditing for Shared Data with Large Groups in the Cloud, Proc. 10th Int'l Conf. Applied Cryptography and Network Security, PP.507-525, June.
- [7] C Wang, Q Wang, K Ren, and W Lou.[2010] Privacy-preserving public auditing for data storage security in cloud computing, in InfoCom 2010, *IEEE*.
- [8] Q. Wang, C Wang, J Li, K Ren, and W Lou.[ 2009] Enabling Public Verifiability and Data Dynamic for Storage Security in Cloud Computing, Proc. 14th European Conf. Research in Compute Security(ESORICS'09
- [9] G Ateniese, RD Pietro, LV Mancini, G Tsudik.[2008] Scalable and Efficient Provable Data Possession, Proc. Fourth Int'l Conf. Security and Privacy in Comm. Networks (ICST SecureComm'08)
- [10] G Ateniese, R Burns, R Curtmola, J Herring, L Kissner, Z Peterson, and D Song.[ 2007] Provable Data Possession at Untrusted Stores, Proc. 14th ACM Conf Computer and Comm Security (CCS'07), 598-610.
- [11] Parimala Raghavan, Dr. S Subasree.[2016] performance analysis of public auditing for shared data with efficient user revocation in cloud using RSA and AES algorithms, *International Journal of Future Innovative Science and Engineering Research (IJFISER)* ISSN (Online): 2454- 1966, 2(1) : 257.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# PERFORMANCE EVALUATION OF TRUST MODAL WITH INVASIVE WEED OPTIMIZATION

Loheswaran<sup>1</sup> and Premalatha<sup>2</sup>

<sup>1</sup> Sasurie College of Engineering, Tirupur District, INDIA

<sup>2</sup> Kongu Engineering College, Erode, INDIA

## ABSTRACT

*In the recent decade, the tremendous growth in IT venture had evolved a new term known as Cloud Computing. Cloud computing is the concept of sharing the environment and in turn sharing the infrastructure which may lead to the risk of illicit access of data. As the cloud computing involves sharing of infrastructure, scheduling plays the vital role. Another big challenge is providing Trust to the users in the commercial cloud environments. In the available heterogeneous infrastructure, users tend to opt of the best viable resources only based on the trust. Invasive Weed Optimization (IWO) is a population-based algorithm inspired from the process of weeds colonization and distribution. In this study, it is demonstrated that, with trust modal –IWO performs better results than other techniques.*

Published on: 28<sup>th</sup>– August-2016

### KEY WORDS

Cloud computing, Scheduling, Trust, Invasive Weed Optimization (IWO)

\*Corresponding author: Email: [loheswarank79@gmail.com](mailto:loheswarank79@gmail.com)

## INTRODUCTION

Cloud computing enables the centralized storage of data and access from anywhere using mobile device or thin client. It provides the easy access of data and resilience. In case of individuals, rather than data storage, data security is expensive. In cloud computing, the cloud provider is affordable to endow state of art security techniques than the individuals can. Nevertheless, as the cloud computing keeps the data control away from the data owner, data security issues are anticipated.

In cloud computing, significant facets of security of the associated technologies [1] are observed from the statements of the existing users and researchers. Security plays a vital part in adopting a service provider. As cloud computing involves the accessing the computing resources through internet, the data stored or transferred are prone to the security risks such as data confidentiality, reliability and accessibility which are the common threats of internet. In addition to these widespread risks, cloud computing is done globally and hence the users and service providers should act with bound to the international rules and regulations.

Most of the security risks of computing such as design of the architecture, access control privileges, malware attacks, surface attacks are common in the cloud computing too. Providing secure Cloud computing includes all these aspects. Data stored in cloud are exposed to the attacks irrespective of its storage layout. Apart from these most general risks, there are few specific security features are that explicit for cloud computing domain [2, 3]:

- Many tenants (Sharers) use one cloud and normally one tenant can be the attacker for the other.
- Cloud computing includes the service providers, subcontractors and employees who can access the data.
- Data stored in the cloud may be prone to be lost or modified by the cloud service provider.
- As the cloud is globally accessible by the tenants, the data may travel through various protocols and public networks which are widely insecure.

Cloud computing is the imminent technology being grown radically. As cloud computing involves the task of sharing the common available resources, it is important to achieve the effective resource utilization, quick response for the queries and optimum minimal time for the task completion. These intentions could be attained through a proper sequencing of jobs called job scheduling [4] which is a primary activity in the computing environments. Job scheduling is the most popular solution for optimization problems which enhances the integrity and flexibility of the system. Specific to cloud computing, to increase the efficiency and maximize the profit, an appropriate job scheduler is required. The main purpose of the job scheduling is to assign the jobs to the available resources without time delay and also in the proper sequence such that the jobs are executed in the exact transaction logic constraint. The effective attainment of this purpose will lead to the situation where the processors are equally loaded and their utilization is maximum by cutting the idle time and also the total execution time is minimized. In order to attain these objectives, a new scheduling algorithm need to be proposed that is applied for the task scheduler to surpass suitable allocation design of tasks on the available resources.

Max-min strategy is organized such that it selects the large task for executing first and hence it is noted that the small tasks are made to wait for long time.

It is demonstrated that trying to solve NP-hard scheduling problem [5] with a single objective is a complex task. At the same time solving the problem with more objectives is also difficult. Unlike the single objective optimization which has the evidently defined optimal solution, multi-objective optimization problem clasps a set of best conciliated solution called as Pareto optimal solution. Solution such devised from the Pareto optimal solution is considered to be of better quality than any other available solution in the search space. For multi-objective optimization, it is predicted that Invasive Weed Optimization (IWO) matches appropriately

Trust plays a vital role in the commercial cloud setting. Managing the Trust of the users define the commercial success of the cloud technology [6,7]. Performance trust could be achieved through the belief. Software as a service (SaaS), Platform as a Service(PaaS), and Infrastructure as a Service(IaaS) are the service models supported by cloud infrastructure. Cloud Service providers are expected to offer Trust and financially worthy services in all the three aspects such as software, platform and infrastructure.

In this study techniques such as with trust modal –IWO are proposed. Section 2 explains the literatures, section 3 explains in detail about the techniques used in methodology, section 4 discusses about the obtained results and finally section 5 concludes the work.

## RELATED WORKS

In the fast growing internet usage era, cloud computing is the best hopeful way to access the widely spread resources. Younis&Kifayat [8] investigated the issues related to probable security risks and in turn to construct a secure cloud infrastructure. The investigation emphasized on the security confronts related to the cloud and highlighted the requirements to be established for the construction of safe and secure cloud infrastructure by the service providers.

Swift scaling of data, remote storage and sharing of services in vibrant environment appears to be the advantage of cloud computing but at times they become disadvantage of cloud as these aspects tend to decrease the level of confidence in the imminent customers. The conventional mechanisms devised to solve these security and privacy issues are no longer effective in providing the rapid solution. Pearson [9] assesses the privacy, trust and security risks that are tend to occur and also thrash out the ways and means that could be used to handle them in the cloud environment.

Enhancement of the number of cloud deployments requires to root out the cloud related specific issues in addition to the consecutive risks exposed in the regular usage of Internet. Fernandes et al [10] surveyed the cloud security issues, and made the absolute review of related literature. The survey includes a comprehensive review of major aspects related to the security of the cloud environments, predicts that the key concepts of security such as open access of data, malware threats, pervasive attacks, categorization of the data and also confers various futuristic research topics.

Existent scheduling algorithms are brushing off the individual dependent and independent tasks. BrarRao [11] in his study proposes on Max- Min algorithm that could be implemented and focused for scheduling of job sequencing of dependent and independent jobs and ensures the profit of minimal computation time by enabling the parallel processing of independent tasks.

Fister et al [12] proposes a cloud model based Invasive Weed Optimization (IWO) algorithm. In this algorithmic model the weeds are categorized into excellent group, the normal group and the poor group. Through CR the algorithm fine tunes the standard deviation and each subgroup adopts diverse standard deviations to disperse. The group with modest standard deviation that recognizes the fine-grained search is the excellent group. The normal group recognizes the adaptive search as it applies cloud model's fuzzy logic and uncertainty to correct the standard deviation dynamically. Finally, as the poor group is liable to do the global search, it gets the higher standard deviation. This methodology brings better convergence rate, evade the trapping of

algorithm in local optimum search and also improves the equilibrium between the global and local search. New devised algorithm is experimented on various test functions and its performance proves to be better optimal.

Enhanced Invasive Weed Optimization (EIWO) algorithm is formulated to address the issues related to feasible rework, transfer time with an transmitter between the consecutive stages and predictive sequence of tasks. The ultimate criterion for the optimization is to minimize the task span. In the previous researches, Invasive Weed Optimization (IWO) is proved to be a proficient meta-heuristic algorithm, in view of enhancing its competence, Jolai et al [13] included mutation operation which improves the exploration of the global optimum instead of local optimum.

Added to this, a similarity function is enclosed to hinder impulsive junction. IWO exploration and exploitation could be balanced with this. An investigational methodological design named Response Surface Method (RSM) is applied to determine the performance of the proposed algorithm EIWO. The proposed algorithm is assessed primarily by the generation of random test problems and the results obtained are compared with three well known bench mark algorithms and the result analysis is done with the help of various statistical tools. Finally it is proved that the proposed EIWO (Enhanced Invasive Weed Optimization) algorithm provides better solution to the problem.

Modified invasive weed optimization (IWO) algorithm is proposed by Mkeni&Fayech [5]. The algorithm proposed by them is designed to optimize the multi-objective flexible job scheduling problems(FJSSPs). Its performance is based on the three criteria, viz., Minimization of maximum completion time (makespan), the workload distributed to the machines, and workload on the specific decisive critical machine in the cloud. IWO, which is a well known bio-inspired meta-heuristic algorithm that imitates the weed behavior of inhabiting in a right ecological atmosphere which is best suited for its growth and reproduction. Unlike other algorithms IWO is specifically designed and developed to solve continuous optimization problems. Hence the discrete job sequences are converted in to continuous position values by using Smallest Position Value (SPV). Successful achievement of these performance criteria using this proposed Modified invasive weed optimization algorithm demonstrates the highest competence of the algorithm to discover the best optimal solution when compared with other existing algorithms.

## METHOD

### Max-Min Scheduling

The situation where there exists a distributed environment that commences with a sequence of impulsive unscheduled jobs, generally there the Max-min algorithm is deployed. The deployment of this Max-min algorithm is followed by the estimation of the expected execution flow matrix and the maximum and minimum expected time of individual task completion in the available existing resources. As a next step of action, among the entire task, the task which has maximum overall expected time of completion is selected and it is assigned to the other source which has minimum overall time of execution. Finally, the task which is scheduled recently is removed from the set of meta-tasks, and the calculated times are updated. These steps are repeated until the meta-tasks set are empty [14].

In max-min protocol, as given in [Figure 1],  $r_j$  denotes the ready time of resource  $R_j$  for executing a job, and  $C_{ij}$  as well as  $E_{ij}$  denote the anticipated completion as well as execution time correspondingly. As displayed,  $T_k$  with maximal anticipated completion time is selected to be designated for related resource  $R_j$  which provides minimal execution time.

For every job in meta - jobs;  $T_i$

For every resource  $R_j$

$$C_{ij} = E_{ij} + r_j$$

While meta - job is not empty

identify job  $T_k$  that takes up maximal completion time

designate  $T_k$  to resource  $R_j$  that provides minimal execution time

discard  $T_k$  from meta - jobs set

update  $r_j$  for chosen  $R_j$

update  $C_{ij}$  for  $i$

**Fig:1. Pseudo code for Max-Min Algorithm**

.....



## Trust in Cloud Computing

In the recent decades, apart from various other fields, health care service has emerged as one of the service that is extended globally. At the same time, it is the field which contains the most sensitive data too. Hence, the global health care providers are in search of an appropriate and well secured cloud provider available in the market. In spite of the assurance of all the cloud providers who describe themselves as the best secured environment, and declare to act as per the requirements of the health care provider's, the health care provider's need to do the burden task of comparing the quality of the services offered by various cloud service providers to choose the trust worthy. The process of this comparison involves the procedure of analysis of SLAs and recording the clauses as per the requirement and legal follow ups of the audit standards regarding the security task [15].

Another significant process of security conformity in cloud is audit. Generally the term refer to secretarial. Actually, audit is the process of evaluating a person, system, firm, company or development of a product. In the context of cloud, auditing could be used to ensure the reliability of the data that is stored in the cloud. Corresponding audit and the Field audit are the two types of audit that are commonly taken in the process of execution. Audit is used to categorize the received data for storage, evaluate and authenticate the level of integrity of the data. On these basis the user would be responded. File Storage: Once the files are received by the cloud for storage it can be stored first and then the above mentioned audit process is performed on the file. The advantage of this file storage method are (i) Save Time: This is because the conventional methods of auditing are time consuming which involves large number of interferences and includes many manual method for the further proceedings of the storage of data. (ii) Savings: The main aim of any business to have cost cutting technique which leads to the economical savings while looking for the data storage in the cloud is attained with this type of audit process [16].

In a dynamic multi-owner environment of cloud, the secured storage of data is ensured by the utilization of a new proposed scheme based on code generation and public auditing. This scheme assures the data integrity of fresh and renewal data by the establishment of a third party auditor and an individual proxy on the data owner's behalf. Data owners are liberated from the online burden towards the renewal of spoiled blocks by the implementation of High Availability Proxy servers. This High Availability Proxy server can also find and block the unauthenticated activity on the data by any individual or group.

This technique could be made clear with the following example:

The public and private keys are generated by the staff; authenticator regeneration is given to a proxy through sharing of private keys partially. Thus generated encoded blocks and authenticators are uploaded and distributed by the staff on the cloud servers. On any detection of data corruption, the information is sent to the readily existing High Availability Proxy which will regenerate the data blocks and equivalent authenticators by itself in a high secure method. This technique overlay the way to make certain that a team of staff can be under one project and in turn there can be subgroups to access or modify the data.

## Invasive Weed Optimization (IWO)

In 2006, Mehrabian and Lucas [17] proposed an original stochastic optimization model, invasive weed optimization algorithm. This algorithm got motivated and well identified from agricultural experience: invasive weed colonization. It provides easy understanding, straight forward programming technique and also robustness to the process. These advantages have made the programmers to pertain this algorithm in various field of modern engineering [18].

The existing traditional IWO, Weeds and the Population refers to the feasible solutions to the particular problem and the set of such weeds respectively. Thus generated weeds are spread over the complete search area. As per the agricultural concept, which is the root proposal for IWO, each weed would produce new weeds based on its fitness function. Such reproduced weeds are disseminated over the search space using the distributed random number whose mean is ensured to be equivalent to zero. The regeneration iteration process goes on till it reaches the maximum number of weeds or otherwise maximum population. In the regeneration competition, as per the survival of the fittest theory, those weeds which are with best fitness will survive and the remaining will get terminated. This process will proceed till the maximum numbers of iterations are reached. With this apparent process, the weed with better fitness function which falls very near to the best optimal solution is achieved.

The procedure is detailed thus:

- **Step 1: Initialize a population** A set of arbitrary solutions are distributed over the D dimensional search space with arbitrary positions.
- **Step 2: Reproduction** Greater the weed fitness, more seeds it yields. The equation of weeds yielding seeds is:

$$weed_n = \frac{f - f_{min}}{f_{max} - f_{min}} (s_{max} - s_{min}) + s_{min}$$

Wherein  $f$  refers to the current weed's fitness.  $f_{max}$  as well as  $f_{min}$  correspondingly denote maximal as well as worst fitness of current population.  $s_{max}$  as well as  $s_{min}$  correspondingly denote maximal as well as worst value of weeds.

- **Step 3: Spatial dispersal** The created seeds are arbitrarily spread across D dimensional search space through arbitrary numbers that have normal distribution and mean zero, but with differing variance. This guarantees that the seeds are arbitrarily spread such that they abide close to parent plant.

- Step 4: Competitive exclusion** Once a certain number of iterations are done, the quantity of weeds in the colony reach their maximum (P MAX) through rapid reproduction. During this time, all weeds are permitted to yield seeds. The yielded seeds are permitted to distribute themselves across the search space. When every seed has discovered its position in the search space, they are ranked together with their parents as a colony of weeds. Then, weeds with least fitness are discarded for reaching maximal permissible population in the colony. In this manner, weeds as well as seeds are ranked and those with greater fitness live and are permitted to duplicate. [Figure -2] shows the pseudo code for IWO.

```

Start{
• Set up population of weeds, set variables; • Current_iteration=1;
  While (Current_iteration<Max_iteration) do
  {
• Calculate the greatest and least fitness in the population
• Calculate the standard deviation stdon the basis of iteration
  For all weeds w in the population W
  {
• Calculate the quantity of seeds for w on the basis of its fitness
• Choose the seeds from the potential solutions near the parent weed w in a neighbourhood with normal distribution possessing mean=0 as well as standard deviation=std;
• Append seeds yielded to the population W
If (|W|>Max_SizePopulation)
{
• Rank the population W as per their fitness
• W=SelectBetter(weed,seed,Max_SizePopulation)
}End if
}End for
Current_iteration=Current_iteration+1;
}End while
}End
    
```

Fig. 2. Pseudo code for IWO

The IWO was originally formulated for solving continuous optimization issues, but it is not capable of being employed to discrete issues in a direct manner. Individuals are to be coded adequately for solving scheduling issues. In the current work, coding which considers all restraints as well as specifics of the issue is implemented. For (n jobs, m machines, O operations) FJSSP, all plants are denoted by four aspects: every aspect consists of  $2 \times O$  quantity of dimensions.

## RESULTS AND DISCUSSION

For experiments, the number of tasks taken are 200, 400, 600, 800 and 1000. The methods such as without trust model- max min and with trust model- max min are used for obtaining the average schedule length and ratios of successful execution. [Table -1], [Table -2] and [Figure -3],[Figure -4]shows the average schedule length and ratios of successful execution respectively.

Table: 1. Average Schedule Length

Number of tasks	Without Trust Model - Max Min	With Trust Model Max - Min	Without Trust Model - Invasive Weed Optimization	With Trust Model- Invasive Weed Optimization
200	359	341	357	338
400	1129	1074	1101	1042
600	1902	1805	1854	1764
800	2610	2484	2611	2479
1000	3374	3201	3300	3118

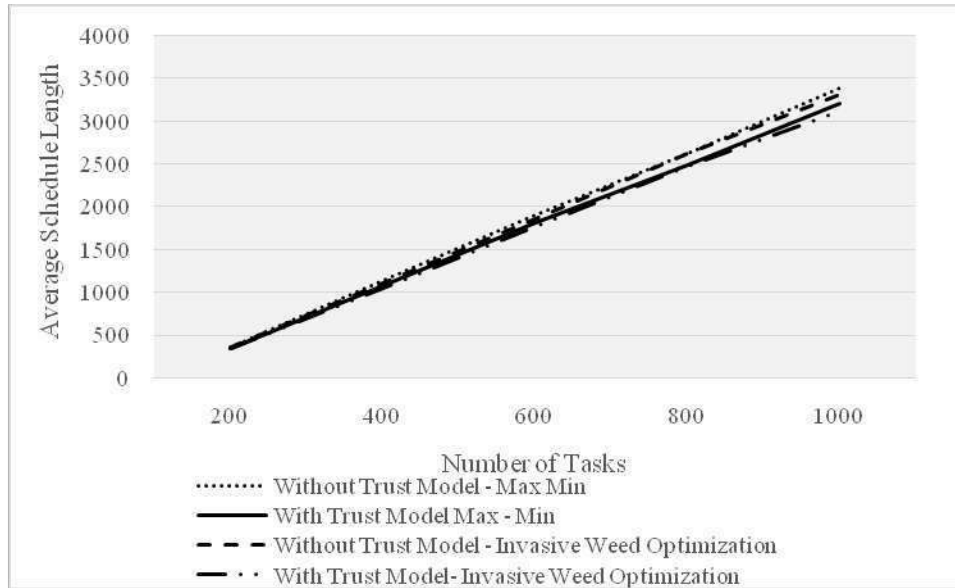
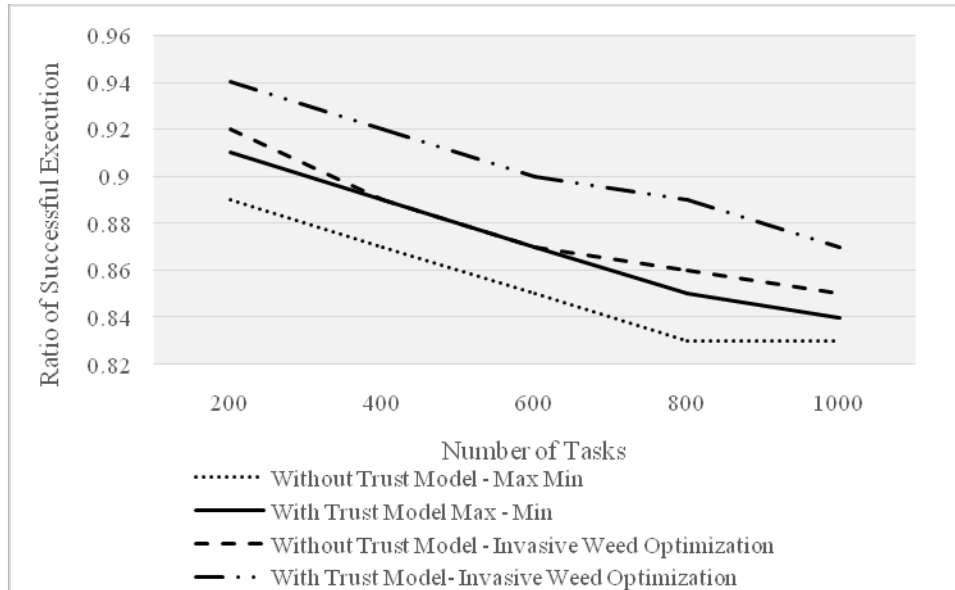


Fig: 3. Average Schedule Length

Table- 1 and Figure- 3 shows that the average schedule length of with trust model Invasive Weed Optimization (IWO) performs better by 6.03% than without trust model max min, by 0.88% than with trust model max min and by 5.47% than without trust model Invasive Weed Optimization (IWO) for number of tasks is 200. For number of tasks 1000, the average schedule length of with trust model Invasive Weed Optimization (IWO) performs better by 7.89% than without trust model max min, by 2.63% than with trust model max min and by 5.67% than without trust model Invasive Weed Optimization (IWO).

Table: 2. Ratios of Successful Execution

Number of tasks	Without Trust Model - Max Min	With Trust Model Max - Min	Without Trust Model - Invasive Weed Optimization	With Trust Model - Invasive Weed Optimization
200	0.89	0.91	0.92	0.94
400	0.87	0.89	0.89	0.92
600	0.85	0.87	0.87	0.9
800	0.83	0.85	0.86	0.89
1000	0.83	0.84	0.85	0.87



**Fig: 4. Ratios of Successful Execution**

[Table -2], [Figure -4] shows that the Ratios of Successful Execution of with trust model Invasive Weed Optimization (IWO) performs better by 5.46% than without trust model max min, by 3.24% than with trust model max min and by 2.15% than without trust model Invasive Weed Optimization (IWO) for number of tasks is 200. For number of tasks 1000, the Ratios of Successful Execution of with trust model Invasive Weed Optimization (IWO) performs better by 4.71% than without trust model max min, by 3.51% than with trust model max min and by 2.33% than without trust model Invasive Weed Optimization (IWO).

## CONCLUSION

Cloud computing presents a rising optimistic expectation among the service providers and data owners in using this highly developed technology. As cloud computing is engaged in decentralization of data, where it is stored in the resources spread globally and also disseminated to the users through various different geographically located data centers, trust is believed to be the most significant concern. The implementation of Max-Min algorithm directly benefits in minimizing computation time.

Results show that the Ratios of Successful Execution with trust model Invasive Weed Optimization (IWO) performs better by 5.46% than without trust model max min, by 3.24% than with trust model max min and by 2.15% than without trust model Invasive Weed Optimization (IWO) for number of tasks is 200. For number of tasks 1000, the Ratios of Successful Execution of with trust model Invasive Weed Optimization (IWO) performs better by 4.71% than without trust model max min, by 3.51% than with trust model max min and by 2.33% than without trust model Invasive Weed Optimization (IWO). The average schedule length for proposed method is reduced than other existing works.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Mell P, Grance T. [2011] The NIST definition of cloud computing.
- [2] Yanpei Chen, Vern Paxson, and Randy H. Katz. What's new about cloud computing security? Technical Report UCB/EECS-2010-5, Electrical Engineering and Computer Sciences, University of California at Berkeley, 2010.
- [3] Ryann MD.[2013]. Cloud computing security: The scientific challenge, and a survey of solutions. *Journal of Systems and Software*, 86(9): 2263-2268.
- [4] Salot P. [2013] A survey of various scheduling algorithm in cloud computing environment. *International Journal of research and engineering Technology (IJRET)*, ISSN, 2319-1163.
- [5] Mekni S, Fayéch BC. [2015] Multiobjective Flexible Job Shop Scheduling Using A Modified Invasive Weed Optimization. *International Journal on Soft Computing*, 6(1), 25.
- [6] Talal H Noor, and Quan Z Sheng.[ 2011] Trust as a Service: A Framework for Trust Management in Cloud Environments, Web Information System Engineering – WISE 2011, *Lecture Notes in Computer Science* , 6997: 314-321.
- [7] Fabrizio Messina, Giuseppe Pappalardo, Domenico Rosaci, Corrado Santoro, and Giuseppe ML Sarné.[ 2013] A Trust-Based Approach for a Competitive Cloud/Grid Computing Scenario, Intelligent Distributed Computing VI, *Studies in Computational Intelligence* 446: 129-138.
- [8] Younis MYA., Kifayat K. [2013] Secure cloud computing for critical infrastructure: A survey. *Liverpool John Moores University, United Kingdom, Tech. Rep.*
- [9] Pearson, S. [2013] Privacy, security and trust in cloud computing. In *Privacy and Security for Cloud Computing* (pp. 3-42). Springer London.
- [10] Fernandes DA, Soares LF, Gomes JV, Freire M M, Inácio PR. [2014] Security issues in cloud environments: a survey. *International Journal of Information Security*, 13(2), 113-170.
- [11] Brar SS, Rao S. [2015] Optimizing Workflow Scheduling using Max-Min Algorithm in Cloud Environment. *International Journal of Computer Applications*, 124(4)
- [12] Fister Jr I, Yang XS, Fister I, Brest J, Fister D. [2013] A brief review of nature-inspired algorithms for optimization. *arXiv preprint arXiv:1307.4186*.
- [13] Jolai F, Tavakkoli-Moghaddam R, Rabiee M, Gheisariha, E.[2014]An enhanced invasive weed optimization for makespan minimization in a flexible flowshop scheduling problem. *ScientiaIranica. Transaction E, Industrial Engineering*, 21(3): 1007.
- [14] Elzeki OM, Reshad MZ, Elsouid MA. [2012] Improved max-min algorithm in Cloud computing. *International Journal of Computer Applications*, 50(12).
- [15] Ming Li, Shucheng Yu, KuiRen and Wenjing Lou, "Securing Personal Health Records in Cloud Computing: Patient - Centric and Fine-Grained Data Access Control in Multi-owner Settings", in Security and Privacy in Communication Networks, Proceedings of 6th International. Conference on Security Privacy Communication Network, 2010.
- [16] Loheswaran K, Remalatha J. Renaissance System Model Improving Security and Third Party Auditing in Cloud Computing. *Wireless Personal Communications*, 1-16.
- [17] Ding S, Huang H, Yu J, Wu F. [2014] Polynomial smooth twin support vector machines based on invasive weed optimization algorithm. *Journal of Computers*, 9(5):1226-1233.
- [18] GR Gourab, D Swagatam, C Prithwish and NS Ponnuthurai. Design of Non-Uniform Circular Antenna Arrays Using a Modified Invasive Weed Optimization Algorithm, *IEEE Transactions on antennas and propagation*, 59(1):110-118.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# USING OPTIMIZED FEATURE SELECTION FOR CLASSIFICATION OF BRAIN MRI

Chellammal<sup>1</sup> and Venkatachalam<sup>2</sup>

<sup>1</sup> Bharathiar University, Coimbatore & Assistant Professor, Dept of Computer Applications, Erode Sengunthar Engineering College, Erode. INDIA

<sup>2</sup> Principal, The Kavery Engineering College, Mecheri, INDIA

## ABSTRACT

Several studies into the detection of brain anomalies have been carried out because of its considerably significant role in the identification of anatomical areas of interest for diagnosing diseases, treating illnesses or even the planning of surgeries. Alzheimer's disease (AD) is the most typical kind of dementia amongst elderly people across the world. Magnetic resonance imaging (MRI) is a method which yields images of excellent quality of the anatomical features of the human body, particularly in the brain and offers clinical data supporting diagnoses as well as for biomedical research. The current study is a features selection as well as classification study for normal brain subjects. During the features selection phase, features are chosen through Bacterial Foraging Optimization (BFO). Experimental evaluation was carried out for several other features selection techniques such as Information Gain (IG), Minimum Redundancy Maximum Relevance (mRMR) as well as classifiers such as Instance-based Learning (IBL), C4.5 as well as Fuzzy Classifiers.

Published on: 28<sup>th</sup>– August-2016

### KEY WORDS

Image retrieval, Feature selection, Information Gain (IG), Bacterial Foraging Optimization (BFO), Fuzzy Classifier

## INTRODUCTION

Humans have an extremely complex brain anatomy because of its complicated structure as well as functions. The brain is an integral part of the Central Nervous System (CNS) and is the center that controls the mental processes as well as physical actions of the human body. Brain abnormalities are symptoms wherein motor impairments or neuropsychological issues impact the CNS [1]. It is typically an anomalous growth of cells in the brain that may or may not be carcinogenic in nature. Several studies into the detection of brain anomalies have been carried out because of its considerably significant role in the identification of anatomical areas of interest for diagnosing diseases, treating illnesses or even the planning of surgeries.

Alzheimer's disease is a typical form of dementia and mostly presents itself in elderly people. Almost thirty million individuals in the world are afflicted with Alzheimer's and due to the rise in life expectancies the figure is expected to be tripled by the year 2050. Identifying efficient biomarkers is critical so as to diagnose as well as treat AD earlier in people who are most vulnerable to the illness. Mild Cognitive Impairment (MCI) is the stage of transition between age-related deterioration in cognition and Alzheimer's or between the earliest identifiable phase of progression toward dementia or Alzheimer's [2]. On the basis of earlier work, it is known that a considerable set of MCI afflicted individuals, around 10-15% will end up with Alzheimer's annually. Alzheimer's is recognized through the formation of intra-cellular neuro-fibrillary tangles as well as extra-cellular  $\beta$ -amyloid plaques and tremendous loss of synapses as well as neuronal deaths (atrophy) in the brain. The growth of the neuropathology in Alzheimer's may be noted for several years prior the appearance of the clinical symptoms of the illness.

In the current work, a new MRI-based method for detecting Alzheimer's conversion earlier on in MCI afflicted individuals is suggested through the usage of sophisticated machine learning protocols as well as combination of MRI information with standard neuropsychological test results. In further detail, the objective is the prediction of whether MCI afflicted individuals will develop Alzheimer's within a three year period through usage of baseline data. The information utilized in the current work is taken from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset that has MRIs as well as neuropsychological test results from standard Normal Controls (NC),

Alzheimer's, and MCI cohorts. In recent studies, the focus has been on the prediction of transformation of MCI to Alzheimer's through the usage of MRI, Positron Emission Tomography (PET), Cerebro-Spinal Fluid (CSF) biomarkers as well as demographic/cognitive data.

Recently there has been a rise in the amount of software as well as hardware technology, several kinds of medical images like MRI, Computer Tomography (CT), X-Ray, ultrasound among others which are being produced in several medical centers. The medical scans provide significant anatomical as well as functional data regarding various body parts for detecting, diagnosing, treating diseases as well as for research and educational purposes [3]. Hence, the issue of efficient storage, processing as well as retrieval of medical image data is an important research area. An effective retrieval system is required for efficient organization as well as retrieval of medical images. Image retrieval is a model that is capable of browsing, searching as well as retrieving images from large datasets in an automated manner.

Image Retrieval refers to the job of looking for images from an image dataset. Queries to the dataset may be of any one of the several kinds given below:

- Query-by-text: Textual descriptions of image being searched for, is given as query.
- Query-by-sketch: Sketch of image being searched for is given as query.
- Query-by-example: Example image that is like the one being searched for is given as query.

In Image Retrieval, interactions are allowed with the users and these permit the users to manage their searches thereby providing assistance in the reaching of excellent results rapidly. One generally utilized method in information retrieval is relevance feedback. That is to say, after initial searches, users are given a list of results. It is quite obvious that certain results will match the queries while some do not. Users then mark the matching results as relevant while the remaining are deemed irrelevant. It is proven that relevance feedback is particularly helpful in information retrieval jobs which excellent results have been attained in image retrieval too.

Diagnostic MRIs are excellent clinical tools for visualization of organs as well as soft tissue in human skulls with no negative effects. It permits the healthcare professional to choose correct image plane for displaying pathological anatomies in an accurate fashion. The key points are that it is safe for handling, not radiological as well as non-invasive. Identification of brain diseases is the most accurate through these grey scale images. Conventionally, the determination of normal or anomalous depends on radiologist expertise [4]. The decisions are also highly reliant on their experience that might be related to particular features from their visual interpretations of the image or certain comparisons with other pathologies.

Images may be characterized in terms of basic attributes known as features, which can be categorized into two [5]: 1) Natural features which are delineated on the basis of visual appearance of the image and 2) Artificial features which are due to manipulation of images.

Features selection is extremely significant as all features are not useful in image retrieval systems. Certain features might interfere and reduce the success rates of the model. The primary goal of features selection therefore is the selection of a set of optimal features from a huge quantity of features. Accuracy is maximized while retrieval is simplified. Features selection may also be described as the choosing of amalgamations which best describe a features set. Features selection is a vast research area ever since the 1970s in the fields of pattern recognition, image retrieval as well as several other research domains.

Extra or repetitive features are discarded through the usage of dimensionality reduction methods and adequate quantity of useful features is extracted. Intrinsic dimensionality is needed for representing image feature values. Because of this reason there is a lot of interest in the reduction of dimensionality of descriptors through the stabilization of unique topologies of higher dimensional spaces. Dimensionality reduction methods in literature include Principal Component Analysis (PCA), Weighted Multi-Dimensional Scaling, Tabu Search (TS).

In recent years, Evolutionary Algorithms (EA) operate on the population of potential solutions through the relation of the presence of fittest yields best estimates to a solution. The primary benefit of usage of Evolutionary Algorithms is the searching of a set of potential solutions in a simultaneous manner for finding optimum features selection with less runs of the protocol. The most commonly utilized Evolutionary Algorithms for optimization of features are Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Gravitational Search Algorithm (GSA)

Ant Colony Optimization (ACO). Generally, algorithms are sorted on the basis of the technique of evolution. For instance, GA evolves through every subsequent population of chromosomes while PSO as well as ACO update as per social behavior.

Classification as well as clustering are important components of image mining. Machine learning methods are utilized for reducing semantic gaps between low-level image features as well as high-level semantic features. Data classification is a two-stage procedure, comprising learning stage as well as classifications stage [6]. Classification algorithms are employed on image datasets wherein images are described for classification into classes. Classification is a problematic task in several fields such as biomedical imaging, video surveillances, vehicular navigation, remote sensing and so on and so forth. Image classification comprises three stages:

Feature extraction – Here, features are extricated from sample images which are pre-labeled and features descriptions for all images are established.

Training – Here, samples of all classes are trained and model descriptions for all classes are set up.

Classification – Model is utilized for classifying as well as indexing images which are not labeled.

Classifying pixels is useful in several areas within the domain of image processing. Particularly speaking, classification of pixels in image is helpful as a pre- or post-processing stage for the issue of image segmentation, that is, for facilitating segmentation or for refining the results. The usage of pixel classifications may also be employed for improving the performance of certain applications in the retrieval of images from the dataset.

In this paper Bacterial Foraging Optimization algorithm is proposed for feature selection and the experiment results compared with the other methods to evaluate the proposed method. The remaining sections organized as: Section 2 reviews the related work in literature. Section 3 explains the methods which are used in the proposed work. Section 4 discusses the experiment results and section 5 concludes the proposed work.

## LITERATURE REVIEW

Veeramuthuet al., [7] utilized Spatial Gray Level Difference Method (SGLDM) features extraction protocol as well as Correlation based Feature Selection (CFS). Projected Classification protocol (PROCLASS) is suggested for brain image data. Experimental evaluation was carried out for comparison of the plug-in protocols with FAST, FCBF features selection protocols.

Spedding et al., [8] looked into the issue of features selection in neuroimaging features from structural MRI brain images for classifying subjects as healthy (controls), MCI-afflicted or AD afflicted. Genetic algorithm wrapper technique for features selection as utilized in conjunction with Support Vector Machine (SVM) classifier. Greatest accuracy attained was during the classification of test data (65.5%) through usage of genetic algorithms for features selection with three-class SVM classifier.

Padilla et al., [9] suggested a new computer-aided diagnosis (CAD) method for diagnosing Alzheimer's earlier on the basis of non-negative matrix factorization (NMF) as well as Support Vector Machine with bounds of confidence. Resultant NMF-transformed datasets contained lesser quantity of features and are sorted through usage of SVM-based classifier with bounds of confidence for decisions. The suggested NMF-SVM technique provides 91% classification precision with extremely high sensitivity as well as specificity rates (greater than 90%). The NMF-SVM CAD tool is an accurate technique for SPECT as well as PET AD classification.

A smart system was formulated by Gopal&Karnan [10] for diagnosing brain tumors through usage of image processing clustering protocols like Fuzzy C Means alongside intelligent optimization techniques like GA as well as PSO. Detecting tumors happens in two stages which are pre-processing as well as improvement in the first and segmentation as well as classification in the second.

Sweety&Jiji [11] suggested Particle Swarm Optimization for reducing features as well as decision tree classifiers for classifications. Detecting Alzheimer's early happens in three stages: 1) features like Eigen vectors, Eigen brain, eman, variance, kurtosis and so on are extricated from MRIs and 2) Number of features are reduced through Particle Swarm Optimization and 3) Decision Tree classifiers are utilized for detecting if brain images are impacted by Alzheimer's or not.

Sasikala&Kumaravel [12] suggested as well as contrasted features selection protocols for detecting glioblastoma multiform in brain images. Textural attributes are extricated from normal as well as tumorous regions (RoI) through usage of spatial grey level dependence methods as well as wavelet transforms. Artificial neural networks are utilized for classification. PCA, traditional sequential techniques as well as floating search protocol are contrasted with genetic algorithms with regard to best recognition rates obtained as well as optimum quantity of features. Genetic algorithms achieve classification performance of 97.3% with optimum features when contrasted with sequential method as well as Principal Component Analysis.



Brain atrophy with certain standard positions was valuated by Alam et al. [13] through usage of dimensionality reduction techniques. Comparisons were carried out between PCA as well as manifold learning through usage of LaplacianEigenmaps for quantifying brain atrophy. Additionally, a new technique has been suggested using both Principal Component Analysis as well as Manifold Learning that valuates brain atrophy with respective age groups. The suggested technique performs better than both dimensionality reduction technique with a score of ( $p < 0.0030$ ). The discoveries indicate that multivariate network analyses of deformation maps identifies generic features of atrophy and offers an excellent tool for predicting brain atrophy with age.

## METHOD

In this section we discussed different feature selection methods and classifiers used in the proposed work.

### FEATURE SELECTION

Features selection is a global optimization issue in machine learning that decreases quantity of features, discards non-relevant, noise-filled as well as repetitive information and leads to adequate recognition accuracy. Though features selection is generally carried out for selecting relevant as well as useful features it also reduces computational overheads.

### INFORMATION GAIN (IG)

The primary objective of IG criterion is the discovery of the quantity of unique data added by one feature to the entire features set. A feature's IGf may be calculated by  $F(S \cup f) - F(S)$ , wherein  $F(.)$  refers to the evaluator criteria while  $S$  represents the chosen features subset. Features with more IG are favored.

IG is a metric based technique utilized for choosing best split features in decision tree classifiers as well as denotes the extent to which data's entropy is decreased. Furthermore, it detects values of every particular feature. All features basis obtain IF values that are utilized for deciding whether features are chosen or discarded. Therefore, threshold values for features selection are to be setup initially, features are selected when IG values are greater than threshold ones. Assume there is a set  $A$  comprising  $s$  samples and a set  $B$  comprising  $k$  classes. If  $P(B_i, A)$  is the fraction of samples in  $A$  which have class  $B_i$ , then, the anticipated information for class membership is expressed through [14]:

$$Info(A) = - \sum_{i=1}^k P(B_i, A) * \log(P(B_i, A))$$

The greater the IG, the greater the chance of obtaining pure classes in target class if splits are based on the parameter with the greatest gain.

### MINIMUM REDUNDANCY-MAXIMUM RELEVANCE (MRMR)

Minimum Redundancy-Maximum Relevance (mRMR) refers to a mutual information based technique and it chooses features as per maximal statistical dependency criteria. Because of the challenges involved in the direct implementation of maximal dependency condition, mRMR is an approximation for maximization of dependency between joint distributions of chosen attributes as well as the classification parameter [15]. Minimization of redundancy for discrete attributes as well as continuous attributes is given by:

$$W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i, j)$$

For Discrete attributes:  $\min W_I$ ,

$$W_C = \frac{1}{|S|^2} \sum_{i,j \in S} |C(i, j)|$$

For Continuous attributes:  $\min W_C$ ,

wherein  $I(i, j)$  as well as  $C(i, j)$  refer to mutual information as well as the correlation between  $i$  and  $j$ , correspondingly. Maximization of relevance for discrete as well as continuous attributes is given by:

$$V_I = \frac{1}{|S|^2} \sum_{i \in S} I(h, i)$$

For Discrete attributes:  $\max V_I$ ,

$$V_C = \frac{1}{|S|^2} \sum_i F(i, h)$$

For Continuous attributes:  $\max V_C$ ,

wherein  $h$  refers to the target class while  $F(i, h)$  represents the F-statistic.

## PROPOSED BACTERIAL FORAGING OPTIMIZATION (BFO)

Features selection is a global optimization issue in machine learning which optimizes/reduces the quantity of repetitive as well as noise-filled attributes, discards inappropriate attributes resulting in adequate accuracy. Bacteria travel in an arbitrary fashion for finding greater amount of nutrients. Therefore the optimization method is helpful when gradients of cost functions are not made known. BFO, a non-gradient optimization method, is excellent due to its relative mathematical simplicity.

The notion of Bacterial Foraging Optimization has its basis in the fact that natural selection leads to the elimination of species with inadequate foraging schemes. After several generations, inadequate foragers are either wiped out or evolved into excellent ones. For instance, the E coli bacteria's foraging scheme is governed by four procedures which are chemotaxis, swarming, reproduction, as well as elimination and dispersal.

### CHEMOTAXIS:

Chemotaxis is attained through swimming as well as tumbling. On the basis of flagella rotation in all bacteria, it is decided whether they should travel in a particular direction (which is known as swimming) or if they should travel in another direction (which is known as tumbling), for the duration of the bacteria's total lifetimes.

### SWARMING:

Bacteria which have obtained optimal route of food attracts other bacteria for ensuring that they also arrive at that place in a fast manner.

### REPRODUCTION:

Those bacteria that are not healthy die while healthy ones are divided into two and situated in one location so as to maintain constancy in the population of bacteria.

### ELIMINATION AND DISPERSAL:

There is a possibility that in local environments, population lives change in a gradual manner through consumption of nutrients or because of arbitrary outside influences.

Extricated features are decreased through the usage of Bacterial Foraging Optimization for removing repetitive as well as non-relevant attributes and the resultant features subset is the most representative one. The positions of bacteria are either 0 or 1 on the basis of whether features are chosen or not in the search space. In the period of Chemotaxis, tumbling results in novel arbitrary positions which determine whether features are chosen or not in the subsequent iteration. Fitnesses are valued for all bacteria and positions are updated if fitnesses are improved. Bacteria with the worst fitness are discarded while those with best fitness are split (reproduction). When the iterations are completed, positions of bacteria represent the most optimal features subset obtained [Table-1].

Table: 1. BFO Parameters

Parameter Name	Description
$J_{cc}$	Cost function value
$J_{Health}$	Health of bacterium $i$
$L$	Counter for elimination-dispersal step
$P_{ed}$	Probability of occurrence of elimination-dispersal events
$S$	Population of the E. coli bacteria
$\omega_{attract}$	Width of attractant
$\omega_{repellant}$	Width of repellent

## CLASSIFIERS

Classification accuracy relies primarily on two elements: accurate features vectors which can identify uniquely the content of a single image; the image classification technique with excellent results which are similar to human beings' perceptions.

### C4.5

C4.5 is a protocol that is utilized for generating decision trees and was formulated by Ross Quinlan. It is an extension of a previous ID3 protocol. Decision trees created by C4.5 may be utilized for classification and for this particular reason, c4.5 is typically known as a statistical classifier. C4.5 protocol utilizes IG as splitting criterion. It may input data with categorical or numerical values. For handling continuous values, it creates threshold and later splits features with values greater than threshold as well as values equal to or lesser than the threshold. C4.5 is capable of handling missing values [16] because missing feature values are not used in gain computations by C4.5.

C4.5 utilizes Gain Ratio as a features selection metric for building decision trees. It discards the bias of IG when there are several outcome values of a feature. Initially, the gain ratios of all attributes are computed. Root nodes are the attributes whose gain ratios are maximum. C4.5 utilizes pessimistic pruning for removing non-required branches in the decision tree for improving classification accuracy.

## INSTANCE-BASED LEARNING (IBL)

IBL is based in a set of machine learning protocols and it popular as memory-based learning as well as case-based learning. IBL protocols store the information and begin processing solely when predictions are demanded, which is why it is known as a lazy learning technique. IBL protocols do not generate extensional concept descriptions. As an alternate, concept descriptions may be determined by how IBL's chosen similarity as well as classification functions utilize the current set of saved distances. These functions are two of the three elements in the model given below which defines all IBL protocols:

**Similarity Function:**It computes similarity between training samples  $i$  as well as samples in the concept description. Similitude is a numeric value.

**Classification Function:**It acquires similitude function's outcomes as well as classification performance records of samples in the concept descriptions. It outputs classification for  $i$ .

**Concept Description Updater:**It maintains records on classification performances and determines the samples to be included in concept descriptions. Input involves  $i$ , classification outcomes, similitude outcomes as well as current concept description. Output is the altered concept description.

Instance-based Learning decreases the quantity of training samples stashed to a minute set of representative ones. A further benefit of Instance-based learning is that it may be utilized in issues other than classification.

## FUZZY CLASSIFIER

Certain problems occur because of the imbalanced classes as well as errors in object segmentation. This is the motivation for usage of fuzzy rule-based classifiers. Fuzzy sets are sets whose elements possess degrees of membership. Elements of fuzzy sets may be full members or partial members. This implies that membership values designated to elements are not constrained to merely two values (0 and 1). It may be 0, 1 or any value between the two. Mathematical functions that define the degree of membership of elements in fuzzy sets are termed membership functions. The linguistic description of the problems rather than precise numerical descriptions is the primary benefit of the fuzzy set theory.

Fuzzy inference refers to the procedure of formulating mapping from particular input to output through usage of fuzzy logic. Procedure of fuzzy inference includes: membership function, fuzzy logic operator as well as if-then rule. Fuzzy sets as well as fuzzy operators are the subjects as well as verbs of fuzzy logic. Typically the knowledge in fuzzy reasoning is described as a rule in the form given below:

If  $x$  is A Then  $y$  is B

wherein  $x$  as well as  $y$  refer to fuzzy parameters while A as well as B refer to fuzzy values. If-component of the rule "x is A" is known as the antecedent or premise and the then-component "y is B" is known as the consequent or conclusion. Statements in the premise or conclusion components of the rule may include fuzzy logical connectives like AND, OR and so on. In the if-then rule, the term 'is' is utilized in two completely distinct ways based on whether it is present in the premise or conclusion component.

## RESULTS AND DISCUSSION

The experiments conducted for 3 different kinds of diseased images for the feature selection methods IG, MRMR and proposed BFO with the 3 classifiers, IBL, C4.5 and fuzzy classifier. [Table-2] shows the results from the experiments for different techniques.

**Table: 2. Experiment Results**

Techniques	IG-IBL	MRMR-IBL	BFO-IBL	IG-C4.5	MRMR-C4.5	BFO-C4.5	IG-Fuzzy classifier	MRMR-Fuzzy classifier	BFO-Fuzzy classifier
Classification accuracy	75.29	77.06	80.59	78.24	79.41	81.18	82.18	82.76	84.71
AD- Sensitivity	0.6909	0.7455	0.8	0.7636	0.8	0.8	0.7797	0.7797	0.8
MCI - Sensitivity	0.8462	0.8462	0.8615	0.8462	0.8462	0.8615	0.8615	0.8615	0.8769
CS-Sensitivity	0.7	0.7	0.74	0.72	0.72	0.76	0.82	0.84	0.86
AD- Specificity	0.8911	0.8911	0.9029	0.901	0.901	0.9038	0.9238	0.9245	0.9346
MCI -Specificity	0.8111	0.8261	0.871	0.8387	0.8602	0.8817	0.8878	0.898	0.9063
CS- Specificity	0.8692	0.8889	0.9009	0.8899	0.8919	0.9009	0.8947	0.8947	0.9099

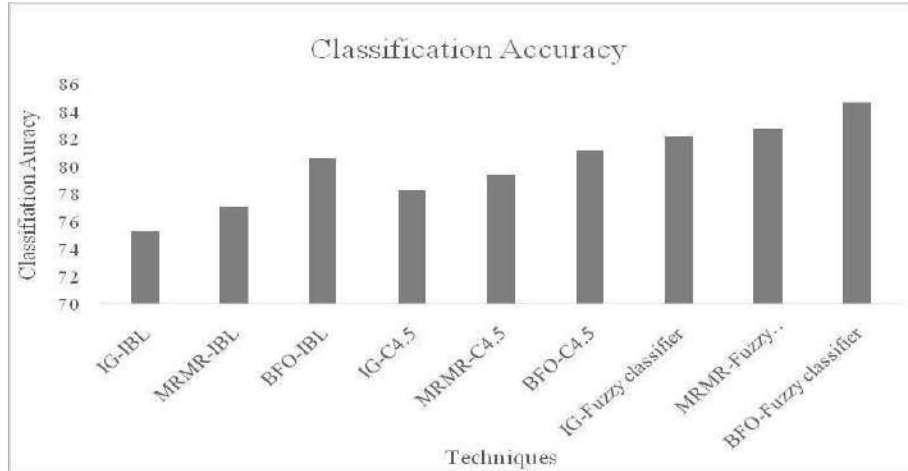


Fig:1. Classification Accuracy

From [Figure -1], it can be observed that the proposed BFO with fuzzy classifier improved accuracy by 4.98% when compared with BFO with IBL method and by 4.26% than BFO with C4.5 classifier.

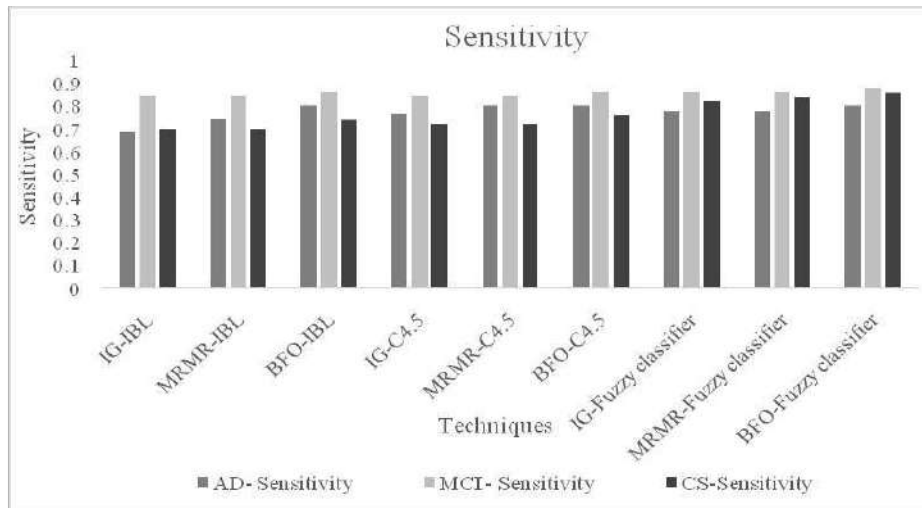
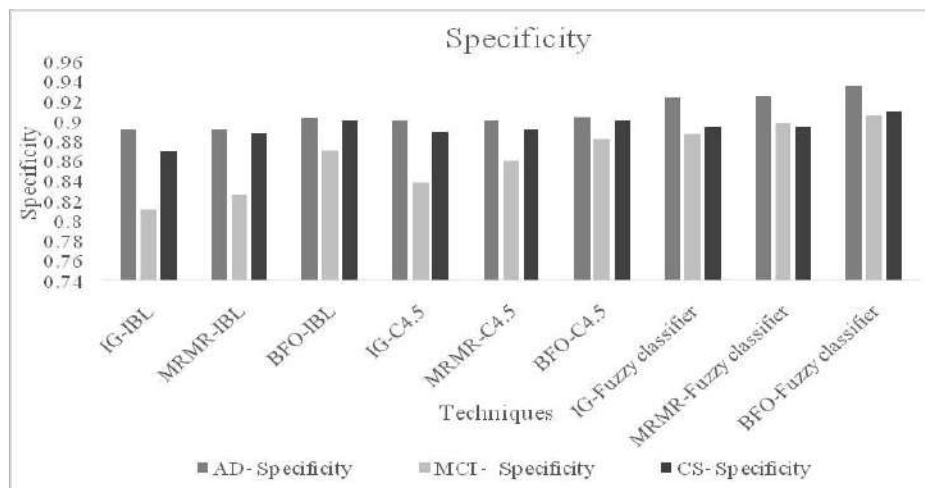


Fig: 2. Sensitivity

It is observed from [Figure -2] that the proposed BFO with fuzzy classifier achieved sensitivity by 15% for CS and by 1.77% for MCI than BFO with IBL method.



**Fig.3. Specificity**

It is observed from [Figure -3] that the proposed BFO with fuzzy classifier achieved specificity by 3.45%, 3.97 and 0.99% for AD, MCI and CS when compared with BFO with IBL method.

## CONCLUSION

MRIs are typically the medical imaging technique used when soft tissue delineation is required. This is particularly true for attempt to sort brain tissue. Current work has proven that classifying human brains in MRIs is possible through supervised methods like ANNs as well as SVMs, and unsupervised classification methods like SOMs as well as Fuzzy C-Means merged with features extraction methods. In this work feature selection is done using bacterial foraging optimization. The experiment is conducted for 3 classifiers with IG, MRMR and the proposed BFO. The results indicated that the proposed feature selection outperformed than other methods.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENES

- [1] Revathy M. [2012] Image classification with application to MRI brain using 2nd order moment based algorithm. *International Journal of Engineering Research and Applications (IJERA)*, 2(3):1821-1824.
- [2] Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Alzheimer's Disease Neuroimaging Initiative. [2015] Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*, 104: 398-412.
- [3] Rohit A, Chitaliya N. [2014] Content Based MRI Brain Image Retrieval-A Retrospective, 2(2):1421-1426.
- [4] Suresh R. [2015] A medical image retrieval system using hybrid FOCS framework, 4(11):256 – 266.
- [5] Castiello C, Castellano G, Caponetti L, Fanelli A M. [2003] Fuzzy classification of image pixels. In *Intelligent Signal Processing, 2003 IEEE International Symposium on IEEE*: 79-82.
- [6] KM Zinzuvadia, BA Tanawala, KN Brahmhatt. [2015] A survey on feature based image retrieval using classification and relevance feedback techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 3(1): 508-513.
- [7] Veeramuthu A, Meenakshi S, Kameshwaran A. [2014] A plug-in feature extraction and feature

subset selection algorithm for classification of medicinal brain image data. In Communications and Signal Processing (ICCSP), 2014 *International Conference on* (pp. 1545-1551). *IEEE*.

- [8] Spedding AL, Di Fatta G, Cannataro M. [2015] A Genetic Algorithm for the selection of structural MRI features for classification of Mild Cognitive Impairment and Alzheimer's Disease. In Bioinformatics and Biomedicine (BIBM), 2015 *IEEE International Conference on* (pp. 1566-1571). *IEEE*.
- [9] Padilla P, López M, Górriz JM, Ramírez J, Salas-Gonzalez D, Álvarez, I. [2012] NMF-SVM based CAD tool applied to functional brain images for the diagnosis of alzheimer's disease. *Medical Imaging, IEEE Transactions on*, 31(2):207-216.
- [10] Gopal NN, Karnan M. [2010] Diagnose brain tumor through MRI using image processing clustering algorithms such as Fuzzy C Means along with intelligent optimization techniques. In Computational Intelligence and Computing Research (ICCIC), 2010 *IEEE International Conference on* (pp. 1-4). *IEEE*.
- [11] Sweety ME, Jiji GW. [2014] Detection of alzheimer disease in brain images using PSO and

decision tree approach. In Advanced Communication Control and Computing Technologies (ICACCCT), 2014 *International Conference on* (pp. 1305-1309). *IEEE*.

- [12] Sasikala M, Kumaravel N. [2005] Comparison of feature selection techniques for detection of malignant tumor in brain images. In *INDICON, 2005 Annual IEEE* (pp. 212-215). *IEEE*.
- [13] Alam SB, Nakano R, Kobashi S, Kamiura N. [2015] Feature selection of manifold learning using principal component analysis in brain MR image. In *Informatics, Electronics & Vision (ICIEV), 2015 International Conference on* (pp. 1-5). *IEEE*.
- [14] Sasi Kumar M., Kumaraswamy YS. [2012] Medical Image Retrieval System Using PSO for Feature Selection, 182-186.
- [15] Tang J, Alelyani S, Liu H. [2014] Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37.
- [16] Patel Brijain, R., &Rana, K. K. (2014, March). A survey on decision tree algorithm for classification. In *International Journal of Engineering Development and Research*,2(1): 1-5 (March 2014)). *IJEDR*.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# FEATURE SELECTION USING IMPROVED SHUFFLED FROG ALGORITHM FOR SENTIMENT ANALYSIS OF BOOK REVIEWS

Madhusudhanan<sup>1</sup> and Srivatsa<sup>2</sup>

<sup>1</sup>Anna University, Chennai, Tamilnadu, INDIA

<sup>2</sup>Department of Computer Science and Engineering, Prathyusha Engineering College, Chennai, TN, INDIA

## ABSTRACT

*Sentiment Analysis refers to a method to identify and mine subjective data from texts, sorted as either positive or negative. Features selection means selecting the best subsets of features for classifications from larger sets which will invariably comprise of unnecessary and repetitive data. Shuffled Frog Leaping Algorithm (SFLA) denotes a metaheuristic optimizing mechanism that imitates the memetic evolutionary activity of frogs searching for the place which possesses most quantity of food. In the current work a hybrid SFLA is utilized for sentiment analysis alongside 2-OPT local search algorithm, for the purpose of reviewing books. Outcomes from experiments reveal the efficacy of the suggested technique.*

Published on: 28<sup>th</sup>– August-2016

### KEY WORDS

*Sentiment analysis, Feature selection, Shuffled Frog Leaping Algorithm (SFLA)*

\*Corresponding author: Email: [ssmadhu80@gmail.com](mailto:ssmadhu80@gmail.com), [profsks@rediffmail.com](mailto:profsks@rediffmail.com)

## INTRODUCTION

Sentiments, opinions and impressions are a major component of most human behavior [1]. They determine the way in which individuals think, take actions and so on. Over the last couple of years, web documents have gained a lot of interest as a resource of opinions and impressions. This has resulted in research into mechanisms for the purpose of automated extraction and analysis of individuals' opinions from online texts like consumer reviews, web blogs, and news sites' comments and so on.

Book reviews may be utilized as a mega citation [2] that comprises two major elements to be appraised. These are (1) credibility, which indicates research quality which the reviewer judges the writer to possess and (2) writing quality, which indicates the writing style quality as assessed by the reviewer. These two features generate a book review credibility-quality scale. The accessibility of several machine-readable documents from the web, has led to an increase in this. Simultaneously, machine learning techniques in Natural Language Processing (NLP) as well as Information Retrieval (IR) were significantly advanced in terms of practical utility, thereby generating easily accessible corpora. Current opinion mining (OM) as well as sentiment analysis denotes a domain of research at the juncture of IR as well as NLP, while sharing certain features with other domains like text mining as well.

Sentiment classification [3] may be developed as a supervised learning method comprising two classes: positive or negative. Reviews are typically utilized for training and evaluating information. Current supervised learning methods may be utilized for classifying sentiments, like K-nearest neighbors (KNN) or fuzzy classifier. A major part in sentiment classification is the choosing of optimal sets of features.

Typically utilized characteristics in sentiment classification include: (1) terms and the frequency that comprise single words and (2) word n-grams and the frequency or presence. These aspects are greatly utilized for classifying sentiments and are proven to be efficient for the job. Part-of-speech data also indicates sentiments in reviews [4]. Opinion/Sentiment words are phrase or words which imply either positive or negative feelings. For example, well, excellent, and great are words with positive connotation while worse, inferior, and awful have negative connotations. Although opinion words are typically adjectives or adverbs, nouns or verbs may be utilized to convey emotions as well. In a similar fashion, negation words are also crucial for evaluating polarities of

sentences as they are capable of transforming sentiments orientations. Syntactic dependency words utilized are dependency based characteristics created from dependency trees or through parsing. The quantity of high-dimensional features is mentioned and to enhance precision of classification, features selection methods are utilized.

Feature Selection (FS) [5] refers to selecting subsets of features from documents. Feature Selection is carried out by retaining words with the greatest score as per predefined metrics of the prominence of the word. High dimensionality of features space is an issue for text classification. Plenty of feature valuation measures that are notable include Information Gain (IG), term frequencies, chi-square, expected cross entropies, odds ratios, weight of evidence of texts, mutual information, Gini Index and so on.

Feature selection is the procedure of choosing the smallest subset of M features from original set of N features, such that features space is best decreased as per specified criteria. When the dimensionality of domains expand, quantity of features N also becomes larger. Discovering the optimal features subset is typically difficult and several issues with regard to features selection are NP-hard [6].

In the previous couple of years, automated text classification has been thoroughly researched and machine learning methods like Bayesian classifiers, decision trees, KNNs, Support Vector Machines (SVM), Neural Networks (NN), Rocchio's have achieved great advancements.

SFLA is a meta-heuristic optimizing approach that is inspired by the activities of frogs during their search for a locale with the most quantity of food [7]. SFLA, initially proposed by Eusuf and Lansey, may be utilized to resolve several complicated optimization issues that are non-linear, not differentiable as well as multimodal [8].

An enhanced SFLA (ISFLA) is utilized to analyse reviews of books. Section 2 deals with relevant literature; Section 3 discusses methodology utilized; Section 4 reveals outcomes of experiments and Section 5 concludes the work.

## RELATED WORKS

Basari et al [9] studied binary classifications having two classes: positive, that shows good opinions or negative, that shows bad opinions. Justifications on the basis of precision degree of SVMs with the evaluation procedure utilized 10-fold cross evaluation as well as confusion matrices. Hybrid Particle Swarm Optimization (PSO) enhanced the choosing of most optimal variable to resolve dual optimizing issue. Outcomes proved enhancement in precision levels from 71.87% to 77%.

Sharma &Dey [10] looked into the implementation possibilities of five typically utilized features selection approaches in data mining as well as seven classification approaches with a basis in machine learning to analyze opinions in a dataset filled with online movie reviews. The study revealed that features selection enhances sentiment based classification's performance, with a dependency on the approach utilized as well as quantity of features chosen. Outcomes from experiments reveal that GR provides most optimal performance in sentiment features selection while SVM outperformed the rest in sentiment based classification.

Dey [11] identified an implementation of features selection approached in analyzing sentiments and studies their execution with regard to recall and precision. Features selection approaches as well as common sentiment features lexicons like HM, GI as well as Opinion Lexicon were utilized to analyze a movie reviews data set comprising around 2000 reviews. Outcomes from experiments reveal that IG provided excellent results consistently while GR provided best results in terms of sentiment features selection. It was also revealed that the a classifier's performance relies on the quantity of representative features chosen from the texts.

Baek et al [12] compiled 75,226 customer reviews on Amazon.com through the usage of a web data crawler. Further data was also acquired through sentiment analysis. Outcomes revealed that peripheral cues like review rating as well as reviewer credibility, as well as central cues like review content determine the usefulness of a review. On the basis of dual process theories, it is revealed that customers observed various data resources of reviews, according to their needs be it merely searching for information or for valuating substitute products.

Xiong et al [13] observed the polarity of Chinese statements utilizing appraisers, degree adverbs, negations and so on, and presented a novel rule-based approach. The model merges three kinds of words in the predetermined rules; it uses the word distances of the rules as restrictions; it uses the strengths of appraisers as well as degree adverbs as items of the rules. It then uses PSO to get optimal variables of the rules like thresholds of restrictions as well as adjustment of strengths. Moreover, it makes use of the Chinese lexicon 'HowNet' as a resource for Chinese sentiments. Results prove that the approach outperforms with regard to better accuracy, recall as well as F1.



## METHODOLOGY

### DATASET

A fresh data set for sentiment domain adaptations by choosing Amazon reviews for books. All reviews comprise a rating (0-5 stars), the name of the reviewer, their location, the product's name, title for the review, date as well as the content of the review itself. A review with a rating over three stars was labelled positive; a review with a rating less than three stars was labelled negative while the remaining were removed due to ambivalent polarity. Once this sorting was completed, there were 800 samples of both positive and negative reviews so that the data set has a balanced composition. [14].

Tokenization refers to the procedure of splitting a set of text into meaningful words (stems), phrases or symbols. The tokens can be used further for parsing (syntactic analysis) or text mining. Tokenization is generally considered easy relative to other tasks in text mining and also one of the uninteresting phases. However, errors made in this phase will propagate into later phases and cause problems.

Stop words are function words like prepositions, articles, conjunctions and pronouns, providing language structure instead of content. These terms do not affect category discriminations. Additionally, common words like 'a' and 'of', may be removed as they recur frequently so that it is not discriminating for a specific class. Generic terms are detected by a threshold on the quantity of documents the term appears in, for instance, if it is present in more than 50% of the texts, or through the provision of a stop word list. Stop words are language as well as field-specific. On the basis of the classification tasks, removal of terms that are crucial predictors may be risked, for instance, the term 'can' discriminates between aluminium as well as glass recycling.

Word stemming is a rough pseudo-linguistic procedure which discards suffixes for reducing words to their stem. For instance, the words searching, searched, searches may be conflated to one stem – search. The common practice of stemming or lemmatizing, combining several word forms like plural or verb conjugation into one singular word decreases features number to be regarded.

Within Parts of Speech (PoS), the total contents of the text are denoted by unigram as well as N-gram, and are split into two categories: the first comprising single terms known as unigram, the second comprising multi-words known as N-gram. Those features with greatest relevance are regarded for sentiment classification.

### FEATURE EXTRACTION

In sentiment analysis, extracting features is the most difficult task as it needs the usage of NLP methods for automated identification of features in the opinions being analysed. The task is more challenging when considering opinions regarding stores as the features in the particular field are not pre-specified. Apart from the most apparent feature (cost of the available product) other significant features are site usability, delivery costs as well as time, dependability of store, customer care as well as packaging of products. Weighted frequency statistic is Term Frequency Inverse Document Frequency (TFIDF) statistic that computes a weight for every term reflecting its importance. The term's relevance to a specific document depends on how many words it has. This is why TFIDF denominator is adjusted for words number in a document. It is also adjusted for number of records (or documents) having the word (terms appearing on many records are down-weighted).

### TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY (TF-IDF)

Every term in the document is designated a weight based on the number of times it recurs in the text. It is known as term frequency [15].  $tf(t,d)$  is a normalized frequency so as to avoid bias toward larger documents. Term frequency possesses a major issue every term is regarded with equal importance during the assessment of relevance in a request. In the real world however, particular terms have almost no power in determining the character of the document.

$$idf(t, d) = \log \frac{N}{1 + |\{d \in D : t \in d\}|}$$

Where,  $|D|$ : Total quantity of documents in the archive,  $|d \in D : t \in d|$ : quantity of documents wherein the term is present.  $idf(t,d) \neq 0$ .

If a term is not available in the archive the formula changes to  $1 + |d \in D : t \in d|$ .

Then  $tf-idf$  is computed as:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, d)$$

Otherwise put,  $tfidf(t, d, D)$  designates to term  $t$ , a weight in document  $d$ .

### FEATURE SELECTION

Features selection techniques minimize the original set through the removal of non-relevant features for sentiment classification in order to enhance classification precision and minimize the runtime of learning models.

## INFORMATION GAIN (IG)

Entropy is a typically utilized information theory metric that qualifies the purity of a random set of samples. It is the basis of IG feature ranking techniques [16]. The entropy metric is regarded as an assessment of a system's unpredictability and if entropy is assumed as  $H(Y)$ , then IG is given by:

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y)$$

IG is a symmetrical metric. The information obtained regarding Y after observation of X is the same as the information obtained regarding X after observation of Y. A shortcoming of the IG criterion is that it is biased in favour of features with more values even when they are not more informative.

## SHUFFLED FROG LEAPING ALGORITHM (SFLA)

SFLA [17] is a population based arbitrary search algorithm that owes its inspiration to natural memetics. Here, a population of probable solutions defined by a set of frogs is split into various communities known as memeplexes. Every frog in the memeplexes performs a local search. In all memeplexes, a single frog's activity may be shaped through the activity of other frogs and it evolves in a memetic evolutionary process. Once a particular set of memetic evolution stages are done, memeplexes are pressured to merge and fresh memeplexes are generated through shuffling. Local searches and the shufflings prevail till a convergence criterion is fulfilled.

## PROPOSED IMPROVED 2-OPT -SFLA FOR FEATURE SELECTION

An enhancement algorithm that is essentially a basic local search heuristic for resolving TSP, 2-OPT is suggested. Though, failings of 2-OPT include the fact that its performance is greatly reliant on the initial solution given and that it does not possess a global search technique to avoid local minima through uphill moves. Hence, hybrid algorithms are required to merge a global search heuristic with 2-OPT local search model. In the current work, a hybrid 2-OPT-ISFLA, is suggested that merges the global search ISFLA with the local search 2-OPT. [Figure -1] below, illustrates the suggested model

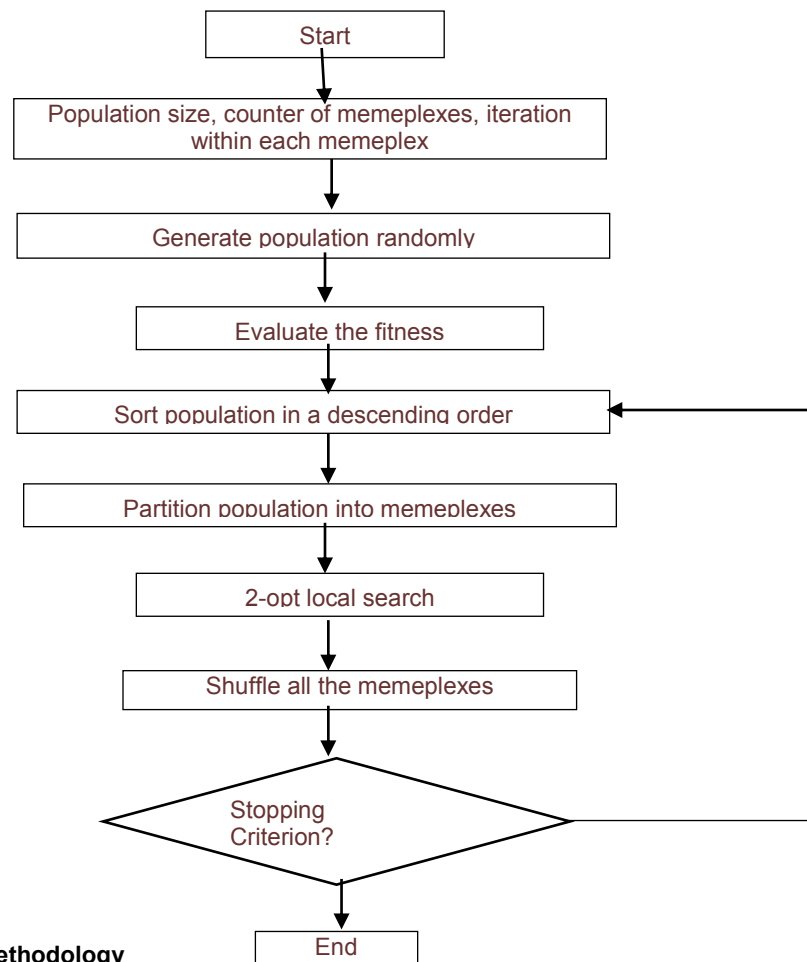


Fig: 1. Flow chart for proposed methodology

ISFLA is suggested alongside 2-OPT local search algorithm and amongst the outcomes, the most optimal solutions are extracted through features selection methods. If the outcome got is 0, then it means that it is not selected and if it is 1, then it means that it is selected.

## CLASSIFIER

### FUZZY CLASSIFIER

Classification [18] is a supervised learning issue that accepts labelled data examples and creates a classifier which sorts the data into several predetermined classes. The classification issue may be resolved through fuzzy logic. Fuzzy logic utilizes fuzzy set theories, wherein a parameter belongs to one or more sets, with a certain level of membership. Fuzzy logic when employed in computers, permits them to mimic human reasoning procedures, quantify non-precise data, take decisions on the basis of ambiguous and partial information, and yet, through the application of a 'defuzzification' procedure, reach definitive conclusions. Fuzzy classifiers comprise interpretable if-then rules denoting features as well as output class of the format:

$$R_j : \text{if } x_{p1} \text{ is } A_{j1} \text{ and } x_{pn} \text{ is } A_{jn} \text{ then class } C_j$$

Wherein  $A_{j1}, \dots, A_{jn}$  refer to antecedent fuzzy sets of input parameter  $x_{p1}, \dots, x_{pn}$  and  $C_j$  is an output class label.

Opinions words are by nature, fuzzy. For instance, the words (as well as the boundaries between them) bad, poor, awful are not apparent. Therefore, fuzzy logic is capable of representing these kinds of subjective terms and assigning them to classes with certain degrees of membership. This implies that these terms are in fuzzification stage. Definition of fuzzy sets for the words requires basis in expert opinions. Because all opinions are fuzzy, the meaning behind opinion terms may be understood in various ways. Fuzzy logic, therefore is an effective method for consideration for the proper extraction, analysis, categorization as well as summarization of opinions. This is because [19]:

- Fuzzy logic is malleable, in a conceptual aspect, simple to comprehend and is built for handling inaccurate information such as opinion terms.
- Fuzzy logic has its basis in natural languages and therefore is adequate for resolving fuzziness in human conveyed opinions.
- Sentiment classification in several recent studies employ supervised machine learning methods such as Support Vector Machine or Naïve Bayes methods.
- Fuzzy logic is an intelligent control method that depends on human-like expertise through usage of if-then rules.
- Already present OM methods as well as mechanisms are capable of classifying opinions into positive, negative or neutral classes only.
- Fuzzy logic permits improved classification of sentiments with appropriate strength designated to every opinion level. This assists in the improvement of classification accuracy.
- Subjective terms are fuzzy by nature and more so when it is with regard to opinion mining. As opinions are typically conveyed in a fuzzy nature, for instance, awesome shoes, pretty dress, cheap bike etc. and in several situations, it is hard to comprehend the level of fuzziness and if it is too awesome, too pretty or too cheap. The idea of fuzzy logic is something that could easily be dismissed by the ignorant or poorly-informed as a trivial issue or even an insignificant one. It does not refer to fuzziness of logic but rather logic of fuzziness and particularly the logic of fuzzy sets.

### K NEAREST NEIGHBOR (KNN)

KNN[20] classifier is the most simple sample based learning model. On the basis of distance functions, KNN designates class of an unknown object to the class of a known training object. Training samples with class labels are needed. The distances between unknown objects and the training samples are calculated. It then designates the class label of the training sample closest to that of the unknown object.

KNN [21] comprises two stages: Training and Classification. In the first, the training samples are vectors with class labels in multi-dimensional feature spaces. Here, feature vectors as well as class labels are stashed. In the next stage, K refers to a user-defined constant, a request or test point which is an unlabeled vector is sorted by designating a label that is the most frequent of the K training sample closest to the request point. Otherwise put, KNN contrasts the inputted feature vector against a library of reference vectors and the request point is labelled with the closest class of the feature vector in the library. This method of sorting request points on the basis of their distance to points in a training dataset is simple and efficient at the same time.

## RESULTS AND DISCUSSION

[Table -1] gives the summary of outcomes. [Figure -2] to 6 depict the results for classification accuracy, sensitivity for positive, sensitivity for negative, Positive predictive value for positive and Positive predictive value for negative respectively.

Table: 1. Summary of results

	IG-KNN	IG-Fuzzy	SFLA-KNN	SFLA-Fuzzy	ISFLA-KNN	ISFLA-Fuzzy
Classification Accuracy	85.25	87.25	89.06	90.94	91.25	92.13
Sensitivity for positive	0.835	0.86	0.8925	0.905	0.9063	0.92
Sensitivity for negative	0.87	0.885	0.8888	0.9138	0.9188	0.9225
Positive predictive value for positive	0.8653	0.8821	0.8892	0.913	0.9177	0.9223
Positive predictive value for negative	0.8406	0.8634	0.8921	0.9058	0.9074	0.9202

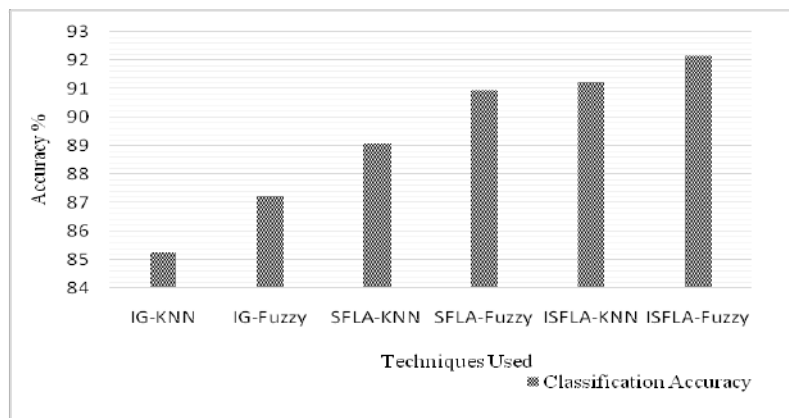


Fig:2. Classification Accuracy

From [Table -1] as well as [Figure -2] it is observed that classification accuracy of fuzzy classifier outperforms KNN classifiers. ISFLA outperforms IG and SFLA. Results show that accuracy of ISFLA-Fuzzy performs better by 5.44% than IG-Fuzzy, by 1.3% than SFLA-Fuzzy.

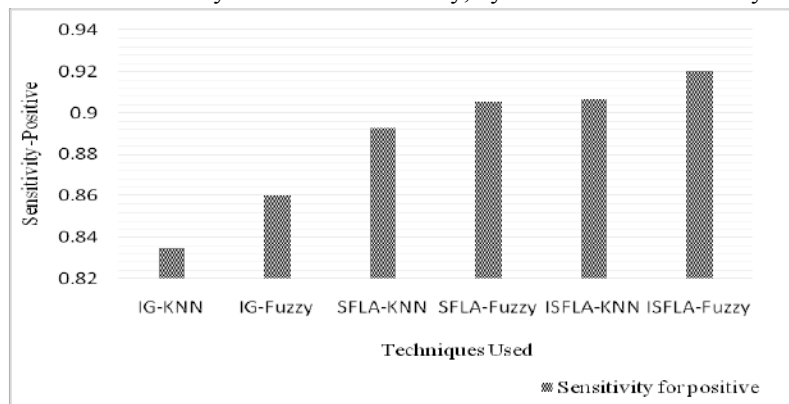


Fig: 3. Sensitivity for Positive

From [Table -1] as well as [Figure -3] it is observed that the sensitivity for positive of fuzzy classifier outperforms KNN classifiers. ISFLA outperforms IG and SFLA. Results show that sensitivity for positive of ISFLA-Fuzzy performs better by 6.74% than IG-Fuzzy, by 1.64% than SFLA-Fuzzy.

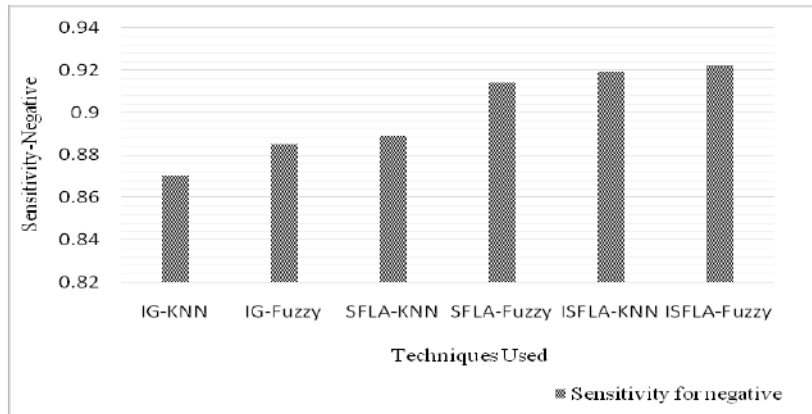


Fig: 4. Sensitivity for Negative

From [Table -1] as well as [Figure -4] it is observed that the sensitivity for negative of fuzzy classifier outperforms KNN classifiers. ISFLA outperforms IG and SFLA. Results show that sensitivity for negative of ISFLA-Fuzzy performs better by 4.15% than IG-Fuzzy, by 0.95% than SFLA-Fuzzy.

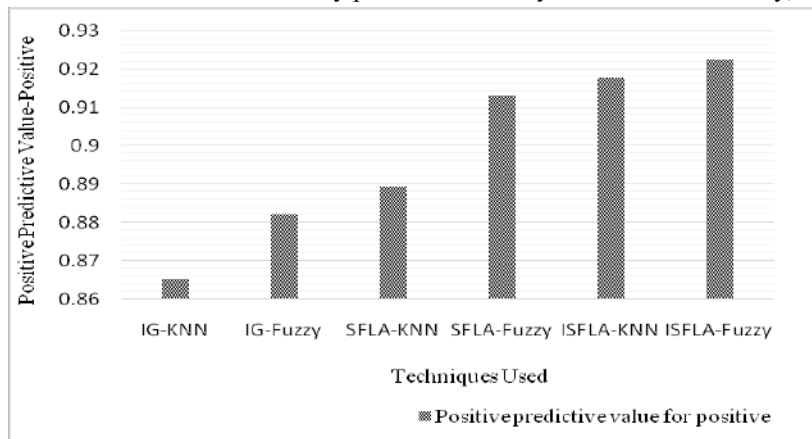
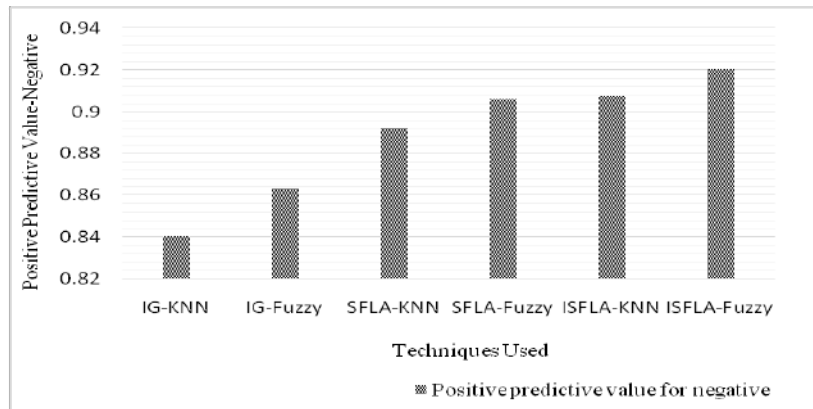


Fig: 5. Positive predictive value for positive

From [Table -1] as well as [Figure -5] it is observed that positive predictive value for positive of fuzzy classifier outperforms KNN classifiers. ISFLA outperforms IG and SFLA. Results show that positive predictive value for positive of ISFLA-Fuzzy performs better by 4.46% than IG-Fuzzy, by 1.01% than SFLA-Fuzzy.



**Fig. 6. Positive predictive value for negative**

From [Table -1] as well as [Figure -6] it is observed that the positive predictive value for negative of fuzzy classifier outperforms KNN classifiers. ISFLA outperforms IG and SFLA. Results show that positive predictive value for negative of ISFLA-Fuzzy performs better by 6.37% than IG-Fuzzy, by 1.58% than SFLA-Fuzzy.

## CONCLUSION

Shuffled Frog Leaping Algorithm (SFLA) is used with local search protocol for book reviews in sentiment analysis. Experimental outcomes reveal that the classification accuracy of fuzzy classifier outperforms KNN classifiers. Improved Shuffled Frog Leaping Algorithm (ISFLA) outperforms IG and SFLA. Outcomes reveal that accuracy of ISFLA-Fuzzy is better by 5.44% than IG-Fuzzy, by 1.3% than SFLA-Fuzzy. In a similar manner, suggested approach performs in better for other metrics as well.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1]. Sadegh M, Ibrahim R, Othman Z.A. [2012]Opinion mining and sentiment analysis: A survey. *International Journal of Computers & Technology*, 2(3): 171-178.
- [2]. Van Bellen M, Intelligentie, BOK., van Someren MW. [2010] Sentiment Analysis on historical book reviews with a Bayesian Classifie.
- [3]. Kumar A, Sebastian TM. [2012]. Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, 9(3):372-378.
- [4]. Prabowo R, Thelwall M. [2009] Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2): 143-157.
- [5]. Korde V, Mahender, CN. [2012] Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications (IJAI)*, 3(2): 85-99.
- [6]. Naseriparsa M, Bidgoli AM, Varace T. [2014] A Hybrid Feature Selection method to improve performance of a group of classification algorithms. *arXiv preprint arXiv:1403.2372*.
- [7]. Venkatesan T, Sanavullah MY. [2013] SFLA approach to solve PBUC problem with emission limitation. *International Journal of Electrical Power & Energy Systems*, 46: 1-9.
- [8]. Ren W, Zhao C. [2013] A localization algorithm based On SFLA and PSO for wireless sensor network. *Information Technology Journal*, 12(3):502-505.
- [9]. Basari ASH, Hussin B, Ananta IGP, Zeniarja J. [2013] Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53:453-462.
- [10]. Sharma A, Dey S. [2012] A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium* (pp. 1-7). ACM.
- [11]. Dey S. [2013] Performance Investigation of Feature Selection Methods. *arXiv preprint arXiv:1309.3949*.
- [12]. Baek H, Ahn J, Choi Y. [2012] Helpfulness of online consumer reviews: readers' objectives and review

cues. *International Journal of Electronic Commerce*, 17(2):99-126.

- [13]. Xiong W, Jin Y, Liu Z. [2014] Chinese sentiment analysis using appraiser-degree-negation combinations and PSO. *Journal of Computers*, 9(6), 1410-1417.
- [14]. Blitzer J, Dredze, M, Pereira F. [2007] Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *In ACL*. 7: 440-447
- [15]. Ratna A, Sawhney A. [2015] Focused Crawler based on Efficient Page Rank Algorithm. *International Journal of Computer Applications*, 116(7): 37-40.
- [16]. Novakovic, J. [2009] Using information gain attribute evaluation to classify sonar targets. In *17th Telecommunications forum TELFOR* (pp. 24-26).
- [17]. Samuel GG, Rajan, CCA. [2015] Hybrid: Particle Swarm Optimization–Genetic Algorithm and Particle Swarm Optimization–Shuffled Frog Leaping Algorithm for long-term generator maintenance scheduling. *International Journal of Electrical Power & Energy Systems*, 65:432-442.
- [18]. Devaraj D, Ganesh Kumar P. [2010] Mixed genetic algorithm approach for fuzzy classifier design. *International Journal of Computational Intelligence and Applications*, 9(01): 49-67.
- [19]. AL-MAIMANI MAQBOOL, SALIM N, AL-NAAMANY AM. [2014] Semantic and fuzzy aspects of opinion mining.

*Journal of Theoretical and Applied Information Technology*, 63(2).

- [20]. Maheswari SU, Ramakrishnan R. [2015] Sports Video Classification using Multi Scale Framework and Nearest Neighbor Classifier. *Indian Journal of Science and Technology*, 8(6): 529-535.
- [21]. Saini I, Singh D, Khosla A. [2013] QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of advanced research*, 4(4):331-344.

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

## AUTHORS BIOGRAPHY



**Mr. S. Madhusudhanan** is working as Scientist 'C' in National Institute of Electronics and Information Technology (NIELIT), Guwahati. He did M.Tech (CSE) in SASTRA University. He is doing research in Opinion Mining at Anna University, Chennai. He is having 8.6 years of engineering experience in teaching. He has published more than 6 papers in National & International conferences. He is interested in the areas of Sentiment Analysis, Data mining, Genetic Algorithm and Neural Networks & Guided number of UG & PG Projects.



**Dr. SK Srivatsa** is retired Senior Professor in Anna University who currently working as Senior Professor (CSE) in Prathyusha Engineering College, Chennai. He received his Bachelor of Electronics and Communication Engineering Degree (Honors) from Javadpur University (Securing First Rank and Two Medals), Master Degree in Electrical Communication Engineering (With Distinction) from Indian Institute of Science and Ph.D also from Indian Institute of Science, Bangalore. He produced 25 PhDs. He is author of 475 publications in reputed journal/conference proceedings. He is a life member/fellow in twenty-four registered professional societies.

# END TO END DELAY IMPROVEMENT IN HETEROGENEOUS MULTICAST NETWORK USING FUZZY GENETIC APPROACH

Chandrasekar V.<sup>1\*</sup>, M. Mahalakshmi<sup>2</sup>

<sup>1</sup>Department of Computer Science, Vimal Jyothi Engineering College, Kerala, INDIA

<sup>2</sup>Department of Computer Science, University of Gondar, Gondar, ETHIOPIA

## ABSTRACT

Routing plays an important part in delivering data from senders to receivers efficiently, in multicast communication. In QoS multicast, receivers receive data within specified QoS constraints. This is a challenge if the heterogeneous network consists of wired and wireless devices. This paper investigates the performance of Protocol Independent Multicast – Sparse Mode (PIM-SM) protocol performance within a heterogeneous network operating a video conferencing application and suggests an improved routing protocol with Fuzzy based Genetic Optimizing techniques to enhance QoS parameters such as delay, speed and packet dropped in bits/sec in the multicasting. The video conference transmission speed in bits/sec was increased in Fuzzy logic genetic approach and it was much better when compared to PIM-SM and Genetic algorithm based approach in a Heterogeneous multicast network simulation. Extensive simulations using the proposed technique and existing PIM-SM were tried out and it was revealed that the proposed Fuzzy based optimization technique both improved network throughput and lowered end to end delay.

Published on: 28<sup>th</sup>– August-2016

### KEY WORDS

Protocol Independent Multicast – Sparse Mode (PIM-SM), Wireless Network, Heterogeneous Network, Genetic Algorithm, Fuzzy Logic

\*Corresponding author: Email: [vchandruskce@gmail.com](mailto:vchandruskce@gmail.com)

## INTRODUCTION

Multimedia communication applications need a source to forward information to many destinations through communication networks. Application examples include:

- In multiparty multimedia teleconference, video and voice from every conference site are sent to other conference sites to ensure that conferees see and communicate in real time [1].
- In remote video lectures for distance education, the instructor's video and voice are sent to students via a communication network.
- In video-on-demand systems, a problem is video retrieval from disks [2]. To overcome this, many customer requests for the same video object are batched and then one I/O stream serves multiple customers [2]. Multimedia information is forwarded from the video server to many customers through this method.

A multicast session is created with a multicast tree through which data is shifted to a multicast group. Multicast routing algorithms construct multicast trees. Quality of service (QoS) requirements like end-to-end delay, delay variation, loss, cost, throughput should be met in group applications for effective network functioning. When a multicast tree goes through wired and wireless devices in heterogeneous networks, resources alongside the path could fail to provide required QoS leading to multi cast tree failure. For efficient multicast communication, the tree should meet all resource requirements. The aim of QoS is provision of specific level of predictability and service control. Delay, Jitter, Bandwidth and Reliability are parameters which measure QoS.

It is important to determine a multicast tree of minimal cost [3], [4] to support such applications for every communication session. The source node sends multimedia information to destination nodes through this tree. As multimedia information should not be delayed too much [5], total delay from a source node to a destination node must be lower than a specific requirement. The issue of selecting multicast trees is called multimedia multicast routing and is NP-complete [3]. Hence heuristic algorithms are of great practical interest here.



Protocol Independent Multicast (PIM) is generally used in multicasting [6] which is a multicast routing architecture to create trees in various sparsely represented groups. PIM's robustness, flexibility, and scaling properties ensure that it is suited to large heterogeneous networks [7]. PIM is a protocol collection optimized for various scenarios. PIM Sparse mode (PIM-SM) and PIM Dense mode (PIM-DM) are common multicast protocols. PIM-SM uses source-based and core-based trees while PIM-DM uses only source-based trees. PIM-SM is mainly used in all networks and PIM-DM is mainly in small domains. In PIM routing protocols, join and prune messages join and leave a multicast distribution tree.

The Protocol Independent Multicast - Sparse Mode (PIM-SM) architecture:

- Maintains conventional IP multicast service model of receiver-generated membership;
- Uses specific joins which propagate hop-by-hop from members' linked directly to routers toward distribution trees.
- Constructs a shared multicast distribution tree in the middle of centered at a Rendezvous Point, building source-specific trees for sources whose data traffic requires it.
- Independent of specific unicast routing protocol; and
- Uses soft-state mechanisms to adapt to network conditions/group dynamics.

Evolutionary algorithms are good to solve optimization and search problems [8, 9]. Two genetic algorithms were suggested for solving multicast routing problem without delay [10]. A step is computing shortest paths between all network node pairs. When algorithm include delay constraints, the resulting algorithms solve delay-constrained shortest path problem (i.e., solve an NP-complete problem). So a genetic algorithm is not required to solve any hard sub-problem should be so designed which can also give nearly optimal solutions for multimedia multicast routing.

This work proposed implementation of Fuzzy Logic in Genetic Optimization techniques to improve QoS for heterogeneous environments including wired/wireless nodes. The rest of the paper is organized as follows: Section 2 reviews related works, Section 3 details the methodology, Section 4 gives the results and Section 5 concludes the paper.

## RELATED WORKS

Biswas and Izmailov[11] suggested a PIM-SM based IP-multicast routing framework to deliver heterogeneous QoS. Two tree construction algorithms: TIQM and NUQM were suggested. TIQM depended on total availability of tree-specific information on a multicast group while NUQM did not need tree specific information. Pseudo-optimal QoS-constrained trees are computed using TIQM but face control-scalability problems. NUQM overcomes this issue by restricting information used in tree computation. A QoS-extended intra-domain PIM-SM framework was proposed.

For multimedia multicast routing, Qingfu Zhang et al., [12] proposed an orthogonal genetic algorithm. The incorporation of an experimental design approach known as orthogonal design into the crossover operation is its salient feature. The solution space can be searched in a statistically sound manner and it is the suitable one for parallel implementation and execution. To solve two sets of benchmark test problems, the orthogonal genetic algorithm is implemented. For practical problem sizes, the results reveal that near-optimal solutions within moderate numbers of generations are capably found by the orthogonal genetic algorithm.

Ren-Hung Hwang et al., [13] proposed a novel multicast routing algorithm based on Genetic Algorithms (GA). To evaluate the performance of the proposed algorithm, computer simulations were performed on a random graph. The numerical results that are demonstrated by the proposed algorithm provide a better solution to the Steiner tree problem. As the multiple QoS constraints, like delay, jitter, and loss probability are required by many multimedia applications, the proposal extends to address the application of the proposed algorithm to get multiple constrained multicast trees. Along with the increase in the number of nodes of the network or the number of destinations of a multicast request, the time complexity of the proposed algorithm also increases linearly. Hence the novel multicast routing algorithm proposed emphasis to reduce multicast cost while maintaining a reasonable path delay and provides an effective solution to the issue of multicast routing.

Shangchao Pi, et al., [14] proposed a fuzzy controller based multipath routing algorithm in MANET (FMRM). The aim of developing the FMRM algorithm is to construct the fuzzy controllers by assisting to decrease reconstructions in the ad hoc network. The results obtained by simulation of the proposed algorithm shows that it effectively outperforms in applications to the MANETs. For multipath routing decision, the FMRM algorithm is an efficient routing protocol. Sukhvinder singh [18] proposed the genetic algorithm may be employed for heuristically approximating an optimal solution to a problem, in case of quality of service routing finding the optimal route based on QoS constraints.

To find routes which gratify the multiple independent QoS constraints simultaneously are addressed with multi-constrained QoS routing. Santhi et al., [15] proposed a Fuzzy cost based Multi constrained Quality of service Routing (FCMQR) protocol in order to choose an optimal path with regarding to multiple independent QoS metrics like number of intermediate hops, bandwidth and end-to-end delay. This is on the basis of multi criterion objective fuzzy measure. Every available resources of the path are changed into a single metric fuzzy cost in this proposed technique. To find the lifetime of the path, mobility prediction is performed. The optimal one concerned is the path with the maximum lifetime and minimum fuzzy cost which is implemented in transmission. The performance of the proposed FCMQR is obtained by simulation. The results illustrated improved packet delivery ratio, increased path success ratio and incurred less end-to-end delay. Hence an exact and effective technique to evaluate and estimate the QoS routing stability and cost in dynamic mobile networks is provided. And also, the proposed method outperforms than the existing FLWMR and FLWLAMR protocol. Sara Aliabadi et al., [19] proposed the Quality of Service in routing needs to ensure the chosen path has less traffic, less packet loss, optimum length and the most possible bandwidth together.

## METHOD

### GENETIC ALGORITHM

Gradient search techniques are meant for local search, getting solutions around its starting point [16]. Global search techniques get more optimal solutions, though dependent on ideal setting of starting values. The Genetic Algorithm (GA) is a population-based optimization algorithm, based on biological evolution. Gene sets are replicated, varied and mutated in natural evolution. So, mechanisms like selection, reproduction, mutation are used in GA to get better solutions. Solutions to optimization problem as seen in bit-strings is the population used in GA. Fitness functions evaluate every solution. An initial population is evolved to the next generation on application of operators such as selection, reproduction, crossover and mutation. Better solutions are produced in every generation. The evolution process continues forming new generations until a correct solution is reached or specific number of generations obtained. The pseudo-algorithm of GA is shown in [Figure -1].

For each  $n_i \in N_r$  requires a new routing table. Consisting of the R shortest, R cheapest and R least used paths. R is an algorithm parameter. A chromosome is represented by a lengthy string.  $|Nr|$  where each element (gene)  $g_i$  represents a path between source s and destination  $n_i$ . Route selection s dependent on the following parameters.  
 Discard individuals: Set P might have duplicate chromosomes. Hence, randomly generated individuals replace duplicated chromosomes.  
 Evaluate individuals: The P individuals are evaluated using objective functions. Then, non-dominated P individuals are compared with individuals in  $P_{nd}$  to update the non-dominated set, removing  $P_{nd}$  dominated individuals.  
 Compute fitness: Fitness is computed for every individual, through use of a SPEA procedure.

Selection: A roulette selection operator is applied over the set  $P_{nd} \cup P$  to generate a subsequent evolutionary population P.

Crossover and Mutation: This work includes a two-point crossover operator over selected pair of individuals with some genes in each chromosome of the new population being randomly changed (mutated), to get a new solution and this process continues until a stop criterion/ specific maximum number of generations, is satisfied.

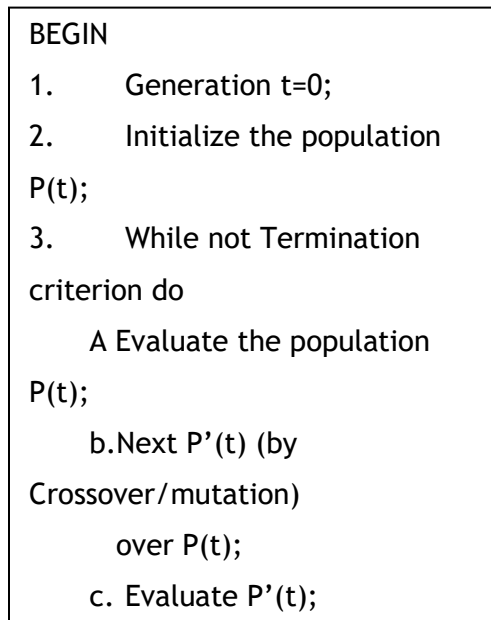


Fig: 1. The Genetic Algorithm process

## FUZZY LOGIC

Fuzzy Logic (FL) is a multi valued logic which handles imprecision and approximate reasoning with precise logic [17]. FL allows defining of intermediate values between conventional evaluations like true/false, yes/no, etc. Concepts/ideas like very slow or rather fat can be formulated mathematically. FL helps formulating problems in order to apply a more human-like way of thinking in the programming of computers. Fuzzy Logic methodology is a problem-solving control system. It can be implemented into small, simple systems to large networked systems. It can also be incorporated into hardware or software or in both. FL computes definite conclusion when presented with ambiguous, imprecise, noisy input information. FL forms simple rules based on "IF X AND Y THEN Z" to solve control problem rather than modeling a system mathematically. The FL model is empirically-based, formulated more on operator's experience rather than its technical understanding of the system.

The simplest fuzzy model consists of a set of rules with an "if – then" structure:

*If < condition 1 > and ... and < condition n > then < conclusion >*

Nodes in an actual dynamic network find it hard to maintain globally accurate network state information. Hence it is unreasonable/inefficient to express QoS constraints with deterministic values. This paper proposes a QoS multicast routing problem through use of generalized fuzzy-constrained fuzzy-optimization model. Network state information imprecision and QoS fuzziness constraints account for the need to invoke fuzzy set theory. A new fuzzy genetic algorithm for QoS multicast routing is presented and simulation experiments prove its efficiency.

## RESULTS

[Figure -2] shows the experimental test bed used for simulation of the proposed system, which is implemented as a layer over PIM-SM. The network is a heterogeneous environment. The throughput for video traffic using the proposed system and compared with PIM-SM is shown in [Figure -3]. Figure 4 and 5 show the end to end delay and the overall data dropped in the network respectively.

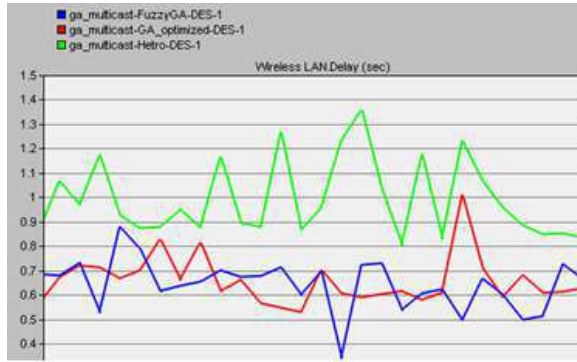


Fig: 2. The experimental test bed used in this work

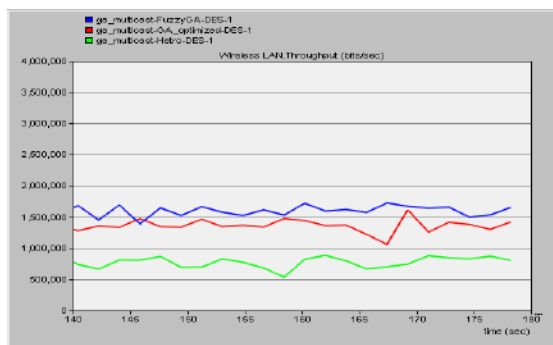


Fig: 3. shows the throughput in WLAN.

Figure- 3. The throughput obtained in bits/sec in the WLAN section. (\_\_\_ the proposed fuzzy based optimization multicast, \_\_\_ genetic algorithm optimized multicast and \_\_\_ PIM-SIM)

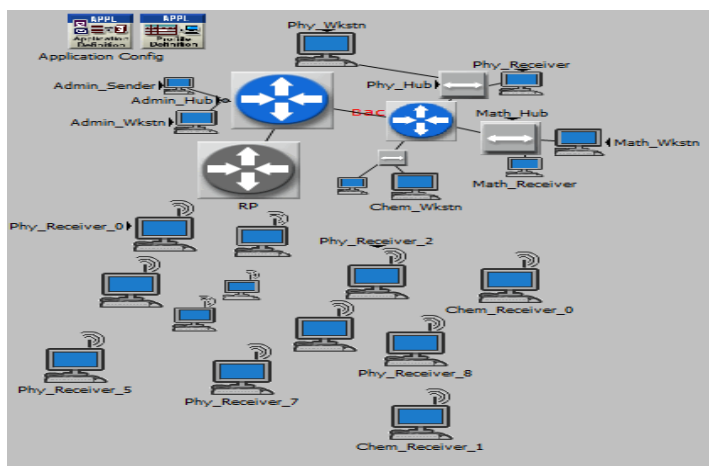
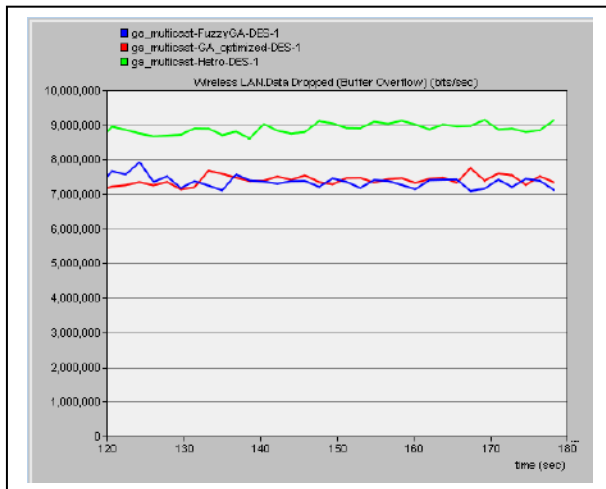


Fig:4. shows the overall end to end delay in the network.

Figure- 4: The overall end to end delay in the heterogeneous network (\_\_\_ the proposed fuzzy based optimization multicast, \_\_\_ genetic algorithm optimized multicast and \_\_\_ PIM-SIM)

Figure 3 shows the throughput in the wireless section of the proposed optimization technique. It can be seen the proposed fuzzy based optimization technique (blue graph) increases the throughput for video transmission.

It can be seen that the end to end delay for the video conferencing decreases. The packet dropped in the wireless LAN section drops considerably due to the fuzzy based genetic optimization as shown in [Figure-5](#).



**Fig: 5. The data dropped in bits/second**

**Figure- 5:** The data dropped in bits/sec in the heterogeneous network (\_\_\_ the proposed fuzzy based optimization multicast, \_\_\_ genetic algorithm optimized multicast and \_\_\_ PIM-SIM)

## DISCUSSION

This paper proposes a QoS multicast routing problem through use of generalized fuzzy-constrained fuzzy-optimization model. Network state information imprecision and QoS fuzziness constraints account for the need to invoke fuzzy set theory. A new fuzzy genetic algorithm for QoS multicast routing is presented and simulation experiments prove its efficiency. In this work, the effectiveness of PIM-SM in heterogeneous environment using a Fuzzy based Genetic optimization is studied. Video conferencing traffic was used in this study. Critical QoS parameters for PIM-SM and the proposed routing is investigated. A novel fuzzy based optimization technique over the PIM-SM stack was proposed using fuzzy logic and genetic optimization. The proposed method reduced the overall end to end delay in the network and increased the throughput of the network.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1]. YW Leung and TS Yum.[ 1996] Connection optimization for two types of videoconferences, ” *IEE Proc. Commun*, 143(3): 133–140,
- [2]. DP Anderson.[ 1993]Metascheduling for continuous media, *ACM Trans. Computer Syst.*, 11(3): 226–252.
- [3]. VP Kompella, JC Pasquale, and GC Polyzos.[ 1993] Multicast routing for multimedia communication, *IEEE/ACM Trans. Networking*, 1(3): 286–292.
- [7]. ZeyadM. Alfawaer GuiWeiHua and Noraziah Ahmed. [2007] A Novel Multicast Routing Protocol for Mobile Ad Hoc Networks. *American Journal of Applied Sciences*, 4: 333-338.
- [8]. M Mitchell.[ 1996] An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press,.
- [9]. BA Julstrom.[1993] A genetic algorithm for the rectilinear Steiner problem, in *Proc. ICGA*, S. Forrest, Ed. San Mateo, CA: Morgan Kaufmann, , pp. 474–480.
- [10].H Esbensen.[1995] Computing near-optimal solutions to the Steiner problem in a graph, *Networks*, 26: 173–185.
- [11].Biswas SK, Izmailov R.[2000] A QoS-aware routing framework for PIM-SM based IP-multicast, Global Telecommunications Conference, 2000. GLOBECOM '00. *IEEE* ,1(1):376-381
- [12].Qingfu Zhang and Yiu-Wing Leung.[ 1999] An Orthogonal Genetic Algorithm for Multimedia Multicast Routing, *IEEE Transactions On Evolutionary Computation*, 3(1)
- [13].R-H Hwang, W-Y. Do and S-C Yang.[ 2000] Multicast Routing Based on Genetic Algorithms, *Journal of Information Science and Engineering*, 16: 885-901.
- [4]. Q Zhu, M Parsa, and JJGL Aceves.[ 1995] A source-based algorithm for delay-constrained *minimum-cost multicasting*, in *Proc IEEE INFOCOM* , 377–385.
- [5]. F Fluckiger, Understanding Networked Multimedia: Applications and Technology. Englewood Cliffs, NJ: Prentice-Hall, 1995.
- [6]. Deering S, Estrin DL, Farinacci D, Jacobson V, Ching-Gung Liu, Liming Wei. [1996] The PIM architecture for wide-area multicast routing, *Networking, IEEE/ACM Transactions on* , 4(2):153-162
- [14].Shangchao Pi, Baolin Sun, Fuzzy Controllers Based Multipath Routing Algorithm in MANET” *Physics Procedia*, 24( Part B):1178-1185 *International Conference on Applied Physics and Industrial Engineering* 2012
- [15].G Santhi, AlameluNachiappan.[2012] Fuzzy-cost based multiconstrainedQoS routing with mobility prediction in MANETs, *Egyptian Informatics Journal*, 13(1): 19-25
- [16].AllamMaalla, Chen Wei and Haitham J Taha.[2009] Optimal Power Multicast Problem in Wireless Mesh Networks by Using a Hybrid Particle Swarm Optimization. *American Journal of Applied Sciences*, 6: 1758-1762.
- [17].LA Zadeh.[1975] Fuzzy logic and approximate reasoning, *Synthese* 30 :407–428.
- [18].Sukhvinder Singh, Manoj Yadav,Sanjeev Kumar [2015] A Review on Quality of Service Multicast Routing Techniques based on Genetic Algorithm in Manet, *International Journal of Advanced Research in Computer Engineering & Technology* 4,( 6)
- [19].Sara Aliabadi and Mehdi Agha Saram [2014] An approach to DSR Routing QoS by Fuzzy – Genetic Algorithms, *International Journal of Wireless & Mobile Networks (IJWMN)* 6(2)

\*\*DISCLAIMER: This article is published as it is provided by author and approved by guest editor. Plagiarisms and references are not checked by IIOABJ.

# AN INTELLIGENT FEATURE SELECTION APPROACH FOR GENE EXPRESSION DATA USING HYBRID BIOGEOGRAPHY BASED OPTIMIZATION (BBO) WITH ARTIFICIAL BEE COLONY (ABC) ALGORITHM

S Venkata Krishna Kumar<sup>1\*</sup> and R Nedunchezian<sup>2</sup>

<sup>1</sup>Dept. of Computer Science, PSG College of Arts and Science, Coimbatore, Tamil Nadu, INDIA

<sup>2</sup> Dept. of Computer Science and Engineering, Kalaignar Karunanidhi Institute of Technology, INDIA

## ABSTRACT

Micro array data's plays major role in the bio medical fields which is used to analyze and predict the various diseases effectively. Classification is a most concerned data mining strategy that can be used to predict the disease by classifying the micro array data in terms of their feature values. However, this would be more difficult task due to presence of irrelevant information present in the micro array data such as noises, redundant genes, and uninformative genes and so on. This research work focus on implementing the methodology which can be used to classify the micro array data accurately with increased performance result. Methods: In the proposed research methodology optimal feature selection is focused to avoid the irrelevant information given as input to the classifier, thus the performance can be optimized. It is achieved by introducing the novel mechanism namely Hybridized biogeography based optimization (BBO) and Artificial bee colony algorithm (ABC) in which migration operator process is combined with the ABC algorithm. In the proposed research method, preprocessing is done by using NLLS impute algorithm which would select the top most 60 genes from the set of gene with replacement of missing values. After pre-processing SVM-REF and BBF strategies are used for the classification and gene selection purpose that is finished with the assistance of leave-one-out cross validation methodology (LOOCV). Results: The evaluation of the proposed research methodology is done in the matlab simulation environment under varying size of micro array data set. The results are compared against the different performance metrics with various existing research works for gene expression dataset bench marks. Conclusions: The finding of the research work proves that proposed method outperforms the existing methodologies. Applications: This approach would be more useful in the fields such as bioscience, bioengineering, and medical fields for the appropriate detection and analysis of the various diseases and their patterns.

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Gene expression data, gene selection, hybrid approach, multi-objective binary biogeography based optimization

\*Corresponding author: Email: [leenasilvoster@gmail.com](mailto:leenasilvoster@gmail.com) Tel.: +91-7385644304; Fax: 020 24351308

## INTRODUCTION

Gene selection is one of the most important technique to find the informative genes from the cells. However it is most difficult to predict the genes from the large volume of data's that are irrelevant to the cancer disease. This can be resolved by integrating the gene selection approach which is used to find the informative genes and omit the irrelevant gene data's [1].

The gene expression data like non-relevant data, noise data and comparison of gene samples (high dimensional data) causes tedious problem in the gene selection process [2]. To avoid this problem, the efficient classifier's are required to classify the exact samples of genes in a particular set of instructive genes from the part of a larger set of gene samples by using a gene selection method. The gene selection is called feature selection in the computational intelligence domain [3]. The advantages of gene selection are

- 1) It improves the accuracy of the classification,
- 2) Shorter process time.
- 3) It can reduce the over fitting of the data.
- 4) It can take away the not related and noise genes.

The feature selection method is mainly used to decrease the more number of genes in classification in the classification accuracy [4,5]. It means this method mainly involved increasing the classification performance and decreasing the more number of features into small number of features selected. Generally the feature selection

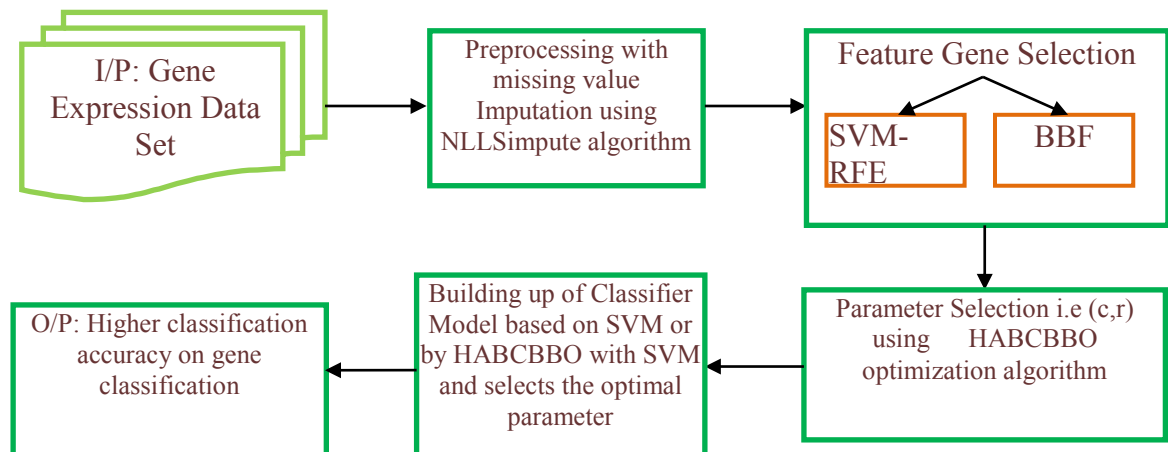
based on single objective problem based evolutionary algorithms is not as much of common and computationally costly as multi objective approached. For that reason the evolutionary algorithm is used to develop a multi objective approach is required and at the same time it make the most of the classification rate and reduce the number of features selected.

The NSGA-II [6] and SPEA [7] are currently planned for the feature selection in multi objective based methods. To get small set of non-redundant disease related genes by using the multi objective particles swarm optimization from the hybrid multi-objective optimization method [8]. Salcedo-Sanz et al uses multi objective genetic algorithms used to common uniqueness of the samples as feature correlation and search the subset of features by combining different filter approaches criteria for feature selection [9]. The multi-objective approach in hybrid GA and support vector machine classifier (GASVM) is proposed for gene selection and the classification of gene expression data [10]. An archived multi-objective simulated annealing (AMOS) is developed for predicting miRNA promoters using an SVM with RBF kernel in a feature selection method [11, 12].

A neural network is working to let the use of a representative database to calculate suitability. A multi-objective evolutionary algorithm is projected to solve the difficulty of gene selection in the gene subset size minimization and performance maximization [13]. To establish the benchmark problems by using the optimization of split modified radius-margin model selection criteria [14]. The a multi-objective genetic algorithm is used to gene selection in the microarray datasets [15]. This process is consummate by support vector machines. A multi objective genetic algorithm used to increase the classification accuracy rate in the number of features [16]. And also the multi objective genetic algorithm useful for image processing especially for edge detection and the complete comparison on the canny edge detector [17]. In the bioinformatics and computational biology field using multi objective optimization technique [18]. This technique is mainly used to elucidate the reasons after the use of multi objective optimization in each application area and also to direct the possible futures.

## MATERIALS AND METHODS

In the proposed research methodology, an improved classifier for the micro array data to diagnose the disease is introduced which is achieved by introducing the optimal feature selection mechanism that can avoid the irrelevant features from the set of micro array data set. In this approach gene expression data set is given as a input. Initially infomax ICA mechanism is used to select the top 60 genes from the input data. Then the preprocessing is done to replace the missing values using NLLS imputation approach. Then the SVM-REF and BBF is used to perform gene selection by adapting leave one at last out cross validation (LOOCV) method in the support vector machine. At last HABCBO is adapted to perform optimal feature selection with the concern of discrete problems. It is called BBHK and is based on migration model. The overall proposed methodology shown in the [Figure-1].



**Fig.1: Overall Proposed Methodology**

### Preprocessing of the dataset with missing value imputation

The gene expression microarray data is preprocessed to reduce the larger amount of genes to the small set of genes. It is used to increase the competence of classification accuracy rate in the high dimensional data set. For various tentative reasons the microarray data might contains a missing values. Before applying the microarray data in data analysis algorithm it is very important to impute the missing values. In this proposed work a New Local Least Square imputation based algorithm (NLLS impute) implemented.



This algorithm is similar as K-nearest neighbor imputation algorithm. In this research work it is used to select the K-nearest gene from the complete set of obtained candidate gene as a substitute of the complete genes and also reuse the imputed information available. Subsequently the highest missing rate will be imputed in the target gene with the available information. The New Local Least Square impute algorithm proceeds as follows:

Algorithm: NSLLSimpute

Input: The m number of genes and n number of samples from the microarray gene expression matrix T it contains some artificial missing entries.

Output: The missing entries are predicted.

- 1) At first from the complete gene expression matrix X to replace the missing positions of the given gene expression matrix.
- 2) According to their missing rate and imputation one after another the m number of genes in X are sorted in ascending order.
- 3) For each target gene  $gt$  in X do:
  - a) For each missing position  $p$  in target gene  $gt$  do:
    - i) The Pearson Correlation coefficient is used to select the k number of candidate genes nearest to the target gene. for the period of the computation of Pearson Correlation coefficient the missing positions filled with row average are ignored [9].
    - ii) The conception of local least squares is imputing the missing position  $p$  in the target gene beneath deliberation [9].
    - iii) The result of imputed values are positioned in the missing positions of the considered target gene.
- 4) End

Gene selection

The integrated SVM-RFE having an excellent result on classification but it having the poor performance on separability in redundant class labels. BBF is used to avoid redundant class labels in selection. Both combined achieve as good as possible. The Fig.1 shows the schematic view of overall process carried. The SVM-RFE and BBF are used to select the feature from dataset after selection the features are going to the process of classifier for training. Finally in the training session the evaluation is conducted with testing data.

Feature Selection Method

The recursive elimination procedure of SVM-RFE [3] is implemented as follows:

1. Start: ranked set  $R=[ ]$ ; picked feature subset  $S=[1, \dots, d]$ .
2. Repeat until all features in subset gets ranked:
  - a. Train the features with SVM from set S as input variables.
  - b. Calculate the weight vector for each feature.
  - c. Calculate the ranking score for features in set S.
  - d. Identify the feature with the smallest ranking score.
  - e. Update.
  - f. Eliminate smallest ranking feature.
3. Result: Ranked feature set  $R[ ]$ .

After ranking genes by SVM-RFE we eliminate redundancy by applying Based Bayes Error Filter (BBF). The BBF initially processed the relevant candidate genes that are selected by a criterion function and second the criterion controlling the upper bound of the Bayes error is applied to the relevant candidate genes in order to remove the redundant genes. By this way, the genes and the subset of the gene are selected for further processing.

### HABCBO for optimizing the objective function

The stages of ABC algorithms are categorized into three phases. Those are employe bee phase, onlooker bee phase, scout bee phase. Initially, The population of bee  $N_p$  would be created randomly by using the bees that are present in the search space  $\Omega$ . This population is nothing but the group of food sources. Based on the volume of the food space, bees attraction would differ. For example, food space with volume would be attracted more by the bees and less food volume would be less attracted by bees. Thus in ABC approach, objective function is find a food source with more volume of food. The fitness function based on this objective function is calculated after initialization of the bee population. The equation used to calculate the fitness function is given as follows:

$$fit(X_i) = \begin{cases} \frac{1}{1 + f(X_i)} & \text{iff}(X_i) \geq 0 \\ 1 + |f(X_i)| & \text{Otherwise} \end{cases}$$

In employee phase, swarm of employees bees would be sent to find a food source that contains more volume of food. In this swarm of employees bees, at least one employee bee would find the food source where the bee that find the food is denoted as  $i$  and the food source that was found is  $v_i$ . After finding the new food source with more volume of food, the old food source position would be updated by substituting the new  $x_i$  in old value. Onlooker bees are used to filter the food sources that are found by the employee bees based on the food quality and the volume. This is calculated by using the roulette wheel selection approach which finds a food source based on probability values. Scout bees are used in the emergency situation where the optimal solution cannot be found. It will send out to find a food source when the employee and onlooker bees cannot find an optimal solution for the particular period of time [10, 11, 12]. It is employed based on the parameter called 'limit' in terms of time unit. When the

optimal food source cannot be obtained in the limit time period then the scout bees would be employed. The ABC algorithm lacks in solution updation where the solution are not updated in the intermediate level like cross over, mutation. Thus the solution identification would not produce optimal solution [16]. This can be improved by integrating the biogeography-based optimization (BBO) approach for the migration of solution in the intermediate level.

### Information migration scheme

The BBO scheme is based on the biological behaviour of the sea species isolation from the one location to another location. The measure that is used in the BBO process for finding the optimal solution is habitat suitability index (HSI). Here the solution that have more HSI is defined as good solution, less HSI would be taken as poor solution. In BBO approach, habitats with more emigration rate and the less immigration rate would be taken as best solution. Based on these HSI values, solutions would be sorted out and the calculation procedure of  $\mu$  and  $\lambda$  is calculated using the following equation.

$$\lambda_i = I \left(1 - \frac{i}{n}\right)$$

$$\mu_i = E \left(\frac{i}{n}\right)$$

### Hybrid ABC with migration operator

As mentioned in the introduction, ABC does not do well in exploiting the existing information of solutions. On the other hand, BBO is good at exploiting the information of existing solutions. Based on these observations, we propose to hybridize ABC with migration operator of the BBO algorithm to combine their strengths, called HABCBBBO. This algorithm combines the migration operator with the employed bee phase of ABC. In addition, Rechenberg's one-fifth success rule is employed to effectively control the adaptation of immigration probability.

#### Algorithm 1

**Input:**  $f(\cdot)$ ,  $D$ ,  $x^{\min}$ ,  $x^{\max}$ ,  $N_p$ , limit

**Output:** the best solution obtained from the algorithm

1. Randomly create  $N_p$  solution  $x_1, x_2, x_3, \dots, x_{N_p}$
2. Evaluate function values of solution and their fitness by (1).
3. Set counter  $l_i, i=1, 2, \dots, N_p$  to 0;
4. Repeat
5. Send out employed bees (see algorithm 2);
6. Send out onlooker bees depending on their nectar
7. Amounts by (2);
8. Evaluate candidate solution and their fitness by (1).
9. Do greedy selection by (3) and update  $l_i$  by (5)
10. Send out scout bees if  $l_i$  reaches limits
11. Evaluate candidate solution and their fitness by (1).
12. Reset  $l_i$  to 0.
13. Until termination criterion are met

#### Algorithm 2: Pseudo code of the employed bee phase in HABCBBBO

Set  $S_c = 0$ ,  $\sigma = 0.5 D$  and  $c_d = 0.82$

Calculate  $\lambda_i$  and  $\mu_i$  for each solution in the bee colony

For  $i=1$  to  $N_p$  do

Randomly choose  $i_1 \in [1, D]$  and  $r_1 \in [1, N_p]$

For  $j=1$  to  $D$  do

If  $j=i_1$  then

$$v_{ij1} = x_{ij1} + \varphi(x_{ij1} - x_{r_1j1})$$

Else if  $\text{rand}(0,1) < \lambda_i$  then

Choose a solution  $x_{r_2j}$  with  $r_2 \neq i$  using roulette wheel selection method based on emigration rates  $\mu_i, i=1, 2, \dots, N_p$

$$v_{ij} = x_{r_2j}$$

else

$$v_{ij} = x_{ij}$$

else

else for

evaluate candidate solution and their fitness by (1)

if  $f(v_j) < f(x_j)$  then

replace  $x_j$  by  $v_j$

```

Sc++
End if
End for
If  $\frac{S_c}{N_p} < \frac{1}{5}$  then
 $\sigma = c_d \cdot \sigma$ 
Else If  $\frac{S_c}{N_p} > \frac{1}{5}$  then
 $\sigma = \frac{\sigma}{c_d}$ 
Else if
Reset Sc = 0

```

HABCBO is described in Algorithm 1. This algorithm is same as ABC algorithm in which procedure differs in the phase of employee. Initially immigration and the emigration rate would be calculated by using equations qw and 13. And the optimal solution updation procedure is represented in the following equation.

$$v_{ij} = \begin{cases} x_{i,j} + \varphi(x_{i,j} - x_{r1,j}) & \text{if } j = j_1 \\ x_{r2,j} & \text{if } \text{frand}(0,1) < \frac{\lambda_i}{\sigma} \text{ and } j \neq j_1 \\ x_{i,j} & \text{Otherwise} \end{cases}$$

The above procedure is repeated for the  $N_p$  times to obtain the optimal solution. And the more opt best solution is found by comparing the best population with the location of the populations. If  $f(v_i) < f(x_i)$  then the solution is said more optimal, else optimal solution is not obtained. During this iteration, number of successful operation would be counted and it is incremented by  $S_c$  and the one fifth rule is applied.

$$\sigma = \begin{cases} c_d \cdot \sigma & \text{if } \frac{S_c}{N_p} < \frac{1}{5} \\ \frac{\sigma}{c_d} & \text{if } \frac{S_c}{N_p} > \frac{1}{5} \\ \sigma & \text{Otherwise} \end{cases}$$

Where  $c_d \rightarrow$  decay factor which is assumed as 0.82 for the adaptation [18].

In the hybridized employed bee phase,  $\sigma$  is used to control the adaptation of immigration rate. In Algorithm 2,  $\sigma$  is initialized to 0.5D, which is determined based on experiment on toy functions. The final result of the optimization is the best individual of the last iteration.

## HABCBO and support vector machines

The issues that are found in the SVM methodology are optimal selection of input feature subset and the kernel parameters. This issue is resolved in the proposed research method by integrating the HABCBO algorithm with the SVM approach which is called as HABCBO-SVM to optimize the parameters C and r. By doing so, feature subset selection can be done optimally and the testing accuracy of SVM can be improved. In the first phase, the BBKH algorithm provides a binary encoded individual where each bit represents a gene. If a bit is 1, it denotes this gene is kept in the subset; else if at bit is 0, it represents a non-selected feature. Therefore, the individual length is equal to 2+D in the initial microarray dataset. Then, the fitness of each individual is assessed by the accuracy of leave-one-out cross-validation method (LOOCV). The leave-one-out cross-validation method can be described as follows: when there are n data to be classified, the data are divided into one testing sample and n-1 training samples. Each individual will be selected as a testing sample in turn. The other n-1 individuals serve as the training data set to determine the prediction parameter of the model. The proposed research methodology HABCBO\_SVM is demonstrated by using the data set that contains 7 records namely A1, A2, A3, A4, A5, A6, and A7 that contains four features. Among these records, six records are used for training process and the remaining one record is taken for testing process. The SVM method fixes the parameter values as  $C = 2^5$  and  $r = 2^{-2}$ , thus the more classification accuracy can be obtained. The fitness calculation is calculated as like follows:

$$f_1 = \text{SVM\_accuracy}$$

$$f_2 = \left( \frac{D - R}{D} \right)$$

$$f = [f_1, f_2]$$

where SVM\_accuracy  $\rightarrow$  SVM classification accuracy, D  $\rightarrow$  total number of the genes, and R  $\rightarrow$  number of selected genes.

The process of the algorithm is as follows

Step 1: The gene expression data are pre-processed by the Infomax ICA method. The 60 top genes with the highest scores are selected as the crude gene subset. The corresponding subsets in the testing parts are also selected at the same time.

Step 2: Converting genotype to phenotype. This step will convert parameter C, r and feature subset from its genotype to a phenotype.

Step 3: the two objectives functions,  $f_1$  and  $f_2$  in (9) are calculated using ABCBBO optimization.

Step 4: Multi-objective binary biogeography based optimization. In this step, the algorithm searches for better solutions by binary migration model and binary mutation model.

Step 5: Checking the termination criterion. The final feature subsets are selected, and then output the feature subset and the parameters C and r.

The System architecture of the proposed HABCBBO based feature selection and parameters optimization for support vector machine is shown in [Figure- 2].

## RESULTS

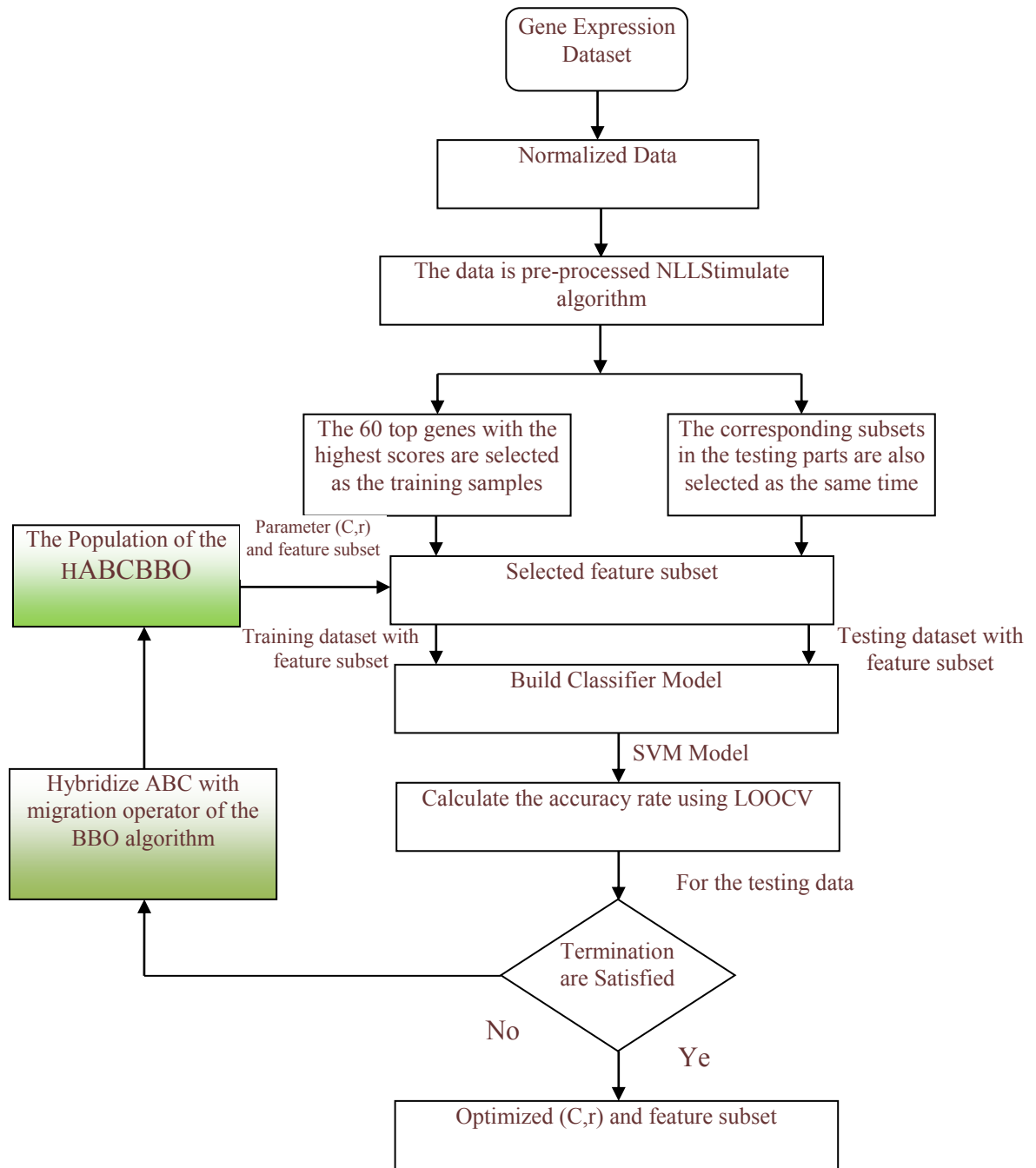
The relevant gene selection is the most complex task in the gene expression classification method. In this research work, ten gene expression data is selected for the classification that includes micro array data of tumor samples, brain tumor, leukemia, lung cancer, and prostate tumor samples. This misro array data can be downloaded from the website <http://www.gems-sytem.org>. The data sets are partitioned into one testing sample and n-1 training samples. Note that each individual will be selected as a testing sample in turn. The remaining n-1 individuals will then serve as the training data set.

**Table: 1. Format of gene expression classification data**

Dataset Number	Dataset Name	Number of Samples	Number of Genes	Number of Classes
1	11_Tumors	60	9	5726
2	9_Tumors	174	11	12533
3	Brain_Tumors 1	90	5	5920
4	Brain_Tumors 2	50	4	10367
5	Leukemia 1	72	3	5327
6	Leukemia 2	72	3	11225
7	Lung_Cancer	203	5	12600
8	SRBCT	83	4	2308
9	Prostate_Tumor	102	2	10509
10	DLBCL	77	2	5469

The proposed research work namely HABCBBO-SVM is implemented in the Matlab-7.9 language development environment which is extended using the libsvm that was originated by Chang and Li [28]. The computing processor characteristics are Pentium 3.0 GHz Processor with 1.0 GB of memory. The proposed research work is evaluated and compared with the existing research works namely SVM Grid search, IBPSO, HPSOTS, PSO/GA [29], and a multi-objective algorithm NSGA-II [6]. The parameter values that are set are, "Population size:50, Number fo generation:100, HIS:1, Mutation rate:0.5".

To make the experiments more accurate, each algorithm will be run ten times for each gene data. After that, an average result of the ten independent runs is obtained and compared. As stated earlier, the main objective of the proposed research work is increasing classification accuracy and the number of genes selected. Thus in this multi objective problem, the final solution is taken by using the pareto front method in which any solution is not dominated by any other solutions. Pareto front optimal procedure is used to select the most optimal solution among the various number of solutions that can lead to high classification accuracy.



**Fig. 2:** System architectures of the proposed HABCBBO -based feature selection and parameters optimization for support vector machine

[Figure- 3, 4] represents the experiment results of MOBBKH SVM on ten gene datasets. In these figures, #acc denotes the testing accuracy, and #selected gene denotes the number of genes selected for these gene expression data. From the figures, we can see that the results of the proposed algorithm are consistent on all datasets. For the Leukemia1, Leukemia2, SRBCT, and DLBCL datasets, MOBBBO, the proposed MOBBKH can achieved 100% LOOCV accuracy with less than 10 selected genes. For the Brain\_Tumor2 dataset, MOBBBO can obtain the 100% LOOCV accuracy for nine times. For the other datasets of Lung\_Cancer, Prostate\_Tumor, and Brain\_Tumor1, the BBKF algorithm can also provide more than 96% classification accuracies except for the 11\_Tumors dataset (92.414%) and 9\_Tumors dataset (80.5%).

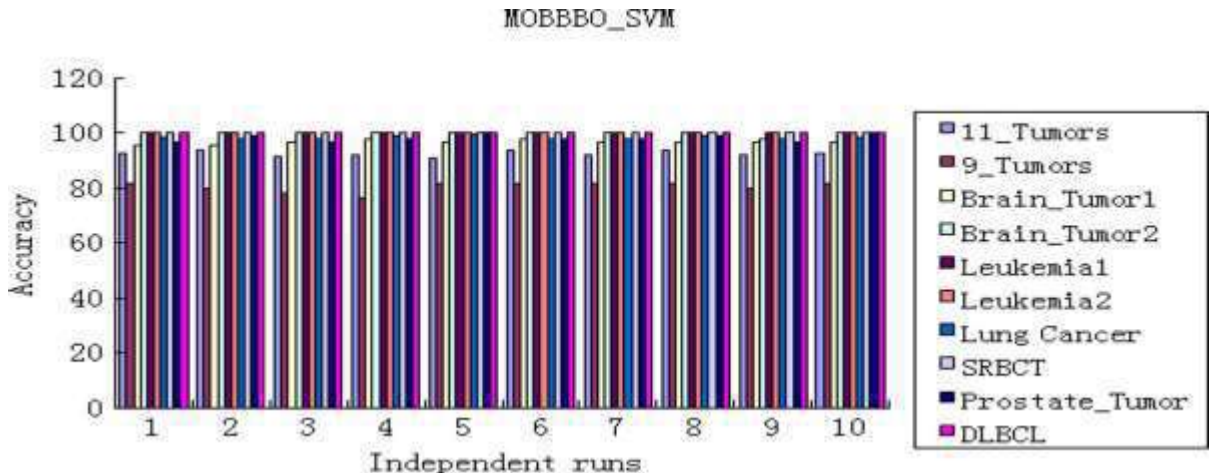


Fig. 3: The accuracy obtained by MOBBKH in each independent runs

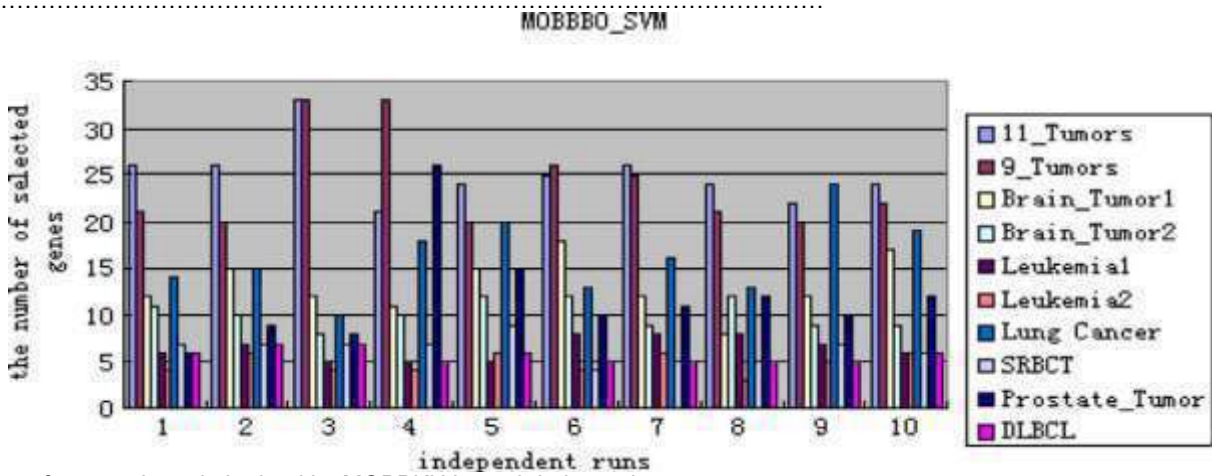


Fig. 4: The number of gene selected obtained by MOBBKH in each independent runs

The characteristics of the genes, selected genes, and percentage of genes selected percentage are shown in Table II. The average percentage of genes selected is 0.0017. For the Leukemia1, Leukemia2, SRBCT, and DLBCL dataset, our algorithm can obtain 100% LOOCV accuracy even though the percentage of genes selected is reduce to 0.0012, 0.0004, 0.0028, and 0.0010. So we can conclude that not all features are necessary for achieving the better classification accuracy.

Table 2: The genes, selected genes, and percentage of gene selected percentage

Dataset Name	Total Number f Genes	Number of Genes Selected	Percentage of genes selected
11_Tumors	5726	25.1	0.0044
9_Tumors	12533	24.1	0.0019
Brain_Tumors 1	5920	13.2	0.0022
Brain_Tumors 2	10367	10.2	0.0010
Leukemia 1	5327	6.5	0.0012
Leukemia 2	11225	4.8	0.0004
Lung_Cancer	12600	16.2	0.0013
SRBCT	2308	6.4	0.0028
Prostate_Tumor	10509	11.9	0.0011

In order to show the effective of each part, two experiments are conducted. The first experiment is to show the effective of the Fisher-Markov selector, and the second experiment is to show the effective of MOBBKH. In [Table 3](#), the better results between two algorithms are highlighted using boldface. It is easy to see that both the classification accuracy and the number of selected genes of HABCBBO SVM are superior to grid search SVM. This also demonstrates the effective of HABCBBO.

**Table 3:** Compared With mobbbo SVM With Other Previous PSO

Dataset Name	Evaluation	Proposed HABCBBO	MOBBKH	MOBBBO	IBPSO [8]	IBPSO [6]
11_Tumors	Accuracy	96	95	92.414	95.06	93.10
	Genes	26	26	25	240.9	2948
9_Tumors	Accuracy	82.5	81.23	80.5	75.50	78.33
	Genes	25	25	24.1	240.6	1280
Brain_Tumors 1	Accuracy	98	97.12	96.667	92.56	94.44
	Genes	14	14	13.2	11.2	754
Brain_Tumors 2	Accuracy	99.94	99.92	99.8	92.00	94.00
	Genes	11	11	10.2	9.10	1197
Leukemia 1	Accuracy	100	100	100	100	100
	Genes	7	7	6.5	3.5	1034
Leukemia 2	Accuracy	100	100	100	100	100
	Genes	5.2	5.2	4.8	6.7	1292
Lung_Cancer	Accuracy	99.5	99.1	98.473	95.86	96.55
	Genes	17	17	16.2	14.90	1897
SRBCT	Accuracy	100	100	100	100	100
	Genes	7.5	7.5	6.4	17.50	431
Prostate_Tumor	Accuracy	99.2	99	98.33	97.94	92.16
	Genes	13	13	11.9	13.60	1294
DLBCL	Accuracy	100	100	100	100	100
	Genes	6	6	5.7	6	1042

From the [Table-3](#) we can see for Leukemia1, Leukemia2, SRBCT, and DLBCL, both MOBBBO SVM and MOBBBO SVM without Fisher-Markov selector can obtain the 100% classification accuracy, while MOBBBO SVM can provide lower gene numbers. For 9\_Tumors, MOBBBO SVM, and MOBBBO SVM without Fisher-Markov selector can obtain the 80.5% classification accuracy, and MOBBBO SVM can also provide lower gene number. For Brain\_Tumors1, Brain\_Tumors2, Lung\_cancer, and Prostate\_Tumor datasets, MOBBBO SVM can not only provide better classification accuracy, but also lower gene numbers. Only for the 11\_Tumors, the MOBBBO without Fisher-Markov selector can generate the better classification accuracy, which also demonstrates that the Fisher-Markov selector is not suitable for all the situations. For the second experiment, we compare our approach with grid search SVM.

Based on the above analysis, the experimental results can demonstrate the flexibility and robustness of the proposed HABCBBO in feature selection. This algorithm can provide positive results when applied to the gene expression data with the limited number of features and samples. The reason may be that the multi-objective binary biogeography based optimization tends to share their features with low HSI solutions, which can accept a lot of new features from high HSI solutions. In HABCBBO, a habitat is a vector which follows binary migration and binary mutation step to the optimal solution. The new candidate habitat is generated from all the solutions in population by using the binary migration and binary mutation model. Following these rules, the HABCBBO algorithm finally produces a better subset for the gene classification.

## CONCLUSION

In this paper, a hybrid multi-objective binary biogeography based optimization with support vector machine is proposed for gene selection on ten gene expression datasets. Experimental results show that the algorithm can simplify feature selection by finding a smaller number of features needed effectively and a higher classification accuracy compared with other previous methods. The proposed algorithm can obtain the highest accuracy in nine of the ten microarray dataset problems since the multi-objective approach in it can find a diverse solution in Pareto optimal set. Moreover, the results show that there are many irrelevant genes in gene expression data and some of them are not relevant to a given cancer. For further work, the proposed algorithm can be applied to some problems in other fields.

## CONFLICT OF INTEREST

Authors declare no conflict of interest

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

No financial support was received to carry out this project

## REFERENCES

- [1] Liang Q, Cheng X, Samn S.[2010] NEW: Network-enabled electronic warfare for target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 46(2):558–568.
- [2] Liang Q, Cheng X, Huang S, Chen D. [2014] Opportunistic sensing in wireless sensor networks: theory and application. *Comput. IEEE Trans.*63(8):2002–2010.
- [3] Singh S, Liang Q, Chen D, Sheng L.[2011] Sense through wall human detection using UWB radar, *Journal on Wireless Communications and Networking* 2011(1):503–520.
- [4] Liang Q.[2011] Situation understanding based on heterogeneous sensor networks and human-inspired favor weak fuzzy logic system. *IEEE Systems Journal*. 5(2):156–163.
- [5] Liang Q.[2011] Radar sensor wireless channel modeling in foliage environment: UWB versus narrowband. *IEEE Sensors Journal* 11(6):1448–1457.
- [6] Xu L, Liang Q.[2012] Zero correlation zone sequence pair sets for MIMO radar. *Aerospace and Electronic Systems*, *IEEE Transaction*, 48(3):2100–2113.
- [7] Del Ser J, Gil-Lopez S, Perez-Bellido A, Salcedo-Sanz S, Portilla-Figueras JA. [2011] IEEE 73rd Vehicular Technology Conference (VTC Spring). On the application of a novel hybrid harmony search algorithm to the radar polyphase code design problem (IEEE Computer Society Budapest, Hungary. pp. 1–5.
- [8] Gil-Lopez S, Ser JD, Salcedo-Sanz S, Perez-Bellido AM, Cabero JM and Portilla-Figueras JA.[2012] A hybrid harmony search algorithm for the spread spectrum radar polyphase codes design problem, *Expert System Application* 39(12):11089–11093
- [9] Perez-Bellido AM, Salcedo-Sanz S, Ortiz-Garcia EG, Portilla-Figueras JA, Lopez-Ferreras F.[2008] A comparison of memetic algorithms for the spread spectrum radar polyphase codes design problem, *Engineering Applications of Artificial Intelligence*.21(8):1233–1238.
- [10] Karaboga D, Basturk B, [2008] On the performance of artificial bee colony (ABC) algorithm. *Applied soft computing*, 8(1):687–697.
- [11] Diwold K, Aderhold A, Scheidler A and Middendorf M.[2011] Performance evaluation of artificial bee colony optimization and new selection schemes. *Memetic Computing*. 3(3):149–162.
- [12] Zhang X, Zhang X, Ho SL and Fu WN.[2014] A modification of artificial bee colony algorithm applied to loudspeaker design problem. *IEEE Transactions on Magnetics*, 50(2): 737–740.
- [13] Karaboga D and Gorkemli B.[2014] A quick artificial bee colony (QABC) algorithm and its performance on optimization problems. *Applied Soft Computing*. 23:227–238.
- [14] Zhang X, Zhang X, Yuen SY, Ho SL, Fu WN.[2013] An improved artificial bee colony algorithm for optimal design of electromagnetic devices. *IEEE Transactions on Magnetics* .49(8):4811–4816.
- [15] Zhang X, Wu Z.[2015] Advances in Swarm and Computational Intelligence. Lecture Notes in Computer Science, 9140, ed. by Y Tan, Y Shi, F Buarque, A Gelbukh, S Das, and A Engelbrecht. An artificial bee colony algorithm with history-driven scout bees phase. pp. 239–246.
- [16] Karaboga D, Gorkemli B, Ozturk C and Karaboga N.[2014] A comprehensive survey: artificial bee colony (ABC) algorithm and applications. *Artificial Intelligence Review* 42(1):21–57.
- [17] Simon D. [2008] Biogeography-based optimization. *IEEE transactions on evolutionary computation*, 12(6):702–713
- [18] Schwefel HP. [1981] Numerical Optimization of Computer Models (Wiley, Chichester).
- [19] Dukic, ML Dobrosavljevic, ZS. [1990] A method of a spread-spectrum radar polyphase code design. *IEEE Journal on Selected Areas in Communications*. 8(5):743-749.



# BLURRED FACE RECOGNITION USING LOW DIMENSIONAL LINEAR MODEL

M.B. Sudhan<sup>1</sup>, V.S. Dharun<sup>2</sup> \*

<sup>1</sup>Research Scholar, Dept. of Computer Science Engineering, Noorul Islam University, Kumaracoil, Tamilnadu, INDIA

<sup>2</sup>Archana College of Engineering, Palamel, Azhappuzha District, Kerala, INDIA

## ABSTRACT

*Aim: Face recognition in image processing technique remains to be the major technique in identifying a person and also for authentication purpose. This is the main scheme to concentrate more about security issues and identifying the respective person. Many schemes have been evolved to provide better solution to the face recognition issues but the complexity presented in the schemes was a challenging one. A new model is proposed with the hands of three main processes such as (i) Set of all Blurred Images (ii) Blur Kernel Identification and (iii) Blur Removal. One input image is provided to the system for processing and 20 or more sample images are taken. Then the input image is multiplied with the convolution operator. For Blur coefficients, we use Gaussian Kernel algorithm, which produces the estimation of blur content, once it completes we count the blur pixels, after that the analyzed value, should be removed. Next we need to identify the outlier and misalignment. We need to calculate weight for misalignment using Local Binary Pattern Algorithm. Low dimensional linear model performs dimensionality reduction. For all the entire system its efficiency to analyze the face estimation scheme more perfectly compare to the existing results and the final scenario of these kind of implementation clearly explains the nature of image processing and explain its efficiency more perfectly.*

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Face Recognition, Illumination,  
Kernel, Blurred Images,  
Biometric Scheme.

\*Corresponding author: Email: [sudhan\\_mb@yahoo.com](mailto:sudhan_mb@yahoo.com)

## INTRODUCTION

Face gratitude has been a powerfully investigated scheme for a combination of approaches. Although important paces have been complete in undertaking the difficulty in forbidden fields, important confronts lingers in resolving it in the unimpeded provinces. Such situation occurs at the same time as be familiar with faces obtained from far-away web cameras or a regular camera. The major issues that create this a demanding difficulty are image squalor owing to blur and noise as well as differences in exterior due to lighting and pose. In this paper, we purposely take in hand the difficulty of distinguishing faces crossways blur and enlightenment.

## RELATED WORK

Normal method to deal with blurred faces is to deblur the picture initially as well as distinguishes it by means of conventional face appreciation procedures. On the other hand, this advance technique engrosses resolving the demanding difficulty of sightless image Deconvolution. We demonstrate that the deposit of cumulative image getting hold by blurring a agreed picture appearances a rounded position as well as additional particularly, we demonstrate so as to this bunch is the rounded hull of transferred descriptions of the innovative picture.

Thus with every colonnade picture it is able to correlate a matching rounded position. Supported on this set theoretic categorization, a blur robust face gratitude algorithm is being proposed. In the description of the resulting methodology, we work out the coldness of an agreed investigate picture [which is desired to be familiar with] from every of the curved collections, and allocate it the individuality of the neighboring colonnade picture. The detachment multiplication footsteps are invented as rounded convolution difficulties in excess of the breathing space of haze essential part and all of us are do not take for granted some attribute oriented otherwise sequential structure for the blur kernels. On the other hand, stipulation in sequence is obtainable; this able to be with no trouble included into our methodology, resultant in enhanced appreciation presentation. Additional, we construct our technique vigorous to unwanted layers as well as diminutive pixel mismatching arrangements by reinstates the Euclidean detachment by prejudiced L1 standard and evaluate the imagery in the LBP [Local Binary Pattern] liberty.

It has been exposed that all the imagery of a Lambertian rounded entity, beneath all probable enlightenment situations, be positioned on a short measurements [just about nine measurements) linear associate liberty. Although countenances are not precisely rounded or Lambertian, they are to be able intimately estimated by single. Consequently every countenance can be typified by a near to the ground measurements associate liberty and this

description has been second-hand for scheming enlightenment vigorous countenance acknowledgment techniques. Supported happening this enlightenment representation, we demonstrate that the collection of all pictures of a face beneath all haze and clarification differences is a bi-curved place. If we fasten the haze most important part then the collective of pictures get hold of by unreliable the enlightenment situations appearances a curved put as well as stipulation we fasten the lighting situation after that the position of every one in distinct pictures be too bowed. The remoteness calculations footsteps can be prepared as “Quadratically Constrained Quadratic Programs [QCQP]”, in that we resolve by compensating generalization in excess of the blur kernels as well as the enlightenment co-efficient. Comparable to the haze merely container, we create our technique vigorous to outliers as well as minute pixel disarrangements by reinstating the Euclidean model by the subjective L1 standard detachment and evaluate the picture in the LBP breathing space.

To abridge, the major technological involvements of this organization be as follows:

- (i) We demonstrate so as to the put of the entire pictures getting to be hold by hazing a known picture appearances a shaped situations. Additional purposely, we illustrate that this put is the bowed portion of budgeted accounts of the unique picture.
- (ii) According to this set theoretic description, we proposition a haze forceful face acknowledgment technique, in which it keeps away from resolving the demanding as well as superfluous difficulty of sightless picture de-convolution.
- (iii) Stipulation contains supplementary data on the category of haze touching the investigate picture, we can with no trouble integrate this information into our methodology, resulting in improved recognition performance and speed.

## MATERIALS AND METHODS

Our primary appraisal of the difficulty replica is for vague impression. After that, we demonstrate that the position of the entire pictures attained by blurring a known picture is bowed as well as in conclusion in attendance our technique is familiar with indistinct face sequences.

### A. Model of blurred convolution vector

The weighted average ratio of the blurred pixels of the image is nothing but a pixel of blur image ratio, which is the environs pixel ratio in the innovative pointed picture. Therefore, vague impression is a representation of complication procedure flanked by the innovative picture as well as a vague impression sieve is most important fraction in which it stands for the heaviness. Allow  $I$  is the innovative representation and  $H$  be the haze most important part of dimension  $[2k + 1] \times [2k + 1]$ , after that the in distinct picture  $I_b$  be agreed through

$$I_b(r, c) = I * H(r, c) = \sum_{k_i=-k}^k \sum_{k_j=-k}^k H(i, j) I(r - i, c - j)$$

Where “\*” symbolizes the complication operative  $a$ .  $r, c$  are the line and feature index of the picture. Vague impressions are most important part also gratify the subsequent possessions their co-efficient are positive, that is  $H \geq 0$ , in addition to totting up to 1 [that is  $\sum_{k_i=-k}^k \sum_{k_j=-k}^k H[i, j] = 1$ ].

The figure clearly illustrated the complete process and working of this system that is the initial stage of works begins with the input feeding procedure such as providing the test image with single or multiple faces. The input contains full of RGB color coefficients, we extract the structure and coefficients along with respective features such as structure, color [RGB], shape, and texture and so on. Once the features are extracted the data of the image is refined by means of rows and columns, each image contains lots of blocks and sub-blocks, which is mentioned by means of pixel values. Once the Pixels are analyzed the details will be compared to the train dataset which is created already. The resultant of the previous step will be blurring free and illumination free coefficient constraints. The exact matching of images will be the resultant of the final face recognition process.

### B. Pseudo code for low dimension linear model

The rationale of this respective algorithm is to recommend advanced techniques for arithmetical supposition of low dimensional constraints with high dimensional information. We create a center of attention on assembling self-assurance intermissions for personality co-efficient as well as linear amalgamations of more than a few of the respective individuals in a linear deterioration representation, even though our thoughts are appropriate in a great

deal extensive background. The hypothetical consequences obtainable at this time make available enough circumstances for the asymptotic ordinariness of the planned manipulations by the side of with a dependable manipulator for their restricted structured co-variance attributes. These adequate circumstances consent to the numeral of variables to distant go beyond the example dimension. The replication consequences obtainable at this time make obvious the correctness of the reporting likelihood of the planned self-assurance periods, powerfully at the bottom of the hypothetical consequences.

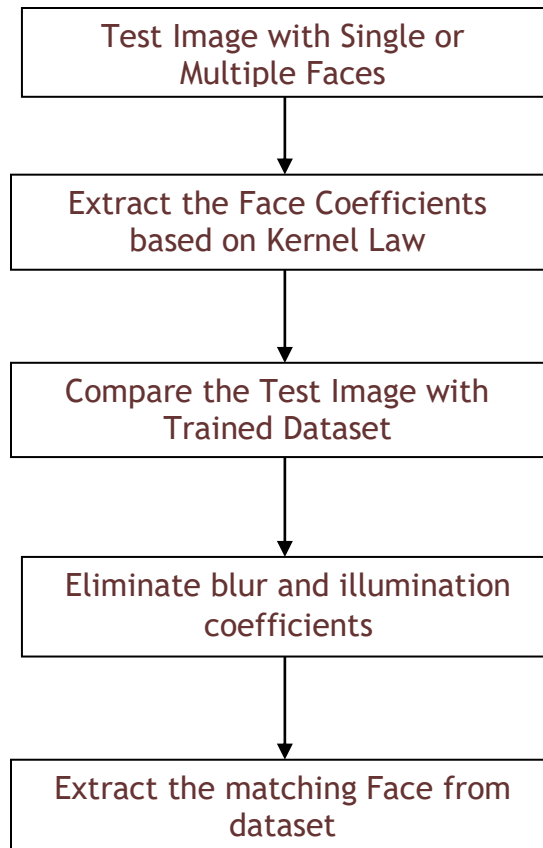


Fig. 1: System Design

## RESULTS

Input blurred image is initially taken and pixel values are determined. The original image is to be extracted from the blurred image by process of detection and segmentation.

## DISCUSSION AND CONCLUSION

Aggravated by the difficulty of inaccessible face appreciation, we tackled the difficulty be familiar within distinct and inadequately illumine faces using low dimensional linear model. The position of all representations get hold of by vague impression, agreed demonstration is a rounded position agreed by the rounded position of changed descriptions of the picture. Supported on this position descriptive categorization, we planned a vague impression vigorous face appreciation algorithm DRBF. In this technique we can with no trouble include preceding acquaintance on the category of haze as restrictions. By means of near to the ground dimensional linear sub-space representation for enlightenment, we illustrated that the position of all pictures acquired from the agreed picture by hazing and altering its enlightenment circumstances is a bunch of a gain, stands on this set theoretic classification, we projected a haze as well as enlightenment strong technique IRBF. We also established the effectiveness of our methodologies in undertakes the difficulty of countenance gratitude in unrestrained surroundings. Our technique is supported on a generative replica by adjacent fellow categorization flanked by the inquiry representation as well as the colonnade space, which creates it hard to balance it to real-life datasets with

numerous amounts of pictures. Therefore we consider that picture integrating with a discriminative acquaintance having a supported move towards SVM into this manipulation would be extremely hopeful bearing for potential effort in the future.

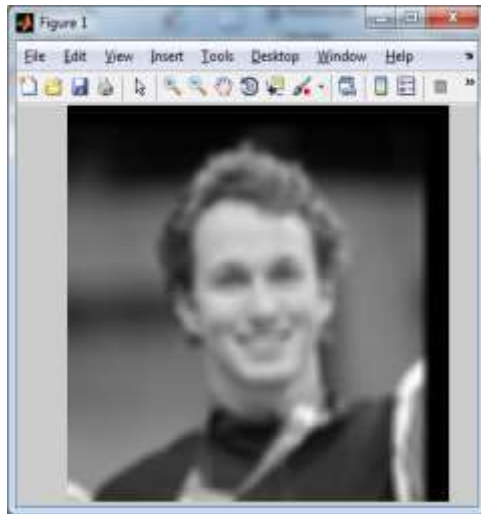


Fig: 2. Input Image with Blurred Pixels

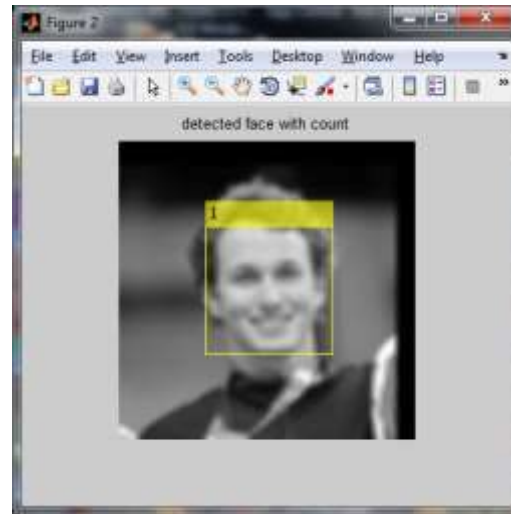


Fig: 3. Detected Face with Count



Fig: 4. Face Features Detection

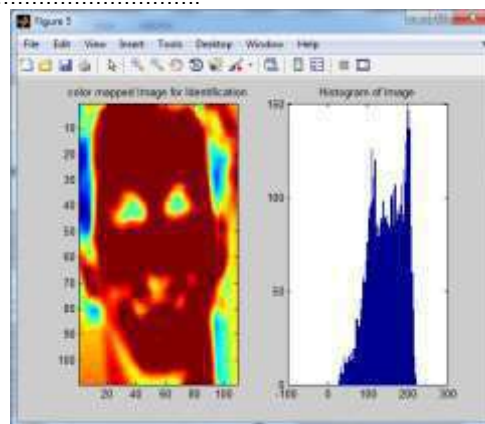


Fig: 5. Color Mapping and Image Histogram Estimation



Fig: 6. Face Segmentation



Fig:7. Equivalent Image Estimation and Extraction

## DISCUSSION AND CONCLUSION

Aggravated by the difficulty of inaccessible face appreciation, we tackled the difficulty be familiar within distinct and inadequately illumine faces using low dimensional linear model. The position of all representations get hold of by vague impression, agreed demonstration is a rounded position agreed by the rounded position of changed descriptions of the picture. Supported on this position descriptive categorization, we planned a vague impression vigorous face appreciation algorithm DRBF. In this technique we can with no trouble include preceding acquaintance on the category of haze as restrictions. By means of near to the ground dimensional linear sub-space representation for enlightenment, we illustrated that the position of all pictures acquired from the agreed picture by hazing and altering its enlightenment circumstances is a bunch of a gain, stands on this set theoretic classification, we projected a haze as well as enlightenment strong technique IRBF. We also established the effectiveness of our methodologies in undertakes the difficulty of countenance gratitude in unrestrained surroundings. Our technique is supported on a generative replica by adjacent fellow categorization flanked by the inquiry representation as well as the colonnade space, which creates it hard to balance it to real-life datasets with numerous amounts of pictures. Therefore we consider that picture integrating with a discriminative acquaintance having a supported move towards SVM into this manipulation would be extremely hopeful bearing for potential effort in the future.

### CONFLICT OF INTEREST

Authors declare no conflict of interest.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

## REFERENCES

- [1] W Zhao, R Chellappa, PJ Phillips, A Rosenfeld.[ 2003]Face recognition: A literature survey, *ACM Comput. Surv.*, 35(4): 399–458.
- [2] J Ni, R Chellappa.[2010]Evaluation of state-of-the-art algorithms for remote face recognition, in *Proc. IEEE 17th Int. Conf Image Process. Sep.*, pp. 1581–1584.
- [3] M Nishiyama, A Hadid, H Takeshima, J Shotton, T.Kozakaya, O Yamaguchi.[ 2011]Facial deblur inference using subspace analysis for recognition of blurred faces,*IEEE Trans Pattern Anal. Mach. Intell.*, 33(4): 838–845.
- [4] D Kundur, D Hatzinakos, Blind image deconvolution revisited.
- [5] A Levin, Y Weiss, F Durand, WT Freeman. [2011] Understanding blind deconvolution algorithms, *IEEE Trans Pattern Anal Mach Intell.* 33(12): 2354–2367
- [6] T Ojala, M Pietikäinen, T Mäenpää. [2002] Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 24 (7): 971–987.
- [7] R. Basri and D. W. Jacobs.[ . 2003] Lambertian reflectance and linear subspaces,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 25( 2):218–233
- [8] R. Ramamoorthi and P. Hanrahan. [2004] A signal-processing framework for reflection,” *ACM Trans. Graph.*, 23(4):. 1004–1042.
- [9] K.-C. Lee, J Ho, DJ Kriegman.[2005]Acquiring linear subspaces for face recognition under variable lighting, *IEEE Trans. Pattern Anal. Mach. Intell.* 27(5): 684–698
- [10] H Hu, G De Haan.[ 2007] Adaptive image restoration based on local robust blur estimation, in *Proc. Int. Conf. Adv. Concep.Intell. Vis. Syst.* pp. 461–472.
- [11] WH Richardson,. [1972] Bayesian-based iterative method of image restoration, *J Opt Soc Amer.*62: 55–59.
- [12] A Levin.. “Blind motion deblurring using image statistics,” in *Proc. Adv. Neural Inform. Process. Syst. Conf.*, pp. 841–848.
- [13] R Fergus, B Singh, A Hertzmann, ST Roweis, WT Freeman, [2006]Removing camera shake from a single photograph, in *Pro. ACM SIGGRAPH Conf.* pp. 787–794.
- [14] T Ahonen, E Rahtu, V Ojansivu, J Heikkilä, [2008] Recognition of blurred faces using local phase quantization,” in *Proc. 19th Int. Conf. Pattern Recognit.*, pp. 1–4.
- [15] R Gopalan, S. Taheri, PK Turaga, R Chellappa. [2012]A blur-robust descriptor with applications to face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 34 6): 1220–1222
- [16] V Ojansivu, J Heikkilä,[2008]Blur insensitive texture classification using local phase quantization,” in *Proc. 3rd Int. Conf. Image Signal Process.* pp. 236–243.
- [17] AC Bovic, *The essential guide to image processing*,” Elsevier, 2009.
- [18] H Hu , G de Haan. [ 2006] Low cost robust blur estimator, in *International Conference on Image Processing*.
- [19] WH Richardson. [1972]Bayesian-based iterative method of image restoration.
- [20] A Levin.[ 2006] Blind motion deblurring using image statistics, in *In Advances in Neural Information Processing Systems (NIPS)*.
- [21] R Fergus, B Singh, A Hertzmann, ST Roweis, WT Freeman.[ 2006] Removing camera shake from a single photograph, *ACM Trans. Graph.*, 25.
- [22] Q Shan, J Jia, A Agarwala.[ 2008] High-quality motion deblurring from a single image, *ACM Transactions on Graphics (SIGGRAPH)*,

- [23] C Likas, N Galatsanos,. [2004] A variational approach for bayesian blind image deconvolution, IEEE Transactions on Signal Processing, 52(8): 2222–2233.
- [24] J Jia. [2007] Single image motion deblurring using transparency, in IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.
- [25] J Deng, AC Berg, K Li, F-F. Li. [2010]What does classifying more than 10 000 image categories tell us? in Proc. Eur. Conf. Comput. Vis, pp. 71–84.

# A REVIEW ON ATTRIBUTE BASED ACCESS CONTROL SCHEME IN CLOUD ENVIRONMENT

S. Divya\*, B. Ananthi, V. Shanmugavalli

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Namakkal, T.N, INDIA

## ABSTRACT

**Aim:** Data security is main concern in cloud computing for protecting the confidentiality of the stored data. Nowadays Computer Science technologies have pulled more and more people to store and share their sensitive data on third party servers. To share the personal data on third party servers, it is essential to obtain an efficient encryption system. Ciphertext-policy attribute-based encryption (CP-ABE) is a promising cryptographic algorithm, where the encryptor can decide the access structure that will be used to protect the sensitive data. Here the survey is made in order to re-examine the attribute-based data sharing proposals. In future an improved Decisional Bilinear Diffie Hellman key exchange protocol has been proposed in CP-ABE scheme. It can guarantee that either key authority or cloud service provider can compromise the entire secret key of a user separately. It also enables dynamic modification of access policies and supports efficient on-demand user/attribute revocation. **Conclusion:** To overcome the drawbacks of ABE scheme in future the attributes can be constructed with weight which also reduces the complexity of access policy, so that the storage cost of cipher text and time cost in encryption can be saved.

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Cloud Security, Secure data sharing, ABE, KP-ABE, CP-ABE, HIBE, and Removing Escrow.

\*Corresponding author: Email: [divyasnk@gmail.com](mailto:divyasnk@gmail.com); Tel.: +91 9486029378

## INTRODUCTION

Cloud computing is a promising a computing paradigm which recently has drawn from both the academic and industry. By combining a set of a different techniques from research areas such as Service Oriented Architecture (SOA) and virtualization, cloud computing has become a widely adopted paradigm for delivering services over the internet. Cloud computing provides services according to three primary service models: Infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS). The five characteristics of cloud computing are: on demand service, self service, location independent, rapid elasticity and measured scale service

### Data Security

With the rapid growth of sensitive information on cloud, security is getting more important than even before. Most of the organization and institute utilized of this characteristics of the cloud computing and take benefit to gain profit. Hence, industries are shifting their businesses towards cloud computing. Cloud computing uses increased day by day, however, Data security is main concern in cloud computing. For protecting the confidentiality of the stored data, the data must be encrypted before uploading to the cloud by using some cryptographic algorithms. A data owner [13] [15] (DO) is usually willing to store large amounts of data in cloud for saving the cost on local data management. Without any data protection mechanism, cloud service provider (CSP), however, can fully gain access to all data of the user. This brings a potential security risk to the user, since CSP may compromise the data for commercial benefits.

Accordingly, how to securely and efficiently share user data is one of the toughest challenges in the scenario of cloud computing [15],[12],[1].

### Attribute based Encryption

Public-Key encryption is a powerful mechanism for protecting the confidentiality of stored and transmitted information. Traditionally, encryption is viewed as a method for a user to share data to a targeted user or device. In 2008, Sahai and Waters [11] introduced fuzzy identity based encryption (IBE), which is the seminal work of attribute based encryption (ABE). Recently, much consideration has been attracted by a new public key primitive called Attribute-Based Encryption (ABE) [4]. ABE has significant advantage over the traditional PKC primitives as

it achieves flexible many-to-many encryption instead of many-to-one. ABE is envisioned as an important tool for addressing the problem of secure and fine-grained data sharing and access control.

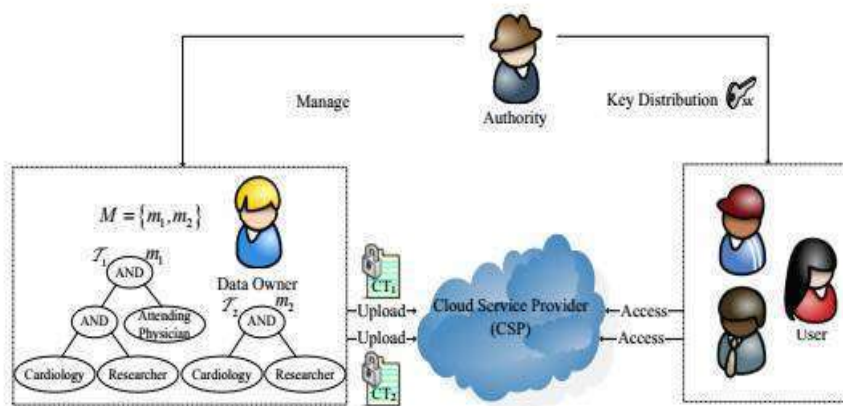


Fig. 1. An example of secure data sharing in cloud computing

In an ABE system, a user is identified by a set of attributes. For example, one can encrypt a recruitment related document to all recruitment committee members in the Computer Science Department. In this case the document would be encrypted to the attribute subset {"Faculty", "CS Dept.", "Recruitment Committee"}, and only users with all of these three attributes in the University can hold the corresponding private keys and thus decrypt the document, while others cannot. There are two variants of ABE: Key-Policy based ABE (KP-ABE) [12] and Ciphertext-Policy based ABE (CP-ABE) [13].

In KP-ABE, the ciphertext is associated with a set of attributes and the secret key is associated with the access policy. The encryptor defines the set of descriptive attributes necessary to decrypt the ciphertext. The trusted authority, who generates user's secret key, defines the combination of attributes for which the secret key can be used. In CP-ABE, the idea is reversed: now the ciphertext is associated with the access policy and the encrypting party determines the policy under which the data can be decrypted, while the secret key is associated with a set of attributes. DO (Data Owner) is allowed to define access structure over the universe of attributes. A user can decrypt a given ciphertext only if his/her attribute set matches the access structure over the ciphertext. A CP-ABE system can be used into a cloud application may cause some open problems. Firstly, all users' secret keys need to be issued by a fully trusted key authority (KA). This brings a security risk that is known as key escrow problem. By knowing the secret key of a system user, the KA can decrypt the entire user's ciphertexts, which stands in total against to the will of the user. Secondly, the expressiveness of attribute set is another concern. As far as we know, most of the existing CP-ABE schemes [1], [4], [19], [12], [15].

## LITERATURE SURVEY

Since all the data is transferred using Internet, data security is of major concern in the cloud. They shift their data from server to service based technology brought a significant change in computing technology. However these developments have created new security vulnerability. There are several security mechanisms are available. Sahai and Waters [12] proposed fuzzy Identity-Based Encryption (IBE) in 2008, which was the prototype of ABE. Latterly, a variant of ABE named KP-ABE, CP-ABE [11], [12], [16], was proposed. The following literature reviews attempts to demonstrate the different ABE schemes to provide a data security.

In [1] authors K. Liang, J. K. Liu, D. S. Wong, and W. Susilo in this paper "Cipher text policy Attribute based Encryption with anonymous access policy" described a scheme for constructing a Ciphertext Policy Attribute based Encryption with hidden access policy and provide security under the Decisional Diffie-Hellman assumption. In this scheme access policy can be expressed using AND, OR Boolean operators, so that it is possible to express the access policy effectively. The access policy can be represented in an n-ary tree, the leaf nodes represents the attribute present in the access policy, interior nodes represents the AND, OR operators. Each attribute in the leaf node can take multiple values. The value assigned for the leaf node by the secret sharing method will be distributed to these multiple values. The disadvantage of this scheme is key escrow problem.



In [2] authors S. Roy, M. Chuah in this paper “Secure Data Retrieval Based on Ciphertext Policy Attribute-Based Encryption (CP-ABE) System for the DTNs” proposed an access control scheme which is based on the Ciphertext Policy Attributed-Based Encryption (CP-ABE) approach. That provides a flexible fine-grained access control such that the encrypted contents can only be accessed by authorized users. The advantage of this scheme is that the incorporation of dynamic attributes value may change over time, and can be revoked its feature. This methodology makes use of groups with efficiently computable bilinear maps, and it is the key to our security proof, which gives a generic bilinear group model. The demerits of using this scheme are high computational overhead.

In [3] authors Tsz Hon Yuen, Joseph K. Liu, Man Ho Au, Xinyi Huang, Willy Susilo, and Jianying Zhou in this paper “k-times Attribute-Based Anonymous Access Control for Cloud computing” explained a new notion called k-times attribute-based anonymous access control. This provides a k-times limit for anonymous access control. That is, the server may limit a particular set of users to access the system for maximum k-times within a period or an event. Further the additional access will be denied. It provides an option for service provider to make it linkable or unlinkable. Yet even if it is unlinkable, the service provider knows whether the user has exceeded the k-times access limit. The security of this system is instantiated by using q- Decisional Bilinear Diffie-Hellman Inversion (DBDHI). The disadvantages of the paper is the user revocation is not possible. The advantage of this paper is limit access control.

In [4] authors Chun-I Fan, Vincent Shi-Ming Huang, and He-Ming Ruan in this paper “Arbitrary-State Attribute-Based Encryption with Dynamic Membership” proposed an ABE scheme that aims at dynamic membership management with arbitrary states, not only the binary states for every attribute. That also keeps high flexibility of the constraints on attributes and makes users to dynamically join, leave, and update their attributes. It is not necessary for those users who do not change their attribute statuses to renew their private keys when some user updates the values of her/his attributes. It consists of six algorithms which are Setup, Enrollment, Leaving, Updating, Encryption, and Decryption. The disadvantage of this scheme is it takes more time to process a retrieval of documents.

In [5] authors Fuchun Guo, Yi Mu, Willy Susilo, Duncan S. Wong, and Vijay Varadharajan in this paper “CP-ABE With Constant-Size Keys for Lightweight Devices” proposed a novel CP-ABE scheme with constant-size decryption keys independent of the number of attributes. Normally CP-ABE schemes suffer from the issue of having long decryption keys, in which the size is linear to and dependent on the number of attributes. This drawback was prevented by the use of lightweight devices in practice as storage of the decryption keys of the CP-ABE for users. The size of the device is 672 bits. The authority generates decryption keys of users and stores them in an RFID tag, which is embedded within a user’s ID card.

Since the key size is constant and small, the user can extract the key from his/her ID card for a security purpose. The advantages of this CP-ABE scheme allow all applications with key storage in lightweight devices. The main disadvantage of this scheme is when the key size is exceeded more than 672 bits, the cost of the device will get increased.

In [6] authors A. Balu, K. Kuppusamy in this paper “An expressive and provably secure Ciphertext-Policy Attribute-Based Encryption” proposed a new type of Ciphertext-Policy Attribute-Based Encryption based on linear integer secret sharing scheme. That scheme is very expressive and provably secure under the Decisional Bilinear Diffie-Hellman assumption. So the encryptor can specify the access policy in terms of LISS matrix  $M$ , over the attributes in the system. LISS focus on the advantages of secret sharing over integers opposed to secret sharing over finite groups or fields in LSSS. In LISS, the cost of the secret sharing is less.

In [7] authors Aparna C Bhadrans, and Maria Joy in this paper “Enhanced Large Universe Ciphertext Policy Attribute Based Encryption” proposed the traceability and blocking properties to the large universe based on CP-ABE scheme. Traceability property traces the malicious user who tries to access the encrypted data without proper decryption key and pin code. Blocking property blocks the illegal user who has been traced as a malicious user. Large universe property supports a flexible number of attributes to the system. The decryption key and pin code which are needed to decrypt the data are sent to the receiver via email. When the receiver gives the invalid pin code that receiver will be blocked temporarily. Admin can activate the blocked user if wanted for one time. But if the user again gives invalid pin code the corresponding user is permanently blocked. MiM attack is possible in this scheme. This is the main disadvantage of this scheme.

In [8] authors Shulan Wang, Junwei Zhou, Joseph K. Liu, Jianping Yu, Jianyong Chen, and Weixin Xie in this paper “An Efficient File Hierarchy Attribute-Based Encryption Scheme in Cloud computing” proposed an efficient file hierarchy attribute-based encryption scheme. The shared data files generally have the characteristic of multilevel hierarchy, particularly in the area of healthcare and the military. The layered access structures are integrated into a single access structure, and then the hierarchical files are encrypted with the integrated access structure. The ciphertext components related to attributes could be shared by the files. Therefore, both ciphertext storage and time cost of encryption are saved. The demerits of this scheme is when more files are stored in hierarchal way there occurs a high computational overhead.

In [9] authors Zhihua Xia , Liangao Zhang , and Dandan Liu in this paper “Attribute-Based Access Control Scheme with Efficient Revocation in Cloud Computing” introduced the access controller and designs an escrow-free key generation protocol between the attribute authority and the access controller to generate user’s secret keys in order to remove the key escrow problem. An efficient attribute revocation mechanism is presented using the version key. They adapt Hur’s secure key generation protocol to construct our escrow-free key generation protocol. This scheme consists of five phases: system initialization, key generation, data encryption, data access, and attribute revocation. The advantage of this scheme is the security has been proven by the random oracle model and the user revocation mechanisms.

In [10] authors Chaudhari Swapnil and Mandre in this paper “Secure Data Retrieval based on Attribute-based Encryption in Cloud” described the Hierarchical attribute base encryption scheme implementation on cloudsim tool. In this the Rijndael algorithm are used to encrypt the data. To perform encryption the string value of attribute and data file are converted into a bytes. Rijndael algorithm performs encryption on byte arrays. The issue addressed by Hierarchical CP-ABE scheme is time require for encryption and decryption overhead and reduce the generation of complex key. The advantages of this proposed work provides easy and simple to and understandable key structure. The disadvantage of this scheme is, it takes more time to convert a string value to bytes for large files.

### COMPARATIVE ANALYSIS OF DIFFERENT ABE SCHEMES

This section presents the comparison of different Attribute Based Encryption schemes such as HABE, FP-ABE and CP-ABE. This also gives the advantage and disadvantage of different techniques based on the algorithm.

**Table: I. Comparison of different ABE schemes**

Papers	Techniques	Algorithm	Advantage	Disadvantage
[1]	CP-ABE	Decisional Diffie Hellman(DDH)	Access policy can be expressed using AND, OR Boolean operators	Key escrow problem
[2]	CP-ABE	Decisional Diffie Hellman(DDH)	The incorporation of dynamic attributes whose value may change over time, and the revocation feature.	High Computational overhead
[3]	CP-ABE	Decisional Bilinear Diffie-Hellman Inversion (DBDHI)	It provides a k-times limit for anonymous access control	User revocation is not possible
[4]	CP-ABE	Decisional Bilinear Diffie-Hellman (DBDH)	It makes users are able to dynamically join, leave, and update their attributes	Occurs a collision
[5]	CP-ABE	Decisional Diffie Hellman(DDH)	It allows all applications with key storage in lightweight devices.	It takes a more cost

[6]	CP-ABE	Decisional Bilinear Diffie-Hellman (DBDH)	Secret sharing over integers opposed to secret sharing over finite groups	High computational overhead
[7]	CP-ABE	Not Mentioned	Traceability and Blocking property	Man in the middle attack is possible
[8]	FH-ABE	Decisional Bilinear Diffie-Hellman (DBDH)	Ciphertext storage and time cost of encryption are saved.	Average Computational overhead
[9]	CP-ABE	Hur's secure key generation protocol	Efficient revocation mechanism are presented	Key escrow problem
[10]	HABE	Rijndael algorithm	Which provides easy and simple to and understandable key structure	It's getting more time to convert a string value to bytes for large files

From the above survived we got some information on the performance achieved by the different ABE schemes. [Figure- 2] shows the storage cost of ciphertext with fixed attribute N=50. [Figure- 3] displays measurements of key generation time, encryption time, and decryption time on a range of different attribute size.

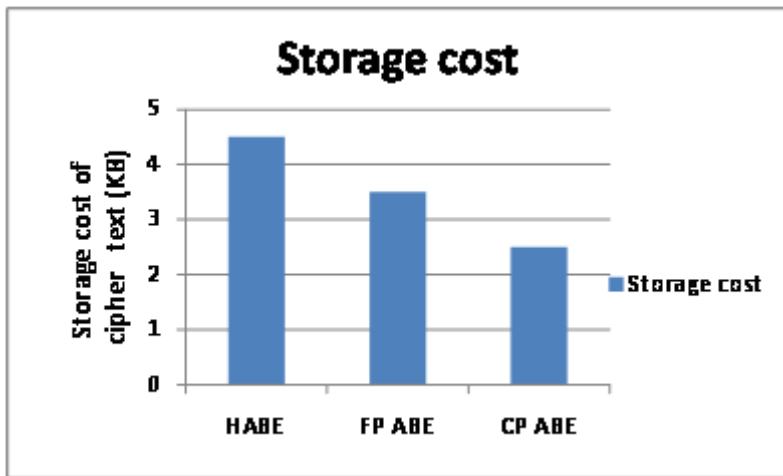


Fig: 2. The storage cost ciphertext to different techniques

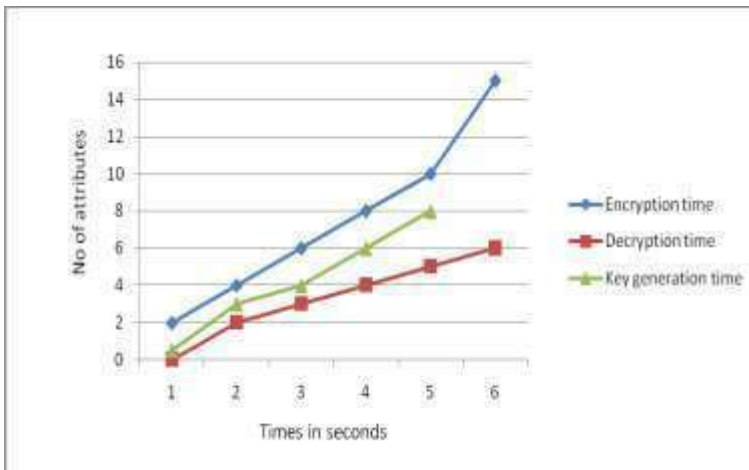


Fig. 3. The Encryption, Decryption, and Key generation time for existing CP-ABE method

## POSSIBLE SOLUTION

Several security issues such as scalability in key management, flexible access and efficient user revocation that has been the most important challenges toward achieving fine-grained data access control. From the above comparison the storage cost of CP-ABE is 60%. And the encryption, decryption and key generation time is 70%. So for improving the efficiency of the above CP-ABE scheme we propose a ciphertext-policy weighted ABE scheme with removing escrow (CP-WABE) by using improved Decisional Bilinear Diffie Hellman key exchange protocol. By using this protocol the storage cost of ciphertext and time cost in encryption can be saved. It also enables dynamic modification of access policies and supports efficient on-demand user/attribute revocation.

## CONCLUSION

Hence the survey is made up on a various attribute based access control schemes such as ABE, CP-ABE, HABE and KPABE. Where the ABE scheme provides more security but does not provide more resistance and key escrow problem. To overcome the drawbacks of ABE scheme in future the attributes can be constructed with weight which also reduces the complexity of access policy, so that the storage cost of cipher text and time cost in encryption can be saved. In order to improve the efficiency of encryption we can use an improved Decisional Bilinear Key Exchange protocol. It also enables dynamic modification of access policies and supports efficient on-demand user/attribute revocation.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] A Balu, K.Kuppusamy.[2010] Ciphertext policy Attribute based Encryption with anonymous access policy, International Journal of peer-to-peer networks (IJP2P) 1(1)October.
- [2] S Roy, M Chuah. "Secure Data Retrieval Based on Ciphertext Policy Attribute-Based Encryption (CP-ABE) System for the DTNs" Lehigh University Bethlehem, PA, USA – 18015.

- [3] Tsz Hon Yuen, Joseph K. Liu, Man Ho Au, Xinyi Huang, Willy Susilo, Jianying Zhou. "k-times Attribute-Based Anonymous Access Control for Cloud Computing" *IEEE Transactions on Computers*.
- [4] Chun-I Fan, Vincent Shi-Ming Huang, and He-Ming Ruan. [2014] "Arbitrary-State Attribute-Based Encryption with Dynamic Membership" *IEEE Transactions on computers*, 63(8).
- [5] Fuchun Guo, Yi Mu, Willy Susilo, Duncan S. Wong, and Vijay Varadharajan, "CP-ABE With Constant-Size Keys for Lightweight Devices" *IEEE Transactions on information forensics and security*, 9( 5)May 2014.
- [6] A. Balu, K. Kuppusamy. [2014] "An expressive and provably secure Ciphertext-Policy Attribute-Based Encryption" *Information Sciences* 276:354–362.
- [7] Aparna C Bhadrans, Maria Joy. [2016] "Enhanced Large Universe Ciphertext Policy Attribute Based Encryption" *International Journal of Advanced Research in Computer Science and Software Engineering*, 6(1), ISSN: 2277 128X.
- [8] Shulan Wang, Junwei Zhou, Joseph K. Liu, Jianping Yu, Jianyong Chen, Weixin Xie. [2015] "An Efficient File Hierarchy Attribute-Based Encryption Scheme in Cloud Computing" *IEEE Transactions on Information Forensics and Security* 1556-6013
- [9] Zhihua Xia, Liangao Zhang, Dandan Liu. [2016] "Attribute-Based Access Control Scheme with Efficient Revocation in Cloud Computing" *China Communications*
- [10] Chaudhari Swapnil, [2016] "Secure Data Retrieval based on Attribute-based Encryption in Cloud" *International Journal of Computer Applications* (0975 – 8887) 134 (13)
- [11] A Sahai and B. Waters. [2009] "Fuzzy identity-based encryption" *Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 457–473.
- [12] V Goyal, O Pandey, A Sahai, B Waters. [2008] "Attribute-based encryption for fine-grained access control of encrypted data" *Proceedings of the 13th ACM conference on Computer and communications security*, pages 89–98.
- [13] J Hur. [2013] "Improving security and efficiency in attribute-based data sharing," *IEEE Transactions on Knowledge and Data Engineering*, 25(10): 2271-2282
- [14] J. Bethencourt, A. Sahai, B. Waters, "Ciphertext policy attribute based encryption" *IEEE Symposium on Security and privacy*, 2007, pp. 321–334.
- [15] K. Liang, JK Liu, DS Wong, W Susilo. [2014] "An efficient cloud based revocable identity-based proxy re-encryption scheme for public clouds data sharing." *Proceedings of the 19th European Symposium on Research in Computer Security*, pages 257–272.
- [16] K Yang, X Jia, K Ren et al., "Dac-macs: Effective data access control for multi-authority cloud storage systems," pp. 2895-2903.

# A STUDY ON SECURITY ISSUES IN HEALTHCARE APPLICATIONS USING MEDICAL WIRELESS SENSOR NETWORK AND IOT

E. Sowmiya\*, L. Malathi, A. Thamarai selvi

Dept. of Computer Science and Engineering, Vivekanandha college of Engineering for Women, Elayampalayam, Tiruchengode-637205, Tamilnadu, INDIA

## ABSTRACT

**Aim:** Now a day's, due to the change in the human food habits, environmental changes, and the industrial consequences the cause of the chronic disease was increasing rapidly. People want to check and monitor their health data regularly. It is a big annoyance and sometimes not possible to check their health condition by spending their time in hospitals. Similarly, the doctors also need to take care of their In-Patients (IP) and Out-Patients (OP) data. In order to reduce these, many healthcare smart devices were found and used by the people. The integration of these smart devices leads to Internet of Things (IoT). Already some secure IoT- based solutions were suggested by the researchers where the smart devices hold by the patients will deliver the patient health condition to the remote doctor securely by using the Internet. This study analyzes that system in terms of integration and security analysis. **Result:** The performance parameters like rate of malicious node detection and number of successful packets transfer were analyzed. Here IoT application is used to fields and it is very success full. Finally, the security loopholes of the existing system and proposed solution for that security threat were presented.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Healthcare, MWSN, Security, Authentication, DoS attacks, Internet of things, chronic disease, RFID

\* Email: [eswaramoorthysowmiya@gmail.com](mailto:eswaramoorthysowmiya@gmail.com); Tel.: +91 9486088962

## INTRODUCTION

In the present era, the IoT (Internet of things) plays an important role in almost add fields importantly Home automation, Agriculture, Environmental monitoring and especially in medical fields. For example, IoT has lively capabilities to connect D2M (Device-to-Machine), O2O (Object-to-Object), P2D (Patient-to-Doctor), P2M (Patient-to-Machine), D2M (Doctor-to-Machine), S2M (Sensor-to-Mobile), M2H (Mobile-to-Human), T2R (Tag-to-Reader). This astutely connects humans, machines, smart devices, and dynamic systems and other devices to reduce the human workload [12].

In the area of healthcare is not only used in medical alliance and hospital, but also reachable by persons who are not actually present in the hospital. This is possible by the in-home patient monitoring system which gives good results with more efficiency in terms of healthcare. This in-home patient monitoring system has a high demand for age population and the ripened people. As the age population is more prone to chronic diseases and is in need of an effective in-home health monitoring system, recently, a Wireless Body Area Network (WBAN) with wearable medical sensors was developed [11]

**IoT in Healthcare:** Prevailing works in communication technologies have motivated the development of telemedicine to a great amount. Telemedicine reimbursement not only helps the customers who are able to receive health care more efficiently but also benefits the doctors to refashion their efforts to take care of more patients. To a great amount of computer- based monitoring has become an economical stipulation requiring deployment on a vast scale in various intensive care units around the world. **Figure-1** shows the patient monitoring. The system design shows that the sensors of two types one is wearable and another one is an in-planted sensor using any of this sensor the data about the patient is sensed and send it to the patient's mobile phone through Bluetooth from there it will send to the database from there it will send to the web server. The doctor can view the updates of patients detail by a web server. And else the doctor will suggest the treatments by observing the regular updates of patients through mobile or other devices.



Fig. 1: The System design

Without using PC and to support the patient monitoring in outdoor environment GPS enabled mobile phone is used in the proposed system. Security is the major issue when storing Personal Health Records (PHR) because PHR contains responsive data and they have to be securely stored and accessed [12]. If it is not possible to obtain the authentic and correct data by the medical experts, it leads to wrong and ineffective treatment and it turns to fatal predicament [13].

**Security Threats in Health care applications:** Security violates in healthcare applications of sensory networks is a major concern. It is also worth to mention that since healthcare applications of sensory networks are almost similar to WSN application environment, most of the security issues. Some of them are Malicious Node in Routing, Data modification, Impersonation attack, Eavesdropping, and Replaying.

## PROBLEM DEFINITION

Up to now, we have talked about different methods and techniques used in health Monitoring system all the above discussion is about the data transferring using different techniques such as ZigBee, cloud storage etc., in this only few authors, discussed the securities and authentication. Each and every node has to cross more levels to reach the destination. In between it needs to cross different access points, base station; gateways etc., in that time there will be a chance of occurring the attacks by the hacker. Because in healthcare system the data are very sensitive even one wrong information can lead to the tedious end. There are different types of Denial of services are there to attack the data. The most of the attacks are occurs while routing. In that, they discussed about different attacks such as black hole and selective forwarding Even though it has a solution with the minimum of 86% accuracy of detecting the malicious node .It also has some drawbacks while distributing the sequence number to access point and there is no trustworthiness in the routing algorithm. In the existing work, they provide the security up to the central data storage. To overcome this drawbacks the solution were found in the proposed system.

## PROPOSED SYSTEM

In the existing work some drawbacks has been raised to overcome those drawbacks the former solution has been found by changing the pre-deployment and routing phase algorithms, in the existing it focused up to the central storage. In this proposed work the security and authentication is provided up to the other end receiver (Doctor) which means providing security and authentication from the Patient to doctor (P 2 D).In this the different modules has been discussed in detail.

## MODULES

Our project, Health care application using medical wireless sensor network and IoT is made up of four modules. They are:

- Pre-deployment
- Clustering
- Secure Routing
- Secure access by doctors

## MODULE DESCRIPTION

### Pre-deployment

In this module, during this module every access point (AP) should be in-range with its base station (BS) which make easy for the distribution of exclusive random numbers from the BS to APs. Where the final use of the exclusive node ID as an initial seed to create a final seed. After the creation of numbers, the base station distributes them using unicast messages. That message distributed only to the access points to its own hardware devices. That avoids the malicious node which takes place in the beginning.

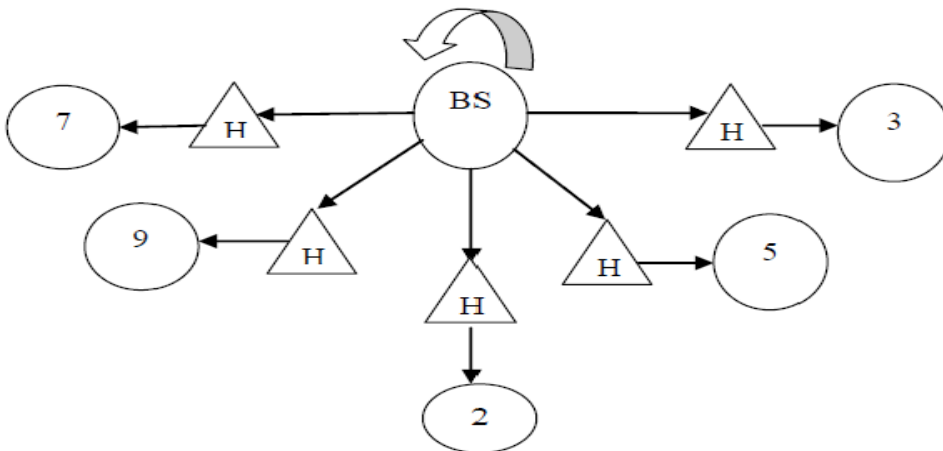


Fig. 1: The Pre-deployment unique code transmissions to its own hardware device

### Clustering

This module It is not feasible for all nodes to transmit data independently to the nearest access point (AP). This is because packet collisions and battery exhaustion leads to data loss. One solution involves the sensor nodes to electing the cluster head (CH) among themselves by using a weight computation procedure after the unbeaten election of the cluster head the nodes transmit their data to the cluster head. Then it aggregates the encrypted data and transmits it to the nearest access point. This access point in turn routes the data to the base station using a customized version of mesh routing this prevents the data transmission from collision, traffic and data loss.

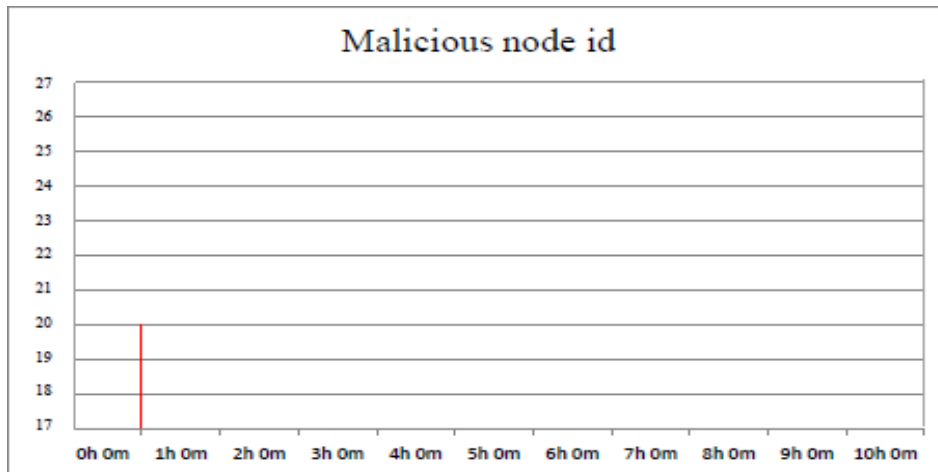
### Secure Routing

Routing is very important in any type of data transmission and most of the attacks occur during the routing phase. To avoid this, the trust worthy between each node should be included before the route reply and route request has been confirmed. If trustworthy is added to each node then the data transmission will be on trusted way so there is no chance of passing the information through the malicious node thereby the attack can avoid. If any un-trusted node has been identified in between the transmission it automatically select the another trusted node for the data transmission this helps to avoid the malicious act in the routing phase.



### Secure Access by Doctors

In the medical health care monitoring system the security and authentication is very important because it carries very sensitive data about the patient. That much security and authentication should be provided to the data which is going to be accessed by the doctor from the central storage. Only the authorized doctor should access the patient data from the central storage by using the password or by using other cryptography techniques.



Graph-1 Node 15 turns to malicious node in time 1000s

## LITERATURE SURVEY

### Introduction

In this work, the detailed study of the recent health care monitoring system has been studied and analyzed for the remote data delivery and security of the data in HMS. Some papers based on different technologies has been discussed their transmission of data and methods and others focused on the security and authentication of the remote data. From the different sensors which are placed in the patient body. In this different mechanisms has been analyzed briefly in the below sections.

### Defence against Black Hole and Selective Forwarding Attacks for Medical WSNs in the IoT

In this they proposed the solution for black hole and selective forwarding attack in PMS (Patient monitoring system) by introducing the pre-deployment and routing algorithm which is to find the malicious node in data transmission. This algorithm shows high accuracy compared to its existing work. Even though it gives high accuracy it has some draw backs in the pre-deployment and in routing phase because routing is important to transmit the data and attacks also frequently occur in the routing phase, so it needs more security in routing [1].

### Secure authentication for remote patient monitoring with wireless Medical sensor networks

In this work the protocol is based on the Rabin authentication algorithm, which is modified in this paper to improve its signature signing process, making it suitable for delay-sensitive medical sensor network applications. To advance the competence of the Rabin algorithm, they implemented the algorithm with different hardware settings and encoding of an FPGA to appraise its design and presentation. It is considered an individual security of data called of RSA. However, Rabin's system was faster and lighter than RSA. This gives it an estimable candidate for our distant patient monitoring system with medical sensor network. Compared to earlier hardware platforms, FPGA implementation provides better features, including hustle, lower cost, faster growth time and elasticity [2].

### Raspberry Pi Based patient health status observing method using internet of things

In this a temperature, respiration, patient's body actions and heartbeat analysis are monitoring using Raspberry Pi. These sensory signals send to the Raspberry Pi via amplifier circuit and to the signal conditioning unit (SCU), because its signal levels are low (gain), so amplifier circuit is used to boost up the signal and transmit the signals to the Raspberry Pi. The Raspberry pi is a Linux -based operating system it will work as a small PC processor system. The patient's body temperature, body movements, respiration and heart rate is measured using own sensors and it can be monitored on the monitor screen of a computer using Raspberry Pi and it can be monitor through anywhere in the world using internet source [3].

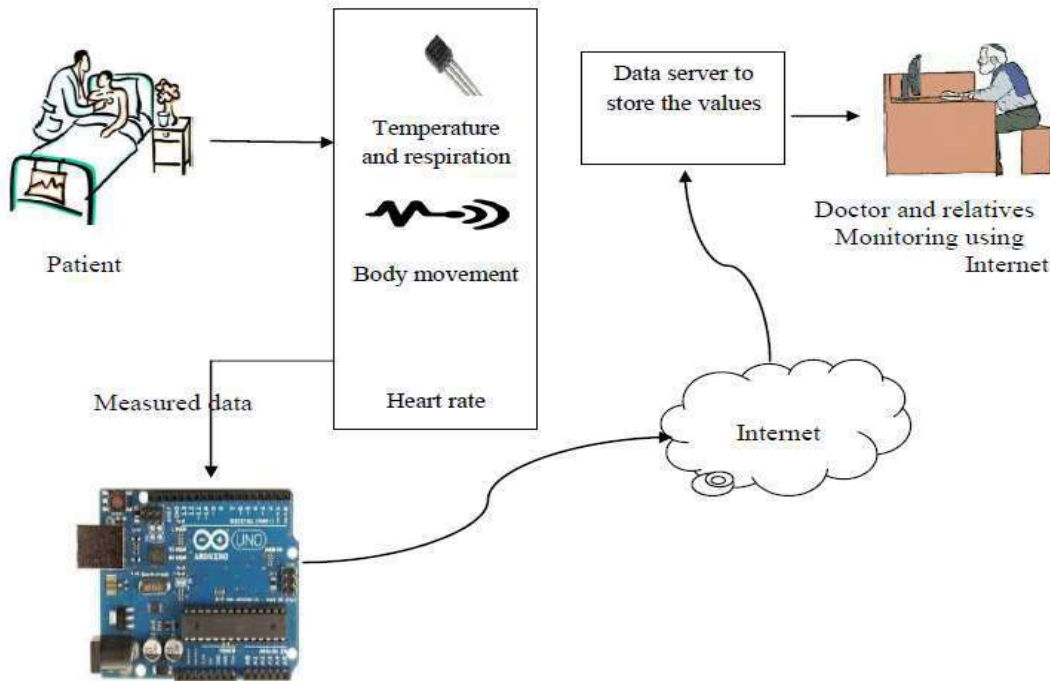


Fig.4. Architecture of Healthcare using Raspberry Pi

#### Secured smart healthcare monitoring system based on IoT

This project has PIC18F46K22 microcontroller used as a entry to converse the various sensors such as warmth sensor and pulse sensor which was placed in the human body .The microcontroller act as bridge to transfer the collected information and stored in the PIC18F46K22 by using the WiFi that data will transfer to the particular doctors web page from there the doctor can view the patients update regularly. And it also has additional feature such as alerting technique. When the sensor observes any emergency values beyond the normal level the buzzer alerts will be sent to the doctor and patients care taker. Even though it has some advantages it also has drawbacks because it only focused on in-patient monitoring system [4].

#### A healthcare monitoring system using wireless sensor network with GSM

This paper presents a monitoring system to monitor the physiological parameters such as BP, ECG, Body Heat and Respiration etc., the manager node has to attach to the body of patients for collecting a signal from wireless sensors. The wireless sensors send this signal message to base station or control room of physician. The wireless sensors from wireless body sensor network. The node of each WSN collected of healthcare sensors and RF trans-receiver it will send data back to end server. Sensors can be chosen in the range of WSNs, while RF trans-receiver is implemented as a manager which manages WSN other than forwarding the data. The sensory data of each patient are stored in the back-end of the server with each has its own ID. The data analysis, database inquiry, data manning and the system managing are the process in a web page of a server. The system can detect an abnormal condition of the patient and send it through the SMS or e-mail to the physician [5].

### Real time wireless health monitoring application using mobile Devices

This real time monitoring system uses the mobile phone as a device to display the updates to the doctor it uses the LabVIEW technique. The same process is used in this by sensing the data from the human body using the sensor and transmits its psychological information using the WiFi and to display its update by using the LabVIEW technique. In this it used the alerting system such as buzzer and email to the doctor and patients relatives who cared for the patient. This also focused only on the in-patient monitoring system [6].

### Patient monitoring system using MSP430 microcontroller

It is a tough job for the doctor and medical staff to monitor each patient for 24 hours. In this paper, it introduces a system in which it collects physiological signals of patient transfer this signals to a personal computer using Ethernet controller. So hear patient can be analyzed by the doctor or other paramedical team from central observation centre or from their PC. And also this bodily signal displayed on LCD screen [7].

### Security and privacy issues in wireless sensor networks for healthcare applications

There is a possibility of serious communal conflict due to the fright on such devices may be used for monitoring and tracking individuals by government agencies or other private organizations. In this paper, the discussion is about the issues and analyze in detail. The problems and their possible measures. All communications over wireless networks and internet are essential to be encrypted to protect the data from the attack. Some countries have added this type of section in their active legal acts or enacted new laws. Another important gauge is to create awareness in general public. It will be extremely beneficial if people are educated regarding security and privacy issues and there is an implication from now on. It is mentioned by authors in that common people do not understand the technology and therefore may not be in a position to make balanced judgments concerning the extent to which it will have a negative impact on their own values of privacy. Therefore educating a common people will greatly help in this regard [8].

### Wireless health care monitoring

Doctor need to monitor the patient's health condition regularly but it is tedious job to take care of regular update of both in-patient and out-patient. To make it easy the different sensors are placed in the patient's body by using the GSM module the sensed data was transmitted using the GSM. Now-a-days the use of GSM in different fields becomes very familiar. Based on that technique in healthcare monitoring system GSM plays a vital role by using that the data will be transmitted to the webpage and doctor can retrieve the stored information anywhere and suggest the treatment at any time [9].

### Detection of insider selective forwarding attack based on monitor node and trust mechanism in WSN

In this, the protected routing protocol is based on observing of different nodes and trust mechanism. The eminence value is made up of packet forwarding rate and node's lingering energy. So this exposure and routing device is worldwide because it can take account of both the safety and lifespan of a network. In this, they discussed different techniques such as watchdog, Trust Mechanism to identify unruly nodes in wireless ad hoc networks [10].

This proposal can be applied in WSN and it was the earliest trust mechanism which is the foundation of many defense methods. In their approach, each sensor node has its own watchdog that checks and records the behaviors by its one-hop neighbors. The watchdog system of each node provisions the routing table which is about the act good or bad of the neighbor nodes. In proposed work the OPNET simulation is to appraise the custom of our trust monitor model and routing system. In our simulation, 100 sensor nodes and 10 monitor nodes are erratically dispersed over a 100 m × 100 m area. We only do the imitation of the local network. A monitor node controls 10 common nodes in each region. Firstly, the nodes will be numbered to meet the design needs of the software.

## ANALYSIS

This section presents the comparison of different methods used and compared its advantage and disadvantages of on patient health monitoring system which we reviewed detail and based on this the solution will be found in the finest scheme of the patient health care monitoring system.

**Table:1. Comparative Analysis of Health Care Applications**

PAPER No.	METHODS USED	ADVANTAGE	DISADVANTAGE
1	- Cryptographic hashes -pre-deployment -Routing phase	- It shrink the further possibility of different attacks because it has different time intervals	-In pre-deployment there is no chance to find the malicious attack before the distribution of unique code. -Needs trust worthiness
2	-Rabin algorithm	-Rabin signature generation drastically reduced the delays	- Rabin scheme is only focused on the in-patient monitoring system and its performance was measured using with and without parallel execution. It doesn't have any standard execution.
3	-Raspberry Pi, putty Software, putty Software, apache server	-It reduces patient's money and time. -Using this technology development, patient's record their health status in Their own mobile phone and then store the data. -it will automatically send alert message to doctors and relatives mail.	-Only few Parameters have been measured.
4	-PIC microcontroller, ESP8266 Wi-Fi module, AES128	-It is capable with low power utilization capability, easy setup, high routine and time to time response.	-It won't connect all the sensed data easily through WiFi
5	-RF trans-receiver , GSM modem	- Decrease the power utilization	-Security issues are very high
6	-ZigBee, LabView	-Low cost, low power, easy functioning, trustworthy, and high sanctuary.	-Few parameters has been measured
7	-Ethernet controller	-Patient can be analyzed by doctors from central observation centre.	-For hospitalized patients
8	-WBAN, Privacy and security	-Improves the security by encrypting the data adequately	- All infrastructure over wireless networks and internet are required to be encrypted to protect the user's time alone.
9	-GSM, ZigBee.	-Low cost for monitoring	-Need more sensors -Vulnerable Security threats -Investment Cost is high
10	-Secure routing protocol, monitor node	-Improves the finding of malicious node in beneficial way	-Need to improve the trustworthy

**Table: 2. Performance Evaluation**

Technology and algorithm	Accuracy of remote data delivery	Malicious node detection rate
- Cryptographic hashes -pre-deployment -Routing phase[1]	86%	93%

-Rabin algorithm[2]	88%	92.5%
-Raspberry Pi, putty Software, putty Software, apache server [3]	92%	-
-PIC microcontroller, ESP8266 Wi-Fi module, AES128 [4]	90%	-
-RF trans-receiver, GSM modem [5]	93%	-
-ZigBee, LabView [6]	92.5%	88%
-Ethernet controller [7]	91%	-
-WBAN, Privacy and security [8]	89%	90%
-GSM, ZigBee [9]	90%	-
-Secure routing protocol, monitor node [10]	89%	93%

## CONCLUSION

The Importance of the automated health care application and the existing available health care applications are discussed in this article. Then the security threads imposed on those applications and their performances are compared for remote data delivering from patient to doctor database by using Sensor ,Bluetooth, Mobile phones ,Cloud storage etc., All this is to reduce the human workload .This type of techniques is used in different fields for regular monitoring and secure data delivery. In the Healthcare it plays the major role and it is very useful for the patients to manage their health conditions frequently with help of doctors being in the home. In future security and privacy issues while transmitting the remote data delivery can be protected using a cryptographic technique.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Avijit Mathur, Thomas Newe and Muzaffar Rao.[2016] Defence against Black Hole and Selective Forwarding Attacks for Medical WSNs in the IoT , Published on 19 January 2016 Received- 17 December 2015, Accepted- 14 January 2016, Academic Editor- Rongxing LuSensors 2016, 16:118, doi:10.3390/s16010118
- [2] Thayer Hayajneh, Bassam J Mohd, Muhammad Imran, Ghada Almashaqbeh, Athanasios V.[2016] Vasilakos Secure Authentication for Remote Patient Monitoring with Wireless Medical Sensor Networks” Published: 24 March 2016, Sensors 16, 424: doi:10.3390/s16040424
- [3] R Kumar, M Pallikonda Rajasekaran. Raspberry Pi Based Patient Health Status Observing Method Using Internet of Things” *International Conference on Current Research in Engineering Science and Technology (ICCREST-2016)* E-ISSN :2348 – 8549.
- [4] BK Bhoomika, Muralidhara K N. Secured Smart Healthcare Monitoring System Based on IoT’ *IJRITCC*, ISSN:-2321-8169 3(7):4958 – 4961.
- [5] L Sunil Rahane, S Ramesh Pawase, A Healthcare Monitoring System Using Wireless Sensor Network with GSM’ *IJAREEIE*, An ISO 3297:- 2007 Certified Organization 4(7)-July 2015 ISSN: 2320 – 3765:ISSN (Online)-2278 – 8875.
- [6] Ali Abou-ElNour ,Asma Ismael, Aisha Rashid, Amna Abdullah.[2015] Real Time Wireless Health Monitoring Application Using Mobile Devices’ *IJCNC- 7(3)* DOI - 10.5121/ijcnc.2015.7302.
- [7] Pradip S. Bhendwade, Ms. Sarika M. Patil,” Patient monitoring system using MSP430 microcontroller”, *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)* 4(1), January 2015,ISSN: 2278 – 909X All Rights Reserved © 2015 IJARECE
- [8] Kyungsup Kwak,Moshaddique Al Ameen ,’Security and Privacy Issues in Wireless Sensor Networks for Healthcare Applications’ Received- 18 December 2009 ,Accepted- 16 February 2010 Published online- 12 March 2010: *J Med Syst* (2012) 36:93–101 DOI 10.1007/s10916-010-9449-4
- [9] Bazila Banu A, Kavitha KC.[2014] Wireless Health Care Monitoring’ *IJRSET*, 3(SPL.3) ISSN (Online) :- 2319 – 8753 ;ISSN (Print) :- 2347 – 6710.
- [10] Hongshuai Wang ,Yu Hu, ‘ Detection of Insider Selective Forwarding Attack Based on Monitor Node and Trust

- Mechanism in WSN' WSN; 2014; 6; 237-248 Published Online November 2014 in SciRes.
- [11] M Li, W Lou. Data Security and Privacy in Wireless Body Area networks", IEEE Wireless Comm., Feb. 2010, pp. 51-58.
- [12] Luan Ibraimi, Muhammad Asim, Milan Petko Vic.[2010] Secure Management of Personal Health Records by Applying Attribute-Based Encryption, *IEEE* 2010.
- [13] N Jaisankar, R Saravanan, KD [2010]'A novel security approach for detecting black hole attack in manet';GaoI, F.L., Thankachan, N, Eds- Springer:- Berlin: Germany:.
- [14] An Internet of Things. Last access Mar. 14, 2015. Available on: <http://postscares.com/internet-of-things-examples/>.
- [15] Wassnaa AL-mawee. [2012]Privacy and Security Issues in IoT Healthcare Applications for the Disabled users. A Survey" Western Michigan University Scholar Works at WMU Master's Theses, Graduate College2-2012
- [16] P Thangaraj, K Geetha.[2015] FGT2- ABR: Fuzzy Game Theory Trust Associativity Based Routing to Mitigate Network Attacks in Pervasive Health Monitoring Systems, *Journal Of Pure And Applied Microbiology*, 9:161-168.
- [17] S Brand.[ 1994] What happens after they are built, in *How Buildings Learn*, London, U.K.: Penguin.

# A SURVEY ON MAPREDUCE USING FREQUENT SUBGRAPH MINING

M. Gokilavani\*, B. Anitha, R. Jayanthi

Dept. of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Elayampalayam, Tamilnadu, INDIA

## ABSTRACT

Graph mining approaches have become more popular, especially in certain domains such as bioinformatics, chemo informatics and social networks. In data mining applications, mining frequent sub graph from a large number of small graphs is an important task. For mining frequent sub graphs many algorithms have been proposed. To overcome this some distributed solution using Map Reduce is becoming important paradigm for computation on massive data. In experimented work, we investigate how to economically perform extraction of frequent sub graph over a large datasets using Map Reduce. However, as the real-world graph data grows in amount and mass, result could not be met. To overcome this, a few graph database-centric methods have been projected in real problem for solving FSM. however, a distributed solution using Map Reduce paradigm has not been explored extensively. Since Map Reduce is flattering the de-facto paradigm for computation on massive data, an efficient FSM algorithm on this paradigm is of huge demand. The efficiency of extracting the frequent subgraph is experimentally investigated over a large datasets using map reduce.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

FSM, database-centric method de-facto paradigm, mapreduce

\*Corresponding author: Email: [greenvani88@gmail.com](mailto:greenvani88@gmail.com); Tel.: +91 9442041624

## INTRODUCTION

The big data is a huge amount of data, when compared to other datasets. The large number of data is stored in memory, the stored data will be retrieved (wanted data) in that time we are facing the many number of problems. And also the problems resolved means, in problem solving operation using the data mining process. In data mining, large amount of data, the wanted data only extracted from big data, i.e. ,stored large amount of data. Some storage area only using the data mining concepts and also using the map reduce concepts. The map reduce concepts are mainly focused on the frequent set of items, the frequent path finding and then the map reducing process is executed. In map reduce process, having two type of functionalities. They are mapping and reducing functions used in map reduce process. In map reduce function of operation, the large amount of graph data will be divided into sub graphs, divided sub graphs are merged. For the use of finding the frequent set of path, time is reduced for retrieving the wanted set of data.

## BIGDATA

Big data is a big term for data sets so large or complex that conventional data dispensation applications are incompetent. In the bigdata Challenges are investigation, arrest, hunt, contribution, storage, relocate, apparition, and information isolation. The term often refers merely to the use of extrapolative analytics or other definite advanced methods to mine value from large dataset, and seldom to a particular size of data set. Correctness in big data may lead to more convinced judgment making. And enhanced decisions can mean greater operational competence, cost diminution and reduced risk. Analysis of data sets can find new correlations, to spot business trends, put a stop to diseases and so on. Scientists, business executives, practitioners of media and marketing and governments alike recurrently meet difficulties with large data sets in areas including Internet rummage around, economics and industry informatics. Work with big data is inevitably uncommon; most investigation is of personal computer size data, on a desktop personal computer or notebook that can switch the obtainable data set [13].

Big Data processing chiefly depends on equivalent programming models like MapReduce, as well as on condition that a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented similar computing model. There is still a convinced gap in presentation with relational databases. Civilizing the presentation of MapReduce and attractive the concurrent scenery of large-scale data processing have acknowledged a momentous amount of consideration with MapReduce similar programming being sensible to many machine knowledge and data mining techniques. Data mining techniques habitually need to scan through the instruction data for obtaining the statistics to solve or diminish model parameters. It calls for exhaustive computing to admittance the large-scale data habitually. To amplify the efficiency of algorithms, Chu et al. projected a universal-purpose parallel programming method, which is applicable to a more number of machine learning algorithms based on the usual MapReduce programming model on multi-core processors. Classical data mining algorithms are realized in the skeleton includes in the vicinity prejudiced linear regression, k-Means, naive Bayes, linear sustain vector machines, the self-regulating variable investigation, Gaussian discriminant study, anticipation maximization, and flipside-broadcasting neural networks.

FSM-H is distributed frequent subgraph mining method in excess of MapReduce. Given a graph database, and a fewest amount support threshold, FSM-H generates a unqualified set of frequent subgraphs[14]. To ensure completeness, it produces and retains all patterns in a partition that has a non-zero hold up in the map phase of the extracting, and then in the reduce phase, it decides whether a pattern is recurrent by aggregating its support computed in all partitions from dissimilar compute nodes. To triumph over the reliance among the states of a mining process, FSM-H runs in an repeated process fashion, where the output from the reducers of iteration  $i-1$  is used as an input for the mappers in the iteration  $i$ . The mappers of iteration  $i$  produce candidate subgraphs of size  $i$  (number of edge), and also calculate the local support of the candidate pattern. The reducers of iteration  $i$  then compute the true frequent subgraphs (of size  $i$ ) by aggregating their.

## FREQUENT SUBGRAPH MINING

The FSM is to haul out all the frequent subgraphs, in a given data set, whose incident counts are above a individual threshold. The immediately forward idea behind FSM is to “grow” candidate subgraphs, in either a breadth first or depth first manner (candidate production), and then institute if the documented candidate subgraphs crop up recurrently enough in the graph data set for them to be painstaking interesting (support counting) [15]. The two main research issues in FSM are thus how to efficiently and effectively (i) produce the candidate regular subgraphs and (ii) bring to a close the frequency of occurrence of the generated subgraphs.

Effective candidate sub graph production given that the generation of photocopy or superfluous candidates is to defeat. Occurrence counting given cyclical comparison of candidate subgraphs with subgraphs in the input data, a process known as isomorphism inspection. FSM, in many compliments, can be viewed as an extension of Frequent Itemset Mining (FIM) popularised in the context of group rule mining. Consequently, many of the proposed solutions to addressing the main research issues completing FSM are based on comparable techniques found in the sphere of FIM.

The existing system involves mining frequent sub-graphs from the given graph database (A database with ‘N’ graphs). Here MapReduce approach is used which is a programming model that enables disseminated computation over massive data. In extraordinary, Iterative MapReduce is used here which can be defined as a multi theatrical execution of map and reduce function pair in a cyclic fashion, i.e. the production of the period  $i$  diminish is used as an input of the phase  $i + 1$  mappers. An peripheral condition decides the standstill of the job. Graph isomorphism is also measured.

A elementary feature for truthful search based algorithms is that the taking out is absolute i.e. the mining algorithms are distinct to find all frequent subgraphs in the input data. Complete mining algorithms execute imaginatively only on bare graphs with a large amount of labels for vertexes and edges. Due to this completeness constraint these algorithms take upon physically extensive subgraph isomorphism assessment either overtly or absolutely substantial in a outstanding computational visual protuberance The usefulness of integrating constraint into the FSM process is liable by many aspects, including the belongings of the data and pruning cost. limitation based mining algorithms therefore necessitate to take into account the trade-off flanked by the pruning cost and any credible assistance[11].



- Before sending input graph data to nodes, they are not neutral For example, one node may be assigned with more big graphs and other node with small graphs.
- Nodes have to stay for Reduce phase and can start the process only after all the mapper processes are finished in all nodes.
- Overall time efficiency is poor.
- Candidate generation strategy poor.
- The mechanism for traversing the search space and low occurrence counting process.
- Single graph based FSM applied.
- Graph isomorphism is neither known to be solvable in polynomial time nor NP-complete.

## LITERATURE SURVEY

Penetrating plays an significant role in Map Reduce algorithm. It helps in the combiner phase (elective) and in the Reducer phase. Let us try to appreciate how penetrating works with the help of a lot of papers. The papers are includes, map reduce concepts and their operational details. Many process will based on map reduce operations, the following papers are surveyed.

In [1] authors Jeffrey Dean and Sanjay in the paper “Map reduce: Simplified Data Processing On Large Clusters” describe the Map Reduce and its a brainwashing model and an connected functioning for executing and producing large data sets. Users identify a map function that processes a key/value pair to produce a set of transitional key/value pairs, and a reduce function that joins all intermediate values connected with the identical intermediate key, Many real world tasks are expressible in this model. Programs written in this functioning style are automatically parallelized and executed on a huge cluster of examination machines. The run-time arrangement takes care of the particulars of separating the input data, scheduling the program’s implementation crosswise a set of machines, managing machine failures, and behaviour the obligatory inter-machine announcement.

In [2] authors Jie Tang, Jimeng Sun, Chi Wang and Zi Yang in the paper “ Social Influence Analysis in Large-scale Networks” illustrate the big public networks, nodes are prejudiced by others for an assortment of reasons. For example, the colleagues have brawny pressure on one’s work, while the friends have strong influence on one’s every day life. How to make a distinction the social influences from different angles (topics). How to enumerate the strength of those social influences and the model on real large networks and implement to address these fundamental questions, Topical Affinity Propagation (TAP) to model the topic-level social authority on hefty networks.

In [3] authors U Kang, Charalampos in this paper “Pegasus: A Peta-Scale Graph Mining System - Implementation and Observations” describe PEGASUS, an unlock foundation Peta Graph taking out library which performs attribute graph removal everyday jobs such as compute the diameter of the graph, manipulative the radius of every node and finding the associated mechanism. As the amount of graphs reaches quite a few Giga-, Tera- or Peta-bytes, the predictability for such a documentation grows too. To the brilliant of our knowledge, PEGASUS is the primary such documentation, implemented on the pinnacle of the HADOOP display place, the unbolt foundation version of MAPREDUCE. Numerous graph extracting operations (PageRank, ghostly clustering, diameter judgment, associated components etc.) are fundamentally a recurring matrix-vector multiplication.

In [4] authors Siddharth Suri Sergei in this paper “Counting Triangles and The Curse of The Last Reducer” describe the grouping coefficient of a node in a social network is a elementary estimate that quantifies how tightly-knit the community is in the region of the node. Its calculation can be reduced to counting the number of triangles incidence on the meticulous node in the network. In case the graph is too far above the ground to fit into storage, this is a non-trivial job, and foregoing researchers showed how to guesstimate the clustering coefficient in this circumstances. A different avenue of research is to carry out the computation in parallel, distribution it across many machines. In recent years Map Reduce has emerged as a de facto indoctrination paradigm for parallel totalling on huge data sets. The main sympathy of this work is to give Map Reduce algorithms for including triangles which we use to subtract cluster coefficients.

In [5] authors Rasmus pagh in this paper “Colorful Triangle Counting And A Map reduce implementation ”describe a new randomized algorithm for including triangles in graphs. The underneath gentle circumstances the

calculation of new randomized algorithm is strappingly concerted around the factual figure of triangles. Completely if  $p \geq \max(\log n / t, \log n / \sqrt{t})$ , where  $n, t, i$  denote the numeral of vertices  $G$ , the figure of triangles  $G$ , the most advanced digit of triangles an edge of  $G$  is embarrassed then for any invariable  $> 0$  our unbiased approximation  $T$  is determined around its prospect. Ultimately present a Map Reduce implementation of new randomized algorithm.

In [6] authors Foto N. Afrati in this paper “Enumerating Sub graph Instances Using Map-Reduce” describe to and find all instances of a given “sample” graph in a larger “data graph,” using a solitary about of map reduce. For the standard sample graph, the triangle, we augment upon the superior known such algorithm. Inspect the universal case, bearing in mind together the announcement cost sandwiched between mappers and reducers and the whole functioning cost at the reducers. To diminish statement cost, we take advantage of the techniques of for computing multiway joins (evaluating conjunctive query) in a solitary map-reduce surrounding. Each methods are shown for translating illustration graphs into a union of conjunctive queries with as hardly any queries as probable.

In [7] authors Bahman Bahmani in this paper “Densest Sub graph In Streaming And Map reduce” they are studied the difficulty of judgment dense sub graphs, a original primordial in pretty a few statistics administration applications, in streaming and Map Reduce, two computational models that are all the time more being adopted by significant data processing applications. A simple algorithm that make a small number of passes over the graph and obtains a  $(2+i)$  rough calculation to the densest sub graph. The obtained several extensions of this algorithm: for the case when the sub graph is prescribed to be more than a convinced size and when the graph is absorbed To the best of our knowledge, these are the sample algorithms for the densest sub graph problem that truly scale yet over demonstrable guarantees.

In [8] author Jun Huan in this paper “Mining Protein Family Specific Residue Packing Patterns From Protein Structure Graph” report on the submission of the recurrent sub graph withdrawal algorithm to protein structure correspond to as graphs. The aspire of this consideration was to be memorable with normal subgraphs widespread to each and every one (or the mainstream of) proteins belonging to the indistinguishable structural and well-designed family in the SCOP catalogue and explore these sub graphs as folks specific amino acid set down signature of the necessary family unit. even though protein graphs are talented this submission has be converted into possibly will or maynot, appreciation to every highly developed rationalization of the habitual subgraph withdrawal algorithm in employment in this paper.

In [9] authors Bay Vo and Bac Le in this paper “OO-FSG: An Object-Oriented Approach to Mine Frequent Subgraphs” present a fresh algorithm for withdrawal alliance rules. The progress the algorithm which scans database one time only and use Tidset to estimate the support of comprehensive item set faster. A tree structure called GIT-tree, an glasshouse of IT-tree, is developed to store database for extracting frequent item sets from hierarchical database. Fresh algorithm is often more rapidly than MMS\_Cumulate, an algorithm extracting frequent item sets in hierarchical database with more bare minimum supports, in tentative databases.

In [10] authors authors Srichandan B and R. Sunderraman in this paper “OO-FSG: An Object-Oriented Approach to Mine Frequent Subgraphs” describe a Frequent sub graph mining (FSG) and has all time been more popular issue in data mining. Several frequent subgraph removal methods have been superior for taking out graph data. However, most of these are chief memory algorithms in which scalability is a most well-liked issue. A hardly any algorithms have opted for a relational approach that provisions the graph data in relational table.

## COMPARITIVE ANALYSIS

S.NO	Paper Title	Algorithm	Advantages	Disadvantages
1	Map Reduce: Simplified Data Processing on Large Clusters	Map Reduce	Execution time	Worker node crash
2	Social Influence Analysis in Large-scale Networks	Topical Affinity Propagation (TAP)	Efficiency and Effective	Scalability
3	PEGASUS: A Peta-Scale Graph Mining System Implementation	Generalized Iterated Matrix-Vector multiplication	Decrease running time	Computation transparency

	and Observations	(GIM-V)		
4	Spectral Analysis for Billion-Scale Graphs: Discoveries and Implementation	Eigen solver (HEIGEN)	Accuracy	Convergence trouble
5	Counting Triangles and the Curse of the Last Reducer	Sequential triangle counting	No triangle is lost	Increase disk space
6	Colourful triangle counting and a map reduce implementation	Map Reduce colourful triangle counting	Total count is scaled	Low efficiency
7	Enumerating Sub graph Instances Using Map-Reduce	Partition Algorithm	Lowering the number of reducer	Communication and computation cost
8	Densest Sub graph in Streaming and Map Reduce	Largest densest sub graph	Achieve quality and performance	Scalability
9	Mining Protein Family Specific Residue Packing Patterns From Protein Structure Graphs	Delaunay tessellation	Computational competence	Robust
10	OO-FSG: An Object-Oriented Approach to Mine Frequent Sub graphs	Frequent pattern withdrawal	Stability, computation load	Chunk amount and duplication factor

### MAPREDUCE

MapReduce is a programming model that enables dispersed computation over large amount of data. The model provides two nonfigurative operations: map, and reduce. Map corresponds to the “map” operation and reduce corresponds to the “fold” function in functional programming. Based on its role, a member of staff node in MapReduce is called a mapper or a reducer. A mapper takes a collection if (key, value) pairs and apply the map process on each of the pairs to manufacture an subjective number of (key,value) pairs as halfway output. The reducer aggregates all the values that have the comparable key in a minimize list, and applies the reduce operation on that list. It also writes the output to the output file. Iterative MapReduce: Iterative MapReduce can be defined as a multi staged execution of map and reduce function pair in a returning fashion, i.e. the output of the stage i reducers is used as an input of the juncture i + 1 mappers. An exterior situation decides the standstill of the job. Pseudo code for iterative MapReduce algorithm is obtainable in [Figure- 1].

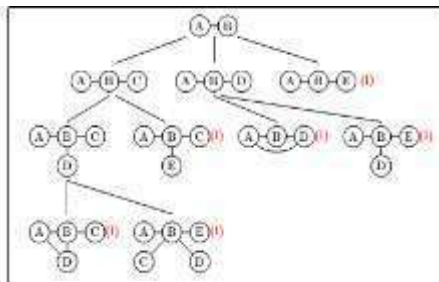


Fig. 1: Map Reduce Mining

```

Algorithm: Iterative MapReduce():While(Condition)
Execute MapReduce Job
Write result to DFS
Update condition
// G is the database
// k is initialize to 1
Mining Frequent Subgraph(G, minsup):
Populate F1
    
```

```

while  $F_k \neq \emptyset$ 
 $C_{k+1}$  = Candidate generation( $F_k, G$ )
for all  $c \in C_{k+1}$ 
  if isomorphism checking( $c$ ) = true
  support counting( $c, G$ )
  if  $c.\text{sup} \geq \text{minsup}$ 
   $F_{k+1} = F_{k+1} \cup \{c\}$   $k = k + 1$ 
return  $S_i = 1 \dots k-1 F_i$ 
  
```

In this paradigm [12], the mining task starts with recurrent patterns of size one (solitary edge patterns), denoted as  $F_1$ . Then in every one of the iterations of the at the same time as loop, the method gradually more finds  $F_2, F_3$  and so on until the entire recurrent pattern set ( $F$ ) is obtained. If  $F_k$  is non-empty at the end of an iteration of the above while loop, from every one of the frequent patterns in  $F_k$  the removal method creates probable candidate recurrent patterns of mass  $k+1$ . These candidate pattern are represented as the set  $C$ . For each of the candidate patterns, the extracting method operates the pattern's support against the dataset  $G$ . If the holdup is well-built than the smallest amount support doorstep (minsup), the agreed pattern is recurrent, and is stored in the set  $F_{k+1}$ . Before bear including, the method also ensures that dissimilar isomorphic forms of a fashionable candidate patterns are amalgamated and only one such reproduction is processed by the algorithm. Once all the persistent patterns of size  $k + 1$  are obtained, the while loop is continued. Thus each repetition of the while loop obtains the set of recurrent patterns of a fixed size, and the development continues until all the recurrent patterns are obtained. The FSM algorithm earnings the union of  $F_i : 1 \leq i \leq k - 1$ .

## POSSIBLE SOLUTION

The following [Table- 1] describes investigational result for obtainable and predictable system investigation. The table contains search node, recurrent sub graph mapping node count and average recurrent sub graph amalgamation sub graph details are shown. The table contains search node, recurrent sub graph amalgamation node count and touchstone recurrent sub graph amalgamation sub graph details are shown.

**Table:1. Frequent Sub-Graph Mining Performances Analysis(Weight of Node)**

S.NO	Search Node	Mapping Sub Frequent Graph Node Count (n)	Average of Mapping Frequent Sub graph Node [%]
1	200	155	77.5
2	250	220	88.00
3	300	272	90.66
4	350	322	92.00
5	400	383	95.75
6	450	429	95.33
7	500	468	93.60
8	550	523	95.05
9	600	578	96.33
10	650	633	97.74

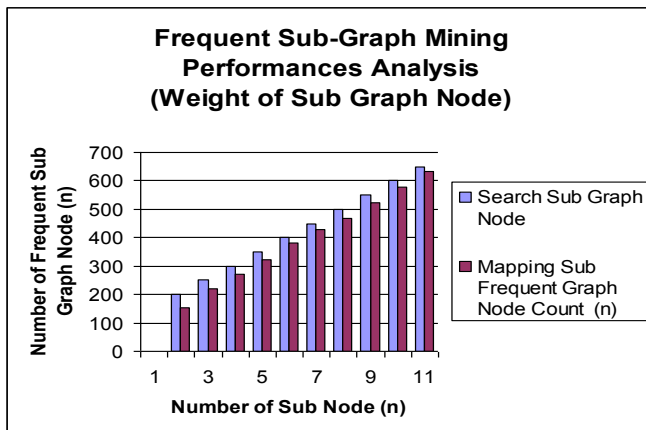


Fig. 2: Frequent Sub-Graph Mining Performances Analysis (Weight of Node)

The following [Figure- 2] describes investigational result for existing and proposed system psychiatry. The figure contains search node, frequent sub graph mapping node count and average recurrent sub graph mapping sub graph details are shown.

## CONCLUSION

The paper achieves solving the task of frequent sub graph mining on a distributed platform like Map Reduce is challenging for various reasons. An FSM method is proposed which computes the support of a candidate sub graph pattern over the entire set of input graphs from a set of graphs (Graph Database). 'N' number of nodes is given with their capabilities. Then the graph details with vertex count are also given. Then graph with minimum vertex count and maximum vertex are found. Then the difference between maximum and minimum is also found out. Then all the groups are grouped such that a) minimum vertex to minimum vertex + 1/3<sup>rd</sup> of difference 'G<sup>a</sup>', b) minimum vertex + 1/3<sup>rd</sup> of difference to minimum vertex + 2/3<sup>rd</sup> of difference 'G<sup>b</sup>' and c) the remaining 'G<sup>c</sup>'. Then the nodes are classified as 1) low capability are assigned with 'G<sup>a</sup>' graphs, 2) medium capability with 'G<sup>b</sup>' graphs and high capability with 'G<sup>c</sup>' graphs. Thus the paper presented a novel iterative MapReduce based frequent subgraph mining algorithm, called FSM-H. The proposed system shows the performance of FSM-H over numerous graph records. This paper shows that FSM-H is significantly better than the existing method.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Afrati F, D Fotakis, J Ullman.[ 2013] Enumerating subgraph instances using map-reduce, in Proc. IEEE 29th Int. Conf. Data Eng, pp. 62–73.
- [2] Agrawal R and R. Srikant. [1994] Fast algorithms for mining association rules in large databases, in Proc. 20th Int. Conf. Very Large Data Bases, pp. 487–499.
- [3] Bahmani B, R Kumar, S Vassilvitskii.[2012] Densest subgraph in streaming and mapreduce,” *Proc. Very Large Data Bases Endow*, 5(5): 454–465
- [4] Chaoji V, M. Hasan, S. Salem, and M. Zaki.[2008] “An integrated, generic approach to pattern mining: Data mining template library, *Data Min. Knowl. Discov. J*, 17( 3): 457–495.
- [5] Cook DJ, LB Holder, G Galal, R Maglothin. [2001]Approaches to parallel graph-based knowledge discovery, *J Parallel Distrib. Comput.*, 61: 427–446,

- [6] Dean J, S. Ghemawat.[ 2008] Mapreduce: Simplified data processing on large clusters,” *Commun. ACM*, 51: 107–113.
- [7] Handan B, R Sunderraman. [2011] Oo-FSG: An object-oriented approach to mine frequent subgraphs” in *Proc. Australasian Data Mining Conf*, pp. 221–228
- [8] Huan J, W Wang, D Bandyopadhyay, J Snoeyink, J Prins.[ 2004] Mining protein family specific residue packing patterns from protein structure graphs,” in *Proc. Int. Conf. Res. Comput. Mol Biol*.pp. 308–315
- [9] Jie Tang, Jimeng Sun, Chi Wang and Zi Yang. [2010] Social Influence Analysis in Large-scale Networks, in *Proc. Int Conf Comput Aspects Soc Netw*, pp. 487–490.
- [10] Kang U, C E Tsourakakis, C Faloutsos. [2009] Pegasus: A petascale graph mining system implementation and observations, in *Proc. 9th IEEE Int. Conf. Data Mining* , pp. 229–238.
- [11] Nijssen S, J Kok.[2004] A quickstart in frequent structure mining can make a difference” in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, pp. 647–652.
- [12] Pagh R, CE Tsourakakis,[ 2012] Colorful triangle counting and a mapreduce implementation, *Inf. Process. Lett.* 112(7): 277–281
- [13] Srichandan B, R. Sunderraman, Oo-FSG:[2011] An object-oriented approach to mine frequent subgraphs,” in *Proc. Australasian Data Mining Conf*221–228.
- [14] Suri Sand S. Vassilvitskii.[2011] Counting triangles and the curse of the last reducer, in *Proc. 20th Int. Conf. World Wide Web*, pp. 607–614.
- [15] wang C, S. Parthasarathy.[ 2004 ]Parallel algorithms for mining frequent structural motifs in scientific data, in *Proc. 18th Annu. Int. Conf. Supercomput*, 31–40.

# A REVIEW ON EFFECTIVE TRANSFER OF DATA PACKETS USING MULTICAST ALGORITHM

SN Ranjini\*, AS Renugadevi, P Brindha

Dept. of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Elayampalayam, Tamilnadu, INDIA

## ABSTRACT

This paper deals with the secure routing with multicast algorithm. To increase the lifetime of network, a key design factor is being improved. Because of this increased network lifetime, power metrics of the neighboring node table, routing table and group table can be improved. The data transfer can be done in an effective way by storing the neighboring nodes information within a particular network and the routes information which is useful in transferring the data from one node to the other without congestion. By using the routing information, expire time of the data packets can be known.

Published on: 2<sup>nd</sup> -December-2016

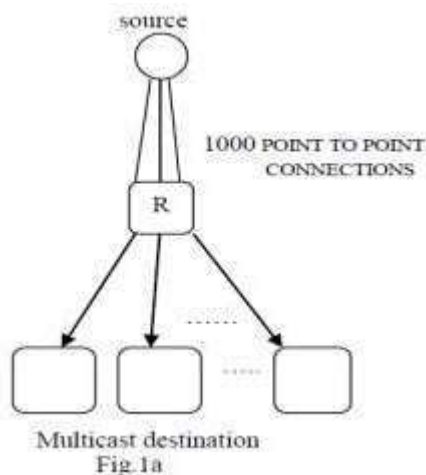
### KEY WORDS

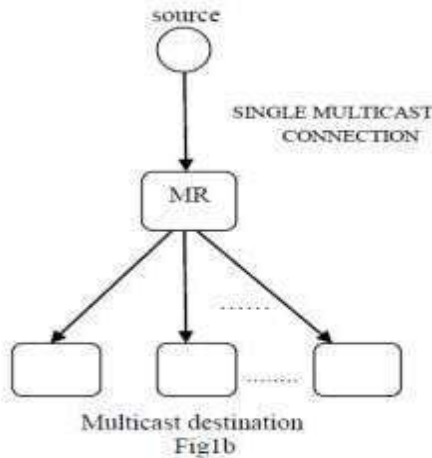
transfer of data packets,  
multicast routing.

\*Corresponding author: Email: [ranjiniramesh177@gmail.com](mailto:ranjiniramesh177@gmail.com); Tel.: +91 8220624546

## INTRODUCTION

Wireless sensor networks are the emerging technique which is mainly nowadays useful to transfer the data packets in the form of sound, light, etc. through a waveguide of air medium. The main impact of this paper is to deal with incorporating the multicast algorithm within a network to increase the lifetime and congestion free transmission of data packets to the destined address. The objective of the WSN is to collect specific data and send it to the required destination. WSNs were studied for many purposes, but research is now focused on a wide range of consumer industries, giving rise to the notion of ubiquitous computing.





Multicasting is the ability of a communication network to accept a one message from an application and to release copies of the message to many recipients at alternate locations. One of the challenges is to reduce the amount of network resources employed by multicasting. To illustrate this point, assume that a video server wants to transmit a movie to 1000 recipients [Figure- 1a]. If the server were to employ 1000 divide point-to-point connections (e.g., TCP connections), 1000 copies of the movie may have to be sent over a single link, thus making poor use of the available bandwidth. An efficient implementation of multicasting permits much better use of the available bandwidth by transmitting at most one copy of the movie on catch link in the network, as shown in [Figure- 1b]. Recently, there has been a lot of research in the area of multicast communication.

## LITERATURE SURVEY

Over the years, there have been several methods in enhancing the technique of data transfer through different techniques. Initially, separate technique is being used to increase the speed of data transfer without any conflict in achieving the data at the destination point. For example, distance vector routing is being used to find the distance between the source node and destination node. This algorithm can find only the routing distance. But nowadays, several techniques are being implemented as a package where all the details can be utilized at a time. Ad Hoc is the wireless technique which uses the secure transfer of data. As there are limited resources in MANET so it faces many problems such as security, limited bandwidth, range and power constraints. Due to this, many new routing protocols are proposed.

The Distributed Multipath Routing Protocol for hybrid wireless networks which establishes multiple paths between source and destination. Reveal the efficient use of watchdog technique in existing trust systems, and propose a suite of optimization methods to minimize the energy cost of watchdog usage, while keeping the system's security in a sufficient level. It reduces overhead and path loss. It also has a congestion control algorithm to avoid traffic among the base stations.

Clustering is a technique which is more effective to exalt system performance. SET IBS and SET IBOOS are two Secure and efficient data transmission protocol for cluster based WSNs. They are used Identity Based digital Signature (IBS) method and Identity Based Online/Offline digital Signature (IBOOS) methods, where the clusters are formed dynamically and periodically.

Even though SET IBS and SET IBOOS are secured routing protocols, it has some demerits like energy efficiency, trustability and elongating network lifetime. To overcome these issues, Reliable Minimum Energy Cost Routing (RMECR) and Reliable Minimum Energy Routing (RMER) are two routing algorithms proposed for Mobile Ad-hoc network in which trustability is ensured either hop by hop or end to end retransmissions. It considers the energy level in a battery of a node and also through send a links to find energy efficient and trustable paths that improves the working longevity of the network



MANETs require privacy and communication security in routing protocol. In this paper present the type of attacks and operation on network layer with routing protocol technique i.e. based on an on-demand location based on anonymous MANET routing protocol called SMRT (secure MANET routing technique) with trust model that achieves security and privacy against insider and outsider adversaries.

In [1], "ActiveTrust: Secure and Trustable Routing in Wireless Sensor Networks", authors "Yuxin Liu, Mianxiong Dong, Kaoru Ota and Anfeng Liu", A review describes the High successful routing probability, security and scalability. The ActiveTrust scheme can quickly detect the nodal trust and then avoid suspicious nodes to quickly achieve a nearly 100% successful routing probability.

In [2], "Trust Based Energy Aware Reliable Reactive Protocol in Mobile Ad Hoc Networks", authors "M.Pushpalatha, Revathi Venkataraman, and T. Ramarao", explain the High energy efficiency. The ActiveTrust scheme fully uses residue energy to construct multiple detection routes.

In [3], "A Reputation-based Trust Management System for P2P Networks", authors "Ali Aydin Selcuk, Ersin Uzun and Mark Resat Pariente", says the simple and efficient in design and can be integrated into most first generation P2P systems easily. A diverse set of simulation experiments conducted to test the performance of the system exhibit that it can be highly effective in preventing the spread of malicious content.

In [4], "Performance Analysis of Mobile Ad-hoc Network Using AODV Protocol", authors "Dr.Aditya Goel and Ajai Sharma", describes the mobility behaviour has been controlled by the approach Optimized-AODV Protocol, it uses the ant colony optimisation algorithm. They have increased the lifetime of the network, by the means of discovering new shortest routes from path accumulation. This method entirely depend upon artificial intelligence algorithms like ACO, to find the best routes which may result into high process and load on the device. This difficulty in this approach, affect the performance of the system.

In [5], "A Reverse AODV Routing Protocol in Ad Hoc Mobile Networks", authors "Chonggun Kim, Elmurod Talipov and Byoungchul Ahn", explains the reverse AODV method has been used. It tries multiple route replies when compare to AODV protocol. The main motto of this work is to reduce the communication delay and power consumption, but this have given efficiency in reduces path fail correction message.

In [6], "Comparative Performance Analysis of DSDV, AODV and DSR Routing Protocols in MANET using NS2", authors "Asma Tuteja, Rajneesh Gujral and Sunil Thalia", describes the comparative performance has been done for Mobile Ad-Hoc network routing protocols like DSDV, AODV and DSR. This paper deals only about the improvement throughput, End to End Delay, Routing overhead, Packet Delivery Ratio and network lifetime. Eventhough the network lifetime is not increased up to the mark.

In [7], "Trust management in mobile ad hoc networks for bias minimization and application performance maximization", authors "Ing-Ray Chen, Jia Guo, Fenyue Bao and Jin-Hee Cho", describes the effectiveness of proposed approach with an integrated social and quality-of-service (QoS) trust protocol (called SQTrust) with which we identify the splendid trust aggregation setting under which trust bias is minimized despite the presence of malicious nodes performing slandering attacks.

In [8], "Trust Based Routing in Mobile Ad-Hoc Networks", authors "Vinesh H. Patel, Mukesh A. Zaveri, and Hemant Kumar Rath", explains the Routing protocols are vulnerable to routing attacks like packet dropping and delayed packet forwarding. The proposed scheme is implemented in QualNet simulator with modification in the AODV routing protocol by incorporating the trust model based approach increase network lifetime and improves network performance in presence of malicious activities.

In [9], "A Light-Weight Trust based Mobility Aware Routing Algorithm for Mobile Ad Hoc Networks", authors "Saurin J. Choksi and Nikhil N. Gondaliya", explains the applying security the packet forwarding behaviour of the nodes is used and for mobility speed and the relative direction of the node is taken. The algorithm is implemented on AODV protocol and checked the final simulation results against the normal AODV and trust based AODV (that uses only forwarding behaviour of the nodes) using NS2.

In [10], "Mobile Target Detection in Wireless Sensor Networks With Adjustable Sensing Frequency", authors "Yanling Hu, Mianxiong Dong", Kaoru Ota, Anfeng Liu and Minyi Guo, describes the important issue of the

balance between the quality of target detection and lifetime in wireless sensor networks. Two target-monitoring schemes are proposed. One scheme is Target Detection with Sensing Frequency K (TDSFK), which distributes the sensing time that currently is only on a portion of the sensing period into the whole sensing period. That is, the sensing frequency increases from 1 to K. The other scheme is Target Detection with Adjustable Sensing Frequency (TDAF), which adjust the sensing frequency on those nodes that have residual energy.

## COMPARITIVE ANALYSIS

S.No	Title	Techniques	Advantages	Disadvantages
1	Load-balancing in MANET shortest-path routing protocols	Multipath Routing Protocol	-Multi-path routing can balance the load better than single-path routing, only if we use a very large number of paths between any source-destination pair of nodes.	-No secret share from the source to destination. -Security issue on multipath routing protocol.
2	An Efficient Countermeasure to the Selective Forwarding Attack in Wireless Sensor Networks	multi-path DSR	-The multi-Path topologies (DSR) scheme to defend against the selective forwarding Attack. -The DSR scheme has some advantages. -First, the base station can receive information sensing from sensor nodes continuously under the selective forwarding attack. Second, this scheme is lightweight and simple. -Third, the dropped packets do not need to be re-sent when detecting malicious sensor nodes. Finally, this scheme can defend several kinds of attacks.	-The communication distance can be Increased and waste of communication cost. -It's seriously affect the network lifetime.
3.	H-SPREAD: A Hybrid Multipath Scheme for Secure and Reliable Data Collection in Wireless Sensor Networks.	SPREAD	-The advantage of this algorithm is that through multi-path routing, each path routes only one share, and the attacker must capture at least $T$ shares to restore nodal information, which increases the attack difficulty.	-This mechanism will improve both reliability and security at the same time. -This weakness opens the door for various attacks if the routing algorithm.
4.	A Reliability-Oriented Transmission Service in Wireless Sensor Networks	proliferation routing capability-based path finder, randomized disparity, and reproduction	-The basic idea is that we estimate the packet loss in a several-hop manner and generate new packet copies after certain steps. -The packet loss and packet generation so that the E2E service quality can be maintained with any length of the data paths and the scale of the network.	-the energy efficiency of proliferation routing is not encouraging yet. -The capability-based path finder only works for transmission from sensors to sink. For general communications, say sensor-to-sensor, more investigations are needed.
5.	Design and Implementation of TARF: A Trust-Aware Routing Framework WSN's	Trust-aware routing framework protocol (TARF)	-A robust trust-aware routing framework for dynamic WSNs. Without tight time synchronization or known geographic information, TARF provides trustworthy and energy-efficient route. -Using trust and energy cost for route decisions, to prevent malicious nodes from misleading network traffic	-This approach affect network lifetime. -Affects system performance. The trust route mechanism has high costs and is difficult to obtain trust, so the guiding significance is limited
6	Performance Analysis of Mobile Ad-hoc Network Using AODV Protocol	AODV Protocol	-Two important mechanisms, Route Discovery and Route Maintenance. -AODV is chosen for the obvious reason that it is simple and has a low overhead and its on-demand nature does not unduly burden the networks.	- It is designed to be self-starting in an environment of mobile nodes, withstanding a variety of network behaviors such as node mobility, link failures and packet losses.
7	DSDV Routing	DSDV protocol	-In proactive protocols, routes to all	-This increases the overhead

	Protocols in MANET using NS2		the nodes in the network are discovered in advance. - broadcasts after a fixed interval of time independent of any route changes or not.	and so decreases the throughput of network using DSDV protocol. In DSDV Protocol, every node stores one or more routing tables.
8	AODV Routing Protocols in MANET using NS2	AODV protocol	Reactive protocol as AODV protocol.	-AODV only stores address of next hop to the destination. -It is a reactive protocol as it only requests a route when needed and does not require nodes to maintain routes to the destinations that are not actively used in communication.
9	DSR Routing Protocols in MANET using NS2	DSR protocol	-DSR (Dynamic Source Routing) is a source initiated. -It stores complete route from source to destination including all the intermediate nodes.	-Sender of the packet determines the complete sequence of nodes through which to forward the packets; the sender explicitly lists this route in the packet's header, identifying each forwarding "hop" by the address of the next node to which to transmit the packet on its way to the destination host.
10	A Reverse AODV Routing Protocol in Ad Hoc Mobile Networks	R-AODV Routing protocol	to avoid RREP loss - improve the performance of routing in MANET. - R-AODV prevents a large number of retransmissions of route request messages, and diminishes the congestion in the network.	RREP Delivery Fail

### PROPOSED SYSTEM

This paper aims at using the technique of multicast algorithm. In this algorithm, three main properties have been included to improve the power aware metrics and congestion free routing. The information of neighboring node is helpful in transferring the data packets to the destined point with the destination address. The group name as well as address of the data packets of any particular network is provided in the group node. The details of the neighboring node and routes are provided in the neighboring node and routing table. By using all these information, the data can be sent to the required destination point in a faster manner.

### RESULT AND DISCUSSION

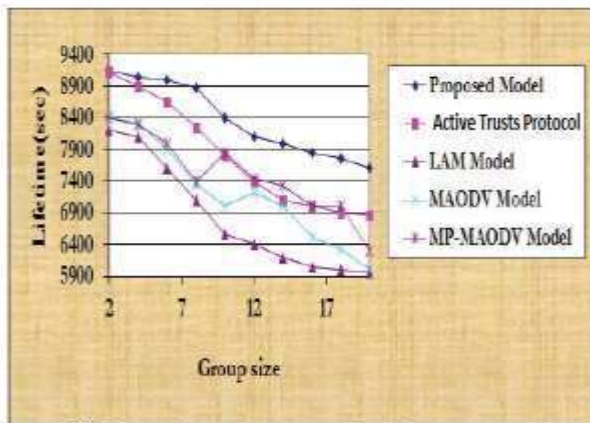


Fig. 2: Group size node versus lifetime

The proposed system and along with the protocols like Active Trust protocol, LAM Model, MAODV model, MP-MAODV has been simulated to show the performance of time along with the group size of wireless nodes. It is clear from the simulated result that the network lifetime is increased for the proposed system when compare to all other methods. During the initial stage, the value starts at 9000sec, for a group size of 2. The power sources have a very low negative gradient during the initial stages, and reach the value of 8900sec at group size 7. Then after that if the groups size has been increasing, the network lifetime decreasing steadily, and it reaches the saturation level of 7500sec for group size greater then 17, but for other networks the value reached around 5900sec.

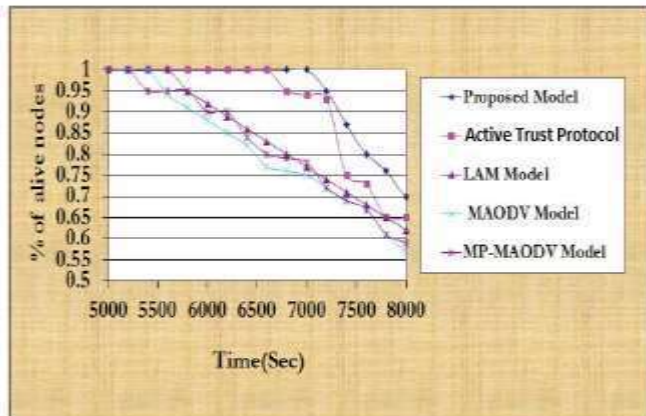


Fig. 3: Time against percentage of alive-nodes

The next parameter shows the characteristics curve of percentage of alive nodes along with time. Even in this parameter too, the proposed method excels the other existing methods. Till 7000sec, all nodes are alive, it begin to decrease at after 7000sec, and reaches 75% at 8000sec.

From the both performance curve, it shows the proposed system performs better than the all other existing system. If we incorporate this technique, we will have increase in lifetime of the nodes. This proposed system mainly deals with the secure transmission of data packets to the destination point of multiple recipients in a particular network. This transmission should be secure and congestion free technique.

## CONCLUSION

This whole paper mainly focuses on delivery of data packets to multiple users with the help of multicast algorithm. This technique also utilizes the node configuration for the secure transmission through wireless sensor network. The Ad Hoc is also one of the method to transfer the data. Some of the disadvantages are faced in that technique mainly the security of data transfer. But in this proposed system encryption is done with in that particular network. Moreover , in this technique the neighboring node details as well as distance and expire time are well known. The main advantage of this system is the increment of network lifetime.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Yuxin Liu, Mianxiong Dong, Kaoru Ota and Anfeng Liu.[ 2016] ActiveTrust: Secure and Trustable Routing in Wireless Sensor Networks, IEEE Transactions on Information Forensics and Security,11(9)
- [2] M. Pushpalatha, Revathi Venkataraman, and T Ramarao.[ 2009] Trust Based Energy Aware Reliable Reactive Protocol in Mobile Ad Hoc Networks. *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, 3(8): 1535-1538.

- [3] Ali Aydin Selcuk, Ersin Uzun and Mark Resat Pariente.[2008] A Reputation-based Trust Management System for P2P Networks, *International Journal of Network Security*, 6(3):235-245
- [4] Aditya Goel, Ajaii Sharma. Performance Analysis of Mobile Ad-hoc Network Using AODV Protocol", *International Journal of Computer Science and Security (IJCSS)*, 3(5): 334-343.
- [5] Chonggun Kim, Elmurod Talipov and Byoungchul Ahn.[2006] A Reverse AODV Routing Protocol in Ad Hoc Mobile Networks, *IFIP International Federation for Information Processing ,EUC Workshops, LNCS 4097*, pp. 522 – 531,
- [6] Asma Tuteja, Rajneesh Gujral and Sunil Thalia.[2010]Comparative Performance Analysis of DSDV, AODV and DSR Routing Protocols in MANET using NS2, 2010 International Conference on Advances in Computer Engineering, pp 330-333
- [7] Ing-Ray Chen, Jia Guo, Fenyue Bao, Jin-Hee Cho, "Trust management in mobile ad hoc networks for bias minimization and application performance maximization", *Adhoc Networks*, Elsevier, 19: 59-74 , 2014.
- [8] Vinesh H Patel, Mukesh A Zaveri, and Hemant Kumar Rath.[2015] Trust Based Routing in Mobile Ad-Hoc Networks, *Lecture Notes on Software Engineering*, 3(4) November 2015.
- [9] Saurin J Choksi, Nikhil N Gondaliya, [2014] A Light-Weight Trust based Mobility Aware Routing Algorithm for Mobile Ad Hoc Networks", *International Journal of Computer Applications*, 97(14).
- [10] Yanling Hu, Mianxiong Dong, Kaoru Ota, Anfeng Liu and Minyi Guo, "Mobile Target Detection in Wireless Sensor Networks With Adjustable Sensing Frequency, *IEEE systems journal*.
- [11] Mianxiong Dong, Kaoru Ota, Anfeng Liu and Minyi Guo, Joint Optimization of Lifetime and Transport Delay under Reliability Constraint Wireless Sensor Networks, *IEEE transactions on parallel and distributed systems*, tpd-2013-12-1250.
- [12] Shibo He, Jiming Chen, Fachang Jiang, David KY Yau, Guoliang Xing and Youxian Sun.[2013] Energy Provisioning in Wireless Rechargeable Sensor Networks, *IEEE transactions on mobile computing*, 12(10)
- [13] Chunsheng Zhu, Hasen Nicanfar, Victor CM Leung, Laurence T Yang. [2014]An Authenticated Trust and Reputation Calculation and Management System for Cloud and Sensor Networks Integration, *IEEE transactions on information forensics and security*,
- [14] Zhongming Zheng, Anfeng Liu, Lin X Cai, Zhigang Chen, and Xuemin (Sherman) Shen, "Energy and Memory Efficient Clone Detection in Wireless Sensor Networks, 10.1109/TMC.2015.2449847, *IEEE Transactions on Mobile Computing*.
- [15] C Rajan, N Shanthi.[ 2015] Genetic based Optimization for multicast Routing algorithm for Manet' Sadhana - Academy Proceedings in Engineering Science, 40(7): 2341-2352.
- [16] Shigen Shen, Hongjie Li, Risheng Han, Athanasios V. Vasilakos, Yihan Wang, Qiyang Cao.[2014] Differential Game-Based Strategies for Preventing Malware Propagation in Wireless Sensor Networks, *IEEE transactions on information forensics and security*, 9(11).

## ABOUT AUTHORS

**S.N. Ranjini** : Graduate student, Department of Computer Science and Engineering, Vivekanandha college of engineering for women, Elayampalayam Tamilnadu, India

**A.S. Renugadevi** : Assistant Professor, Department of Computer Science and Engineering, Vivekanandha college of engineering for women, Elayampalayam Tamilnadu, India

**P. Brindha** : Assistant Professor, Department of Computer Science and Engineering, Vivekanandha college of engineering for women, Elayampalayam Tamilnadu, India

# A SURVEY OF MULTI KEYWORD SEARCH OVER THE ENCRYPTED DATA IN CLOUD

S. Sathya\*, J. Gayathri, D. Radhika

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for women, Namakkal, T. N., INDIA

## ABSTRACT

*Aim: Cloud computing [11] is a tremendous growth in every years and it can be utility the computing and large storage capability to the public users. The data owner can store the data in the cloud server is called data outsourcing and then the cloud data access for public users through the cloud server. The outsourced data are contains sensitive privacy information and it can be encrypted before uploaded to the cloud server and then the search user can access to the data through the cloud server is some difficulty of searching over the encrypted data in cloud. In this paper address this problem by developed the fine-grained multi-keyword search scheme over encrypted data in the cloud. There are three contribution of this paper. First one is, to provided relevance scores and preference factors upon keywords which enabled the precise keyword search. Second one is, to developed a complicated logic search the mixed AND, OR and NO operations of multi-keyword search scheme. And finally, auxiliary employ the classified sub-dictionaries technique to accomplish the index building, trapdoor generating and query. By using this experiments to the real-world dataset, so easily retrieve the result from dataset.*

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

*Fine grained multi keyword encryption, searchable encryption, and index and trapdoor generation.*

\*Corresponding author: Email: [sathyashree.mecse@gmail.com](mailto:sathyashree.mecse@gmail.com); Tel.: +91 9655883123

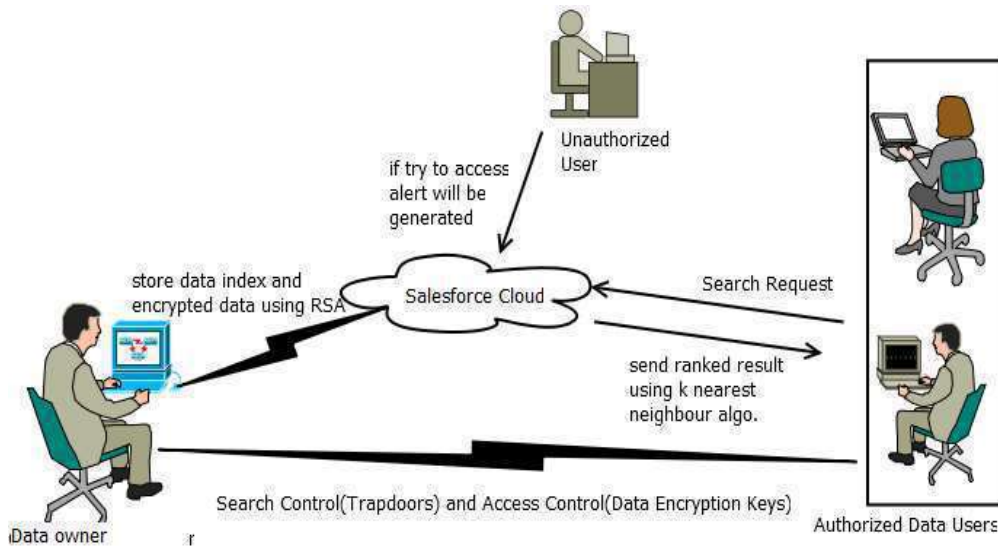
## INTRODUCTION

In cloud computing [11], data owners are moving their data to the cloud server to access the multiple public users. And then the cloud data open for public usage through the cloud server. The data owner uploaded the file to the cloud before that it can be encrypted is called as outsourcing data. The benefits of this kind of data outsourcing such as low-cost and flexible data access. It can also cause some privacy problems since the outsourced data had some sensitive information. It is necessary to encrypt the sensitive data before updating them to the cloud server. Moreover, data owners would be like to allow multiple search users to search over the encrypted data while access the control policies.

### Multi Keyword Search

Multi keyword search technique [17] allows users to selectively retrieve files of interest and has been applied in plaintext search scenarios. Unfortunately, data encryption, which restricts users able to perform keyword search and further demands the protection of keyword privacy, makes the traditional plaintext search methods fail for encrypted cloud data. Ranked search [17] greatly improves system usability by normal matching files in a ranked order regarding to certain relevance criteria.

Considering three different framework entities in **Figure-1**, Such as Data owner, Search user, and cloud server. Data owner have a collection of data documents to be sent to cloud server in the encrypted format. To activate the searching efficient over encrypted data, data owner, before sending data, data owner can encrypt the data and then outsource both the index and the encrypted document collection sent to cloud server. It can be search the document, an authorized user require a corresponding trapdoor through search mechanisms to receiving from data users, cloud server is take over to search the index and returns the matched set of encrypted documents.



**Fig. 1: An example of search over encrypted cloud data in sales force cloud**

To improve the search efficiency, the searchable encryption schemes must support multi-keyword search [12]. And search users would like the cloud server to return results in a specific order, so they can achieve the more relevant results quickly. Furthermore, to make the searchable encryption schemes suitable for more practical situations, such as the situation that the data is contributed from many data owners and can be searched by many search users. And this scheme to support search authorization, It means the cloud server would only return the authorized results to the search users.

### System Model

To provides the relevance scores and the preference factors of keywords for searchable encryption over the encrypted data. The relevance scores of keywords can enable more precise returned results and the preference factors of keywords denote the importance of keywords in the search keyword set specified by search users. To realize the AND, OR and NO operations in the multi-keyword search for searchable encryption over the encrypted cloud data and it compared with schemes in [16], [3] and [4], the proposed scheme can achieve more broad functionality and lower query complication.

### Data Owner

The data owner outsources the data to the cloud for reliable data access to the search users. To protect that data privacy and it means unauthorized person doesn't modify the content of original data. And the data owner encrypts the outsourced data through symmetric encryption. To improve the search efficiency, the data owner generates some keywords for each outsourced data. The corresponding index is then created according to the keywords and a secret key.

The data owner has a collection documents  $F = \{f_1, f_2, f_3, \dots, f_n\}$ . And the data owner wants to outsource the data to the cloud in encrypted format at the same time he wants to keep the capability to search on them for effective utilization. What is all data owner's process? The Data owner first builds a secure searchable keyword index  $I$  from document collection  $F$  and then generates an encrypted document collection  $C$  for Afterwards, the data owner outsources the encrypted collection  $C$  and the secure keyword index  $I$  to the cloud server. Securely distributes the key information of trapdoor generation (including IDF values) and document decryption to the authorized data users. Data owner is responsible for the update operation of his documents stored in the cloud server while updating. The data owner generates the update information and sends it to the server.

### Cloud server

Cloud server stores encrypted documents collection  $C$  and encrypted searchable keyword indexes  $I$  that are received from the data owner. Data owner provides data access and search services to search users upon receiving the trapdoor  $TD$  from the data user, the cloud server excites the search over the keyword index  $I$ . Finally returns the corresponding collection of collection of matching documents based on certain operations such as AND, OR and NO operation of keywords.

### Search user

Search users are authorized one's to access the document from cloud server with following three steps. First step is, the search user receives both the secret key and symmetric key from the data owner. Second step is, according to the search keywords, the search user using the secret key to generate trapdoor according to search control mechanisms and sends it to the cloud server. Finally, the search user receives the matching document collection from the cloud server and decrypts the encrypted data by using symmetric key.

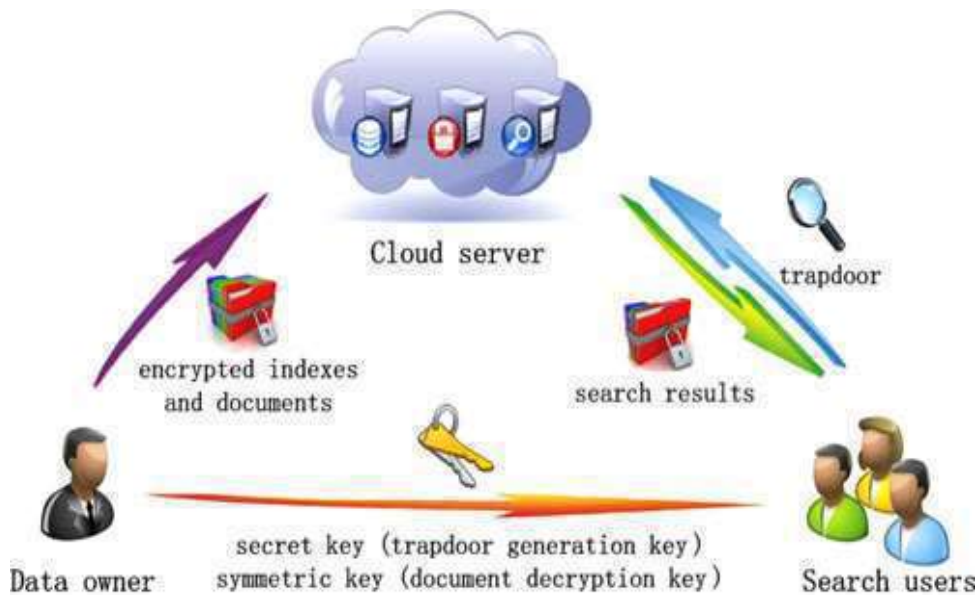


Fig. 2: Key generate of both data owner and search user over the cloud

## LITERATURE SURVEY

It deals with the information by using multi keyword search scheme [2] to identify the original encrypted data and its major role for the user to collect original data.

In [1] authors D. X. Song, D. Wagner, and A. Perrigin the paper “**A Practical and Secure Multi-Keyword Search Method over Encrypted Cloud Data**” presented a well-organized privacy-preserving multi keywords search method over encrypted cloud data by used minhash functions. A multi-keyword search technique can be combining of several keywords in a single query. By dint of increasing the search constraints and also fetched the most relevant items returned to the search user. Since a multi-keyword search method that returns the matching encrypted data in a ranked ordered manner and it can hold three steps is like, First step is, to present a minhash based privacy-preserving multi keyword search method that provides high precision rates. Second step is, to provides the security requirements and formally prove that the proposed method satisfies adaptive semantically security. Third step, to use a ranking method depends on term frequencies and inverse document frequencies (tf-idf) of keywords and demonstrate that it is efficient and effective by providing the implementation results. One of the main advantages of this paper is the enable of multi keyword search in a single query.

In [2] authors H. Li, D. Liu, Y. Dai, T. H. Luan, and X. Shen in this paper “**Enabling Efficient Multi-Keyword Ranked Search Over Encrypted Mobile Cloud Data Through Blind Storage**” describes the searchable



encryption for multi-keyword ranked search over the storage data. There are large numbers of outsourced documents (data) in the cloud. In this paper utilize the relevance score and  $k$ -nearest neighbor techniques to develop an effective multi-keyword search scheme and it can be returned to the ranked search results based on the accuracy. Security analysis established that scheme can achieve confidentiality of documents and index, trapdoor privacy and concealing access pattern of the search user. Finally, using extensive simulations can achieve much improved efficiency in terms of search functionality and search time.

In [3] authors BoyangWang, YantianHou, Ming Li in this paper “**Practical and Secure Nearest Neighbor Search on Encrypted Large-Scale Data**” presented a new searchable scheme which can effectively and securely enable nearest neighbor search over encrypted data in clouds. Particularly modify the search algorithm of nearest neighbors with tree structures (R-trees), where the modified algorithm adapts to lightweight cryptographic primitives (Order-Preserving Encryption) without affecting the linear search complication. Moreover, which can be used for secure  $k$ -nearest neighbor search and it is compatible with another similar tree structures. In this paper results on Amazon EC2 show that scheme is extremely practical over massive datasets.

In [4] authors H. Li, D. Liu, K. Jia, and X. Lin in this paper “**Achieving Authorized and Ranked Multi-keyword Search over Encrypted Cloud Data**” presented an authorized and ranked multi-keyword search scheme (ARMS) over encrypted cloud data by included the cipher text policy attribute-based encryption (CP-ABE) and SSE techniques. To motivate the research on the searchable encryption technique, it can be allows the search user to search over the encrypted data in cloud. In this paper, particularly focus the symmetric searchable encryption (SSE) techniques. However, they do not conceive the search authorization problem that requires the cloud server to return the search results to authorized users. Security analysis demonstrates that the ARMS scheme can achieve confidentiality of documents collusion resistance and trapdoor unlinkability.

In [5] authors Y. Yang, H. Li, W. Liu, H. Yang, and M. Wen in this paper “**Secure Dynamic Searchable Symmetric Encryption with Constant Document Update Cost**” presented to leverage the secure  $k$ -nearest neighbor to proposed a secure dynamic searchable symmetric encryption scheme. In this scheme can achieve two important security features is backward privacy and forward privacy which are very challenging in Dynamic Searchable Symmetric Encryption (DSSE) area. And also to evaluate the performance of proposed scheme compared with other DSSE schemes. The comparison results demonstrate the efficiency of the scheme in terms of the storage, search and update complexity.

In [6] authors Neelam S. Khan, Dr. C. Rama Krishna, Anu Khuranain this paper “**Secure Ranked Fuzzy Multi-Keyword Search over Outsourced Encrypted Cloud Data**” to solve the problem of effective Secure Ranked Fuzzy Multi-Keyword Search over Outsourced Encrypted Cloud Data (RFMS). RFMS improves user searching experience by returning the matching files when users input query to retrieve any exactly matches the predefined keyword dictionary or closest possible keywords in the dictionary depended on similarity semantics when exact match fails. Information discovery has been made effective by searching with multiple keywords with ranking so as to eliminate false positives. Keyword dictionary has been made dynamically. Overhead of updating the dictionary when new files need to be uploaded has been minimized. And also by using one-to-many mapping between plaintext and cipher text, the method guarantees security.

In [7] authors J. Yu, P. Lu, Y. Zhu, G. Xue, and M. Li in this paper “**Towards Secure Multi-Keyword Top-k Retrieval over Encrypted Cloud Data**” an addressing data privacy problems by using searchable symmetric encryption (SSE). For the first time to formulated the privacy issues from the aspect of similarity relevance and schemes robustness. To observe that server-side ranking depended on order-preserving encryption (OPE) unavoidably leaks data privacy. To eliminate the leakage, to provides a two-round searchable encryption (TRSE) scheme that supports top- $k$  multi-keyword retrieval. In TRSE, service a vector space model and homomorphic encryption. The vector space model helps to deliver sufficient search accuracy and the homomorphic encryption enables users to include in the ranking and it's done on the server side by operations on cipher text. The result, information leak can be eliminated and data security is ensured.

In [8] authors W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li in this paper “**Verifiable Privacy-Preserving Multi-Keyword Text Search in the Cloud Supporting Similarity-Based Ranking**” it presents the privacy-preserving multi-keyword text search (MTS) scheme used similarity-based ranking over encrypted data in cloud. For supporting the multi-keyword search and search result ranking to construct the search index based on term frequency and the vector space model with cosine similarity measure to suggest higher search result

accuracy. To increase the search efficiency for developed the tree-based index structure and different adaptive methods for multi-dimensional (MD) algorithm. They enhanced the search privacy scheme to construct the two secure index schemes to meet the difficult privacy requirements under strong threat models is called cipher text and background model. Advantage of this scheme depends upon the index tree structure to enable authenticity check over the returned search results.

In [9] author Zihua Xia in this paper “**A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data**” describes a secure multi-keyword ranked search scheme over encrypted cloud data, which consecutively supports dynamic update operations like deletion and insertion in the documents. Particularly the vector space model and the broadly used TF×IDF model are combined in the index construction and query generation. In this paper construct a special tree-based index structure and suggest a Greedy Depth-first Search algorithm to provide well-organized multi-keyword ranked search. The secure kNN algorithm is exploited to encrypt the index and query vectors, and meanwhile ensure accurate relevance score calculation between encrypted index and query vectors. Advantage of this paper is to use the special tree-based index structure, the proposed scheme can succeed sub-linear search time and deal with the deletion and insertion of documents flexible in encrypted cloud.

In [10] authors B. Zhang and F. Zhang in this paper “**An efficient public key encryption with conjunctive-subset keywords search**” presented a Public Key Encryption Keyword Search (PEKS) scheme with Subset keywords search and it means that the receiver could query the subset keywords embedded in the cipher text. In this paper to solve the problem of conjunctive with subset keywords search function, deliberate the demerits about the existed schemes and then give out a more effective construction of Public Key Encryption with Conjunctive-Subset Keywords Search (PECSK) scheme.

## COMPARATIVE ANALYSIS OF DIFFERENT MKS SCHEMES

This section presents the comparison of different Multi Keyword Search (MKS) schemes the user can identify the information are reviewed in the comparative analysis section.

**Table: 1. Comparison of different MKS schemes**

Papers	Search Techniques	Advantage	Disadvantage
[1]	Privacy Preserving Multi Keyword Search	1)Most relevant items are retrieved by the user 2)Preventing unnecessary Communication	1)Low efficiency compared Randomized 2)Order Preserving Low privacy for securing data.
[2]	Multi keyword Ranked Search scheme	1)It enable accurate and secure search Over encrypted mobile cloud data. 2) Security analysis can effective documents and index and trapdoor privacy access pattern of the search user.	1)It not give accurate results 2) This scheme doesn't consider the importance of the different keywords.
[3]	Nearest neighbor search (or)k-nearest neighbor scheme	Very secure nearest neighbor search practical on massive Datasets.	To retrieve the nearest data from datasets.
[4]	Authorized And Ranked Multi Keyword Search Scheme(ARMS)	ARMS scheme can achieve confidentiality of documents, trapdoor unlink ability and collusion resistance.	ARMS is not explore the dynamic searchable Encryption in dataset.
[5]	Similarity Based Ranking Multi Keyword Search	This search scheme achieves better than linear search efficiency	The results in precision loss
[6]	Ranked Fuzzy Multi-Keyword Search	It can be more efficient by reducing the searching time	1)Un trusted server to search for a secret word 2)Low bandwidth

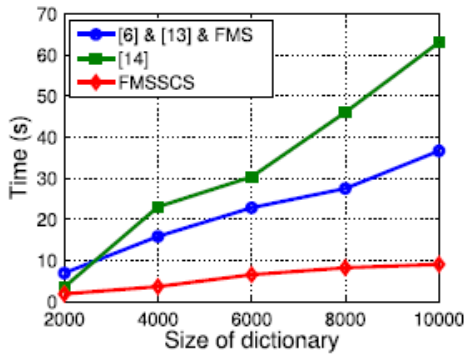
[7]	Top-k multi-keyword retrieval search	To retrieve the result very effective from cloud	1)Low level protections only 2)Computation cost high
[8]	Privacy-Preserving Multi-Keyword Text Search	It depends upon the index tree structure to enable authenticity check over the returned search results	Index problem is occurred during multiple search user access the same data
[9]	Dynamic Multi-Keyword Ranked Search (DMRS) Scheme	It special tree-based index structure can achieve sub-linear search time and deal with the deletion and insertion of documents flexible	1) Low efficiency 2)Dynamic search scheme doesn't realize the multi keyword ranked search functionality
[10]	Conjunctive-Subset Keywords Search	It can only return the results Which match all the keywords simultaneously	It cannot Provides acceptable results ranking functionality

### Performance evaluation

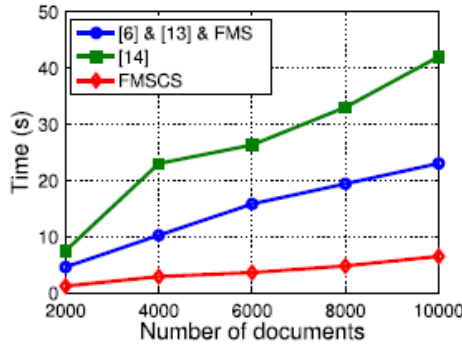
The outsourced documents are encrypted by the symmetric encryption algorithm. In also, the data owner generated the secret key and it sent to the search user through a secure model. Since symmetric encryption algorithm is secure, unauthorized can't recovery the encrypted documents without the secret key. So it can be achieved by confidentiality of encrypted documents. Compare with other multi keyword search scheme, the fine grained multi keyword search is best performance to retrieving matching results from datasets. In below graphs represented the storage and communication to the overall process model in encrypted retrieval results from the cloud server. And also, in (b) number of documents increasing, the retrieving process time will be increased. In (c) the number will be increasing, the searching process is best compare with previous techniques by using AND, OR and no operation. Every keyword is encrypted by using symmetric encryption algorithm to store in cloud server and it takes lot of space. Avoid this issue; in future to developing the compression techniques and it means first encrypted the keyword after that before storing the encrypted data in the cloud by using compression techniques to reduce the space in cloud.

### POSSIBLE SOLUTION

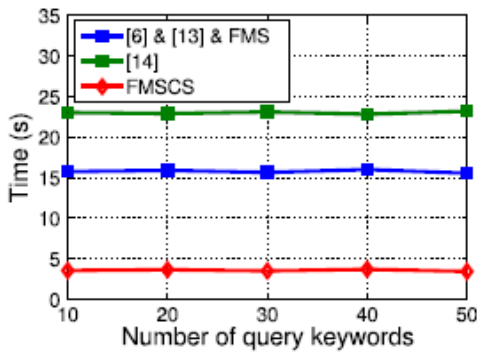
The relevance score and k-nearest neighbor techniques [3] to develop an effective multi keyword search scheme that can return the ranked search results depended on the accuracy. Within fine grained multi keyword search leverage an efficient index to further improve the search efficiency and adopt the blind storage system to conceal access of the search user. An authorized and ranked multi-keyword search scheme (ARMS) [4] over encrypted cloud data by exploit the cipher text policy attribute-based encryption (CP-ABE) and SSE techniques. Security analysis demonstrates that the proposed ARMS scheme can achieve collusion resistance. In this paper presents the FMS schemes which not only support multi-keyword search over encrypted data, but also achieve the fine-grained keyword search to investigate the relevance scores and the preference factors of keywords and the logical rule of keywords.



(a)



(b)



(c)

## CONCLUSION

In this paper, to investigated on the fine-grained multi-keyword search (FMS) problem in encrypted data and it can be describes the two FMS schemes. First scheme is FMS I take in both the relevance scores and preference factors of keywords to give efficient search to the search users. Second schemes is FMS II achieve secure and effective search with practical functionality like as AND, OR and NO operations of keywords. In this scheme, multiple user can access the same encrypted data at same time is very hard situation and it means data owner to send the symmetric and secret key to multiple search user is very difficult. For the future work, can be consider the extensibility of the file set and the multi-user cloud environments. One more future is developing the highly scalable searchable encryption to enabling effective search on large databases.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] DX Song, D Wagner, A Perrig.[2000] Practical techniques for searches on encrypted data, in Proc. S&P, IEEE, pp. 44–55.
- [2] H Li, D Liu, Y Dai, TH Luan, X Shen .[2014] Enabling efficient multi-keyword ranked search over encrypted cloud data through blind storage, IEEE

- Trans. Emerging Topics Comput., 2014, DOI:10.1109/TETC.2014.2371239.
- [3] Boyang Wang, Yantian Hou, Ming Li. [2016] Practical and Secure Nearest Neighbor Search on Encrypted Large-Scale Data, The 35th Annual IEEE International Conference on Computer Communications, IEEE INFOCOM.
- [4] [4] H. Li, D. Liu, K. Jia, and X. Lin. [2015] Achieving authorized and ranked multi-keyword search over encrypted cloud data, in *Proc IEEE Int. Conf Commun.*, to be published.
- [5] Y Yang, H Li, W Liu, H Yang, M Wen. [2014] Secure dynamic searchable symmetric encryption with constant document update cost,” in Proc. IEEE GLOBECOM, 775–780.
- [6] Neelam S Khan, C Rama Krishna, Anu Khurana. [2014] Secure Ranked Fuzzy Multi-Keyword Search over Outsourced Encrypted Cloud Data, 2014 5th International Conference on Computer and Communication Technology, 978-1-4799-6758-2/14/\$31.00 ©2014 IEEE.
- [7] J Yu, P Lu, Y Zhu, G Xue, M Li. [2013] Towards secure multi keyword top-k retrieval over encrypted cloud data, *IEEE Trans. Dependable Secure Comput.*, 10(4): 239–250, Jun..
- [8] W Sun, B Wang, N Cao, M Li, W Lou, YT Hou, H. Li. [2014] Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking, *IEEE Trans. Parallel Distrib Syst*, 25(11): 3025–3035
- [9] Zhihua Xia. [2016] A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data,” *IEEE transactions on parallel and distributed systems*, 2016 to be published.
- [10] B. Zhang and F. Zhang, “An efficient public key encryption with conjunctive-subset keywords search,” *J. Netw. Comput. Appl.*, vol. 34, no. 1, pp. 262–267, 2011.
- [11] H Liang, LX Cai, D Huang, X Shen, D Peng. [2012] “An smdp-based service model for inter domain resource allocation in mobile cloud networks,” *IEEE Transactions on Vehicular Technology*, 61(5): 2222–2232.
- [12] W Sun, S Yu, W Lou, YT Hou, H Li. [2014] “Protecting your right: Attribute based keyword search with fine-grained owner-enforced search authorization in the cloud,” in *Proceedings of INFOCOM*. IEEE, 2014.
- [13] Y Yang, H Li, W Liu, H Yang, M Wen. [2014] Secure dynamic searchable symmetric encryption with constant document update cost ”in *Proceedings of GLOBECOM, USA*
- [14] Y Yang, H Li, M Wen, H Luo, R Lu. [2014] Achieving ranked range query in smart grid auction market,” in *Proceedings of ICC*, pp.951–956.
- [15] N Cao, C Wang, M Li, K Ren, W Lou. [2014] “Privacy-preserving multi-keyword ranked search over encrypted cloud data,” *IEEE Trans. Parallel Distrib. Syst.*, 25( 1): 222–233
- [16] <https://support.google.com/websearch/answer/173733?hl=en.2014>.
- [17] CR Barde, Pooja Katkade, Deepali Shewale, Rohit Khatale. [2014] Secured Multiple-keyword Search over Encrypted Cloud Data, Department of Computer, Pune University, GESRHSCOE, Nashik, India, 2014.

# AN APPRAISAL ON GAS LEAKAGE DETECTION AND CONTROLLING SYSTEM IN SMART HOME USING IOT

S. Sivaranjani\*, D. Gowdhami, P. Karthikkannan

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tiruchengode, Tamilnadu, INDIA

## ABSTRACT

**Aim:** The Internet of Things (IoT) is playing a vital role in most of the fields all over the world, especially in smart home automation system. The IoT smart home system runs on conventional Wi-Fi network, Bluetooth or Internet to be implemented based on Embedded Microcontroller and sensor environment. Now a days the major crisis in residential premises is the leakage of Liquefied petroleum gas. Due to this hindrance gas sensor can be deployed into the smart home environment and after the detection of gas leakage the SMS (short message service) is sent to authorized user by GSM, then electric power supply is turned off with the help of fire sensor through relay control. After the detection of gas leakage in the smart home, window is opened by means of window sensor and the concentration of gas gets reduced slowly. Hence the survey is done here on the gas leakage detection and controlling system to overcome this trouble. **Conclusion:** Hence the occurrence of accidents due to the LPG gas leakage is controlled by using the proposed technique which is listed in the possible solution section.

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

IoT (Internet of Things), smart home system, gas sensor, fire sensor, window sensor, PIC microcontroller, GSM.

\*Corresponding author: Email: [sivaranjinisiva16@gmail.com](mailto:sivaranjinisiva16@gmail.com); Tel.: +91 8220356140

## INTRODUCTION

The inter-connection between the physical devices like buildings, vehicles and other devices which are implanted with electronic devices, software, sensors, actuators, and network connectivity which helps and makes these objects to gather and exchange data is called as Internet of Things. When IoT is augmented with sensors and actuators, it also covers the following technologies such as smart grids, smart homes, intelligent transportation and smart cities so that the technology becomes an illustration for more general class of cyber-physical systems [1].

However, the rate of IoT adoption among home users depends on their willingness to purchase these devices, and convenience and security are identified to be the two key factors influencing their decision. As such the design and implementation of a Wi-Fi Bluetooth and or internet based IoT smart home system that uses a gateway to enable secure communication between IoT devices, and to also allow user to configure, access and control the system through user friendly interface running on mobile devices such as the ubiquitous smart phone. A smart home is also one of the applications of IoT. Even though there is a rapid growth in technologies and improvements in architecture, it comes out with many problems like how to manage and control the whole system, server side security, security in smart homes etc., Smart homes are those where household devices/home appliances could monitor and control remotely [2].

### Gas sensor

A gas sensor which is made on the basis of catalytic principle is called catalytic gas sensor. The output of catalytic gas sensor is measured by a Wheatstone bridge. By monitoring the resistance change of the platinum resistance arising from increase in temperature the concentrations of gases can be detected. Following are some of the gas sensors like Carbon dioxide sensor, Carbon monoxide detector, Electro chemical gas sensor [3].

### Fire sensor

The presence of a flame or fire is detected with the help of flame detector by means of sensor is designed to detect and respond accordingly. Responses to a detected flame are based on the installation, but can include sounding an alarm, switching off a fuel line (such as a propane or a natural gas line), and turning on a fire suppression system [4].

## RELATED WORK ON THE EXISTING SYSTEM

The following section describes about the review which is made on the smart home automation, different sensor's used in smart home automation.

### Related Work

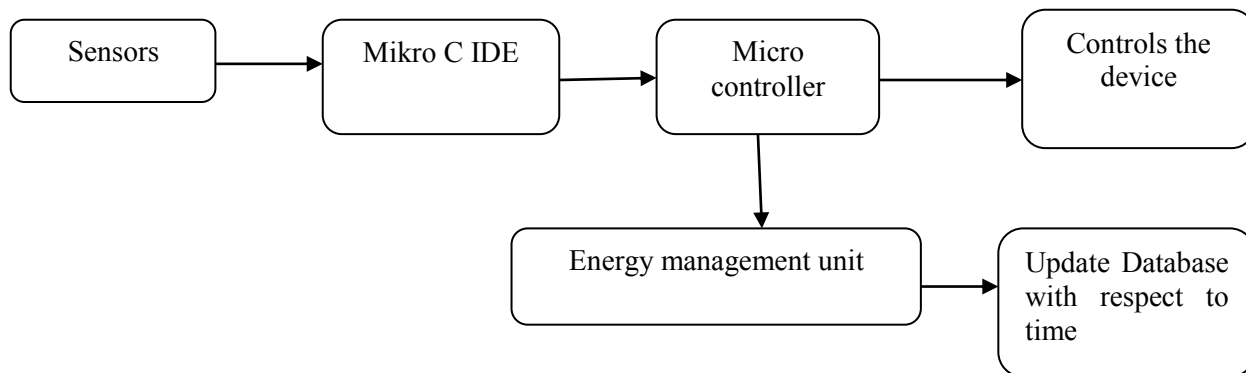
In [5], “**HOME AUTOMATION USING INTERNET OF THINGS**”, the authors “Vinay sagar K N, Kusuma S M” described about the Wireless Home Automation system(WHAS) by IoT system in which it is sometimes called as s called a smart home. The smart home system is comprised of computers or mobile devices that are used to control basic home functions and features automatically through internet from anywhere around the world. Hence with the help of the smart home system can able to save the electric power and human energy. In this method the authors presented a home automation system (has) using Intel Galileo that employed the integration of cloud networking and wireless communication. The main advantage of this scheme is that, it provided the user with remote control of various lights, fans, and appliances within their home and storing the data in the cloud. Based on the data gathered from the sensor the system will be changed automatically and also this system was designed to be low cost and expandable allowing a variety of devices to be controlled.

In [6], “**INTERNET OF THINGS (IOT) BASED REAL TIME GAS LEAKAGE MONITORING AND CONTROLLING**” the authors “Hina ruqsar, Chandana R, Nandini R, Dr. T P Surekha” discussed about xively which is a secured scalable platform that includes directory services, data services, a trust engine for security, and web-based management application. It also helps to provide a general ground through which any external device connected to the internet cloud actually communicates with any other device. Xively is an old fixture within the internet of things ecosystem. Gas sensor senses the gas leakage and alerts the house rescue teams which are buzzer and exhaust fan. From this the gas will be changed and replaced from the interior. A solenoid valve is an electro mechanical device that is used for controlling liquid or gas flow. When the coil is keyed up, magnetic field is created, causing a plunger inside the coil to move. The valve will return to its de-energized state, when electrical current is removed from the coil. Here using this concept one can easily know the exact date and time of the hazard.

In [7], “**EMBEDDED SYSTEM FOR HAZARDOUS GAS DETECTION AND ALERTING**”, the authors “V.Ramya, B. Palaniappan” focused on how the microcontroller of designed and worked based on detecting toxic gas and then alerting the system. Here the embedded system technique was used to determine how the hazardous gases like LPG and propane were sensed and displayed each and every second in the LCD display. Then an alarm is generated immediately and also an alert message (SMS) is sent to the authorized person through the GSM when the gas exceeds its normal level. The benefit of this automated detection and alerting system over the manual method is that the response time and accurate detection of an emergency is done very quickly and in turn it also leads faster diffusion of the critical situation.

In [8], “**GSM BASED LPG LEAKAGE DETECTION AND CONTROLLING SYSTEM**” the authors “Prof.M.Amsaveni, A.Anurupa, R.S.Anu Preetha, C.Malarvizhi, M.Gunasekaran” discussed on how to detect and control the LPG gas which mainly comprised with butane and propane. In this scheme the authors used MQ6 gas sensor to detect the leakage of gas. As soon as the leakage is detected the sensor sends a signal to the microcontroller in which it sends an active signal to other devices which are connected externally. Then the alert message has been sent to the user through GSM module. The advantage of this technique is it reduces the concentration of gases. Even though this technique reduces the concentration of the gas the demerit of this system is efficiency of using the microcontroller which is used here is less and also it requires changes in program whenever multiple SMS is to be sent at a time.

In [9], “**AUTOMATION AND ENERGY MANAGEMENT OF SMART HOME USING Lab VIEW**” the authors “J.Ashley Jenifer, T.Sivachandrabanu, A.Darwin Jose Raju” discussed about how the information from the sensors has been fed into the pc. Here in this system the photo voltaic installation is made to tackle the energy. Then the Mikro C IDE was used to bind the sensors actuators and devices. Also the Lab VIEW (laboratory virtual instrumentation engineering) was done to visualize the home automation like lighting, temperature, security, gardening and energy management.



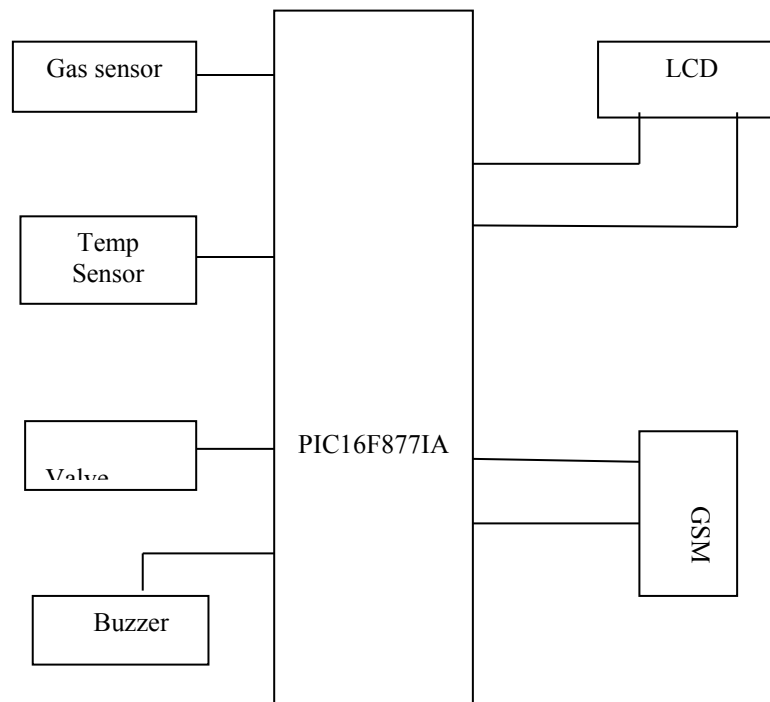
**Fig. 1: Diagrammatic representation of the smart home automation using Mikro C IDE**

In [10], “SMART HOME MONITORING AND CONTROLLING SYSTEM USING ANDROID PHONE” the authors “Gowthami.T, Dr. Adiline Macriga G” described about the home monitoring and controlling by using android phone. In this method the author used both PC and android phone to monitor and control the home for safety, security and human health care. They also used zigbee module for the Lab VIEW. The personal computer is used to monitor the various parameters in the proposed system. In this system, the various parameters are monitored and controlled with the help of android phone. Android phone has the more benefit when compared to personal computer for using at any place. Then here the information from the sensors which is from various appliances are first passed to the home monitoring server and from there the data is transferred to the local monitoring server. Hence the sensor node is responsible for collecting, processing and transmitting the data into the local monitoring server which is embedded with the web server.

In [11 &12], “LPG GAS LEAKAGE DETECTION & CONTROL SYSTEM AND DESIGN AND DEVELOPMENT OF KITCHEN GAS LEAKAGE DETECTION AND AUTOMATIC GAS SHUT OFF SYSTEM” the authors discussed on how they dealt on home security based on LPG gas leakage in the kitchen. In both these techniques the authors mainly concentrated on the controlling of detected LPG gas as well as on the security in home against intruders and fire accident. When the gas leakage was detected the sensor sends signal to PIC microcontroller and the system more like a first aid and a normally closed solenoid valve for the shutting off of the gas valve is used before calling for help via visual display and audible alarm to those within the environment. Since the system is an intelligent system that does not create a noise nuisance by continuously sounding alarm but the alarm stops beeping once the concentration of the gas in the atmosphere after leakage goes below the set point and opens the valve again for normal operations. The advantage of this work is it minimized injuries/losses occasioned by explosions due to gas leakages, which also improved safety of life and property while using domestic cooking gas. The following block diagram shows the gas leakage detection system which uses PIC microcontroller.

In [13], “SMART HOME AUTOMATION BASED ON DIFFERENT SENSORS AND ARDUINO AS THE MASTER CONTROLLER” the author “Subhankar Chatteraj” focused on how arduino, graphical user interface controls the smart home with the help of different sensors. Here the idea proposed was with a low cost solution using off the shelf components to decrease its cost and open source software to obtain around licensing requirements of software. An arduino controls sensors and actuators that monitors a defined location and take action based on specified parameters like ambient light, temperature etc. It was also focuses on sensing the temperature, gas leakage and smoke, light detection.





**Fig.2: Block Diagram of Gas Leakage System using PIC Microcontroller**

In [14], “SMART HOME FOR ELDERLY CARE, BASED ON WIRELESS SENSOR NETWORK” the authors “Rasika S. Ransing, Manita Rajput” said that in a smart home, sensors are used for monitoring general parameters like temperature, humidity, LPG leakage, etc. Thus, with the development of wireless network technology, the data rate will be low, battery life will last for long time, complex protocols are less in number for such applications as an alternative to wasting bandwidth of high data rate protocols. Then here the authors used WI-FI, Zigbee which are the short distance communication technologies. For the system proposed in this system, Zigbee technology was employed to detect the above mentioned parameters. Zigbee is a worldwide open standard for wireless radio networks in the monitoring and control fields. The development of Zigbee technology was done by the IEEE 802.15.4 committee. So with the help of Zigbee technology information from the sensors are gathered and controlled. Hence by using Zigbee technology in the battery-powered applications the usage of data rate will be lower, cost is less, and also battery life is long. The lower data rate of the Zigbee devices allows for better sensitivity and range, but it offers less throughput only which is considered as the main disadvantage of this technology.

In [15], “DESIGN AND IMPLEMENTATION OF A SMART FIRE ALARM SYSTEM USING GSM TECHNOLOGIES VIA SHORT MESSAGES SERVICE” the authors discussed on how the fire has been controlled in smart home. The gas smoke sensor keeps on sensing and if there occurs any fire the SMS will be sent to user through GSM with the help of (T-BoxN12R device). Then after the alarm notification has been sent, the T-BoxN12 device turns on the alarm buzzer and the door will be closed. After this the water pumping motor will be turned on. If so any fire occurred also it will be controlled by this system.

This data transfer may be done easily through two communication protocols namely, the short messaging service (SMS) and wireless application protocol (WAP). These two technologies are complementary. SMS messages have long been familiar to the mobile phone users and are affordable. The use of SMS extends the data transfer to a larger number of telephone sets. WAP is an open protocol for wireless messaging. It provides the same technology to all vendors regardless of the network system. This means that there will be WAP compliant terminals from several manufacturers.

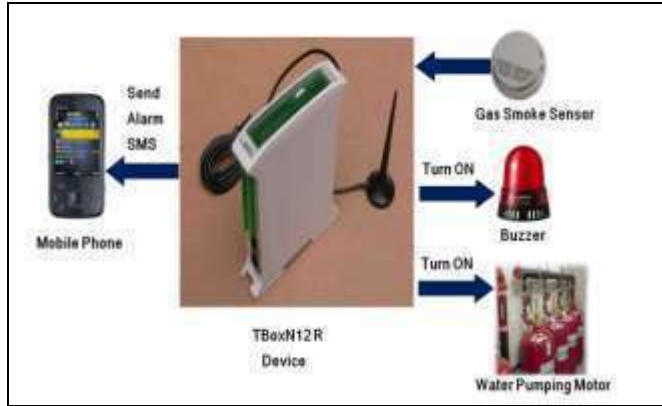


Fig. 3. The System Block Diagram

GRAPHICAL REPRESENTATION FOR EXISTING SYSTEM ANALYSIS

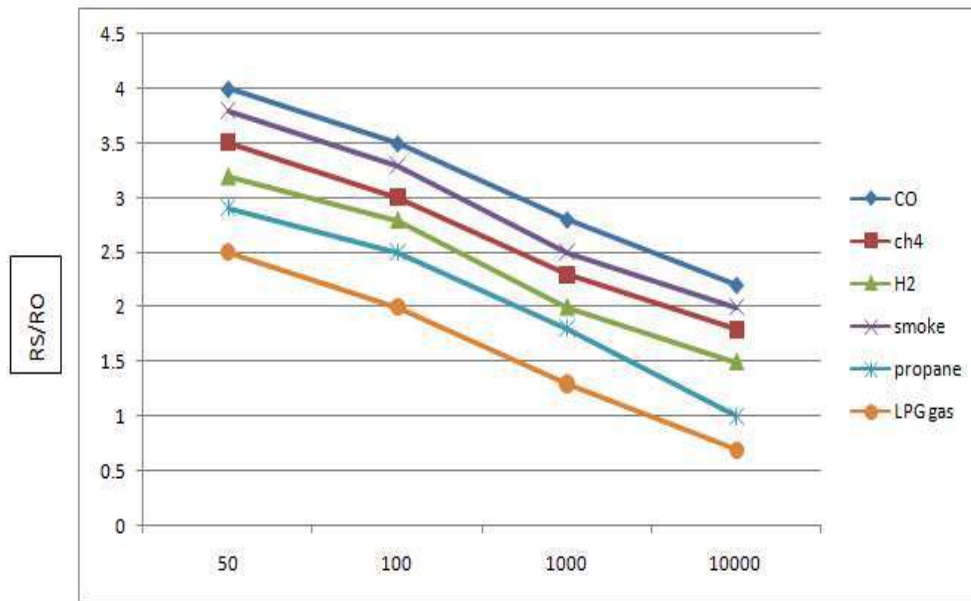


Fig. 4: Sensitivity Characteristics of MQ2 Sensor

The above graph represents the graphical representation for the existing system analysis. This graph shows the typical sensitivity characteristics of MQ2 sensor for various gases which are used in home as well as in industries. Then MQ2 sensor is used to detect LPG gas, carbon monoxide, hydrogen, smoke, propane. Based on the response time measurements are taken as soon as possible. The graph is drawn between the resistance value and the concentration of the gas.

Where as in graph, RO represents the sensor resistance at 1000 ppm of h<sub>2</sub> in clean air and RS represents the sensor resistance at various concentration of gas. Then this MQ2 sensor has different resistance value in different concentration.

COMPARATIVE ANALYSIS ON EXISTING SYSTEM

In this section the analysis is done to make a view on the different sensors, controller devices and their purposes which helps to overcome the hindrance occurred during the LPG gas leakage in smart home system.

**Table: 1. Purpose of Various Sensors, Controller Devices Used in Smart Home System**

S.NO	Title of the Paper	Platform/ controller device	Sensor's/device used	Purpose
1	Home Automation Using Internet of Things <sup>[5]</sup>	Cloud networking, Wi-Fi network, Intel Galileo Microcontroller	Temperature sensor, Gas sensor, Motion sensor.	The parameters and threshold of the sensors are read. Those data are sent to web server and stored in the cloud which can be analyzed anywhere at any time.
2	Internet of things (IoT) based real time gas leakage Monitoring and controlling <sup>[6]</sup>	Xively	Electronic sensor	Secure, scalable platform which stores data. Stored data helps to detect the root cause of the gas leakage by electronic sensors with accurate time and date.
3	Embedded system for hazardous gas detection and alerting <sup>[7]</sup>	PIC16F877 Microcontroller	Combustible gas sensor	The hazardous gases like LPG and propane has been detected. With the sensor used, offers quick response time and accurate detection of an emergency and in turn leading faster diffusion of the critical situation by alerting the authorized user's via GSM.
4	GSM based LPG leakage detection and controlling system <sup>[8]</sup>	PIC Microcontroller (PIC16F877A)	Gas sensor MQ-6	The leakage of LPG gas is detected by the MQ6 gas sensor. Its analog output is given to the microcontroller. Then the SMS will be sent to the user through GSM and the leakage is controlled up to 0.001%.
5	Automation and Energy management of Smart home using Lab VIEW <sup>[9]</sup>	MICRO C IDE. Micro controller	Energy management sensors	The complete LABVIEW of the smart home has been done in order to display the energy management of fan, air conditioner, light.
6	Smart Home Monitoring and Controlling System Using Android Phone <sup>[10]</sup>	ZigBee	PC, android phone, home appliance sensors.	Data acquisition has been done using sensor nodes. A node contains three sensors. The sensors are used for monitoring the physical parameter measurements. In the home monitoring, the parameters such as kitchen temperature, gas and obstacle are monitored.
7	LPG Gas Leakage Detection & Control System <sup>[11,12]</sup>	RISC Microcontroller( version of PIC)	MQ 6 Gas sensor	LPG gas leakage is detected accurately with the quick and fast response time. For safety from gas leakage in heating gas fired appliances like boilers, domestic water heaters and also for cooking gas fired appliances like oven, stoves, etc.
8	Smart Home Automation based on different sensors and Arduino as the master controller <sup>[13]</sup>	Arduino board	Temperature Sensor (LM35), LPG and Smoke Sensor (MQ2), Temperature and	The temperature sensor controls the smart home. Gas sensor detects the voltage level of the smoke compared with the threshold.

			Humidity sensor (DHT11)	Humidity sensor and a thermistor to measure the surrounding air.
9	Smart Home for Elderly Care, based on Wireless Sensor Network <sup>[14]</sup>	ZigBee , Arduino MEGA 2560	Temperature sensor LM 35, Door sensor	Precise the centigrade of the temperature which in turn says about the chance of fire occurrence. LM 35 measures the temperature level from the range of -55° c to +150° c. Door sensor is to detect the position of the door, whether it is opened or closed.
10	Design and Implementation of a Smart Fire Alarm System Using GSM Technologies Via Short Messages Service <sup>[15]</sup>	T-BoxN12R microcontroller	Gas smoke sensor, Fire sensor	It detects the presence of smoke, the door will be turned closed and the water pumping motor is turned to put off the fire occurred.

## POSSIBLE SOLUTION

By using the stepper motor which can be connected to microcontroller will slowly reduce the concentration of the LPG gas in the room of the smart home when it leaks.

With the help of the window sensor in the smart home, when the leakage of LPG gas is detected window sensor will automatically open the window as soon as the alert(SMS)message has been sent to the authorized user .then, when the window is opened the exhausting fan also turned on which leads to push out the smoke outside through opened window .when the alert (SMS)message is sent to the authorized user through GSM the fire sensor also turned through relay in order to check the occurrence of fire.

## CONCLUSION

Internet of things is nothing but the interconnection of physical devices which are embedded with the electronic devices, sensors, etc. In which security is very important in the smart home system and here our work mainly concentrated on the controlling of LPG gas in smart home system. Hence with the help of the survey made, we came to know about the different sensors and their purposes used in the smart home automation. From the above analysis, we can introduce window sensor in smart home so that when the gas leakage is detected the information is given to window and the window will be opened automatically through the sensor which reduces the concentration of the LPG gas and therefore prevents the occurrences of accident.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Luigi Atzori, Antonio Iera, Giacomo Morabito. The Internet of Things: A survey. *Journal homepage: www.elsevier.com/locate/comnet.*
- [2] Mikael Asplund, (Member, IEEE), AND Simin Nadjm-Tehrani, (Member, IEEE). "Attitudes and Perceptions of IoT Security in Critical Societal Services". Received February 28, 2016, accepted April 5, 2016, date of publication April 29, 2016, date of

- current version May 23, 2016. Digital Object Identifier 10.1109/ACCESS.2016.2560919.
- [3] Xiao Liu, Sitian Cheng, Hong Liu, Sha Hu, Daqiang Zhang and Huansheng Ning. "A Survey on Gas Sensing Technology". ISSN 1424-8220 www.mdpi.com/journal/sensors. 12:635-9665; doi: 10.3390/s120709635.
- [4] [https://en.wikipedia.org/wiki/Flame\\_detector](https://en.wikipedia.org/wiki/Flame_detector)
- [5] Vinay sagar K N, Kusuma SM. [2015] Home Automation Using Internet of Things". International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056, 02(03) | June-2015 www.irjet.net p-ISSN: 2395-0072.
- [6] Hina ruqsar, Chandana R, Nandini R, TP Surekha.[2014] Internet of Things (IoT) based real time gas leakage monitoring and controlling. *International Journal of Electronics and Communication Engineering & Technology (IJCET)*. 5(8) : 208-214.
- [7] V.Ramya, B Palaniappan.[2012] Embedded system for hazardous gas detection and alerting". *International Journal of Distributed and Parallel Systems (IJDPS)*.3(3)
- [8] M Amsaveni, A Anurupa, R.S.Anu Preetha, C.Malarvizhi, M.Gunasekaran.[2015] GSM Based LPG Leakage Detection And Controlling System. *The International Journal Of Engineering And Science (IJES)* ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805 : 112-116
- [9] J.Ashley Jenifer, T.Sivachandrabanu, A.Darwin Jose Raju. "Automation and Energy management of Smart home using LabVIEW". 978-1-4673-9925-8/16/\$31.00 ©2016 IEEE.
- [10] Gowthami.T, Adiline macriga G.[2013] Smart Home Monitoring and Controlling System Using Android Phone". *International Journal of Emerging Technology and Advanced Engineering*. Website: www.ijetae.com (ISSN 2250-2459, ISO9001:2008 Certified Journal, 3 (11)
- [11] Hitendra Rawat, Ashish Kushwah, Khyati Asthana, Akanksha Shivhare. "LPG Gas Leakage Detection & Control System". National Conference on Synergetic Trends in engineering and Technology (STET-2014). *International Journal of Engineering and Technical Research* ISSN: 2321-0869, Special Issue.
- [12] Apeh S.T, Eramah K.B and Iruansi U."Design and Development of Kitchen Gas Leakage Detection and Automatic Gas Shut off System". *Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS)* 5(3): 222-228. © Scholarlink Research Institute Journals, 2014 (ISSN: 2141-7016).
- [13] Subhankar Chattoraj. "Smart Home Automation based on different sensors and Arduino as the master controller". *International Journal of Scientific and Research Publications*, Volume 5, Issue 10, October 2015 ISSN 2250-3153.
- [14] Rasika S. Ransing, Manita Rajput." Smart Home for Elderly Care, based on Wireless Sensor Network". 2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2015).
- [15] Jayashri Bangali and Arvind Shaligram. [2013] Design and Implementation of a Smart Fire Alarm System Using GSM Technologies via Short Messages Service". *International Journal of Smart Home* 7(6):201-208. <http://dx.doi.org/10.14257/ijsh.2013.7.6.19>.

# A SURVEY ON VARIOUS APPROACHES FOR VERIFYING CORRECTNESS AND COMPLETENESS OVER THE CLOUD DATA

S. Suvetha\*, V. Arul, G. Sasikala

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for women, Namakkal, T.N., INDIA

## ABSTRACT

**Aim:** Cloud computing is popularizing the computing archetype in which data is outsourced to a third-party service provider (server) for data mining Outsourcing. However, it raises a serious security issue: how can the client of weak computational power verify that the server returned correct mining result or not. By using homomorphic encryption algorithm can check the completeness and correctness of retrieved data from the server.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Verifiable computing, Discovery,  
Data Mining as a Service.

\*Corresponding author: Email: [suvethajan14@gmail.com](mailto:suvethajan14@gmail.com); Tel.: +9500952218

## INTRODUCTION

These days, the IT world is moving towards the pay-per-use paradigm named Cloud Computing. Companies of all sizes reduce their computing assets and shift to a use of computing resources in the clouds [1]. One consequence of this shift is that the IT world outside the clouds is moving to a use of weaker and smaller computer devices, like Virtualized Thin Desktops and Smart phones. Whenever stronger resources are needed, those devices can use the cloud. Data mining-as a service expertise to outsource their data mining needs to a third part service provider [4]. As an example, the operational transactional data from various stores of a supermarket chain can be shipped to a third party which provides mining services [3]. The supermarket management need not employ an in-house team of data mining experts [2]. Besides, they can cut down their local data management requirements because periodically data is shipped to the service provider who is in charge of maintaining it and conducting mining [6] on it in response to requests from business analysts of the supermarket chain. It is generally expected that the paradigm of “mining and management of data as service” will grow with the advent and popularity of cloud computing.

### Cloud Computing

Cloud computing is a type of Internet-based computing that provides shared computer processing resources and data to computers and other devices on demand. Cloud services are given by an outsider entity, has many security and integrity problems [2]. Security in this context means that the client will receive an assurance that the computation performed by the server is correct, with the optional property that the client will be able to hide some of his data from the server [3]. Thus, the client must have some way of verifying the cloud’s computation. However, one basic problem is inherent in the model: How can a weak client verify the correctness of the cloud’s computation? Can the client be assured that the cloud server follows its declared strategy? These questions are not easily answered by the existing tools of security and cryptography [7]. There are many possible reasons for a cloud to cheat on answers. For example When perform a web search, expect that the list of links returned will be relevant and complete. As heavily rely on web searching, an often overlooked issue is that search engines are outsourced computations [13]. That is, users issue queries and have no intrinsic way of trusting the results they receive, thus introducing a modern spin on Cartesian doubt. This philosophy once asked if can trust our senses now it should ask if can trust our search results.

Some possible attack scenarios that arise in this

context include the following:

1. A news web site posts a misleading article and later changes it to look as if the error never occurred.
2. A company posts a back-dated white paper claiming an invention after a related patent is issued to a competitor.
3. An obscure scientific web site posts incriminating data about a polluter, who then sues to get the data removed, in spite of its accuracy.
4. A search engine censors content for queries coming from users in a certain country, even though an associated web crawler provided web pages that would otherwise be indexed for the forbidden queries[5][8].

### Verifiable Computing

The cloud would like to improve its revenue by computing things with minimal resources while charging for more [9]. This problem of verifiable computation was tackled in many previous works in the theoretical computer science community, most notably by using Probabilistically Checkable Proofs[12]. Other recent works use fully homomorphic encryption and get amortized performance advantages. Today, a common way to verify computations is replication. However, replication may not be verifiable. It also requires assumptions about failure independence. Another technique is auditing but if the performer understands the computation better than the requester, the performer can alter strategic bits, undetected by an audit[20]. A final technique is trusted computing, but it assumes that some component the hardware, the hypervisor, a higher layer is not physically altered. Propose homomorphic encryption verification approaches to check whether the server has returned correct and complete frequent itemsets[25]. Our homomorphic encryption approach can catch incorrect results with high probability, while our deterministic approach measures the result correctness with 100 % certainty. It also design efficient verification methods for both cases that the data and the mining setup are updated. It demonstrate the effectiveness and efficiency of our methods using an extensive set of empirical results on real datasets[12]. An interesting direction to explore is to extend the model to allow the client to specify her verification needs in terms of budget (possibly in monetary format) besides precision and recall threshold.

### LITERATURE SURVEY

The following papers are surveyed. It consists of various verification approaches. These approaches are how the client of weak computational power verify that the server returned correct mining results or not. These following approaches are used to verify the retrieved result from the server.

In[1] R. Canetti, B. Riva, and G. N. Rothblum, presented a paper on “**Verifiable computation with two or more clouds**” server using one cloud, the client uses two or more different clouds to perform the computation[1]. The client can verify the correct result of the computation, as long as at least one of the clouds is honest. It believes that such addition suits the world of cloud computing where cloud providers have incentives not to collude, and the client is free to use any set of clouds he wants. Our results are two fold[7]. First, they show two protocols in this model: 1. A computationally sound verifiable computation for any efficiently computable function, with logarithmically numerous rounds, based on any collision-resistant hash family. 2. A 1-round (2-messages) unconditionally sound verifiable computation for any function computable in log-space uniform NC. Second, It show that our first protocol works for essentially any sequential program, and they present an implementation of the protocol, called quin, for Windows executables. Also describe its architecture and experiment with several parameters on live clouds[20].

In[2] F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and W. Hui Wang, presented a paper on “**Privacy-preserving data mining from outsourced databases**”, the problem of outsourcing the association rule mining task within a corporate privacy-preserving framework[19]. Association rule mining has the objective of discovering groups of products, or items, that are repeatedly purchased together by the supermarket’s customers: the expected output of such task, given the sale transaction database as input, is the list of all possible groups of things, such as {milk, beer, diapers}, that occur together in a fraction of the market baskets that is statistically significant[2]. The complexity of this task is evident: there are tens of thousands of distinct products in the variety of a supermarket, and therefore the number of potential candidate groups of products quickly explodes with the size of the group[23]. Our encryption scheme is based on 1–1 substitutions together with addition of fake transactions such that the transformed database satisfies k-anonymity with respect to items and itemsets. This

computational complexity motivates the introduction of an outsourcing model, where the data owner, like our supermarket, gives the data in outsourcing to a service provider to obtain an association rule mining service from it, within a privacy-preserving framework, i.e., without disclosing neither the sale data nor the information deriving from the mining analysis[25].

In[3] R. Liu, H. Wang, A. Monreale, D. Pedreschi, F. Giannotti, and WengeGuo, presented a paper on **“Audio: An integrity auditing framework of Outlier mining- as-a-service systems”** The AUDIO, an integrity auditing framework for the specific task of distance-based outlier mining outsourcing[3]. It provides efficient and practical verification approaches to check both completeness and correctness of the mining results. The key idea of our approach is to insert a small amount of artificial tuples into the outsourced data[18]. The artificial tuples will produce artificial outliers and non-outliers that do not exist in the original dataset. The server’s answer is verified by analyzing the presence of artificial outliers/non-outliers, obtaining a probabilistic guarantee of correctness and completeness of the mining result. Our empirical results show the effectiveness and efficiency of our method [5].

In[4] S. Setty, A. J. Blumberg, and M. Walfish, presented paper on **“Toward practical and unconditional verification of remote computations”** it propose a new line of systems research: using the machinery of PCPs, can build a system that (i) has practical performance, (ii) is simple to implement, and (c)provides unqualified guarantees? Note that (a) and (b)contrast with PCPs as used in the theory literature and (c)contrasts with current systems approaches. To illustrate the promise of this line of research, do the following: (1) Identify work in the PCP literature that provides a base for systems research [4]. First looked to the PCP literature, then observed that efficient argument systems (PCP variants in which the server proves that it has a proof by answering questions interactively) are promising, and then noticed that a particular argument system could lead to a practical solution[19]. (2) Refine the approach of into a design that is practical over a limited domain . It applied refinements to shrink program encoding (via arithmetic circuits instead of Boolean circuits), enable batched proofs (which enhances performance for computations that can be decomposed into parallel pieces), and improve amortization (by moving more of the work to a setup phase). These innovations are essential to practical performance. (3) Implement this design to demonstrate its practicality. To our knowledge, PCP theory has never before found its way into any efficient implementation. Thus, believe that our implementation, though limited, is a contribution. Our implementation is also comparatively simple; it could conceivably be formally verified. (4) Articulate a research agenda for extending the reach of our approach . Our ultimate goal is a practical system for general-purpose verified computation [17]. The four contributions above provide a concrete foundation for our position, which is that PCP-based verifiable computation can be a systems problem, not just a theory problem. They need this foundation because PCPs are thought to be impractical; indeed, our prior designs were too expensive by over 11 orders of magnitude. Even our prototype achieves goals (a)–(c) above only over a limited domain [22].

In[5] S. Benabbas, R. Gennaro, and Y. Vahlis, presented a paper on **“Verifiable delegation of computation over large datasets”** this learn the problem of computing on large datasets that are stored on an untrusted server[5]. They follow the approach of amortized verifiable computation introduced by Gennaro, Gentry, and Parno. This present the first practical verifiable computation scheme for high degree polynomial functions[23]. Such functions can be used, for example, to make predictions based on polynomials fitted to a large number of sample points in an experiment. Our second result is a primitive which call a verifiable database (VDB). Here, a weak client outsources a large table to an untrusted server, and makes retrieval and update queries. For each query, the server provides a response and a proof that the response was computed correctly[26]. The goal is to minimize the resources required by the client. This is made particularly challenging if the number of update queries is unbounded. It presents a VDB scheme based on the hardness of the subgroup membership problem in composite order bilinear groups[19].

In[6] R. Canetti, B. Riva, and G. N. Rothblum, presented a paper on **“Practical delegation of computation using multiple servers”**this demonstrate a relatively efficient and general solution where the client delegates the computation to several servers, and is guaranteed to determine the correct answer as long as even a single server is honest [6]. It show: A protocol for any efficiently computable function, with logarithmically many rounds, based on any collision resistant hash family[10]. The protocol is set in terms of Turing Machines but can be adapted to other computation models. An adaptation of the protocol for the X86 computation model and a prototype implementation, called Quin, for Windows executables[21]. It describe the architecture of Quin and experiment



with several parameters on live clouds. Also show that the protocol is practical, can work with nowadays clouds, and is efficient both for the servers and for the client.

In[7] D. Fiore and R. Gennaro, presented a paper on **“Publicly verifiable delegation of large polynomials and matrix computations, with applications”** The Outsourced computations (where a client requests a server to perform some computation on its behalf) are becoming increasingly important due to the rise of Cloud Computing and the proliferation of mobile devices[7]. Since cloud providers may not be trusted, a crucial problem is the verification of the integrity and correctness of such computation, possibly in a public way, i.e., the result of a computation can be verified by any third party, and requires no secret key { akin to a digital signature on a message. It present new protocols for publicly verifiable secure outsourcing of Evaluation of High Degree Polynomials and Matrix Multiplication[28]. It can be used for amortized model. Optimal Verification of Operations on Dynamic Sets it present a new authenticated data structure scheme that allows any entity to publicly verify the correctness of primitive sets operations such as intersection, union, subset and set difference. Based on a novel extension of the security properties of bilinear-map accumulators as well as on a primitive called accumulation tree, our authenticated data structure is the first to achieve optimal verification and proof complexity (i.e., only proportional to the size of the query parameters and the answer), as well as optimal update complexity (i.e., constant), and without bearing any extra asymptotic space overhead[13]. Queries (i.e., constructing the proof) are also efficient, adding a logarithmic overhead to the complexity needed to compute the actual answer.

In[8] M. T. Goodrich, C. Papamanthou, D. Nguyen, R. Tamassia, C. V. Lopes, O. Ohrimenko, and N. Triandopoulos, presented a paper on **“Efficient verification of web-content searching through authenticated web crawlers,”**It consider the problem of verifying the correctness and completeness of the result of a keyword search[8]. They introduce the concept of an authenticated web crawler and present its design and prototype implementation. An authenticated web crawler is a trusted program that computes a specially- crafted signature over the web contents it visits. This signature enables (i) the verification of common Internet queries on web pages, such as conjunctive keyword searches this guarantees that the output of a conjunctive keyword search is correct and complete[26]; (ii) the verification of the content returned by such Internet queries this guarantees that web data is authentic and has not been maliciously altered since the computation of the signature by the crawler. In our solution, the search engine returns a cryptographic proof of the query result. Both the proof size and the verification time are proportional only to the sizes of the query description and the query result, but do not depend on the number or sizes of the web pages over which the search is performed[15]. As experimentally demonstrate, the prototype implementation of our system provides a low communication overhead between the search engine and the user, and fast verification of the returned results by the user.

In[9] B. Parno, M. Raykova, and V. Vaikuntanathan, **“How to delegate and verify in public: Verifiable computation from Attribute-based encryption”**The outsourcing computation is useful only when the returned result can be trusted, which makes verifiable computation (VC) a must for such scenarios[9]. In this work extend the definition of verifiable computation in two important directions: public delegation and public verifiability, which have important applications in many practical delegation scenarios. Yet, existing VC constructions based on standard cryptographic assumptions fail to achieve these properties. As the primary contribution of our work, establish an important (and somewhat surprising) connection between verifiable computation and attribute-based encryption (ABE), a primitive that has been widely studied. Namely, it show how to construct a VC scheme with public delegation and public verifiability from any ABE scheme[16]. The VC scheme verifies any function in the class of functions covered by the permissible ABE policies (currently Boolean formulas). This scheme enjoys a very efficient verification algorithm that depends only on the output size. Efficient delegation, however, requires the ABE encryption algorithm to be cheaper than the original function computation[17]. Strengthening this connection, It show a construction of a multi-function verifiable computation scheme from an ABE scheme with outsourced decryption, a primitive defined recently by Green, Hohen berger and Waters[18]. A multi-function VC scheme allows the verifiable evaluation of multiple functions on the same preprocessed input. In the other direction, They also explore the construction of an ABE scheme from verifiable computation protocols.

In[10] Justin Thaler, Mike Roberts, Michael Mitzenmacher, and Hanspeter Pfister presented a paper on **“Verifiable Computation with Massively Parallel Interactive Proofs”** It assess the potential of parallel processing to help make practical verification a reality, identifying abundant data parallelism in a state-of-the-art general purpose protocol for verifiable computation[10]. It implement this protocol on the GPU, obtaining 40-120 server-side speedups relative to a state-of-the-art sequential implementation[22]. For benchmark problems, our

implementation thereby reduces the slowdown of the server to within factors of 100-500 relative to the original computations requested by the client. Furthermore, it reduce the already small runtime of the client by 100. Our results demonstrate the immediate practicality of using GPUs for verifiable computation, and more generally, that protocols for verifiable computation have become sufficiently mature to deploy in real cloud computing systems[27].

## COMPARATIVE ANALYSIS OF VERIFICATION APPROACHES

This section presents the comparison of different verification approaches.

**Table: 1. Comparison of different Verification Approach**

S.NO	PAPER TITLE	TECHNIQUES	ADVANTAGES	DISADVANTAGES
[1]	Verifiable Computation with Two or More Clouds	The interactive-proof model, Efficient-players refereed games (epRG).	It is easier to understand and to implement, and therefore might be adopted for real-world uses.	For each experiment ran the protocol several times with one cheating cloud that cheats on one out of three randomly chosen states.
[2]	Privacy-Preserving Data Mining from Outsourced Databases	Association rule mining, Privacy-preserving data mining (PPDM)	Effective, good privacy and accuracy	An individual item, a transaction, or the server can always be controlled to be a threshold chosen by the owner, by setting the anonymity threshold k.
[3]	AUDIO: An Integrity Auditing Framework of Outlier-Mining-as-a-Service Systems	AUDIO: An Integrity Auditing Framework of Outlier	It feasible to efficiently verify the outlier mining results of large databases.	It may not be able to catch the malicious server as it may launch verification-aware cheating
[4]	Toward practical and unconditional verification of remote computations	Probabilistically Checkable Proofs	realism, simplicity, and unconditional assurance	but it uses a limited domain only.
[5]	Verifiable Delegation of Computation over Large Datasets	amortized verifiable computation	It allows the client to insert and delete values, as well as update the value at any cell by sending a single group element to the server after retrieving the current value stored in the cell.	Implementation cost is high
[6].	Practical Delegation of Computation using Multiple Servers	Probabilistic Checkable	It is efficiently computable function, both for the servers	The main downside of this approach is the need for an honest majority of clouds.

		Proofs	and for the client.	
[7]	Publicly Verifiable Delegation of Large Polynomials and Matrix Computations, with Applications	amortized model	It is faster verification, a constant amount of computation, The result published is secure only under a weaker "selective" notion of security, where the adversary must commit in advance to the input point x on which it is going to cheat.	convolution is high
[8]	Efficient Verification of WebContent Searching Through Authenticated Web Crawlers	Web Crawlers, kew word search scheme	Get to gather the data you want	Traffic may be identified as abusive or suspicious and blocked .It may be constrained by limits in bandwidth, processing, or storage
[9]	How to Delegate and Verify in Public:Verifiable Computation from Attribute-based Encryption	ABE scheme attribute based encryption (one-key secure) , non-interactive verifiable computation	generality, efficiency, and adaptive security.	The attacker can easily recognize the key value.
[10]	Verifiable Computation with Massively Parallel Interactive Proofs	Interactive Proofs	It saves space and time for the verifier even when outsourcing a single computation, while saves time for the verifier only when batching together several dozen computations and amortizing the verifier's cost over the batch.	parallel run due to the slow process.

**POSSIBLE SOLUTION**

In existing work probabilistic approach has been used for verifying the result. That approach used the key value as a whole value for the attacker can easily identify that value. They can use homomorphic algorithm for verification approach and it can split the key value. For example consider key value as 100 and split the key value like as 60 and 40. The attacker does not find out those key value. It also visualize and determine the how many truly relevant results are returned.

**CONCLUSION**

It presents an various verification approaches available in cloud computing. But the main challenge in cloud computing is security and integrity problems. In security in this context means that the client will receive an assurance that the computation performed by the server is correct. Thus, the client must have some way of verifying the cloud's computation. They can use homomorphic encryption algorithm for verification approach.

**CONFLICT OF INTEREST**

The authors declare no conflict of interests.

**ACKNOWLEDGEMENT**

None

**FINANCIAL DISCLOSURE**

None

**REFERENCES**

- [1] Boxiang Dong, Ruilin Liu, and Hui (Wendy) Wang "Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-Mining-As-a-Service Paradigm" IEEE TRANSACTIONS ON SERVICES COMPUTING, 9(1) JANUARY/FEBRUARY 2016.
- [2] R Canetti, B Riva, GN Rothblum. [2011] Verifiable computation with two or more clouds, in *Proc. Workshop Cryptography Security Clouds*.
- [3] F Giannotti, LVS Lakshmanan, A Monreale, D Pedreschi, W Hui Wang. [2011] Privacy-preserving data mining from outsourced databases, in *Proc. 3rd Int. Conf. Comput., Privacy Data Protection*, pp. 411–426.
- [4] R Liu, H Wang, A Monreale, D Pedreschi, F Giannotti, Wenge Guo. [2012] Audio: An integrity auditing framework of Outliermining- as-a-service systems, in *Proc. Eur. Conf. Mach. Learning Knowl. Discovery Databases*, pp. 1–18.
- [5] S Setty, AJ Blumberg, M Walfish. [2013] Toward practical and unconditional verification of remote computations. in *Proc. 13<sup>th</sup> USENIX Conf. Hot Topics Operating Syst.*, p. 29.
- [6] S Benabbas, R Gennaro, Y Vahlis. [2011] Verifiable delegation of computation over large datasets, in *Proc 31st Ann. Conf Adv Cryptol*, pp. 111–131.
- [7] R Canetti, B Riva, GN Rothblum. [2011] Practical delegation of computation using multiple servers, in *Proc 18th ACM Conf Comput Commun Security*, pp. 445–454.
- [8] D Fiore, R Gennaro. [2012] Publicly verifiable delegation of large polynomials and matrix computations, with applications, in *Proc. ACM Conf. Comput. Commun. Security*, pp. 501–512.
- [9] MT Goodrich, C Papamanthou, D Nguyen, R Tamassia, CV Lopes, O Ohrimenko, N Triandopoulos. [2012] Efficient verification of web-content searching through authenticated web crawlers, *Proc. VLDB Endowment*, 5: 920–931.
- [10] B Parno, M Raykova, V Vaikuntanathan. [2012] "How to delegate and verify in public: Verifiable computation from Attribute-based encryption," in *Proc. 9th Theory Cryptography Conf*, pp. 422–439.
- [11] Justin Thaler\_, Mike Roberts\_, Michael Mitzenmacher, and Hanspeter Pfister "Verifiable Computation with Massively Parallel Interactive Proofs Harvard University, School of Engineering and Applied Sciences,2014
- [12] M Yiu, I Assent, CS Jensen, P Kalnis.[2012] Outsourced Similarity Search on Metric Data Assets. In *TKDE*, 24
- [13] AR Sadeghi, T Schneider, M Winandy.[2010] Token-based cloud computing: secure outsourcing of data and arbitrary computations with lower latency. In *TRUST*,
- [14] M Bellare, B Waters, S Yilek.[2011]"Identity-based encryption secure against selective opening attack." To appear in *TCC*.
- [15] A Lewko, Y Rouselakis, B Waters.[2011] "Achieving leakage resilience through dual system encryption." To appear in *TCC*.
- [16] B Parno, M Raykova, V Vaikuntanathan.[2012] How to delegate and verify in public: Verifiable computation from attribute-based encryption. *TCC*.
- [17] S Benabbas, R Gennaro, Y Vahlis. Verifiable delegation of computation over large datasets. In P. Rogaway, editor, *Advances in Cryptology { CRYPTO 2011*, volume 6841 of *Lecture Notes in Computer Science*, pages 111{131, Santa Barbara, CA, USA, Aug. 14{18, 2011. Springer, Berlin, Germany.
- [18] P Mohassel. [2011] Efficient and secure delegation of linear algebra. *Cryptology ePrint Archive*, Report 2011/605.
- [19] C Papamanthou, E Shi, R Tamassia.[2011] Signatures of correct computation. *Cryptology ePrint Archive,Report 2011/587*.
- [20] N Bitansky, R Canetti, A Chiesa, E Tromer.[2011] From extractable collision resistance to succinct non-interactive arguments of knowledge, and back again. *Cryptology ePrint Archive*, Report 2011/443.
- [21] D Boneh, A Sahai, B Waters. [2011]Functional encryption: Definitions and challenges. In *Proceedings of the Theory of Cryptography Conference (TCC)*,
- [22] R Canetti, B Riva, GN Rothblum.[2011] Two 1-round protocols for delegation of computation. *Cryptology ePrint Archive*, Report 2011/518
- [23] C Gentry, D Wichs. [2011] Separating succinct non-interactive arguments from all falsifiable assumptions. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*.

- [24] S Goldwasser, H Lin, A Rubinstein.[2011] Delegation of computation without rejection problem from designated verifier CS-proofs. Cryptology ePrint Archive, Report 2011/456.
- [25] S Setty, R McPherson, AJ Blumberg, and M. Walfish. Making argument systems for outsourced computation practical (sometimes). In Proc. NDSS, 2012.
- [26] J Applequist. New assured cloud computing center to be established at Illinois. May 2011.
- [27] G Cormode, J Thaler, K. Yi.[2011] Verifying computations with streaming interactive proofs. In Proc. VLDB Endowment.
- [28] G Cormode, M Mitzenmacher, J Thaler.[2012] Practical verified computation with streaming interactive proofs. In Proc. ITCS.

# DETECTING BREAST CANCER FROM MAMMOGRAM IMAGES USING A HAAR WAVELET FILTER WITH THRESHOLD-BASED SEGMENTATION

<sup>1</sup>B. Regina, <sup>2</sup>R. Nedunchelian

<sup>1</sup>Saveetha School of Engineering, Saveetha University, Chennai, Tamil Nadu, INDIA

<sup>2</sup>Dept. of Computer Science and Engineering, Sri Venkateswara College of Engineering, Chennai, Tamil Nadu, INDIA

## ABSTRACT

Today, image science rules the domain of computer science. Typically, image processing is a technique to transfer an image into a digital structure and process it to obtain an improved image. It includes scaling the image concerned, removing small objects, smoothing, extracting features and techniques such as these. Engineering and medicine utilize image processing, the latter especially in tracking cancer. Breast cancer is a type of cancer that surfaces and expands in a woman's breast ... cells, and is widely acknowledged as the world's leading cause of death. This deadly malady is curable if detected in the early stages. Earlier, breast cancer detection was carried out through X-ray mammography to arrive at a diagnosis. In recent times, however, numerous computer-aided diagnostic (CAD) proposals have been developed to enhance the radiologist's diagnostic skills. The Haar wavelet filter is applied to facilitate early breast cancer detection and diagnosis effectively. The Haar wavelet transform outlines the simplest compression method of this category. It is perhaps the finest approach to segment images using thresholding-based, connected component pixels applied to group similar types of pixels. The results have been taken from a range of breast cancer computerized tomography (CT) scan mammogram images. In this paper, we used an efficient image filtering system named the OpenCV 2.4.9.0 and cvbloblib for implementation. The improved method was tested over several breast cancer mammogram images and achieved good results. Going forward, advances can include an extension of this work, incorporating its implementation with datasets to augment the detection and treatment of breast cancer.

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Breast cancer, connected component pixels, Haar wavelet filter, image processing, mammography image

\*Corresponding author: Email: [regina@saveetha.com](mailto:regina@saveetha.com) Tel.: +91 9566330222

## INTRODUCTION

Breast cancer occurs mostly in females, and a newly-released update says that 25% of all women suffer from this disease worldwide. Of that number, 20% die. A team that is presently at work to eradicate breast cancer in females concludes that it can be prevented by birthing before the age of 30, breastfeeding, limiting alcohol intake, maintaining a healthy weight, and exercising regularly. Breast cancer first attacks the duct tissues (tubes that carry milk to the nipples and lobule glands that make milk), which form a major portion of the breast. Mammography became a reliable diagnostic tool in the 1950s when industrial-grade X-ray film was introduced. Mammography can detect breast cancer in two ways:

- Screening mammography: used as a preventive search for women who have no symptoms of breast disease.
- Diagnostic mammography: used in X-rays to obtain images showing the affected breast from different regions and angles.

Diagnostic mammography is an X-ray testing technique to detect breast cancer. **Figure-1** shows an illustration of the symptoms of breast cancer. A lump in the breast can be discovered during testing or an abnormality found during screening mammography. Diagnostic mammography takes longer and is a more involved process. It is a perfect test that can detect the size and location of the affected tissues and lymph nodes of the breast region accurately, compared to screening mammography. Images are viewed from different angles and interpreted accordingly.

## ENHANCEMENT OF THE IMAGE PRE-PROCESSING TECHNIQUE

Image enhancement considers the process of attenuation and sharpening, as well as image features such as edge, boundaries and outlines to produce a processed image. Image enhancement includes gray-level and contrast manipulation. This process reduces noise, obliterates the background and can be used to take sharp images with

crisp edges, filtering, interpolation, magnification, and pseudo-coloring to tweak the image obtained. This process utilizes procedures such as median filtering, normalization and modified tracking algorithms for enhancing mammograms. The performance of the enhancement technique is evaluated by a signal-to-noise ratio (SNR) measure. Enhancement involves four steps in its processing.

Removing the X-ray label in the digital image of the breast previously obtained by testing with diagnostic or screening mammography.

Step [1]. Eliminating high frequencies with the median filter.

Step [2]. Decreasing contrast and brightness with the normalization process.

Step [3]. Destroying the muscle region with the modified tracking algorithm.

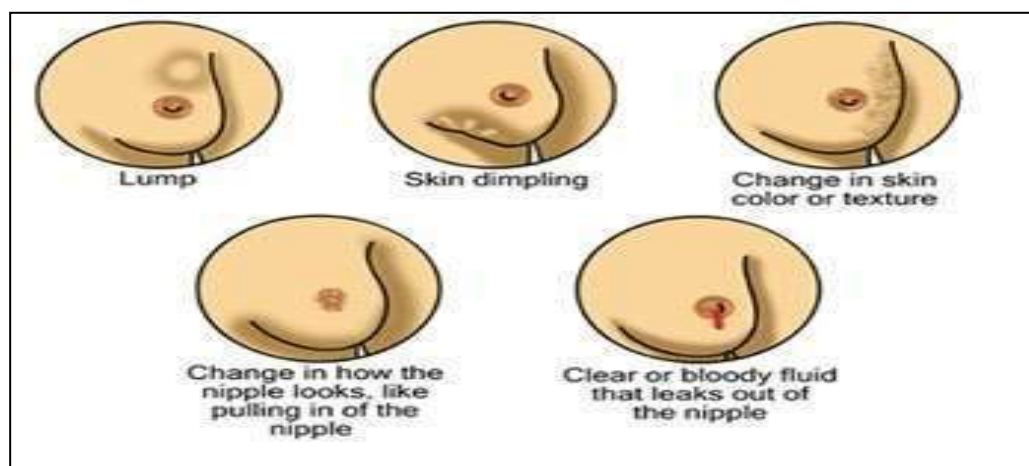


Fig.1: Symptoms of breast cancer

## SEGMENTATION TECHNIQUE OF BREAST CANCER

Image segmentation is a key step in pictorial pattern recognition and analysis applications. The segmentation technique defines, in precise terms, the success or failure of analytic procedures.

Segmentation aims to change or alter an image, rendering it more meaningful and simpler to find. The field of pattern recognition involves the usage of image segmentation, in the early phases, to divide an input digital image into different patterns. Each region has characteristics of its own, on the basis of which regions are grouped. These characteristics may be completely mathematical - for example, based on the area of pixels and their neighbours - or visual, such as colour, intensity, or texture repetitiveness.

Segmenting the breast and non-breast region is a major prerequisite for further bilateral filtering differences. This section presents a border detection method with a genetic algorithm. The breast border can be obtained in segmentation from the image of the breast region. Some authors have developed methods to detect the breast area, based on a global histogram analysis. However, a method that depends on global thresholding alone is based, critically, on the selection of threshold values.

## FILTERING TECHNIQUE OF BREAST CANCER

From the perspective of digital image processing, a filter is a system used in a mathematical operation on an image represented as a sampled, discrete-time signal. The filter is used to decrease certain enhanced aspects of a signal. The notion of filtering has its origins in the use of the Fourier transform for signal processing in frequencies. Filters may be linear or non-linear, based on the relation of the output with the input.

## LITERATURE SURVEY

Mammography is a major image processing technique that diagnoses breast cancer. Mammography establishes the presence of cancer before a physically-evident manifestation. It is also the most sensitive method because it

carries both positive and negative results, and can be done using an analog image or a digital one. A digital image uses a full field detector. Today, however, analog images have come into play. Image processing and intelligent systems are the two mainstreams of computer technologies. The K-means clustering algorithm, i.e., clustering all objects into k distinct groups [14], is being used.

A system theoretic method is presented, based on a Markov chain model to analyse the mammography process of breast cancer detection. Formulas are developed to estimate the patient's length of stay and staff efficiency. An empirical formula is suggested and non-Markovian scenarios are also investigated. Results prove that these methods are highly accurate in evaluating performance. In support, a case study is tabled to demonstrate the application of the model. An increase in patient volume is also examined to show that the increase in capacity is mandatory to meet the demand [1].

Markov models and neural networks comprise a step-by-step process. Dialing out medical images using the data mining method with neural networks delivers very good results. It involves grouping, the chief objective being to detect affected areas [15].

The initial step in most imaging situations is to locate the surface of an object. This information may be essential for reconstruction, or required for the cancellation of the surface reflection and also as a preparatory step prior to imaging. Here, two complementary approaches are developed specifically for the purpose of surface localization. Using the data from phantom measurements and volunteer scans, the recommended approaches are evaluated [3]. The clinical prototype of a microwave imaging system with the monthly scans of healthy patient volunteers is tabled. The purpose is to examine how the system's measurements are affected by numerous issues that cannot be side-stepped in the monthly monitoring of human subjects. These factors include biological and measurement variabilities. The study also quantifies the anticipation level of variability when conducting microwave breast imaging. For frequent monitoring of breast health, this is a key step in establishing the validity of the microwave radar imaging system [2].

CS ideas are utilized to regularize the inversion process and decrease the number of unknowns by limiting the solution of the BI method. By minimizing a cost function given by the mean square error amongst the measured and modelled data, the sparse BI method estimates a contrast function. A solution for the sparse BI method is found, using an iterative algorithm. The sparse BI method is tested successfully on a noise-free and noisy synthetic dataset representing a tomographic scan of a cancerous breast. Results prove that the sparse BI method remains convergent as the number of iterations increases, in comparison to the conventional BI method [4].

A clinical prototype with a wearable patient interface for microwave breast cancer detection is examined. By embedding 16 flexible antennas into a bra, the system is operated with a metastatic time-domain pulsed radar, and is also cost-effective. The resulting data is compared with the data obtained from the table-based prototype. The wearable prototype has enhanced the quality of the volunteer data collected. A wearable breast health monitoring array can be further improved in the near future for patients with various breast sizes and tissue densities [5].

Using mammography, a computer-aided detection and diagnosis system for breast cancer is recommended. The breasts are first partitioned adaptively into regions according to the proposal. Two strategies are examined to define the anomaly detector. The first strategy uses manual segmentations of lesions to train an SVM that assigns an anomaly index to each region. The second strategy uses various MIL algorithms to train local and global anomaly detectors. Results prove that the second approach outperforms the first, demonstrating that anomaly detectors have advantages and can be trained on large medical image archives without manual segmentation [6].

A new concept for learning from crowds that handle data aggregation directly as part of the learning process of the convolutional neural network (CNN) through an additional crowdsourcing layer (AggNet) is discussed. Also, an experimental study is tabled to find answers for training on the CNN, as well as training multiple types of annotation datasets by the CNN, and how accuracy is affected by the choice of annotation and aggregation. The experiment involves Annot8 for realizing image annotation tasks for publicly-available biomedical image databases. The results prove the necessity of data aggregation integration [7].

A dual-photon emission computed tomography (DuPECT) mechanism, an alternative to 3-D imaging systems, is proposed, integrating both preoperative and intraoperative information to trace SLNs using cascade isotopes. SLNs can be located using the line-plane intersection. The Monte Carlo software is used to evaluate performance. The random rate increases with increased initial activities, while the scatter rate is lower than 1.2 count/s for a range of activities. Four injection sites and two LNs are placed at various depths in a simulated study. LNs are



clearly identifiable in the absence of injection sites. Results prove that the suggested three-dimensional imaging system has the potential to identify injection sites and various SLNs [8].

A new computer-aided detection (CAD) system is considered to reduce human involvement and help radiologists in the automatic diagnosis of malignant/non-malignant breast tissues by using polar complex exponential transform (PCET) moments as texture descriptors. The input ROI of a fixed size of  $128 \times 128$  is extracted manually to avoid processing the whole mammogram. Also, a new classifier, ADEWNN, is introduced to improve the classification accuracy of the suggested system. The proposed phase-correction method (PCPCET) is compared with a magnitude-based feature extraction (PCET). The best area is found to be 0.984, with a confidence interval ranging from 0.968 to 0.999 and a  $\pm 0.0108$  standard error under the receiver operating characteristic curve [9].

Magneto-acoustic tomography with magnetic induction (MAT-MI) is a promising technology for the non-invasive detection of breast cancer. This work presents a high-frequency MAT-MI (hfMAT-MI) system, a significant improvement over previous methods. Boundaries between cancerous and healthy tissues, as also the tumours' internal structures, are resolved by hfMAT-MI. For the first time in an in vivo mouse model, a growing tumour was tracked using the hfMAT-MI method. This demonstrates the promise of the magneto-acoustic imaging system for effective detection and diagnosis of early-stage breast cancer [10].

Here, the miRNAs deregulated in breast tumors are identified and the potential of circulating miRNAs in breast cancer detection are examined. miRNA expression profiling of 1919 human miRNAs in paraffin-embedded tissue from 122 breast tumors and 11 healthy breast tissue samples were conducted. From 26 healthy people and 83 patients with breast cancer, the most relevant miRNAs were analyzed in plasma. The results helped in identifying a large number of miRNAs deregulated in breast cancer, and generated a 25miRNA microarray classifier that discriminated breast tumors. Therefore, this supports the use of circulating miRNAs as a method for early breast cancer detection [11].

Actual awareness of breast density, and a knowledge of its impact on breast cancer detection and risk, are unknown. A national cross-sectional study, overseen in English and Spanish, using a probability-based sample of screening-age women was conducted. Around 65% responded out of the 2,311 women surveyed. Overall, 58% of women had heard of BD, 49% were aware that BD affects breast cancer detection, and 53% knew that BD has an effect on cancer. Disparities in BD awareness and knowledge happen by race/ethnicity, education, and income. These findings support continued and targeted efforts to increase BD awareness and knowledge amongst women eligible for mammography screening [12].

The intention behind introducing MRIs is to improve the outcomes of screening programs for women with familial breast cancer. The aim of this study is to assess whether the introduction of MRI surveillance improves 5- and 10-year survival rate of high-risk women and determine the appropriateness of MRI breast cancer detection when compared with mammography. Women with greater breast cancer risk were screened by either mammography alone or with MRIs. Results prove that the rate of survival was higher in the MRI-screened group. Screening with MRIs is more beneficial, particularly in BRCA2 carriers [13].

The author wishes to state that breast cancer is a major issue in this discussion. Two methods for their detection - selection mammography and analytical mammography - are widely used. Selection or screening mammography comprises three approaches. Analytical or diagnostic mammography involves much time because of the number of images and angles to be examined. The drawback of diagnostic mammography is that it is very expensive, though a very effective tool [16].

Image similarity and asymmetry are done after mammography. Image similarity helps categorize breast cancer images, further compared with many others. This technique is specifically used to pinpoint medical imaging on suspicious areas, alongside CAD tools commonly used to improve the detection of breast cancer [17].

An image enhancement algorithm is used to enhance the images in question, after which they are sent for analysis. The Gabor algorithm and several filter algorithms are applied for evaluation and assessment. Image segmentation is done and new features obtained. Developing the quality of the image with the help of MATLAB software is the primary aim of this study, and the Gabor algorithm provides excellent results [18].

Spatial fuzzy clustering is a cluster analysis technique that has to be applied to the segmented image, of use for image processing as well. Pre-processing, a must before segmentation, is used to initialize the seeds in the

growing process. A region-growing algorithm can be used to fetch the primary seeds from the obtained outcome or result. Specific regions are detected in this way and the segmentation problem cleared. This is mostly used to reduce the cost factor and the process can be terminated, if needed [19],[20].

Textural feature studies are extensively used in the field of computer vision and image processing, with both identification and microcalcification utilized to detect breast cancer. Image texture provides for a spatial array of colors. There is no single technique of texture imaging that is sufficient for a range of textures. The offered texture analysis contains both arithmetical and structural methods. GLCM and GLRLM techniques, alongside others, can also be applied [21].

## PROPOSED METHOD

### HAAR WAVELET FOR IMAGE DECOMPOSITION

The Haar wavelet, a sequence of functions, is the simplest wavelet transform. It is a compression process described as

$$H(t) = \begin{cases} 1 & 0 \leq t < 1/2 \\ -1 & 1/2 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}$$

The scaling function  $f(t)$  can be described as

$$f(t) = \begin{cases} 1 & 0 \leq t < 1 \\ \text{otherwise} & \end{cases}$$

### SINGLE-DIMENSIONAL HAAR WAVELET TRANSFORM

The Haar wavelet transform is a 2-element matrix  $[x(1), x(2)]$  and another 2-element matrix  $[y(1), y(2)]$  represented by a relation denoted by the equation that follows

$$\begin{matrix} y(1) & & x(1) \\ & = OM & \\ y(2) & & x(2) \end{matrix}$$

Here, OM is denoted as an orthonormal matrix.

The orthonormal matrix is described as

$$OM = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

### TWO-DIMENSIONAL HAAR WAVELET TRANSFORM

The 2-dimensional Haar wavelet transform  $x$  and  $y$  becomes a  $2 \times 2$  matrix defined by the relation. The transformation can be carried out, at first pre-multiplying the columns of  $x$  by the OM (orthonormal matrix) and then post-multiplying the rows of the result by the OMT.

$$\begin{matrix} y = OM \cdot x \cdot OM^T \\ x = OM^T \cdot y \cdot OM \end{matrix}$$

To compute the transformation of a complete breast image, first divide the breast image into  $2 \times 2$  blocks and apply the equation  $y = OM \cdot x \cdot OM^T$ . Image segmentation is a fundamental step in advanced methods of multi-dimensional signal processing. Haar wavelet functions are generated from a one-dimensional function by means of deletions and translations. The Haar transform forms the simplest compression process of this kind.

## CONNECTED COMPONENT PIXELS TO DETECT BREAST CANCER

Based on the specified heuristic assessment-connected components group, all pixels are subsets. Connected component pixels are utilized in computer vision to detect connected regions in binary digital images, though color images and data alongside higher dimensionalities can additionally be processed. Connected component

labeling can work on a collection of information after being consolidated into an image recognition system or a human-computer interaction interface. Blob extraction, from a thresholding step, is usually provided in the emerging binary image. Blobs can be counted, filtered and tracked. A graph encompassing vertices and connecting edges is crafted from relevant input data. The vertices encompass data needed by the comparison heuristic, as the edges indicate related 'neighbours.' An algorithm crisscrosses the graph, labelling the vertices established on the connectivity and comparative benefits of their neighbors. The algorithm follows this method and creates new region labels whenever necessary.

The key to a fast algorithm depends on how the merging is done. This algorithm uses the union-find data construction that works admirably at keeping track of equivalence relationships. Union-find stores labels that correspond to the lookalike blob in a disjoint-set data construction, making it simplistic to recall the equivalence of two labels by the use of an interface method. Relatively easy to apply and comprehend, the two-pass algorithm iterates across 2-dimensional binary data. The algorithm passes through the image twice: the first pass to allocate provisional labels and record equivalences, and the second pass to substitute every single provisional label with the small label of its equivalence class.

A faster-scanning algorithm for connected-region extraction is given below.

### ONE PASS ALGORITHM

1. Begin the Algorithm
2. {
3. SCAN Each Component Column, THEN through Row
4. IF Component is Not Equal to Background
5. {
6. Acquire the adjacent component of the recent component
7. }
8. IF Neighbor is equal to Zero
9. Mark the present component and continue
10. Else
11. {
12. FIND the neighbor through the least marker
13. Assign it to the present component
14. }
15. Neighboring Marker is Equal to Distance
16. End the Algorithm
17. }

### TWO PASS ALGORITHM

1. Begin
2. {
3. SCAN Each Component Column,
  - a. THEN through Row
4. IF Component is Not Equal to
  - a. Background
5. Relabel the component with the
  - a. Lowest counterpart marker
6. }
7. End

### RESULTS AND DISCUSSION

CT scan images, with a width of 454 pixels, height of 564, and bit depth of 24 and the image type a PNG file, have been used for simulation purposes

For simulating the breast image, OpenCV 2.4.9.0 and cvblobslib library are utilized. The OpenCV (Open Source Computer Vision Library) is an open source-approved library containing lots of computer vision algorithms. Normally, documents are named OpenCV2. x API, which is fundamentally a C++ API. OpenCV has a modular construction and that way the package includes countless public or static libraries. The pursuing modules available are CORE, IMGPROC, WARPING, VIDEO, CALIB3D, FEATURE2D, OBJDETECT, HIGHGUI, and GPU [17].

CvBlobsLib has a built-in Microsoft Visual C++ (6.0) and, additionally, can be utilized in .NET. CvBlobsLib is distributed in a static library (.lib). A .lib file is to be created prior to its use in a project. To create the .lib file, open the MSVC++ undertaking and set the process in motion.

The blob extraction utilized the CvBlobsLib, a library that presents related component labelling of binary images, obtained at the OpenCVWiki page, and is relatively easy to use. This implementation used the OpenCV libraries to open a colour image, change it to a grayscale and then threshold it to change it to a black and white (binary) image.

[Table- 1] represents the Cvconvert of the Haar filter coding.

**Table: 1. Cvconvert of Haar filter**

### CVCONVERT OF HAAR FILTER

```
cvSmooth(image,image,2,3,0,0);
width=image->width;
height=image->height;
step=image->widthStep;
channels=image->nChannels;
data=(uchar *)image->imageData;
printf("%d",image->nChannels);
```

The proposed method for segmentation of breast cancer images is threshold-based filtering. The threshold point can be set as 245 to 255 pixels. Threshold-based connected component pixel breast cancer images are filtered using the Haar wavelet filter technique. Table- 2 describes setting a threshold in the Haar wavelet filter.

**Table: 2. Setting the threshold in the Haar wavelet filter**

### Setting Thresholding

```
cvThreshold( grayimage, originalThr, 245, 255, CV_THRESH_BINARY );
```

The next is the filter method in the CBlobResult to remove all blobs in the breast imaging that conform to a precise size, calculate the number of 'proper' blobs discovered and display them in red in a new window.

To build a CBlobResult, use the breast cancer image constructor on an input 1-channel image. This fills the CBlobResult through all the blobs of the breast cancer image. Blobs from the CBlobResult object can be filtered with the Haar wavelet filter method. The criterion to contain or discard blobs is each object from the classes derived from the COperatorBlob (area, perimeter, gray level, etc, or new classes formed by users). [Table- 3] shows the segmentation operation on connected component pixels.

Table: 4. Segmentation operation

**Segmentation Operation**

```

CBlobResult blobs; int i1, param2=0, param1=0; CBlob *currentBlob;
int height1 = test->height;
int width1 = test->width;
int step1 = test->widthStep;
int channels1 = test->nChannels;
uchar *data1 = (uchar *)test->imageData;
blobs = CBlobResult( originalThr, NULL, 0 );
blobs.Filter( blobs,B_INCLUDE, CBlobGetArea(), B_GREATER_OR_EQUAL, 75 );
printf("%d",blobs.GetNumBlobs());
cvMerge( originalThr, originalThr, originalThr, NULL, displayedimage );
for ( i = 0; i < blobs.GetNumBlobs(); i++ )
{
    currentBlob = blobs.GetBlob(i);
    currentBlob->FillBlob( displayedimage, CV_RGB(255,0,0));
}
cvNamedWindow("BreastCancer",1);
cvFlip(displayedimage,displayedimage,1);
cvShowImage("BreastCancer",originalThr);
cvFlip(image,image,1);
cvNamedWindow("image",1);
cvShowImage("image",image);
cvNamedWindow("Segmented",1);
cvShowImage("Segmented",displayedimage);
cvWaitKey(0);
  
```



Fig.2.1: Original CT Scan Image



Fig.2.2: Segmented Image



Fig.2.3: Breast Cancer / Filtered Image



Fig.3.1: Original CT Scan Image

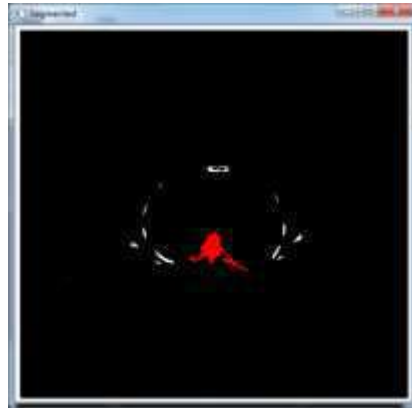


Fig.3.2: Segmented Image



Fig.3.3: Breast Cancer / Filtered

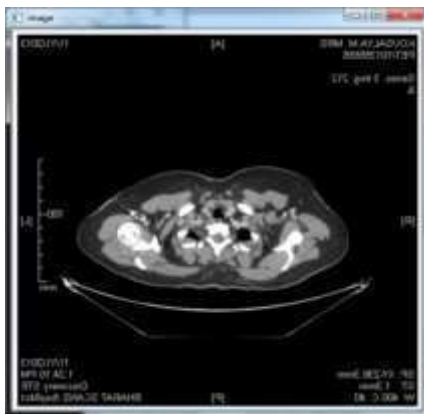


Fig.4.1: Original CT Scan Image

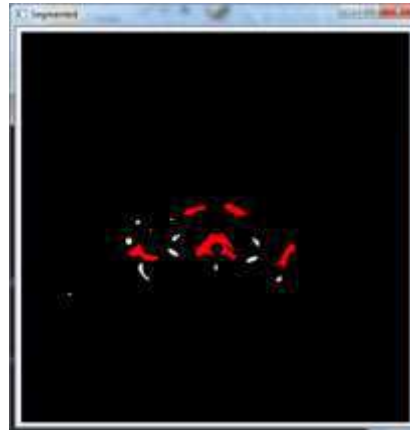


Fig.4.2: Segmented Image

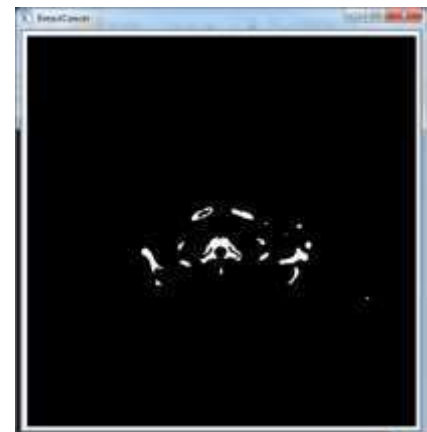


Fig.4.3: Breast Cancer / Filtered Image

Fig 2.1, Fig 3.1 and Fig 4.1 are original CT breast cancer images. While being filtered, they reach an intermediate stage, i.e., that of segmented breast cancer images, as in Fig.2.2, Fig 3.2 and Fig 4.2. The final images are Fig 2.3 and Fig 3.3 and Fig 4.3 comprises filtered images.

### CONCLUSION

Breast cancer is the second-foremost reason for deaths from cancer among women worldwide. A general utilization of screening, subsequently through treatment advances in recent years, has led to a major reduction in deaths from breast cancer. Today, finding a radical cure for breast cancer is a major challenge for scientists, to which end researchers invent technologies to detect breast cancer at the earliest. Using the Haar wavelet filter with connected component pixels through an OpenCV implementation gives excellent results, especially in the initial stages of breast cancer. The major advantage of this paper is that connected component pixels are used, and the mechanism of scanning an image from top to bottom, pixel-by-pixel, as well as sequentially, to recognize connected pixel regions are demonstrated as viable propositions. Additionally, working on binary or gray-level images and disparate measures of connectivity is made possible. During the past decades, the Haar wavelet filter became an essential tool with a collection of requests, normally associated with signal processing, data and image compression. Applying Haar wavelet filtering methods produces four components. The estimate point of the image is defined by low frequencies, horizontal and vertical components by mid-range frequencies, and diagonal

features by high frequencies. In future, this work will be enhanced with datasets for improved detection of breast cancer.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

### REFERENCES

- [1] Zhong Xiang, et al.[2016] A System-Theoretic Approach to Modeling and Analysis of Mammography Testing Process." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46.1: 126-138.
- [2] Porter Emily, Mark Coates, Milica Popović. [2016 ]An early clinical study of time-domain microwave radar for breast health monitoring. *IEEE Transactions on Biomedical Engineering* 63.3 :530-539.
- [3] Sarafianou M, Preece AW, Craddock IJ, Klemm M, Leendertz JA. [2016] Evaluation of Two Approaches for Breast Surface Measurement Applied to a Radar-Based Imaging System. *IEEE Transactions on Antennas and Propagation*, 64(2): 609-617.
- [4] Ramirez, Ana B., and Koen WA van Dongen.[ 2015] Sparsity constrained born inversion for breast cancer detection." *Ultrasonics Symposium (IUS), 2015 IEEE International*. IEEE.
- [5] Porter E, Bahrami H, Santorelli A, Gosselin B, Rusch LA, Popović M. [2016] A Wearable Microwave Antenna Array for Time-Domain Breast Tumor Screening. *IEEE transactions on medical imaging*, 35(6): 1501-1509.
- [6] Queller G, Lamard M, Cozic M, Coatrieux G, Cazuguel G. [016] Multiple-Instance Learning for Anomaly Detection in Digital Mammography.
- [7] Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S, & Navab, N. (2016). AggNet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE transactions on medical imaging*, 35(5), 1313-1321.
- [8] Lin Chang-Shiun, Hsin-Hon Lin, Yu-Ching Ni, Meei-Ling Jan, Kuan-Pai Lu, and Keh-Shih Chuang. "Application of the Intraoperative Dual Photon Emission Computed Tomography System in Sentinel Lymph Node Detection: A Simulation Study." *IEEE Transactions on Nuclear Science* 63, no. 1 (2016): 108-116.
- [9] Singh, Satya P, Shabana Urooj, and Aimé Lay-Ekuakille. "Breast Cancer Detection Using PCPCET and ADEWNN: A Geometric Invariant Approach to Medical X-Ray Image Sensors." *IEEE Sensors Journal* 16.12 (2016): 4847-4855.
- [10] Yu, Kai, Qi Shao, Shai Ashkenazi, John Bischof, and Bin He. In Vivo Electrical Conductivity Contrast Imaging in a Mouse Model of Cancer Using High-frequency Magnetoacoustic Tomography with Magnetic Induction (hfMAT-MI). (2010).
- [11] Matamala, Nerea, et al.[2015] Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection. *Clinical chemistry* 61.8 (2015): 1098-1106.
- [12] Rhodes DJ, Breitkopf CR, Ziegenfuss JY, Jenkins SM, Vachon CM. [2015] Awareness of breast density and its impact on breast cancer detection and risk. *Journal of Clinical Oncology*, 33(10): 1143-1150.
- [13] Gareth ED, Nisha K, Yit L, Soujanya G, Emma H, Massat NJ, Anthony H. [2014] MRI breast screening in high-risk women: cancer detection and survival analysis. *Breast cancer research and treatment*, 145(3): 663-672.
- [14] De Oliveira Martins, L Junior, GB Silva, AC de Paiva, AC, Gattass M. [2009] Detection of masses in digital mammograms using K-means and support vector machine. *ELCVIA: electronic letters on computer vision and image analysis*, 8(2):39-50.
- [15] Ferrero, Gustavo, Paola Britos, and Ramón García-Martínez. "Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks." *Professional Practice in Artificial Intelligence*. Springer US, 2006. 1-10.
- [16] Abinaya S, Sivakumar R, Karnan M, Shankar DM, Karthikeyan M. [2014] Detection of Breast Cancer In Mammograms-A Survey. *Int J Comput Appl Eng Technol*, 3(2):172-178.
- [17] Tahmoush Dave, Hanan Samet.[2006] Using image similarity and asymmetry to detect breast cancer." *Medical Imaging*. International Society for Optics and Photonics.
- [18] Patel Vishnukumar K, Syed Uvaid, AC.[2012] Suthar. Mammogram of breast cancer detection based using image enhancement algorithm. *Int J Emerg Technol Adv Eng* 2.2012: 143-147.
- [19] Shili Hechmi, Lotfi Ben Romdhane, and Bechir Ayeub. An Efficient Model Based on Spatial Fuzzy Clustering and Region Growing for the Automated Detection of Masses in Mammograms. *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCVR)*. The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013.
- [20] Thangaraj P Geetha K. [2015] FGT2- ABR: Fuzzy Game Theory Trust Associativity Based Routing to Mitigate Network Attacks in Pervasive Health Monitoring Systems, *Journal Of Pure And Applied Microbiology*, 9: 161-168.
- [21] Anupa Maria Sabu, D NarainPonraj, Poongodi.[2012] *Textural Features Based Breast Cancer Detection: A Survey*, *Journal of Emerging Trends in Computing and Information Sciences*, 3(9)

# PRIOR ROUND REVEALS RSSI INFORMATION BASED SYBIL DEFENSE IN OPEN WIRELESS NETWORK

S.U. Akshaya\*, D. Thilagavathi

Department Of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, INDIA

## ABSTRACT

*Aim: Open wireless ad-hoc network become harmful by possessing many identity which malicious node gains dis-appropriate influence and information. Many defense based on Sybil attack posed over channel estimation, trusted sources not exposed on the IEEE 802.11 network. The defense against the Sybil attack without any trusted authority is evaluated. Methods: RSSI observation and Sybil classification is performed with MASON TEST protocol with high computation in commodity devices. The method Prior round reveals RSSI information is implemented to reduce the computation time generated by the MASON TEST protocol. Novelty: Specifically, we implement the protocol and the method to defense against the Sybil attack, i.e. 99.99%, without trusted certification in minimum computation time. The performance is illustrated in network simulator and the result is analyzed.*

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Wireless network, Ad-hoc network, security, Sybil attack, Signalprint.

\*Corresponding author: Email: [akshaya.cse004@gmail.com](mailto:akshaya.cse004@gmail.com); Tel.: +91-9677340431

## INTRODUCTION

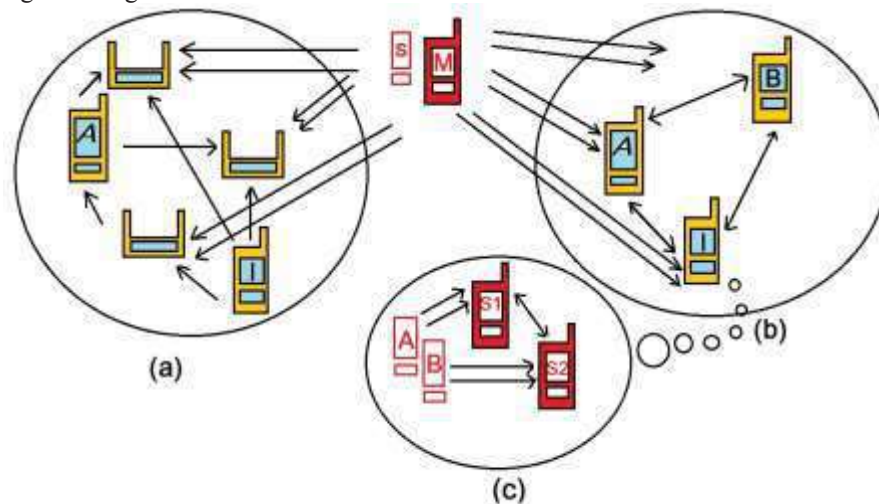
Wireless network technology is one of the hottest topics in network fundamentals. Wireless networks serve many features. In various cases they uses cable replacements, where in other cases they are used to provide access to corporate data from remote location. The main four categories of wireless networks are WPAN (wireless personal area network), WLANs (wireless local area networks), WWANs (wireless wide area networks), and satellite networks. These networks are now commercially available in most of the region. The standards used in WLAN are 802.11 a, b, g, HIPERLAN/2. IEEE 802.11 is a combination of MAC (media access control) and physical layer (PHY) specifications for implementing WLAN (wireless local area network) computer communication in the 2.4 - 60 GHz frequency bands.

Wireless networks turns vulnerable to Sybil attacks, in which Sybil node poses identities in order to gain disproportionate influence. Various defenses based on localization of wireless channels exist, but something not exposed on commodity 802.11 devices. There introduces numerous security concern to defense against the attack, since participants are not vetted this assumption is easily broken by a Sybil attack. Defenses which are proposed falls into categories like trusted certification, social network based technique, misbehavior detection, resource testing, localization techniques. The trusted certification used access point or certification to vet participants, thus not useful in open nature of wireless network. Resource testing methods are most easily defeated in ad-hoc network of resource limited mobile devices by attackers with access to greater resources.

The localization technique supports defense mechanism against open ad-hoc network without trusted certification. RSSI (Received Signal Strength Indication) [1] is a localization technique uses the spatial correlation between the received signal strength and physical location of a node to identify the presence of a Sybil node. It is important to note RSSI does not relay on the quality of signal and usually an action is required for mapping RSSI distance values. In **Figure- 1 (a)** represents the RSSI observation from trusted APs used to identifies the Sybil's, where S is a Sybil presented by attacker M. Trusted RSSI observations, which are not generally available in open ad-hoc networks. In **Figure- 1(b)** represents the participant themselves act as observers. The observations aren't trusted, coming from possible lying neighbors. In **Figure- 1, (c)** represents I believes S1 and S2 are falsified observation and incorrectly accept them and reject A and B as Sybil. A Signalprint [2] is used, as its direction stays unchanged; as RSSI can be changed by varying transmit power. Signalprint are hard to spoof and strongly correlated with physical location of nodes. Signalprints allow a control over Wireless Local Area Network to reliably single out clients. Instead of identifying clients based on MAC addresses or other data,



Signalprints allow the system to recognize the identity based on how clients look like in terms of signal strength levels.



**Fig:1. Trusted RSSI observation and false observations in Ad-hoc networks**

Murat Demirbas and Oguejiofor O.S noted that RSSI is a robust and lightweight solution for Sybil attack issue based client position in both indoor and outdoor environment. The framework naturally evaluates the distance between node hubs by measuring the RSSI (got signal quality marker) at a suitable number of node hubs.

The harmful attack against ad hoc networks is known as the Sybil attack. Sybil nodes refer to a malicious device's additional identities. Open nature of wireless network need a defense against Sybil attack, something exposed on commodity 802.11 devices. Without requiring trust in any other node or authority, RSSI is inherent use true or false RSSI observation reported by one-hop neighbors. The method prior round reveals RSSI information is used to reduce the computation time by comparing the RSSI prior round values. Performing Mason Test protocol with two components: collection of RSSI observations and Sybil classification. The protocol classifies non-Sybil and Sybil by vetting participants without using trusted authority.

## RELATED WORK

Daniel B. Faria, (2006) uses signal print [3] technique to defeat against the sybil attack. The transmitting devices can be robustly identified by its signal print, a tuple of signal strength values reported act as sensors. The signal printcreates signal strength measurement is reliable to client identifiers . The sybil clients can lie about their MAC address, signal print are strongly correlated with the physical location. Therefore, holding nodes with their Signalprints provides the proper matching rules. Signal print is featured in way that wireless network is able to detect a large class of effective DOS based on MAC address spoofing.

Murat Demirbas, (2003)[1][4] uses the RSSI as a solution to the sybil attack in wireless sensor network. The RSSI is said to be lightweight process, the issues like time-varying, unreliable, non-isotropic is over come by using the Received Signal Strength Indication ratio. The RSSI is found to be the robust since it detects the sybil nodes with 100% completeness and less false positive ratio.

Mohamed Salah Bouassida, (2007) [5] reports that by collection of mobile host forming an established infrastructure without aid. By allowing node to verify the authenticity of neighbour

nodes based on the localization. To determine the estimated metric, the nodes are classified between the significance of the node.

Zhuliang Xu, (2013) [6] discuss about RSSI along with Ensemble Empirical Mode Decomposition (EEMD) and evaluate the performance in the indoor and outdoor environment. EEMD normalizes the RSSI value related to the distance and reproduces the movement of the sender. EEMD can effectively ignore the RSSI value that changes in distance equation which is specific for one Wi-Fi device. The EEMD along with RSSI is effective in outdoor than indoor environment. Diogo Monica, (2009) [7] deploys a framework to evaluate the power and performance of radio resource test (RRT), i.e., each node has access to a single radio device, the potential to support protocol that does not require pre-configuration nor pre-shared secret.

Yue Liu, (2013) [8] proposed a method Multiple-Input Multiple-Output (MIMO) [9] in Sybil defense by resource testing. In MIMO the received signal is validated to identify the transmission. The node is identified by multiple identities from same receiver to be a Sybil or malicious node. MIMO gains complete information about the received signal strength.

## MATERIALS AND METHODS

In this segment, we summarize the problem, solution framework and briefly discuss RSSI [5] and Signalprint methods.

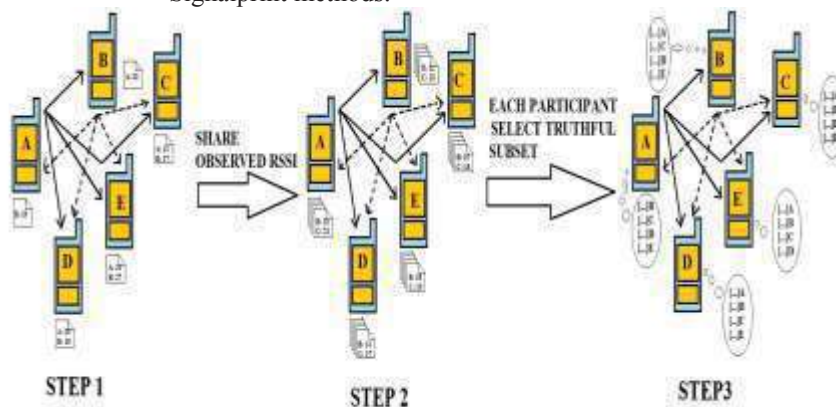


Fig. 2: Trustless truthful subset selection of RSSI observer

### A. Problem Statement

We extend the Signalprint and RSSI based Sybil detection and classification methods to work without any prior detection or observation of participants to determine which of its one-hop neighbors are non-Sybil in an open wireless network. The framework that formed allows us to identify the truthful subset selection of nodes for a secure safe and trustful protocol.

The framework formed, **Figure- 2** illustrates truthful subset selection in three steps:

Step 1: First participant takes turn of broadcasting probe packet and other nodes record observed RSSI

Step 2: All the participants share their observation with their one-hop neighbors, i.e. each and every participant holds the RSSI observation of their one-hop neighbors.

Step 3: Finally each and every participant individually selects a truthful subset for Signalprint based Sybil classification.

### B. RSSI (Received Signal Strength Indication)

Received Signal Strength Indication [1] is a term of measuring the relative quality of the signal of the client nodes. The strength is based on the nodes signal as seen on receiving device, e.g. a smart

phone. The strength of the signal is based on the distance and value of broadcasting power, at maximum broadcasting power the RSSI ranges from 40-50 m distance.

Deploying one node to transmit “hello” messages with constant power (i.e., 0 dBm) and another acts like receiver and capture RSSI then transmit them. The transmitter sends message over 1000 times by setting distance of 15 cm between the transmitter and receiver. But this deployment results to non-uniform nature of RSSI and poor correlation of RSSI value makes it unsuitable for Sybil detection. So, we deploy two receivers to compare ratio of RSSI instead of absolute value of RSSI [6] and observe the time varying of RSSI. By comparing the ratio, RSSI can take care of varied transmission power at sender. By using different transmitting power the sender broadcast 1000 messages. RSSI values are recorded by two receivers and transmit them to base station.

$$P_r(d)_{dBm} = R_{dBm} - 10n \log_{10} \left( \frac{d}{d_0} \right) + Z_{dBm} \quad (1)$$

Where,

R – Received Signal Strength Indication.

$P_r$  – Received signal power.

Z – Gaussian distribution random variable with 0 mean value.

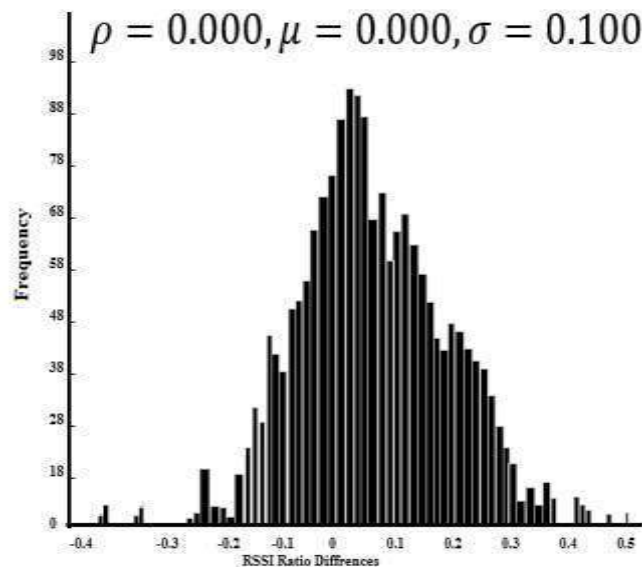
d – Distance difference between receiver and transmitter.

The base station analysis and compute the ratio of two RSSI values it received from the two receivers at time t1 and t2. The difference of RSSI ratio is calculated and logs this value.

This results in uniform distribution of values by following Gaussian Probability Distribution with standard distribution of 0.066 and 0.106, as in [equation \(1\)](#). If D1 and D2 is the difference of RSSI ratio in same location and I1, I2, I3 and I4 are the node identity with a threshold.

$$\left( \frac{R_{I1}^{D1}}{R_{I1}^{D2}} - \frac{R_{I2}^{D1}}{R_{I2}^{D2}} \right) < \sigma, \left( \frac{R_{I3}^{D1}}{R_{I3}^{D2}} - \frac{R_{I4}^{D1}}{R_{I4}^{D2}} \right) < \sigma, \\ \left( \frac{R_{I1}^{D1}}{R_{I4}^{D1}} - \frac{R_{I1}^{D2}}{R_{I4}^{D2}} \right) < \sigma \quad (2)$$

It is safe to set  $\sigma$  as 0.1 and threshold to 0.5 to detected Sybil node 99.999%, i.e. the threshold to be  $5\sigma$ , more specifically 0.1, calculated as in equation (2). [Figure- 3](#) represent the ratio compared



between RSSI.

**Fig. 3: Comparing ratio of RSSI**

### C. Signal print

Signalprint [2] is vector of RSSI median. The properties of Signalprint are: Strongly correlated with the physical location with close proximity of client and Packet violently transmitted by stationary nodes generates similar Signalprint with high probability. Signalprint value can be written as original value or as relative value with respect to high and lower values of RSSI [9][12] levels in dBm. The difference between the value at an appropriate position and maximum values found in the Signalprint, is calculated using the term differential signal strength. When matching two Signalprint (i.e. S1, S2) it should be written with both absolute and differential values. The use of differential values increases the Signalprint operation that varying transmission power between the nodes.

MAX-MATCHES: By comparing the Signalprint (i.e., S1 and S2) the total number of  $\epsilon$  dB is found, denoted by (S1, S2,  $\epsilon$ ), i.e, 10-dB at position I and S1[i] and S2[i] are non-default values, as in equation (3).

If,

$$\text{abs}(S1[i]-S2[i]) \leq 10 \quad (3)$$

MIN-MATCHES: The Signalprint S1 and S2 is compared and the total number of  $\epsilon$  dB is found, denoted by (S1, S2,  $\epsilon$ ), i.e, 10-dB at position I and S1[i] and S2[i] are non-default values, as in equation (4).

If,

$$\text{abs}(S1[i]-S2[i]) \geq 10 \quad (4)$$

## RESULTS AND DISCUSSION

The goal of the research is to defense against the Sybil attack without any trusted authority by minimum computation time by extending the Sybil defense method with Prior Round Reveals RSSI Information.

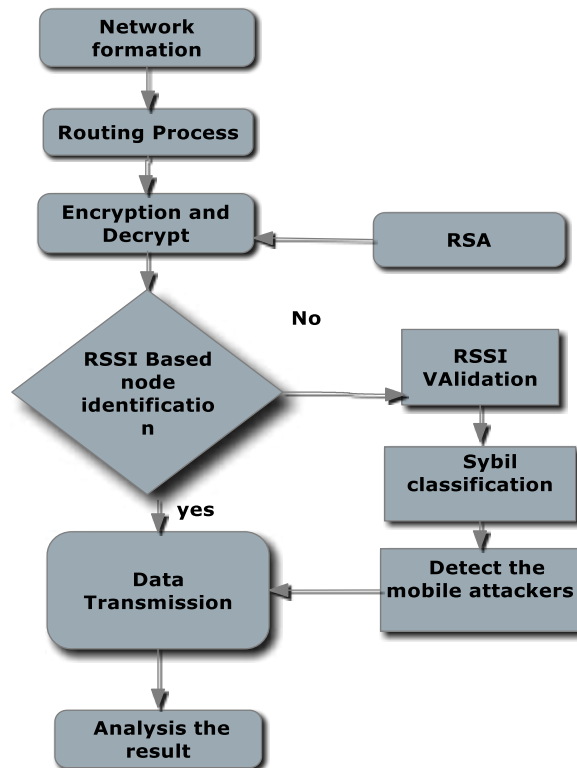


Fig. 4: Sybil Defense Mechanisms

Figure- 4 represent the flow of the defense mechanism. The resultant performance of the Mason Test along with the PRRRI method is evaluated and analyzed based on the simulation result as show in the Figure- 5.

```

simulink.txt *
***** Simulation Result *****
Date: Fri Sep 25 17:12:53 IST 2015

Analyzed File: simple.tr in /root/masontest123/masontest

*****
Total Remained Energy      : 8073.312399
Average Remained Energy    : 807.3312399
Energy Difference          : 179.46624798
Packet Delivery Ratio      : 98.8370188370188
Average End2End Delay      : 0.566106613785712
Average Number of Hops     : 1.48674179648658
Control Packet Overhead    : 22633
Throughput                 : 3621.42857588812
Data Packets Sent          : 6105
Data Packets Received      : 6034
Simulation Endtime         : 99.971597510
Total Delivery Time        : 3536.56851438290
Total Number of Hops       : 8971
Dropped Reply Messages     : 0
Maximum Number of Hops     : 4
Minimum Number of Hops     : 1
*****
  
```

Fig. 5: Simulation Result

### A. Prior Round Reveals RSSI Information (PRRRI)

The method prior round reveals RSSI information is deployed to reduce the high computation time computed during MASON test protocol. The method is not actually the defense mechanism where as it is mechanism to reduce the time of computation time. Three steps of the PRRRI method are:

#### Step 1: Routing Process

The process of selecting the best path to transmit packets between nodes in the open wireless ad-hoc network in the IEEE 802.11. Distance vector routing protocol (DSDV)[10] is the routing protocol used as Routing process. In the 802.11 WLAN network the DSDV[11] operates by having each node  $i$  in the network by maintaining a table, which gives the best distance to each destination and which routes to get information with all its neighbors periodically. Each and every node has a single entry in routing table. The entry node will have following information of the nodes: IP address, lastest know sequence number and the hop count to reach the source node. Along with the details the routing table also holds the track of next hop neighbor to reach the destination node and the timestamp of the last update received for that node i.e., *DSDV\_Agent::Update(int&periodic)*. The updated message of DSDV consist of Destination address, Sequence number and Hop count. *DSDV\_Agent::updateRoute(rtable\_ent \*sequenum, rtable\_ent \*dstadd, rtable\_ent \*nxthop)*. Each nodes deploys two mechanism to send out the DSDV update.s, they are: *Periodic updates*, *Trigger Updates*. When the update with same sequence number is received, the least hop count is given the precedence.

#### Step 2: Node certification

The node is certificated in two different ways: node id, certification id; by RSA cryptography. The generation of node id and certification id deploys the node to be highly secured. RSA generate the public key based on two large prime number must be kept secret. The prime number is large enough, so that someone without the knowledge of prime number cannot decode the message. Figure- 6 represents the Node is certificated by performing RSA Encryption and Decryption, i.e., the certificated node denotes, PRRRI is performed.

#### Step 3: RSSI based Node identification

The prior round RSSI information is made an entry in to hash table and each every time the node is entering the network the Prior round RSSI information is initial step to process the node for data transmission as secure node. In the process of node identification after evaluation of routing process and RSA-Encryption and Decryption, the node is compared in the hash table with RSSI



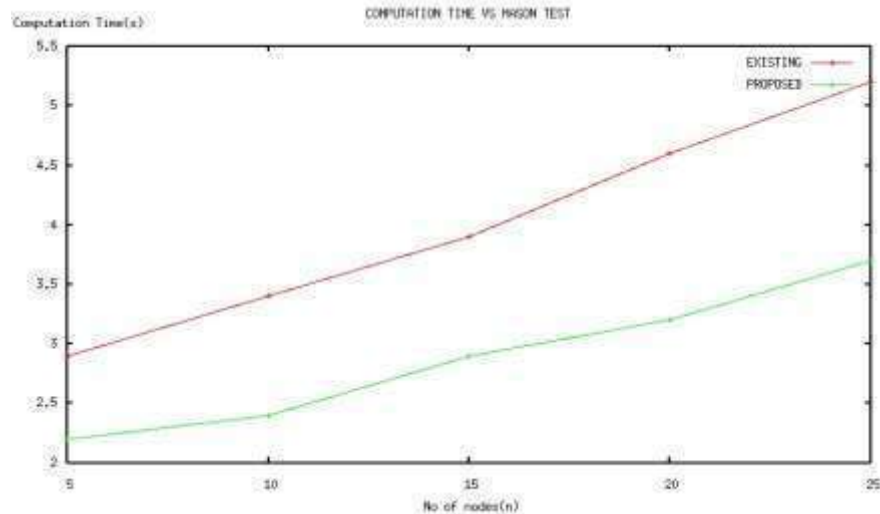


Fig. 7: Comparison of Computation Time

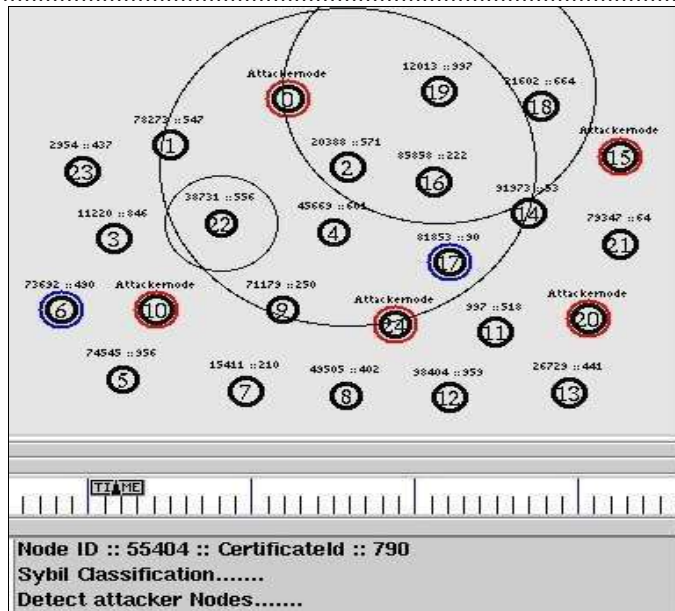


Fig. 8. Detect the Attacker Nodes

The protocol performs two components: RSSI observation [14] and Sybil classification. At the end of the protocol results the nodes are classified in to Sybil and non-Sybil nodes.

1) RSSI observation

The RSSI observation is performed with three phases:

Phase I: Identity collection

The identities participating neighbors ensuring that none of the conforming identities are jammed by attackers are gathered in first phase, e.g. HI message is transmitted each acknowledged with initiator, unacknowledged HI is retransmitted. The process terminated if the channel stays ideal till timeout, all stationary neighbors respond with their own identities.

Phase II: Randomize broadcast request

In second phase the challenge-response protocol RSSI observation and Sybil classification for motion detection. E.g., the participant records the RSSIs of the HI message from the conforming identities. Some identities fails to

responds within minimum duration (i.e., 10ms) might be an attacker attempting to change the physical position and those identities are rejected.

Phase III: Report of RSSI Observation

In third phase first, each identity broadcasts a hash of its observation, then RSSI observation [4][16] values are shared, thus not matching the respective hash values are rejected. To prevent attacker from using the false observation values.

## 2) Sybil Classification

Sybil classification is performed by each participants individually. Correlation between the participants decrease with the RSSI values. The Sybil classification performs only with the current observation uncorrelated with the prior ones. In Algorithm 1, once the receiver set is chosen the set  $S$  contains a truthful receiver set is carried away to examine the -true Sybil classification. The Sybil and non-Sybil nodes are classified and the 99.99% of Sybil nodes are defended in the 802.11 WLAN ad-hoc network.

The goal of the candidate receiver set selection is, at least one of the candidate should be truthful. Size- $n$  is set for desire receiver set,  $S$  is the truthful receiver set,  $R$  is the receiver set identity used to form the Signalprints[2]. Along with  $R$  the random element in the hash table, identities labeled non-Sybil by view  $V$ , i.e.  $V_{NS}(R)$ , is updated to  $R$ . Truthful receiver set id updated with the new set  $\{R\}$ . Updated  $V$  -truthful receiver set is compared with the number of identity whose RSSI [5][6] ratio reported by  $I$  do not match with  $R$ . the view generated by receiver set  $R$   $V_I$  and the view generated by all the participating identities and all Sybil identities i.e.  $V(\{i,s\})$  are not similar. The subset is found with new largest  $V$  - consistent the participating identities are classified as Sybil and non-Sybil identities. **Figure- 8** represents the result of detecting the attacker in open wireless network, named as Sybil nodes.

## CONCLUSION

The Proposed work defense against Sybil attack in ad-hoc network without using any trusted authority. The use of trustless observation is made a significant improvement in detecting the Sybil nodes in the open wireless network. Signalprint method is one among the techniques of untrusted observation is deployed. We have proposed a method Prior Round Reveals RSSI Information to reduce the computation time generated by the MASON protocol. Conforming identities performs classification if their RSSI observations are correlated with the prior rounds. We deployed the RSSI to separate true and false observation of neighbor nodes. The protocol along with the method reduce the computation time compared to protocol deployed alone in IEEE 802.11 WLAN. The protocol robustly defense the Sybil node and method reduce the computation time of the protocol. The performance of the proposed work is analyzed in network simulator. For future work, the method is tested in outdoor and indoor environment and its performance is analyzed.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Murat Demirbas, Youngwhan Song, An RSSI-based Scheme for Sybil Attack Detection in Wireless Sensor Networks, World of Wireless, Mobile and Multimedia Networks, 2006. WoWMoM 2006.
- [2] Daniel B. Faria, David R. Cheriton, Detecting Identity-Based Attacks in Wireless Networks Using Signalprints, WiSe'06, September 29, 2006, Los Angeles, California, USA.
- [3] Daniel B. Faria, David R.[2006] Cheriton, Detecting Identity -Based Attacks in Wireless Networks Using Signalprints, WiSe'06, September29, 2006, LosAngeles, California, USA.



- [4] IyadAldasouqi, WalidSalameh, Detecting and Localizing Wireless Network Attacks Techniques, *International Journal of Computer Science and Security*, 4(1).
- [5] Mohamed Salah Bouassida, Gilles Guette, Mohamed Shawky. [2009] Bertrand Ducourthial, Sybil Nodes Detection Based on
- [7] Advanced Research in Computer Science and Software Engineering 3(8):. 5-10.
- [8] D Monica, J Leitao, L Rodrigues, C Ribeiro. [2009] On the use of radio resource tests in wireless ad hoc networks, in Proc.3rd WRAITS.
- [9] Y Liu, DR Bild, RP. [2013] Dick, Extending channel comparison based Sybil detection to MIMO systems, Tech. Rep. CSE-TR-584-13, Dept. of Electrical Engineering and ComputerScience, University of Michigan, Nov..
- [10] MohitSaxena, Puneet Gupta, Bijendra Nath Jain. [2008] Experimental Analysis of RSSI-based Location Estimation in Wireless Sensor Networks, *Communication Systems Software and Middleware and Workshops*.
- [11] Guoyou He. [2013] Effective Routing Protocol (DSDV) for Mobile Ad Hoc Network, *International Journal of Soft Computing and Engineering (IJSCE)*, 3(5)
- [12] Charles E. Perkins, Pravin Bhagwat, Highly Dynamic Destination-Sequenced Distance Vector Routing (DSDV) for Mobile Computers, SIGCOMM '94 Proceedings of the conference on Communications architectures, protocols and applicationsPages 234-244.
- Received Signal Strength Variations within VANET, *International Journal of Network Security*,9(1):.22-33.
- [6] Zhuliang Xu, KumbesanSandrasegaran, Bin Hu, Cheng-Chung Lin, A Study of WLANRSSI Based Distance Measurement Using EEMD, *International Journal of*
- [13] Madhusudhanan B, Chitra S, Rajan C, Mobility Based Key Management Technique for Multicast Security in Mobile Ad Hoc Networks, *The Scientific World Journal, Hindawi Publishing Corporation*, 2015.
- [14] Yue Liu, David R. Bild. [2015] The Mason Test: A Defense Against Sybil Attacks in Wireless Networks Without Trusted Authorities, *IEEE Transactions On Mobile Computing*, V(99):2
- [15] Giovanni Zanca, Francesco Zorzi, Andrea Zanella and Michele Zorzi, Experimental comparison of RSSI-based localization algorithms for indoor wireless sensor networks, REALWSN'08, April 1, 2008.
- [16] Rajan C, Shanthi N. [2013] Misbehaving attack mitigation technique for multicast security in mobile ad hoc networks (MANET), *Journal of Theoretical and Applied Information Technology*, 48( 3):1349–1357.
- [17] Erin-Ee-Lin Lau, Boon-Giin Lee, Seung-Chul Lee, Wan-Young Chung. [2008] Enhanced Rssi-Based High Accuracy Real-Time User Location Tracking System for Indoor and Outdoor Environments, *International Journal on Smart Sensing and Intelligent Systems*, 1, (2).

# A SURVEY ON MANAGING CLOUD STORAGE USING SECURE DEDUPLICATION

K. Keerthana\*, C. Suresh Gnanadhas, RT. Dinesh Kumar

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Namakkal, T.N., INDIA

## ABSTRACT

*Aim: Cloud is used as a storage platform to store various types of data. Many commercial and sensitive data are available in file format. So, we focus on storing data files in cloud without any replicates. Removing replicates will help the cloud to handle its data efficiently. The process of removing replicates is known as deduplication. But while storing, the files undergo client side encryption to elevate the privacy of user files. This challenges both the user and server to find the duplicate files over encrypted data. Many recent deduplication schemes have been proposed to overcome this challenge. But during client side encryption, all the schemes encrypt the files using symmetric encryption algorithm. In this paper, we survey about various recent deduplication schemes which mainly focus on file – level deduplication and here we analyze the efficiency of existing schemes. Based on the analysis we propose attribute and policy based dedupe system to enhance the security of deduplication process that has efficiency higher than the existing solutions.*

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Cloud, Deduplication, Client side encryption, File – level deduplication, Proof-of-ownership

\*Corresponding author: Email: [k.keerthanakarathikeyan@gmail.com](mailto:k.keerthanakarathikeyan@gmail.com); Tel.: +91 9965727694

## INTRODUCTION

In this Internet era data grows rapidly. Storing huge amount of data locally using Pcs, Pen drive, CD (or) DVD is highly impossible. So the users tend to use the cloud for storage purpose. Cloud is a virtual environment where the user can use remote servers through internet for storing, managing or for processing the data. Cloud storage is one of the services offered by cloud. It provides high security of data, reduce the cost of storage, easy sharing of huge data, data recovery etc. So, using the cloud storage space efficiently is essential for every user. For that purpose data deduplication technique is used.

## OBJECTIVE

Data Deduplication is a method of data compression that eliminates the replicates of data. This process will only save the unique copy of data in a storage media. When a data redundancy occurs it provides a pointer to access that data and ignore the process of saving the redundant data in storage. This process automatically saves the storage space. Reducing the storage space requirements will reduce the cost on disk expenditures, bandwidth usage for transacting data over internet and helps us to use the storage space efficiently. It identifies the redundant data by comparing the data already present in the storage. For preserving privacy of data the user may encrypt a data before uploading it in cloud storage. Then comparing an encrypted data to recognize the redundant data is a challenging task. To overcome this challenge we use convergent encryption algorithm [15]. This algorithm will produce identical cipher text for identical plain text. This algorithm is also known as Content Hash Keying.

## SCOPE

The main aim of this paper is to focuses on File level deduplication which is one type of deduplication. Some types of deduplication which is used in our existing schemes are explained below

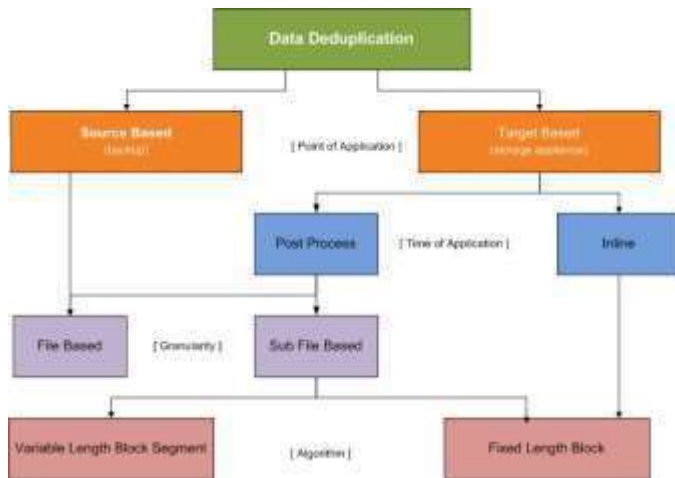


Fig. 1:Types of deduplication

Source based deduplication is also known as client-side deduplication. This process removes the redundant data before transmitting the data to storage (i.e., target). It efficiently decreases the bandwidth usage and storage space where as Target Based deduplication is also known as server-side deduplication. Here the process of removing redundant data takes place within the target system. This method is independent of hardware and software used in client side. Deduplication takes place even when the user is offline.

Post process deduplication is an asynchronous process [11] of removing redundant data after the data is written to the storage and Inline deduplication is a process of removing redundant data before it is written to the storage.

File based deduplication is a process of removing redundant files. It involves calculating single hash value for a file and that hash value is compared against already stored hash value present in cloud storage. If it matches then the file is already present in the storage and then only the pointer will be saved in the storage. The other type is Sub File based deduplication. It is a process of removing redundant blocks present in the files. It split the file into variable size block or fixed size block and calculates the hash value for each block to find the replicates present in the storage.

### ANALYSIS OF EXISTING SYSTEM

This section presents the study on Existing deduplication schemes. This study helps us to find the drawbacks present it in the existing system. Based on this analysis, the proposed system will be designed.

Table 1: COMPARITIVE STUDY ON DEDUPLICATION SCHEMES

SI.N O	TITLE	Type Deduplication	of Algorithms Used	Merits	Demerits
1.	Deduplication on Encrypted Big Data in Cloud	File Level Deduplication	Deduplication Algorithm: PRE (proxy re-encryption) Encryption algorithm: AES Hashing Algorithm: SHA 1 Data Ownership Verification: ECC (Elliptic curve cryptography)	Resist offline brute force attack, low cost and access control over encrypted data	Hash function and symmetric encryptions are not highly secured.

2.	Encrypted Data Management with Deduplication in Cloud Computing	File Deduplication	Level	Deduplication Algorithm: CP-ABE	Low operational and implementation cost. Doesn't depend on third party	Consume more time for key generation.
				Encryption algorithm: AES		
				Hashing Algorithm: SHA 1		
				Data Ownership Verification: RSA		
3.	A Hybrid Cloud Approach for Secure Authorized Deduplication	File Deduplication	Level	Encryption algorithm: AES	Resilient to Insider and outsider attack.	Vulnerable to Brute force attack.
				Tag Generation: SHA-1		
				Token Generation: HMAC-SHA-1		
4.	A Scheme to Manage Encrypted Data Storage with Deduplication in Cloud			Deduplication Algorithm: PRE	Resist offline brute force and dictionary attack	Cannot implemented directly in cloud
				Encryption algorithm: AES		
				Data Ownership Verification: RSA		
5.	Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage	File deduplication	Level and side	Encryption algorithm: AES with electronic code book algorithm	Secured against chosen-plaintext attack, collusion attack and poison attack. It has forward and backward secrecy.	Exposed to data loss attack.
				Key and token generation: MD5		
				Deduplication Algorithm : Randomized convergent Encryption		
6.	Encrypted Data Deduplication in Cloud Storage	Client side / Server deduplication		Hash Function : SHA 2	Secured cipher text.	High Cost
				Encryption algorithm: AES, RSA (Asymmetric encryption) and Elgamal (homomorphic encryption)		
7.	HEDup: Secure Deduplication with Homomorphic Encryption	File deduplication	level	Encryption algorithm: AES	Resilient to Man in middle attack	Resilient to Passive attacks
				Hash Function : MD5 Or SHA		

8.	BDO-SD: An Efficient Scheme for Big Data Outsourcing with Secure Deduplication	Support both File level deduplication and block level deduplication	Encryption algorithm: AES	Resist Brute force attack	communication overhead
			Hash Function : SHA 1		
			Data Ownership Verification: blind signature system		
9.	Secure Deduplication of Encrypted Data without Additional Independent Servers	File deduplication Level	Deduplication algorithm: PAKE	Secure with Online brute force attack.	Exposed to Offline brute force attack, dictionary attack, week passwords and. Block level deduplication will cause additional overhead.
			Key generation: AES		
			Encryption algorithm: Elgamal		
			Hashing Function: SHA 256		
10.	Enhanced Secure Thresholded Data Deduplication Scheme for Cloud Storage	File deduplication Level	Encryption algorithm: AES – 128 – CTR	Resist User collusion attack & Low cost	Vulnerable to attacks based on knowledge of the hash of the plaintext file

### PROPOSED SYSTEM

This section contains a description about the proposed work namely Attribute and policy based Dedupe System. It adds security to the deduplication process by using various factors such as certificates, signature, access control policies, hashing algorithm etc.

#### Attribute and policy based Dedupe System

In this system, we introduce access control policies to secure the user data. Two types of access control policies are used. One is providing read only access to user and the other is providing read & write access to the user. These policies will be provided based on the attributes of the user. Data owner and Data holder are the two kinds of user attributes. For data owner both read & write access will be provided and for data holders only read access will be provided.

#### Terms involved in Attribute and policy based Dedupe System

##### Key generation

The user should generate key based on their attribute and access control policy system [APS]. For E.g., if U1 is a data Owner who has Read Write access then his key is ORW23 where 23 is a random number chosen by user. This key is used as public key of user [PK]. Then the user should generate key pair for hashing algorithm (SHA 2). The dedupe checker will generate its key pair using CP ABE [cipher policy attribute based encryption]. And public key of dedupe checker is distributed to all CSP Users

##### Token and Signature generation

Using SHA 2 algorithm the user can generate data token [hash value of their file]. Then the signature is generated by encrypting the hash value using public key of a user.

### Re encryption Key

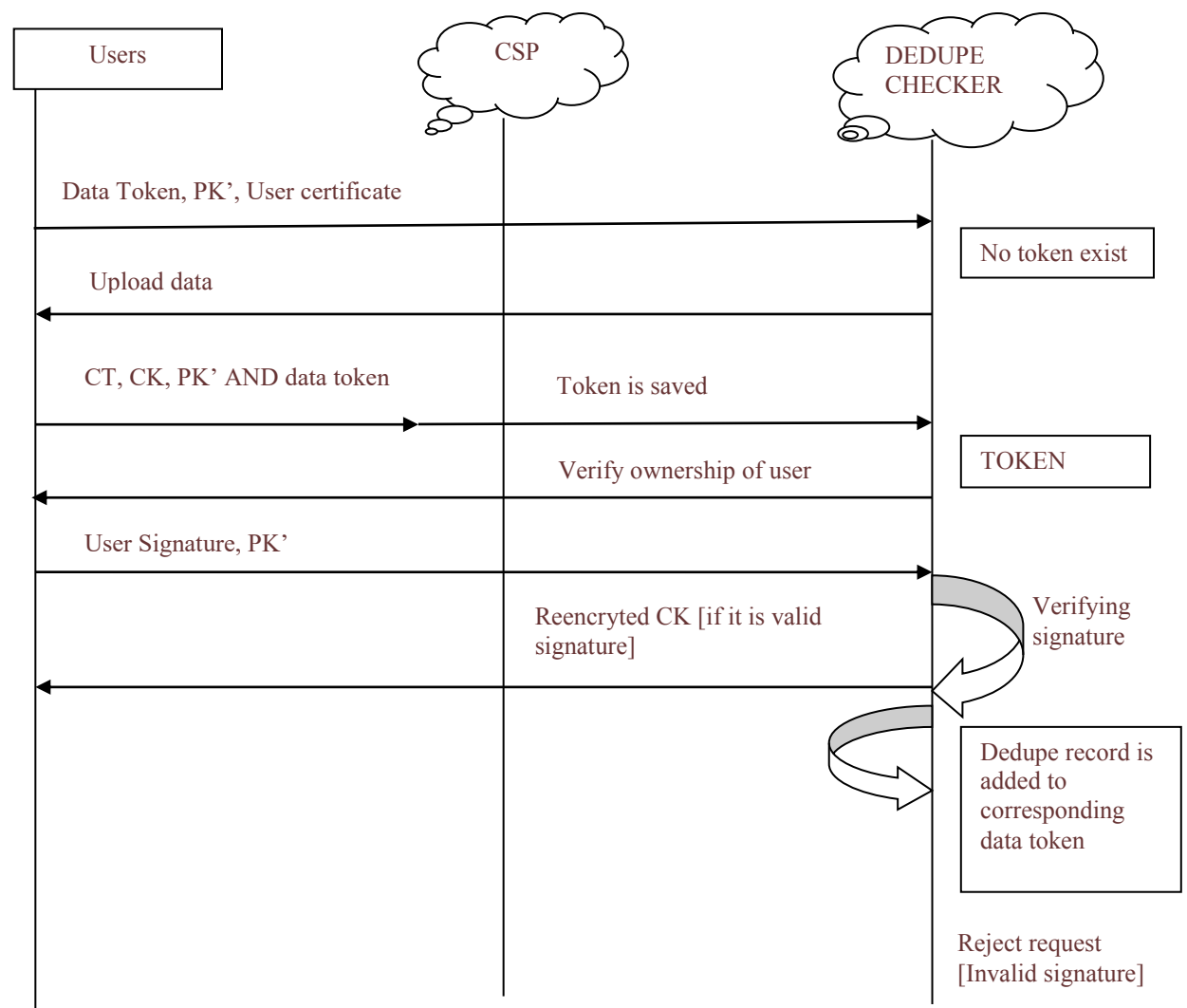
If dedupe occurs and the user is a valid data holder then dedupe checker will generate the reencryption key. The corresponding CK of that data which is stored in CSP is reencrypted by the public key of the data holder who proves their ownership.

### Signature Validation

When dedupe checker receives signature and encrypted public key PK' of user for verification. First the dedupe checker will decrypt the public key using its secret key. Then the signature is decrypted by obtained user public key. Finally a hash value will be produced. If that hash value matches with the data owner hash value then he/she is a valid data holder.

## SYSTEM ARCHITECTURE

The system architecture of Attribute and policy based Dedupe System is given below



Where CT = Encrypted user file

CK = Key used for encryption is encrypted by dedupe checker's public key

PK' = Public key of user encrypted by dedupe checker's public key

**Fig. 2: Deduplication procedure in Attribute and Policy based dedupe system**

## ARCHITECTURE DESCRIPTION

Initially, each CSP user should generate their personal credential such as user key pairs and get the certificate from the dedupe checker which is an authorized third party present in the cloud. To upload the data in CSP, the user should generate data token using hashing algorithm SHA 2. Then the user can request the dedupe checker for duplication check by sending data token, Public key of user encrypted by dedupe checker's public key [PK'] and user certificate.

The dedupe checker will verify the user certificate and then compare the data token with the tokens which is already present in its storage. If no tokens match with the token send by the user then the data is not present in the CSP. Now, the dedupe checker will request the user to upload the data. Then the user send encrypted data [CT], data token, PK' and CK [key used for encryption is encrypted by dedupe checker public key which is known as CK] to the CSP for storage. The CSP will forward only the token [since no subsequent upload of that data occurs] to the dedupe checker for storage.

If the token send by user matches with the existing token in dedupe checker then the duplication occurs. Then the dedupe checker will request the user to prove their ownership of data. The user will generate the signature based on their data token and user public key and send the signature and encrypted public key [ using public key of dedupe checker] to dedupe checker for verification. The dedupe checker will validate the signature. If the signature is valid then the dedupe checker will reencrypt the CK and send it back to the user for accessing the data. Then a dedupe record will be add with the index of data token to the dedupe checker. If the signature is not a valid one then someone is trying to get the data from CSP by sending invalid signature. So, the dedupe checker will reject that user request.

## LITERATURE SURVEY

This section gives a detailed review about various deduplication schemes. Here we reviewed how the deduplication problem is resolved in each scheme. While reviewing a scheme we listed the algorithms and techniques used in that scheme and the merit and demerit of that scheme are also specified. The following papers are survived in this section.

### Deduplication on Encrypted Big Data in Cloud

It has a novel scheme for deduplication integrated with access control. It uses proxy re-encryption algorithm (1024 bit) for deduplication, AES (256 bit) for encrypting and decrypting a file, SHA 1 for hash function and ECC [18] (192 bit) for data ownership verification. This scheme also supports data sharing with deduplication even when the data holder is offline. We apply Elliptic curve cryptography and user certificate to prove the ownership of users. Initially to upload the data user will send a data token along with his public key and his user certificate. The cloud service provider (CSP) will verify the user certificate and check whether the data token already exist or not. If it doesn't exist then csp request the user to upload the data. If data token exist then csp will forward the user request to a trusted third party which is known as authenticated party (AP). The AP will challenge the user to prove that he/she contains the entire data. If the challenge is proved using ECC then a reencrypted key is provided to user to access the data. If else then the user request will be rejected by CSP. In this scheme Data Updating, Data Deletion and Data owner Management is also possible. This scheme has a merit that it resist offline brute force attack caused by convergent encryption. This scheme also saves storage space of CSP which leads to low cost for storage and it support access control on encrypted data. But the demerit in this scheme is same key is used for encrypting and decrypting a file which is not highly secure.

### Encrypted Data Management with Deduplication in Cloud Computing

It proposes a method for deduplication along with secure access control using attribute based encryption [ABE]. It adopts AES (256 bit) for symmetric encryption, RSA for Public Key Cryptography (PKC) and CP-ABE [16] for data deduplication and SHA 1 for hash function. At first the user will create two pair of keys which are RSA key pair generation and ABE key pair generation. And the public keys of user must be certified by authorized third party. When user attempts to save data which is already present in CSP then the deduplication occurs. The user requests the CSP to upload the data by sending DATA PACKET. The data packet contains a cipher text of user ( $CT_u$ ), cipher key of a user ( $CK_u$ ), hash value of the text  $H(M)$ , hash value signed by ABE secret key of a user  $H(M)$  by  $sk_u$  and the two certificates of user  $Cert(PK_u)$  &  $Cert(pk_u)$ . CSP verifies the certificates then check whether the  $H(M)$  by  $sk_u$  is already present. If so then it is from same user it informs the user. If it is from different user then csp contact the data owner. The data owner will verify the eligibility of user. If the result is positive then data owner will issue a secret attribute for the user to access the data. And finally data owner report the CSP about the successful deduplication of a user then the CSP deletes the corresponding user cipher text and cipher key. In this method Data Updating, Data Deletion and Data owner Management is also possible. Some merits of this scheme are low operational and implementation cost and it does not depend on the third party for Key generation. But the disadvantage over here is it consume more time for key generation.

### A Hybrid Cloud Approach for Secure Authorized Deduplication

It describes a scheme for authorized duplication check in hybrid cloud. In this scheme the private clouds maintain a table that contains user's public key and their privileges. It uses 256 bit AES algorithm for encryption, tag generation SHA-1 (256 bit) algorithm and for token generation HMAC [17] is produced using SHA-1. For data upload the user needs to contact private cloud to prove his/her identity. If the user proves then the private cloud find the privilege of the user. When user send the tag of the file to the private cloud it return the file token to the user. User will send the file token to the CSP for deduplication check. If the check is positive then the user has to that he/she is an authenticated data owner. It is done by running Proof of Ownership (PoW) on CSP. If it is proved then a pointer will be provided to user to access the file and Proof (i.e., signature on file token) & time stamp will also be provided. Now, the user uploads the privilege set for the corresponding file along with the proof from CSP to the private cloud. The private cloud first verifies the proof then it compute the file token and return the token to the user. If the check is negative then there is no need to run PoW. The same steps are followed and after the user get the file tokens from the private cloud he/she will compute the encrypted file by using convergent key and finally upload the file with its privilege. Only file uploading and retrieval is specified. This scheme is secured with insider and outsider attack which add merit to this scheme. But it has a disadvantage that files which are predictable are vulnerable to brute force attack.

### A Scheme to Manage Encrypted Data Storage with Deduplication in Cloud

It uses Proxy re encryption (PRE) for data deduplication, AES for symmetric encryption and RSA for PKC. Here the user will create two pair of keys which are RSA key pair generation and PRE key pair generation. And the public keys of user must be certified by authorized third party (AP). When user attempts to save data which is already present in CSP then the deduplication occurs. The user requests the CSP to upload the data by sending DATA PACKET. The data packet contains a cipher text of user ( $CT_u$ ), cipher key of a user ( $CK_u$ ), hash value of the text  $H(M)$ , hash value signed by PRE secret key of a user  $H(M)$  by  $sk_u$  and the two certificates of user  $Cert(PK_u)$  &  $Cert(pk_u)$ . CSP verifies the certificates then check whether the  $H(M)$  by  $sk_u$  is already present. If so then it is from same user it informs the user. If it is from different user then CSP contact the AP. The AP will verify the eligibility of user. If the result is positive then AP will generate a reencryption key for the user to access the data and send the key to CSP. CSP forward the reencryption key to user and finally user report the CSP about the successful deduplication then the CSP records the deduplication information & deletes the corresponding user cipher text and cipher key. In this method Data Updating, Data Deletion and Data owner Management is also possible. This method is resilient to offline brute force and dictionary attack which becomes the advantage of this scheme. But this scheme cannot implement directly in cloud.

### Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage

It has a scheme that uses randomized convergent encryption to manage dynamic changes in ownership of data. This scheme also uses AES with electronic code book algorithm for encryption and decryption, MD5 algorithm for key and token generation. Initially the cloud server will set up a binary tree for universal user to store their key encrypting key (KEK) which are used for the generation of group keys. Here users are at leaf node. The path from



root to leaf is known as path keys. While encrypting a message user will find the hash value for message  $K_i = H(M_i)$  and use that hash value for encrypting the key  $L$  which is used for encryption  $[C_2 = E_{ke}(L)]$ . The hash value of key  $K_i$  is used as tag  $T_i$ . Encrypt the message that is to be uploaded in cloud by using key  $L$   $(C_1 = E_L(M))$  and the cipher text is constructed using  $C_1$  and  $C_2$   $[C_i = C_1 + C_2]$ . Now the user will upload  $T_i$ ,  $C_i$  & ID of the user. Then user deletes the message and retains only its Key for access. The server will insert ID of the user in a group  $G_i$  and maintain the Ownership List  $L_i$  that store the tag value of that group. If  $L_i$  is already exist then cloud will just insert the ID into  $G_i$  to avoid duplication. This is done when the user is a subsequent uploader. Then reencrypt the  $C_1$  by using a randomly selected group key  $(GK_i)$  which is  $C_3$ . Then select the root node that covers the all the user of that group  $[KEK(G_i)]$  and encrypt the  $GK_i$  by root node keys  $E_k(GK_i)$  where  $k = KEK(G_i)$ . Again the Ciphertext will be generated  $[C'_i = C_1' + C_2 + C_3]$ . This ensures that only member of  $G_i$  can decrypt the cipher text. When user send tag and Id for access the data then cloud will check the tag is present in ownership list. If present then CSP will send Tag and  $C'_i$ . Key update and data modification is also possible in this scheme. The merits are it is secure against the chosen-plaintext attack, collusion attack and poison attack. It also ensures forward and backward secrecy of outsourced data. But it can't recover original data under data loss attack because all duplicates are removed from the CSP.

### Encrypted Data Deduplication in Cloud Storage

It proposes a scheme in which the user construct a cipher structure that contains four blocks namely check block, converting block, enabling block and cipher block. In this scheme SHA 2 algorithm is used for hash function, AES 128 bit algorithm is used for Symmetric encryption, RSA algorithm is used for Asymmetric encryption and Elgamal algorithm is used for homomorphic encryption. First, the hash value of the file is calculated and it is saved as Check Block where as Cipher Block contains the encrypted file which is encrypted using AES key shorter than prime modulo  $p$ . Then the encrypted AES key by user public key is saved as Enabling Block and it is encrypted using multiplicative asymmetric homomorphic encryption. At last the user computes the second hash value of the file and it is multiplied by the AES key under the modulus  $p$  which is known as Converting Block. Then the cipher structure will be uploaded to the cloud by user. The cloud server will identify the duplicate files by verifying the Cipher block of the structure. If there is no match found in the CSP then the whole cipher structure is saved in the CSP. If the match exists then CSP will convert the Enabling Block by encrypting the AES key of user1 by public key of user2 and CSP saves only the Converted Enabling Block rather than saving the cipher structure uploaded by user2. In this situation when user2 request for data retrieval then CSP will send the converted Enabling Block along with the cipher block of user1. Then the user decrypt the converted Enabling Block by using his/her secret key to obtain AES key. Finally by using AES key the user2 can decrypt the cipher block. This scheme also specifies only the file uploading and retrieval. The merit of this scheme is it improves the security of the cipher text by using randomized encryption algorithm. But it has a demerit too that is the total cost of this scheme is larger than the existing system [19] [20] [21] [22]

### HEDup: Secure Deduplication with Homomorphic Encryption

It describe about the deduplication will take place with the help of Key Server deployed at CSP. Key Server is mainly used to provide Data Encryption Key to the users. Here AES 128 bit algorithm is used for symmetric encryption and MD5/SHA algorithms can be used for hash functions. When different users upload same file then same data Encryption Key will be provided by Key Server. In this scheme I the file uploading scenario is classified into two types namely First upload and subsequent upload. In First upload the user will authenticate with the Key Server using their credentials. If it is Success then key server will generate public and private key for the user using asymmetric encryption. Then the user will encrypt the file using his/her public key and send it to the key server for duplication check. If the Key Server returns Null then the client will send hash of the file  $H(F)$  and hash part of the file encrypted by  $R$  [random number]  $[E(H(F'), R)]$ . Now the Key Server will save the  $H(F)$  as Key and  $[E(H(F'), R)]$  as value. At last the client authenticate the storage [CSP] and save the cipher text  $C_1$  and cipher Key  $C_2$ . If it is a subsequent upload then during duplicate check the hash value will be present in the Key Server. And Key server will return  $S = E(H(F'), R)$  to the user. Only valid user can obtain  $R$  from the  $S$ . After obtaining  $R$  the user will authenticate with storage and save their  $C_1$  and  $C_2$ . The only difference in Scheme II is during subsequent upload the clients no need to calculate  $C_1$ . The Key server will provide link of  $C_1$  to storage after the client authenticate with storage. In scheme III the homomorphic XOR [14] operation is included to enhance the security. In this three schemes only file uploading and retrieval is specified. The advantage is it resists man in middle attack since we are encrypting the file in client side. The disadvantage is vulnerable to passive attack such as unauthorized reading of the file and traffic analysis etc.

### BDO-SD: An Efficient Scheme for Big Data Outsourcing with Secure Deduplication

It says about outsourcing big data with data deduplication. It is done using convergent encryption and it also has Keyword search over encrypted data. This method support both file level and block level deduplication. It adopt AES (256 bit) for symmetric encryption, SHA 1 for hash function, blind signature system [13] is used for data owner verification. Initially each user will register with trusted authority (TA) and it sets  $(g_{xj}, x_j)$  be user public key and private key. And for each data owner TA will assign ID-based Key pairs. And data owner generate the convergent keys for file by using blind signature system. Then the data owner generates tag [12] for the file and sends it to CSS. CSS verify that tag is already present. If yes then data owner should pass the ownership authentication. If it is a valid data owner then a pointer for the file will be sent to the owner and CSS will abort the upload process. If result of deduplication check is no then block level deduplication will be preceded. The user will generate tag for blocks and send to CSS. If tag matches then ownership of user is verified. If it is valid user then pointer of blocks will be saved. If tag for blocks doesn't match then cipher block is generated using convergent keys. For that purpose, encrypted convergent keys should be generated for each block. Using encrypted convergent keys keyword search is possible. Then data owner uploads the unique blocks  $\{B_i\}$ , all encrypted blocks and encrypted keys  $(C_i, CK_i, S_i, CK_i')$ , as well as  $T(B_i)$  to the CSS. Then CSS will store the signature, cipher key and tag for blocks. If CSS receive the request for retrieval then it verify that user is eligible for retrieval after verification the cipher text and encrypted convergent keys will be provided to user. In this scheme trapdoor generation is also specified. The merit is it resists brute force attack. And the demerit is this scheme has more communication overhead.

### Secure Deduplication of Encrypted Data without Additional Independent Servers

It briefly describe about cross-user deduplication scheme that supports client-side encryption without requiring any additional independent servers. It uses password authenticated key exchange [PAKE] protocol. PAKE Protocol is used for deduplication, AES algorithm is used for pseudo random function and elgamal algorithm is used for additive homomorphic encryption. To upload a file the client C will calculate short hash and hash value for the file. The short hash may be same for many different files. When client send short hash to storage it check for existing clients who have uploaded files with same short hash. Then client should run the PAKE protocol with hash as input function. Then client get the session keys of existing user who upload the same file. Then the existing clients & C use the pseudorandom function to extend the length of session key and split the key into two one is left part of key and other is right part of key. Then left part of key, client public key, and encrypted right part of the key is sent to storage from the existing client and C. then storage check the index value and use homomorphic encryption to calculate e send e to client. Then client calculate the key  $K_f$  from e and encrypt it then it send to the storage. If it is already present in the storage S then s will pointer pointer for the client to access the data. Here the pseudo random function is implemented using AES 128 bit algorithm and homomorphic encryption is done using Elgamal algorithm. Online brute force attack can be prevented which add a merit to this scheme. It is vulnerable to offline brute force attack, dictionary attack and hacking of low level entropy passwords and if block level deduplication takes place then it will cause additional overhead.

### Enhanced Secure Thresholded Data Deduplication Scheme for Cloud Storage

It deals with the new concept known as data popularity. It describes an encrypted scheme that support secure cipher text of a file is downgraded to convergent cipher text of file that support deduplication. It happens when the file become popular. Here two trusted authority is used one is Identity Provider and the other is index repository service [IRS]. Here AES – 128 – CTR used as symmetric encryption, SHA 256 is used for hashing function. The user encrypts the file to generate index and sent it to IRS. If IRS returns the index unchanged then the file is already popular. Then only user is added to the owner list. If IRS return different index then the data is unpopular. Then doubly encrypted file with encrypted random key is send to storage. In this scheme deduplication request is sent to storage using IRS. IRS will send indexes and decryption shares to storage. Then x check the corresponding record for the indexes and decrypt the records using decryption shares. If all the convergent cipher text is equal then it add a new record under index and delete the copies of convergent cipher text. If it is not equal then some clients have cheated S so it abort the deduplication process and sent a failed message to IRS. This scheme also support file retrieval and deletion from the storage. The merit of this scheme is it has low cost and resists user collusion attack. The demerit is it cannot prevent attacks based on knowledge of the hash of plaintext file.

### PERFORMANCE ANALYSIS

This section illustrates the comparison between efficiency of encryption and decryption process present in the existing scheme [2] and the proposed work.

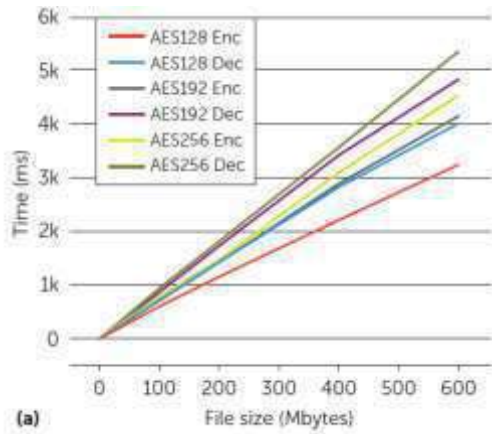


Fig. 3: Operation time of encryption and decryption using AES

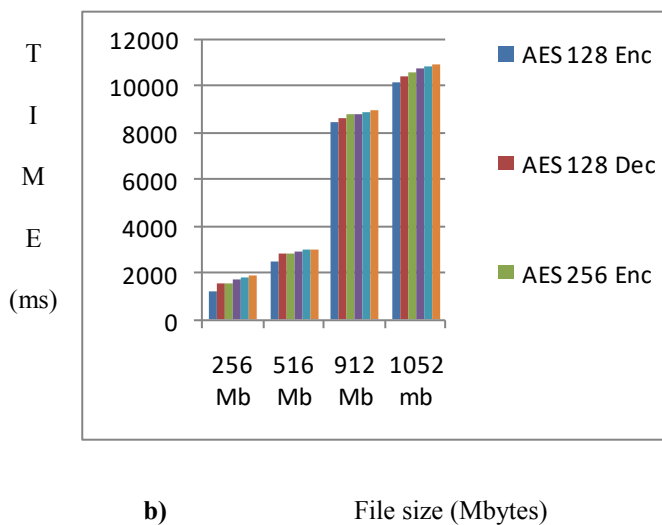


Fig. 4: Operation time of Encryption and Decryption of AES and Attribute and Policy based system

Since all the existing schemes assume that the symmetric encryption is secure and they encrypted the user file using AES algorithm. So, we have analyzed the efficiency of different types of AES in Figure- 3. Among the three types of AES, AES 256 is secure and efficient.

In our proposed system the encryption and decryption is based on attribute and policy. So we make a comparison between AES and Attribute and Policy based system in Figure- 4. Compared to AES 256, Attribute and policy based system is highly secure. The time taken for encryption and decryption increases with increase in data size. But it takes only few milliseconds. So, Attribute and policy based system is efficient and provides high security than the existing schemes.

### CONCLUSION

In this paper, we have survived various deduplication schemes which eliminate the replicates of files from the cloud and also we analyzed those schemes to find the finest method for deduplication. Based on the analysis, many demerits were found which reduce the efficiency and security of deduplication process. To overcome these limitations we have proposed Attribute and Policy based dedupe system to enhance the security of deduplication

schemes and also it provide security against tampering, unauthorized access etc. In future, the proposed method will be implemented with additional features such as data updation, data deletion in cloud etc., which increase the security & efficiency of deduplication process higher than the existing schemes.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

### REFERENCES

- [1] Zheng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng. [2016] Deduplication on Encrypted Big Data in Cloud,” in IEEE TRANSACTIONS ON BIG DATA, 2(2) APRIL-JUNE pp. 138- 150.
- [2] Zheng Yan, Mingjun Wang, Yuxiang Li, and Athanasios V. Vasilakos, Encrypted Data Management with Deduplication in Cloud Computing,” in IEEE Cloud Computing, March/April 2016, pp. 29- 35.
- [3] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, and Wenjing Lou, “A Hybrid Cloud Approach for Secure Authorized Deduplication,” in IEEE Transactions on Parallel and Distributed Systems.
- [4] Zheng yan, Wenxiu Ding, and Haiqi Zhu, “A Scheme to Manage Encrypted Data Storage with Deduplication in Cloud,” in Springer International Publishing Switzerland 2015, pp. 547–561.
- [5] Junbeom Hur, Dongyoung Koo, Youngjoo Shin, and Kyungtae Kang, “Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage,” in IEEE Transactions on Knowledge and Data Engineering, 1041-4347 (c) 2016 IEEE.
- [6] Chun-I Fan, Shi-Yuan Huang and Wen-Che Hsu in “Encrypted Data Deduplication in Cloud Storage,” in 2015 10th Asia Joint Conference on Information Security, pp. 18-25, 978-1-4799-1989-5/15 \$31.00 © 2015 IEEE.
- [7] Rodel Miguel, Khin Mi Mi Aung, and Mediana, “HEDup: Secure Deduplication with Homomorphic Encryption,” 978-1-4673-7891-8/15/\$31.00 ©2015 IEEE, pp. 215 – 223.
- [8] Mi Wen, Kejie Lu, Jingsheng Lei, Fengyong Li, and Jing Li, “BDO-SD: An Efficient Scheme for Big Data Outsourcing with Secure Deduplication,” in The Third International Workshop on Security and Privacy in Big Data (Big Security2015), pp. 214 – 219.
- [9] Jian Liu, N. Asokan, and Benny Pinkas, “Secure Deduplication of Encrypted Data without Additional Independent Servers,” pp. 874 – 885.
- [10] Jan Stanek, and Lukas Kencl. [2012] Enhanced Secure Thresholded Data Deduplication Scheme for Cloud Storage, in IEEE Transactions On Dependable And Secure Computing, 1545-5971 (c) IEEE.
- [11] <http://www.druva.com/blog/understanding-data-deduplication/>
- [12] JR Douceur, A Adya, WJ Bolosky, D Simon, M Theimer.[ 2002] Reclaiming Space from Duplicate Files in a Serverless Distributed File System”, in Proceeding of ICDCS, pp. 617-624.
- [13] M Bellare, S Keelveedhi. [2013]DupLESS: Server-aided encryption for deduplicated storage,” in Proceeding of USENIX Security Symposium, pp. 179-194.
- [14] Shu Qin, Ren., et al. “Homomorphic Exclusive-Or Operation Enhance Secure Searching on Cloud Storage,” in Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on 2014.
- [15] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Message-locked encryption and secure deduplication,” in Proc. Cryptology—EUROCRYPT, 2013, pp. 296–312, doi:10.1007/978-3-642-38348-9\_18.
- [16] <http://crypto.stackexchange.com/questions/17893/what-is-attribute-based-encryption>.
- [17] [http://searchsecurity.techtarget.com/definition/Hash-based-Mes sage -Authentication-Code-HMAC](http://searchsecurity.techtarget.com/definition/Hash-based-Message-Authentication-Code-HMAC)
- [18] <https://bithin.wordpress.com/2012/02/22/simple-explanation-for-elliptic-curve-cryptography-ecc/>
- [19] J Daemen, V Rijmen.The advanced encryption standard (AES),” United States National Institute of Standards and Technology (NIST), vol. Federal Information Processing Standards Publication 197, November 26 2001,<http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf>.
- [20] NSA. (NSA), “Secure hash standard (SHS),” United States National Institute of Standards and Technology (NIST), vol. Federal Information Processing Standards Publication 180-4, March 2012,<http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf>.
- [21] R Rivest, A Shamir, L Adleman.[1978] A method for obtaining digital signatures and public-key cryptosystems,” Communications of the ACM, 21(2):. 120126,, <http://people.csail.mit.edu/rivest/Rsapaper.pdf>.
- [22] T Elgamal.[1985] A public key cryptosystem and a signature scheme based on discrete logarithms,” IEEE Trans. on Information Theory, 31(4): 469–472, 1985.

# INTERNET OF THINGS: A SURVEY ON PRIVACY AND SECURITY FOR SMART HOMES

G. Devi\*, R. Rohini, P. Suganya

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tiruchengode, Tamilnadu, INDIA

## ABSTRACT

**Aim:** Smart Home is one of the applications of Internet of things, in networking side Internet of Things is popular one. The smart home environment is a building block of the future Internet, and many homes are fitting "smarter" by using Internet of Things (IoT) technology to improve home security, energy efficiency and comfort. Latest techniques are used in smart home automation system like automatic door open, automatic light off/on using RFID sensors, etc. The need of home safety is specifically when the elderly person is alone or Children are with baby-sitter and servant. The home attack crimes are done by breaking the door or window and continuously monitoring the home is a challenging one. To Overcome this crimes using door locking sensor is placed in the door and this sensor is to capture the face who are all coming inside the home and the faces are captured. The face will be stored in the cloud and the captured face is verified by the owner of the home. To review the latest technologies are used in smart home automation system security methods in door locking system.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Internet of Things (IoT), Smart Home Automation System (SHAS), Smart Lock System (SLS), Radio Frequency Identification (RFID), Sensors.

\*Corresponding author: Email: [devicse17@gmail.com](mailto:devicse17@gmail.com); Tel.: +91 9095620265

## INTRODUCTION

During the past decades, internet has changed way of the people to communicate with each other by creating a virtual world for both professional and social lives. Internet of Things (IoT) can be considered as an expansion of the internet that will impact our everyday lives and the way we interact with things. IoT can be defined as a world-wide network of interconnected physical things [1]. Smart objects are the building blocks of IoT, which are things we use every day, enhanced with embedded intelligence as well as connectivity to the internet. A smart object can be a lamp that turns on when you get home, or a TV that knows your favorite shows. Objects that are connected to IoT can communicate using a standard protocol.

The Internet of Things (IoT) is the network of physical objects devices, vehicles, buildings and other embedded devices with electronics, software, sensors, and network connectivity that enables these objects to collect and exchange data. [2] When IoT is augmented with sensors and actuators, the technology becomes an instance of the more general class of cyber-physical systems which also encompasses technologies such as smart grids, smart cities, industries and homes, intelligent transportation and smart cities. Each thing is uniquely identifiable through its embedded computing system but is able to interoperate within the existing Internet infrastructure. Researcher estimates that the IoT will consist of almost 50 billion objects by 2020. Internet and its applications have become an integral part of today's human lifestyle.

Smart home is now becoming prevalent with the development of the Internet of things (IoT) techniques. It is aimed at providing the user with a user-friendly method to control the home appliances such as doors, lights, even in a condition of long-distance. This controlling is generally achieved by a mobile phone which can access to the Internet.

## REQUIREMENT OF AUTOMATION

When we talked about automated devices which could start with controller, but today it has become a reality.

i) An automated device can replace good amount of human working force, moreover humans are nears to errors and in intensive conditions the probability of error increases where as an automated device can work with usefulness and almost zero error. [3]

ii) Replacing human operators in tasks that involve hard or uninteresting work. Replacing humans in tasks done in dangerous environments (i.e. fire, space, volcanoes, nuclear facilities, underwater, etc)

## EXISTING SYSTEM

In this work the detailed study of smart home automation security and privacy techniques are analyzed and studied. Different types of security mechanism are used for access the home control and monitoring. Some technologies are used for authentication purpose like people enter to home means get a notification to the user. The various mechanisms are reviewed in the below section.

In [5] authors A. Jacobsson M. Boldt and B. Carlsson “**A risk analysis of a smart home automation system**” says the surveillance camera can also be used for other personal purposes such as to see who is at home with respect to childcare, control of infants sleeping, elderly care, etc. For example if lamps are switched on or off, doors closed or open, cameras showing water leaks, etc. Smart home surveillance cameras can also be used in combination with other connected devices, which together may provide an overly detailed image of the persons living in that home. In [6] authentication procedures are communicated by cryptography technique and the communication is done between the objects. Based on our research to suggest the following steps should be included in the model:

1. Identification and categorization of the personal data in smart homes.
2. Analysis and description of the main risks to privacy and security.
3. Identification and implementation of preventive, detective measures to shrink risks.
4. Strategy for privacy-friendly information management within smart homes.

The analyzed system was divided into the following six parts [7]

1. Connected sensors/devices/actuators
2. In-house gateway.
3. Cloud server.
4. API (Application Program Interface).
5. Mobile device.
6. Mobile device apps.

In [8] authors Hui Suoa, Jiafu Wana,b, Caifeng Zoua, Jianqi Liua “**Security in the Internet of Things: A Review**” says that the IoT will be faced with more challenges in smart home security system. There are the following reasons:

- 1) The IoT extends the Internet through the traditional internet, mobile network and sensor network.
- 2) Every Thing will be connected to this network
- 3) These Things will communicate with each other.

Day by day new security and privacy problems are identified. In [9] the security issues are to concentrate on confidentiality, authenticity, and integrity of data in the IOT.

**Table: 1. Security Algorithms**

S.NO	Algorithm	Purpose
1	Advanced Encryption Standard (AES)	Confidentiality
2	Rivest Shamir Adelman (RSA)	Digital signatures key transport
3	Diffie Hellman (DH)	Key agreement
4	Secure Hash Algorithm (SHA)	Integrity

In [10] authors Andreas Jacobsson ,Martin Boldt and Bengt Carlsson “**On the Risk Exposure of Smart Home Automation Systems**” says that smart home automation, energy services depend on a broad range of hardware and software components for monitoring and controlling an apartment or building. The sensors and actuators record the report metrics like water usage, indoor temperature, and power consumption. Each device runs

independently of each other and communicates using a local mesh network. In this case Zigbee is used with a home gateway and act as the central node. The gateway runs a minimalistic Linux distribution and relays device information to a remote, or cloud, server over the Internet using the protocol.

In [11] the Smart Home Automation System (SHAS) platform used for the connected devices is distributed across multiple hardware solutions, each with its own set of responsibilities and privileges. As the available devices in each apartment differ from each instance of the platform will subsequently be different. The fundamental operations of the platforms are operational capacity the in-house gateway for relaying messages that devices send over the local communication protocol (Zigbee or ZWave) to an Internet-based protocol.

In [12] authors Suvarna Patil, Tanuja Lonhari, Sarika Pati “Internet of Things: Current Research, Trends and Applications” says that the term IoT was initially proposed to refer uniquely identifiable interoperable connected objects with Radio Frequency Identification (RFID) technology. After that researchers relate IoT with more technologies such as actuators, sensors, GPS devices, and mobile devices. Today IoT can be defined as “a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual 'Things' have identities, physical attributes, and virtual personalities and use intelligent interfaces are seamlessly integrated into the information network”.

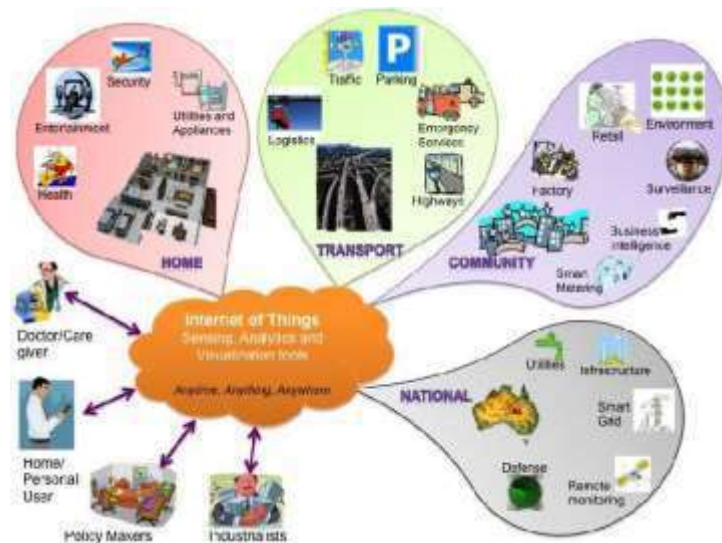


Fig.1: Internet of Things Schematic showing the End users and application areas based on data

Table: 2. A Four Layered Architecture of IoT

S.NO	LAYERS	DESCRIPTION
1	Sensing Layer	This layer is integrated with hardware to sense/control the physical world.
2	Network Layer	This layer provides basic networking support and data transfer over wireless or wired network.
3	Service Layer	This layer creates and manages services. It provides services to

		satisfy user needs.
4	Interface Layer	This layer provides interaction methods to users and other applications.

In [13] authors Sudhir Chitnis, Neha Deshpande, Arvind Shaligram “**An Investigative Study for Smart Home Security: Issues, Challenges and Countermeasures**” says that a home security system to determine the present security status and to find out any extremes of security. It determines the level of protection needed and give suggestions to improve the overall security of home, if required. Traditional techniques of alarm based security have gained much popularity in past decades. Nowadays, embedded system is designed to provide security due to tremendous improvement in microcontroller unit and widespread applications of GSM technology. In [14] most of our homes are still protected by simple lock-and-key mechanisms. Nowadays, most of the families are single type where almost all are working professional. As a result the children at home are left unattended or in the company of servant or a babysitter who are not trustworthy most of the times. Thus, relying on traditional lock-and-key security mechanism is rather risky. Generally, robbery or crimes are committed by low skilled criminals.

In [15] authors Mahnoosh Mehrabani, Srinivas, Benjamin Stern “**Personalized Speech Recognition for Internet of Things**” says that each user was asked to create a set of household devices and select customized names for their appliances. A smartphone application was used by end users as a centralized interface, and speech recognition was performed by connecting to a cloud Application Programming Interface (API). Testers were instructed to issue voice commands including their selected personalized devices. The results presented here are based on a subset of the speech data including 1533 words that was manually transcribed and semantically annotated, and was used as test set. In a smart home scenario each customized devices are connected to home and the user using personalized language model to evaluated the speech recognition.

In [16] authors A. Daramas, S. Pattarakitsophon, K. Eiumtraku1, T. Tantidham N. Tamkittikhun “**HIVE: Home Automation System for Intrusion Detection**” says that the HIVE system deploys three intrusion detection sensors: a passive infrared sensor (PIR), magnetic switch sensor, and load cell sensor. The first sensor, PIR sensor, detects motions in a particular area. The next sensor, magnetic switch, detects the status of doors or windows. There are 2types of magnetic switch: Normally Open (NO) and Normally Close (NC).

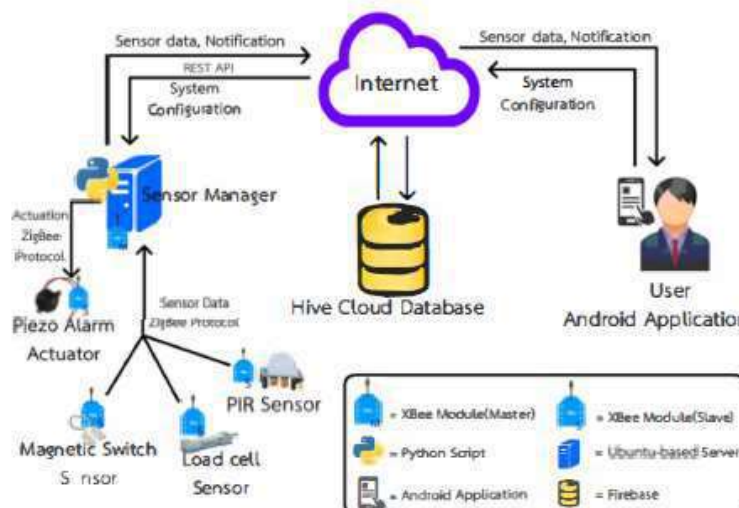


Fig. 2: HIVE System Architecture



In [17] authors Honglei Ren, You Song, Siyu Yang and Fangling Situ “**Secure Smart Home: A Voiceprint and Internet Based Authentication System for Remote Accessing**” says that a smart home architecture, both in hardware and software. It used the NFC and android technologies to enable remote and local accessing. For the local access, users can enter home by making the mobile phone approach to a NFC reader which can recognize the necessary authentication information. A user Personal Identification Number (PIN) is needed to input on the mobile side for safety considerations. The PIN will be encrypted with other information (like MAC address) to produce a fingerprint of the device, which will be authenticated by the server. In the condition of remote access, the security is guaranteed by the PIN and the Virtual Private Network (VPN) tunnel. The VPN can establish a secure connection without the need of specialized software. However the PIN code can still be hacked without lots of efforts. Human put forward a smart home authentication solution based on fingerprint identification, whereas it is still dangerous when it is defrauded with the fingerprint film.

In [18] authors Abhay Kumar, Neha Tiwar “**Energy Efficient Smart Home Automation System**” says that many smart homes are need security and energy saving. We need to implement and optimize the efficiency of the smart home automation system via simulation. [19] We can also implement the smart home technology using VB (visual basic) is being used. Practically we can implement the smart home by many researchers to optimize the better result and to improve the technology for the less consumption of electricity. We observed the deviation in temperature and the speed of fan and light is also varying according to the temperature as they programmed in coding. And the masses which are connected through relays are used to switch on and switch off the loads through sending tones via mobile phone and through serial connection. We can also control the whole system by connection through PCs with server through client PCs. It will work only with the output voltage of +5V. We feed the coding in PIC microchip to run the system according to the feed coding.

In [21] authors Abdallah Kassem and Sami El MurrGeorges Jamous, Elie Saad and Marybelle Geagea “**A Smart Lock System using Wi-Fi Security**” says that smart-Lock-System is a complete recreation of the standard Key-Door lock .Where all the digital keys are kept in a Digital Keychain kept on the owner’s phone. Encrypted and secured Smart-Lock-System can be connected to the Internet through internet cable or Wi-Fi. The Smart-Lock-System consists mainly of three major functions.

- Function 1: Door lock controller
- Function 2: Central Control
- Function 3: Mobile Application

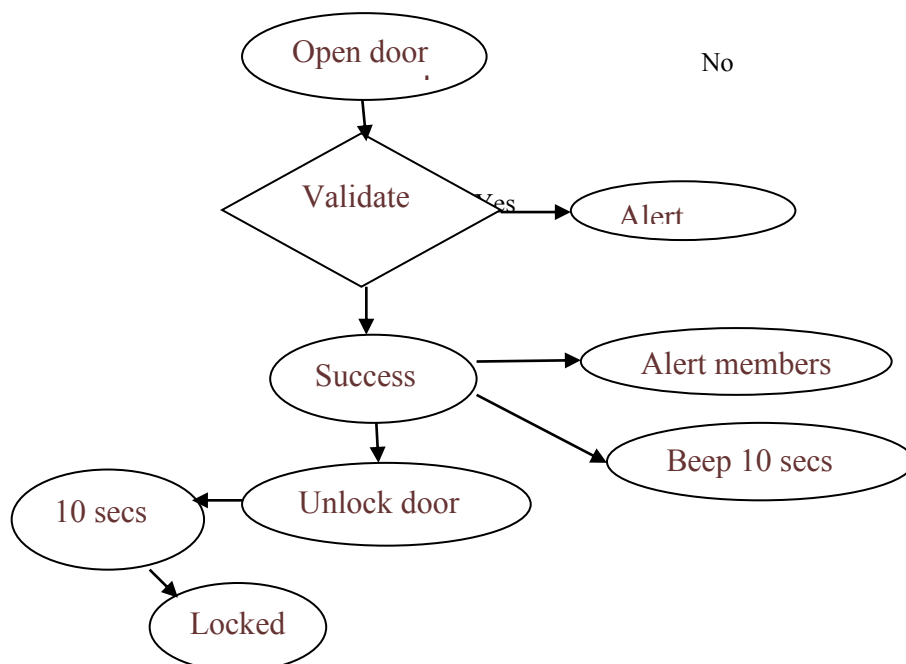


Fig. 3: Flow Chart for Door Open

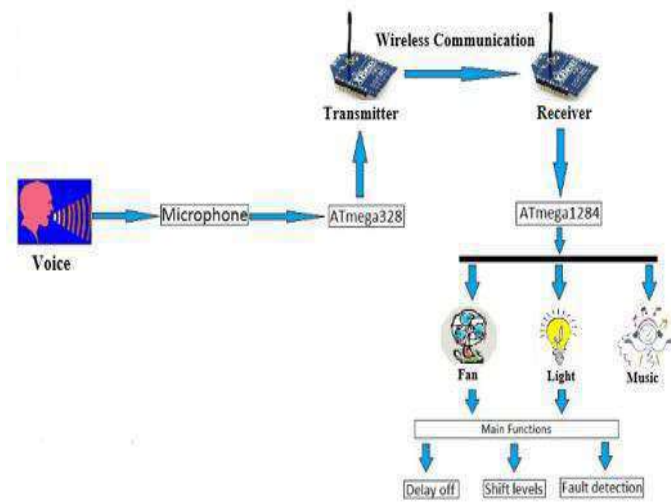


Fig. 2: Overview of logical structure

Existing system having the voice-controlled system is used in smart home automation system. Existing simulation is a product for the future life and the purpose of it is to make people's lives more convenient. The XBee radio module use of a 12V battery for powering the entire device makes it safer one. Similarly, the wireless voice-control system makes it advantageous for disabled people to control the household devices.

### COMPARATIVE ANALYSIS OF EXISTING SYSTEM

This section presents the comparison between the different type of algorithm and sensors used and various purposes for controlling home devices. Different types of sensors are used for sensing the data and monitor the movements from home and various sensors and algorithms are reviewed in the comparative analysis section.

### COMPARATIVE ANALYSIS

S.NO	Paper Title	Algorithm/Risk/Device	Purposes
1	A risk analysis of a smart home automation system	Home Gateway	To provide authentication
2	Security in the Internet of Things: A Review	AES,RSA,DH,SHA	Confidentiality, Digital signatures key , Key agreement, Integrity
3	On the Risk Exposure of Smart Home Automation Systems	Information Security Risk Analysis (ISRA)	Confidentiality, Integrity availability.
4	Internet of Things: Current Research, Trends and Applications	RFID,GPS	To communicate the system and the user.
5	An Investigative Study for Smart Home Security: Issues, Challenges and Countermeasures	CCTV camera	To capture the video.
6	Personalized Speech Recognition for Internet of Things	Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM)	To represent the time-based variance of speech.
7	HIVE: Home Automation System for Intrusion Detection	Zigbee,PIR sensor,	To detect the movement
8	Secure Smart Home: A Voiceprint	Voice ,Virtual private	To prevent

	and Internet Based Authentication System for Remote Accessing	number(VPN)	unauthorized people to enter
9	Energy Efficient Smart Home Automation System	Light ,Fan ,AC	To reduce the energy
10	A Smart Lock System using Wi-Fi Security	AES ,Key	To provide security

## PROPOSED SYSTEM

Overall control of the home is done by smart home automation system using Internet of Things (IOT) In door opening method like Number locking system means user having the pin number to enter the home, the drawback is pin number is misused in other person. The fingerprint system having thumb impression is used to enter the home and the drawback is easy to trace any place. To overcome the number lock and fingerprint system using door locking sensor to capture the face and stored in cloud. Any relations are come to home means to capture the face and send it to the owner. The owner will give the security code and they can enter the home.

## CONCLUSION

Our work mainly concentrates on security system in the existing home automation system. Security factor is most important when it comes from automated home security systems. Such system will definitely provide a security to every person at home. And will also to remember their work, that are not present their home. To give a survey on smart home automation many new technologies are exploring day by day. Smart is the good and beneficial one and to save their electrical energy that is wasted by many people in fixed span of time. With the smart home security system the people are living and will obviously live more comfortable life. All the time home can be saved from automation so that we will have much more time work on the other scenario. In this prose we reviewed on smart security mechanism in the IoT and analyzed security characteristics home automation.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Abdallah Kassem and Sami El Murr Georges Jamous, Elie Saad and Marybelle Geagea "A Smart Lock System using Wi-Fi Security" 2016 3rd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA).
- [2] Abhay Kumar1, Neha Tiwari.[2015 ]Energy Efficient Smart Home Automation System" International Journal of Scientific Engineering and Research (IJSER) www.ijser.in ISSN (Online) 3(1) : 2347-3878
- [3] Andreas Jacobsson ,Martin Boldt and Bengt Carlsson "On the Risk Exposure of Smart Home Automation Systems" 2014 International Conference on Future Internet of Things and Cloud.
- [4] Atzori.L, A. Iera, and G. Morabito, "The internet of things: A survey", Computer networks, vol. 54, no. 15, pp. 2787–2805, 2015.
- [5] Babar S, A Stango, N Prasad, J Sen, R Prasad.[2014] Proposed Embedded Security Framework for Internet of Things (IoT), Int. Conf.on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronics Systems Technology.
- [6] Daramas A, S Pattarakitsophon, K Eiumtraku1, T Tantidham. N Tamkittikhun "HIVE: Home Automation System for Intrusion Detection" 2016 Fifth ICT International Student Project Conference (ICT-ISPC).
- [7] Gan G, Z. Lu, and J Jiang, "Internet of Things Security Analysis", IEEE Conf. on Internet Technology and Applications.
- [8] Hui Suo, Jiafu Wana, Cai Feng Zoua, Jianqi Liua "Security in the Internet of Things: A Review" 2012 International Conference on Computer Science and Electronics Engineering.
- [9] Honglei Ren You Song, Siyu Yang and Fangling Situ "Secure Smart Home: A Voiceprint and Internet Based Authentication System for Remote Accessing".

- [10] Jacobsson A , M. Boldtand B. Carlsson “A Risk Analysis on a Smart Home Automation System”, Future Generation Computer Systems, Elsevier, 2015. DOI:10.1016/j.future.2015.09.003.
- [11] Mantoro.T, MA Ayu, SM. bintiMahmod.[2014]Securing the authentication and message integrity for Smart Home using smart phone”, in 2015International Conference on Multimedia Computing and Systems(ICMCS). IEEE, 2014, pp. 985–989.
- [12] Mahnoosh Mehrabani, Srinivas, Benjamin Stern “Personalized Speech Recognition for Internet of Things” International Conference on Future Internet of Things and Cloud.
- [13] Polkand T., S. Turner. “Security challenges for the internet of things”, <http://www.iab.org/wp-content/IAB-uploads/2011/03/Turner.pdf>
- [14] Shi JH, JF Wan, HH Yan, H Suo. A survey of cyber-physical systems”, in Proc. of the Int. Conf. on Wireless Communications and Signal Processing, Nanjing, China, November, 2015.
- [15] Sudhir Chitnis, NehaDeshpande, ArvindShaligram.[2016] “An Investigative Study for Smart Home Security: Issues, Challenges and Countermeasures” Wireless Sensor Network, 8:61-68.
- [16] SuvarnaPatil, TanujaLonhari, SarikaPati.[2015] Internet of Things: Current Research, Trends and Applications” International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) 3(12)
- [17] System of Monitoring and Environmental Surveillance, <http://www.dimap.es/enviromental-agriculture-services.html> (2011). Oxford University Press, ISBN 0-8218-0531-2, 2014.
- [18] Yang G, J Xu, W Chen, ZH Qi, HY Wang.[2013] Security characteristic and technology in the internet of things, *Journal of Nanjing University of Posts and Telecommunications (Natural Science)*, 30(4).

# TOTAL QUALITY MANAGEMENT ASPECTS OF IMPLEMENTATION AND PERFORMANCE INVESTIGATION WITH A FOCUS ON HIGHER EDUCATION BY USING QFD & STATISCAL ANALYSIS IN MECHANICAL ENGINEERING

Abhishek Soni<sup>1</sup>, Nitin Shrivastava<sup>2</sup>, Sameer Vaidya<sup>3</sup>, Sanjay Soni<sup>4</sup>

<sup>1</sup>PhD Research Scholar(Mechanical) AISECT University Bhopal. (M.P) INDIA

<sup>2</sup> Assistant Professor Mechanical Engineering Department UIT Bhopal. (M.P) INDIA

<sup>3</sup>Professor Mechanical Engineering Department H.C.E.T, Jabalpur (M.P) INDIA

<sup>4</sup>Asstiant Professor Industrial Production Engineering Department JEC JBP. (M.P)INDIA

## ABSTRACT

**Aim:** Education is of numerous types and patterns. There is for example, the arts teaching, the scientific education, the religious education, the physical education. In India, as in other countries, much stress has been laid on the endorsement of technical education since the attainment of independence. India's economic ills are sought to be overcome through a course of Industrialization for which, in turn, technical education is very essential. In other words, technical education is a vital prelude to India's property. The scope of technical education is very comprehensive. It slots in within itself all subjects of study in engineering and technology. Civil engineering, Mechanical engineering, Electrical engineering, Mining engineering, Aeronautical engineering, Metallurgical engineering, Industrial engineering, Chemical engineering, Agricultural engineering, Production engineering, and a host of other fields of engineering form part of technical education. "Quality in technical education is a complex concept that has eluded clear definition". There are a variety of stakeholders in higher education including students, employers, teaching and non-teaching staff government and its funding agencies, a creditors, valuator's, auditors, and assessors (including professional bodies). Each of these stakeholders has a different view on quality, influenced by his or her own interest in higher education. For example, to the committed academic, the quality of higher education is its ability to produce a steady flow of people with high intelligence and commitment to learning that will continue the process of broadcast and advancement of facts. To the government, a high quality system is one that produces skilled scientists, engineers, and architects, doctors and so on in numbers judge to be required by the public. The present work enlightens same path, so as to fulfil the demands of market and to improve quality of education in the present work some quality tools such as linear programming, quality function deployment, with chi square testing and mat lab, have been used. Basic main tool used is QFD which helps in converting demand of customer to action. It helps in understanding understood needs of customer which are desperately needed to be fulfilled. In this upgrading work main focus was on improvement of labs and teaching staff, for maintenance of labs & improvement in teaching, use of quality circle is stressed with concept of TPM and Kaizen approach. Most interesting thing of using these tools was that they helped in achievement of desired target without much added resource, only refinement of procedure; moreover maintenance helps in gaining knowledge with saving extra spending. This also helps in up shade of quality of products which satisfies external client.

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Crowdsourcing, Cross space transferring, OSNs, FlierMeet, OCR, NLP, STA algorithm, intelligent tagging

\*Corresponding author: Email: [abhishek.soni.jbpindia@gmail.com](mailto:abhishek.soni.jbpindia@gmail.com)

## INTRODUCTION

India's higher education system is the world's third major in terms of students, next to China and the United States. Contrasting China, however, India has the advantage of English being the primary language of higher education and research. India educates approximately 11 per cent of its youth in higher education as compared to 20 per cent in China. The main governing body at the tertiary level is the University Grants Commission (India), which enforces its principles, advises the government, and helps coordinate between the centre and the state. Universities and its constituent colleges are the main institutes of higher education in India. At present in 2011, there are 227 government-recognized Universities in India. Out of them 20 are central universities, 109 are deemed universities and 11 are Open Universities and rest are state universities. Most of these universities in India have affiliating colleges where undergraduate courses are being trained. According to the Department of higher Education government of India, 16,885 colleges, including 1800 exclusive women's colleges functioning under these universities and institutions and there are 4.57 lakh teachers' and 99.54 lakh students in various higher education institutes in India. Apart from these higher education institutes there are a number of private institutes in India that offer various specialized courses in India. Distance learning is also a feature of the Indian higher education system.

## REVIEW OF LITERATURE

All of the research review supports the theory that student act depends on different socio-economic, psychological, environmental factors. The findings of research studies focused that student presentation is affected by different factors such as learning abilities because new paradigm about learning assumes that all students can and should learn at higher levels but it should not be measured as constraint because there are other factors like race and gender that can affect student's performance. Some of the researchers even tried to explain the link between students achievement, financial conditions and the risk of becoming a drop-out that proved to be positive (Goldman N., Haney W., and Koffler S., 1988, Pallas A., Natriello G., McDill E., 1989, Levin H., 1986) B.A. Chandrashekhar and A. Mishaeloudis (2001), explain the effects of age, qualification distance from learning place etc. on student presentation. The performance of students on the module is not affected by such factors as age, sex and place of residence but is associated with qualification in quantitative subjects. It is also found that those who live near He University execute better than other students. Yvonne Beaumont Walters, kola sahib, (1998) further elaborated that student performance is very much dependent on SEB (socio economic back ground) as per their statement, "High school students' height of performance is with statistically significant differences, linked to their gender, grade level, school location, school type, student type and socio-economic background (SEB)." Kirby, Winston et al. (2002) alert on student's impatience (his time-discount behaviour) that influences his own academic recital. Goethe found out that weak students do better when grouped with other weak students. (As implied by Zajonc's examination of older siblings (1976) it shows that students' performance improves if they are with the students of their own kind. There are often dissimilar results by gender, as in Hoxby's K-12 results (2000); Sacerdote (2001) finds that grades are higher when students have unusually academically strong roommates. The outcome of Zimmerman (1999, 2001) were somewhat contradictory to Goethe results but again it proved that students performance depends on number of different factors, it says that weak peers might reduce the grades of middling or strong students.

## METHOD

Statistical techniques including regression analysis were used as a methodology. Data collected was primary through a well-defined questionnaire. A sample of private college students was taken where these variables were recognized and response was clear and understandable. Public sector educational institutions were not the focus of the study. A sample of 30 students was taken from a group of colleges. Students were grouped in a classroom they were briefed clearly about the questionnaire and it took on average half an hour to fill this questionnaire. Selection of students was at random. Out of these students only those were selected at random who were voluntarily willing to fill the questionnaires. The data was collected using a questionnaire administrated by the Research team in the 3rd month of 3rd year. The questionnaire dealt mainly with student profile based on his attitude towards Study, Strictness, Attendance, Age, Previous academic achievements, Daily life, etc. All 6 questionnaires were filled with the response rate of 100%.

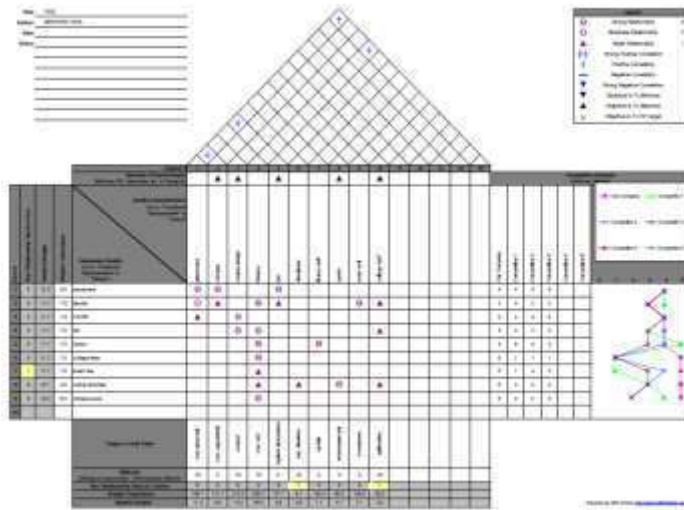
The sample age composition was from 18 years to 22 years of age at maximum because Rajiv Gandhi Technical University of does not allow students over 22 years of age to be admitted in graduate classes.



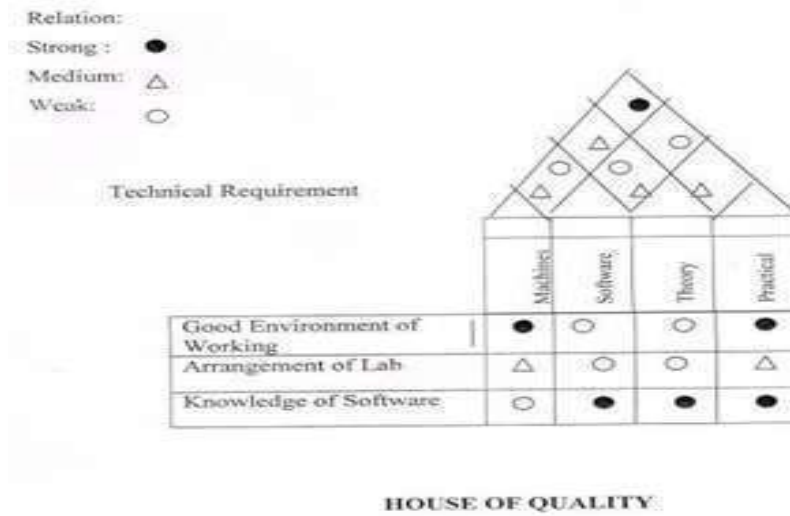
Fig 1\_Basic Ideology

## QUALITY FUNCTION DEPLOYMENT

Quality Function Deployment (QFD) is a style for building the "Voice of the Customer" into product and service design. It is a team tool which captures customer necessities and translates those needs into characteristics about a product or service.



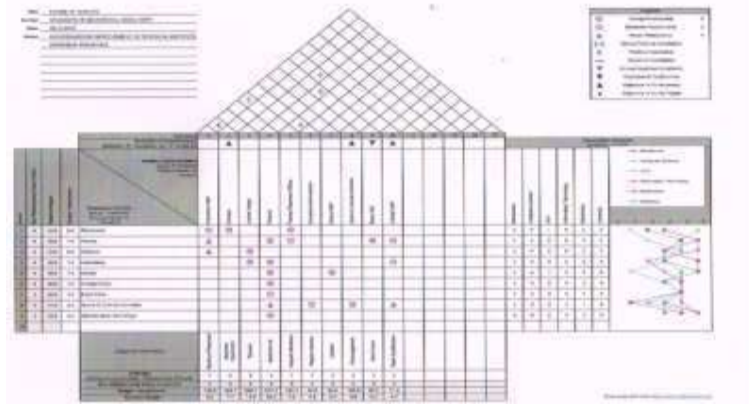
STAGE 1



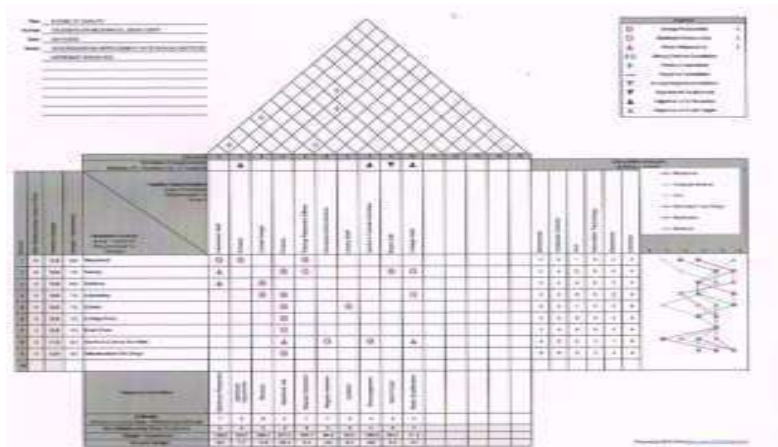
OUTCOME OF STAGE 1



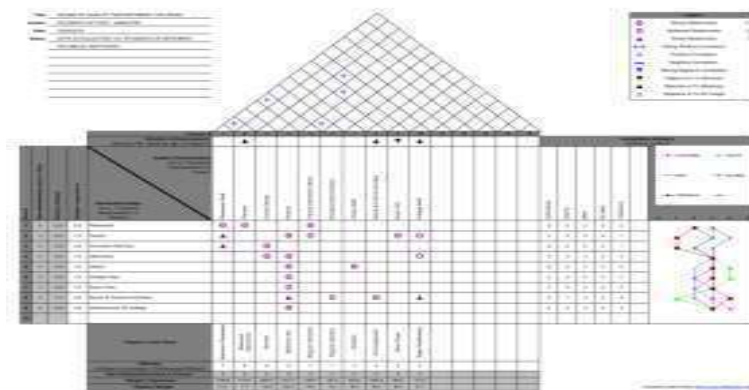
CAMPARISON OF STAGE 1



STAGE 2



STAGE 3



STAGE 4

OUT COME OF STAGE4



## ANALYSIS OF DATA(MECHANICAL IV TH SEM)

N	RESULT	SGPA	CGPA	Compartment (SUBJECT)	Family income (Lakhs)	Attendance	Reference book/ Question Bank	Hostel/Day scholar	Study Hours per day	Type of study
1	FAIL	6.06	6.76	ME403	4.20	82%	Reference book	Day Scholar	1.75	Group
2	FAIL	6.63	6.91	ME404	3.2	90%	Question Bank	Day Scholar	1.75	Group
3	PASS	6.31	6.66		3.4	70%	Reference book	Hostler	2	Individual
4	FAIL	6.13	6.26	ME403	3.3	71%	Question Bank	Hostler	1.5	Group
5	PASS	6.63	6.52		5	75%	Reference book	Day Scholar	1.5	Individual
6	FAIL	3.88	3.88	BE401, ME403, ME405	3	65%	Question Bank	Hostler	near test	Individual
7	FAIL	4.06	5.71	BE401, ME403, ME405	3.5	64%	Question Bank	Day Scholar	near test	Group
8	PASS	6.88	6.89		5	78%	Question Bank	Day Scholar	2	Group
9	PASS	7.75	7.15		3.6	79%	Question Bank	Hostler	2	Individual
10	PASS	7.5	7.70		2.5	82%	Reference book	Day Scholar	2.5	Individual
11	PASS	8.19	7.74		3.6	84%	Reference book	Day Scholar	2.5	Individual
12	FAIL	3.88	5.16	ME403, ME404, ME405	2.8	61%	Question Bank	Day Scholar	near test	Group
13	FAIL	6.19	6.29	ME403	2.3	68%	Question Bank	Hostler	1.5	Group
14	PASS	7.00	7.39		3.7	84%	Question Bank	Hostler	2	Individual
15	PASS	7.31	7.35		3.2	61%	Reference book	Day Scholar	2	Individual
16	PASS	7.31	7.16		3.9	80%	Question Bank	Hostler	2	Group
17	PASS	7.19	6.76		2.3	81%	Question Bank	Day Scholar	2	Group
18	FAIL	4.94	4.82	ME403, ME404	2.1	66%	Question Bank	Day Scholar	near test	Group
19	FAIL	5.19	6.31	ME403, ME405	3.8	67%	Question Bank	Day Scholar	1.5	Individual
20	PASS	7.31	7.03		2.5	84%	Reference book	Day Scholar	2	Individual
21	PASS	7.69	7.43		4.1	88%	Question Bank	Day Scholar	2.5	Individual
22	PASS	5.94	6.71		2.8	60%	Question Bank	Day Scholar	1.5	Group
23	PASS	6.13	5.86		2.3	58%	Question Bank	Day Scholar	1.5	Group
24	FAIL	4.88	5.89	ME403, ME405	2.5	67%	Question Bank	Day Scholar	near test	Individual
25	PASS	7.56	7.66		4	78%	Reference Book	Day Scholar	2.5	Individual
26	FAIL	3.88	4.97	BE401, ME403, ME405	2.4	58%	Question Bank	Hostler	near test	Individual
27	FAIL	5.06	6.01	ME403, ME405	2.7	55%	Question Bank	Hostler	near test	Individual
28	PASS	7.50	7.36		3.4	78%	Question Bank	Hostler	2	Group
29	FAIL	4.13	4.58	BE401,	2.6	55%	Question	Day	near	Individual

				ME403, ME405			Bank	Scholar	test	ual
30	FAIL	6.63	6.08	ME403	2.2	78%	Question Bank	Hostler	1.5	Individ ual

COMPARISON OF EXPECTED RESULTS AND RESULTS OF THE STUDY-

Variable	Expected Relationship	Explanation	Results Of study
Attendance in class	Positive	A regular student is more serious in studies	Positive
Family income	Positive	It is assumed affluence gives more facilities to learn	Negative
Study hours per day after college	Positive	It is assumed that more study hours results in good grade/division/performance	Negative
Book referred	Positive	More books referred results in better grasp of the concept	Positive
Type of study	Positive	Group study results in healthier studying environment, hence better result	Negative
Hosteller/Day Scholar	Positive	Hostellers are found to be more dedication in their studies	Positive

Chi square solution for validation of result

parameter	A	B	C	D	E	Total
Family income	4.2	3.20	3.40	3.30	5	19.10
Study hours	1.75	1.75	2	1.50	1.50	8.50
sgpa	6.06	6.63	6.31	6.13	6.63	6.63
TOTAL	12.01	11.58	11.71	10.93	13.13	59.36

ALCULATION FOR CHI SQUARE ( $\chi^2$ )

$(\chi^2) = (o-e)^2/e$  d.o.f=(m-1)(n-1)  
 Expected frequency is calculated as  
 $E_{r,c} = (n_r * n_c) / n$

o	e	o-e	(o-e) <sup>2</sup>	(o-e) <sup>2</sup> /e
4.2	3.86	0.34	0.1156	0.03
3.2	3.73	-0.53	0.2809	0.075
3.4	3.77	-0.37	0.1369	0.036
3.3	3.52	-0.22	0.0484	0.014
5	4.23	0.77	0.05929	0.014
1.75	1.72	0.03	0.0090	0.005
1.75	1.66	0.09	0.0081	0.004
2	1.68	0.32	0.1024	0.061
1.5	1.57	-0.07	0.0049	0.003
1.5	1.89	-0.39	0.1521	0.08
6.06	6.43	-0.37	0.1369	0.021
6.63	6.20	0.43	0.1849	0.03
6.31	6.27	0.04	0.0016	0.002
6.13	5.85	0.28	0.0784	0.013

6.63	7.03	-0.40	0.1600	0.022
------	------	-------	--------	-------

$(\chi^2) = (o-e)^2/e = 0.536, < \text{TABULATEDVALUE}, (\chi^2)_{0.05} = 15.51$   
 Therefore, hypothesis is accepted

Chi-Square Distribution

Degrees of Freedom(df)	Probability (p)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	N										

#### Validation through matlab(taking 3\*3 matrix)

```

[x,y,z]=solve('4.2*x+3.2*y+3.4*z=10.8','1.75*x+1.75*y+2*z=5.50','6.06*x+6.63*y+6.31*z=19')
[p,q,r]=solve('4.2*x+1.75*y+6.06*z=12.01','3.2*x+1.75*y+6.63*z=11.58','3.4*x+2*y+6.31*z=11.71')
)
%subplot(2,2)
%subplot(2,1,1);
%hold on
x
y
z
p
q
r
%i=[x y z];
%j=[p q r];
%plot(i,j);

%subplot(2,1,2);
%plot(y,q);
%subplot(2,2,1);
%plot(z,r);

```

Value obtained  $x=y=z=p=q=r=1$

## DISCUSSIONS

The objective of this study was to quantify the relationship between the different factors that are considered responsible of affecting the student performance along with providing base for further research regarding student performance. Selecting these combination of variables do have some objectively like, It was expected that relationship between dependent variable and student attitude towards attendance is positive because regularity shows the effort and seriousness of student about his or her studies. It is believed that the relationship between dependent variable and student family income is positive because money can buy you all comfort that you need to concentrate on your studies but the result could not prove this relation because student belonging to more prosperous/affluent family do not give proper weight to studies although this value is very small but still it reflects

the insignificance of affluence that is affluence cannot make a student serious about his studies or if a student want to study then affluence is not a prerequisite but still it requires more research to explain the phenomenon It is still believed strongly that relationship between dependent variable and student attitude towards time allocation for per day after college are positively related but the result could not prove this relation because more study hours are not significant as far as student performance is concerned. It may depend on intelligence level, intellect, memory or method or learning of the student although this value is very small yet it reflects of personal characteristics of student. Further research is required to explore this relation. It is believed that book reference also has great effect on performance of students that if students are referring to books it helps in increase of concepts and deep knowledge about the topic, and if one is studying form question banks then he cannot grasp more knowledge; yes but he can touch every topic with little knowledge. Selecting a type of study .i.e. between Group and Individual affects the student performance. It is believed that Group studies have more impact over individual studies. If a student is studying in group he is scoring better marks that him who is studying individually. One more important attribute is Day scholar or Hosteller. It is found that student that are Hosteller perform better than Day scholar.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None.

### REFERENCES

- [1] Akao Y [1983]‘QUALITY FUNCTION DEPLOYMENT’, QUALITY PROGRESS.BENNETT DC (2001) ‘ASSESSING QUALITY IN HIGHER EDUCATION’, LIBERAL EDUCATION. 87(2): 40-4
- [2] BA CHANDRASHEKHAR, A. MISHAELOUDIS (2001)
- [3] GOLDMAN N, HANEY W, KOFFLER S. [1988]PALLAS A, NATRIELLO G, MCDILL E. [1989], LEVIN H., 1986
- [4] Kirby, Winston Et Al. [2002] Hoxby’s K-12 Results (2000); Sacerdote (2001)zimmerman (1999, 2001
- [5] THAKKAR J, DESHMUKH SG [2006] ‘TOTAL QUALITY MANAGEMENT (TQM) IN SELF-FINANCED TECHNICAL INSTITUTIONS A QUALITY FUNCTION DEPLOYMENT (QFD) AND FORCE FIELD ANALYSIS APPROACH’ QUALITY ASSURANCE IN EDUCATION. 14(1): 54-74.MANAGEMENT. 6(5/6):457-4 67
- [6] SONI ABHISHEK (2013)‘IMPROVEMENT OF QUALITY OF TECHNICAL INSTITUTE THROUGH QFD’. INTERNATIONAL J. OF MULTIDISPL. RESEARCH & ADVCS. IN ENGG.(IJMRAE), VOL. 5, NO. IV (OCTOBER 2013), PP. 133-149
- [7] SONI ABHISHEK. [2013] ‘ TOTAL QUALITY MANAGEMENT IN EDUCATIONAL PROCESS FOCUSED ON QUALITY IMPROVEMENT OF INSTITUTE WITH CUSTOMER SATISFACTION & TEACHING IMPROVEMENT’ INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCHTECHNOLOGY 2(11) PP3195-98
- [8] SONI ABHISHEK. [2014] ‘ APPLICATION OF TQM IN HIGHER EDUCATION FOCUSED ON IMPROVEMENT OF TECHNICAL INSTITUTE VIA QFD’ INTERNATIONAL J. OF ENGG. RESEARCH & INDU. APPLS. (IJERIA).7( 1) : 123-136
- [9] SONI ABHISHEK. [2014]‘ QUALITY IMPROVEMENT OF TECHNICAL INSTITUTE BY USING QUALITY CIRCLES’, NATIONAL JOURNAL OF EXCEL EDUCATION 1(2) PP04-06..

# OPTIMUM CONTROLLING OF THE SAUVE-QUI-PEUT ENERGY USING THE CAFA TECHNIQUE

S. Rajeswari\*, Hemalatha

Department of Computer Science, Karpagam University, Coimbatore, Tamilnadu, INDIA

## ABSTRACT

*Aim: Wireless Sensor Networks (WSNs) consist of spatially spread out unique sensors to monitor several important attend able and untenable environments. The use of data in the network set of sensors is precious in the dangerous mission. Materials and methods: In the updated century, healthcare, architecture, space, mechanical, electrical and electronic industries, and so many other environments, which are small scale or large scale industries, enjoy unforeseen improvements, because of WSN. It is an excellent platform which provides ample opportunities to apply our innovative concept. Results: Real time variables are collected in a predetermined, projected and estimated way, and the observations are channelized towards the base station for further computational use. The paper analyses various real time problems which start from sensing the last process and analyses all the important protocols for the energy saving in the WSN with new CAFA (Constant Analysing Factor Agent) concept. Conclusion: The main contribution is to preserve the existing real time parameters, at the same time increasing the life span of the battery with more accuracy, improving upon the existing algorithms.*

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Crowdsourcing, WSN, Data Collection, Network Constraints, Region Based Data Collection, Dynamic Region Selection.

\*Corresponding author: Email: [rajeshwariphd123@gmail.com](mailto:rajeshwariphd123@gmail.com)

## INTRODUCTION

Sensors can be defined in general as compact battery operated equipment's. In communication process it consumes more energy than all other computation parts. Hence, it plays an important role in conserving energy, and so researchers focus on solving such crucial issues. There should be an energy awareness at all the layers of networking protocol. The issues are common for all kinds of sensor applications. Therefore, the research on these areas mainly focused on the software-level power awareness such as dynamic region selection based CAFA technique with the low duty cycle issues, system network and analysing related issues. At the network layer, the main aim is to find ways for the energy-efficient route set up with the minimum hop, and reliable relaying of data from the sensor nodes to data and from some agent to the sink [1].

Lifetime of the network is maximized when the present algorithm is applied. Routing in the sensor networks is a very vital exercise due to the several real time characteristic values that distinguish them from the other[2]. Generally, if the automated standalone WSN is used for the large scale with limited speed, in the small scale applications like healthcare to monitor patients with complex problems, it is very important to monitor critical cases at a high speed data rate with accuracy. The data received may not be with clear output signal of the concerned category, and so, it is necessary to monitor each and every attribute, big or small, in overall Wireless Sensor Network.

## OBJECTIVE

The present research shows that low utilization of energy has resulted in high output with more accuracy, and is useful under limited resources. So, analyzing the network constraints in Wireless Sensor Network is crucial. Before selecting the set of data sensor nodes, the network constraints are to be considered. Thereby the data collection will be carried out efficiently. Before selecting a sensor node for data collection, the method has to be identified, and the number of supporting nodes present in the network between the sink and sensor node have to be considered.

## SCOPE

Similarly, before using the intermediate sensor nodes for the data collection, the energy constraints of the intermediate nodes have to be verified.

1. Subsequently, the data availability in the sensor node and the amount of data retrieved from the sensor have to be identified before selecting a sensor for the data collection.
2. All the results are routed to the Wireless Sensor Network for a coordinated exercise.
3. Approximate self-adaptive data collection in the Wireless Sensor Networks propose an approximate self-adaptive data collection technique, to approximately collect the data in a distributed Wireless Sensor Network.
4. The data collection investigates the spatial correlations between the sensors to provide an energy-efficient and balanced route to the sink, even though each sensor does not possess any of the global knowledge in the network.
5. Based on the synthetic experiences, the study demonstrates that data collection provides a significant communication savings and equal energy consumption in the sensor nodes.

## EXISTING SYSTEM

A Feedback-Based Secure Path Approach for the Wireless Sensor Networks Data Collection in the novel tracing-feedback mechanism, makes full use of the routing functionality of the WSN, to improve the quality of the data collection [8]. The algorithms of the approach are easy to implement and perform in the WSN [3]. The approach is evaluated with a simulation experiment and the simulation results are analysed in detail. It is illustrated that the approach is efficient to support secure data collection in Wireless Sensor Network.

Optimizing Data Collection for Object Tracking in the Wireless Sensor Networks, proposed optimizing an algorithm of object tracking in Wireless Sensor Network (WSN) [4]. The task under consideration is to control the movements of a mobile sink, which has to reach a target in the shortest possible time. Utilization of the WSN resources is optimized by transferring only the selected data readings (target locations) to the mobile sink [5].

## PROBLEM DEFINITION

1. Formal concept analysis is a data analysis tool, especially investigation and treatment.
2. It provides clues to discover any important information hidden in the data behind.
3. The Design of the Data Collection Methods in the Wireless Sensor Networks Based on the Formal Concept Analysis proposes data collection methods.
4. Experiments show that the proposed FCA-based data collection algorithm in the WSN is more effective than the traditional algorithm.
5. Data Collection with the Multiple Sinks in the Wireless Sensor Networks consider the Multiple-Sink Data Collection Problem in the Wireless Sensor Networks, where a large amount of data from sensor nodes need to be transmitted to one of the multiple sinks.
6. It designs an approximation algorithm to minimize the latency of data collection schedule and it gives a constant-factor performance guarantee.
7. It presents a heuristic algorithm based on the breadth first search for the problem

## PROPOSED SYSTEM

Depending on the application, different structured design goals/constraints are considered for the sensor networks. The performance of a routing revised protocol is a novel idea which differs from the existing standards.

### Current region selection

Region selection entirely depends upon the application. It is a fixed environment like forest fire prevention, disaster rescue and so on. If it is a static period, updating the method is important for a variety of purposes. Basically it adopts to all kinds of regions, static or dynamic [6,7]. The paper considers the dynamic region. The entire geoFigureic area is divided into smaller sections according to the network topology. The technique is applied section-wise. From the architecture, several layers will be created. The research arranges them according to the prescribed number of sensors, placing CAFA Agent node at regular intervals for analysing various real time values. These are sent to the CAFA main node via shortest path (minimum hop) for further computation.

### Deployment of all the nodes

This application is sensitive, and any form of carelessness is sure to affect the performance of the entire routing protocol. The deployment is either projected or self-arranged. In the projected situations, the sensors are manually placed and the data is routed through the projected paths after a thorough investigation. In the self-organizing systems, the sensor nodes are scattered randomly for creating an infrastructure in a dynamic region splitting technique and for application of the researcher's methods. In the model, placing the CAFA node, Agent node, data node and sink node is a vital process. If the distribution of nodes, layering and placing important agent nodes at the prescribed geographical region is carried out accurately, the energy saving rate is automatically raised.

### Constant Analysing Factor Agent Node (CAFA)

There are a large number of selected queries which concentrate on the particular subset and use the same types of query and analyse the route, but the paper concentrates on a different type of query. Depending on the situation, some queries are used from the algorithms grouped for the complete node participation of a network. So, a comprehensive history and current information of all the nodes are stored in an agent node at regular intervals in specific layers. This node supplies the stored information as and when required. It is an important issue because whenever network reconstruction takes place, it has to retrieve all the information about the particular node from the topmost server node. This may lead to delay causing instability in the energy utilizing process. So the paper suggests the CAFA node placement in the prescribed distance to collect information from all the leaf nodes. It practices multilevel experiment at the constant time interval to get information regarding the attributes, using the series of query processing and using the existing data report collected. At last a projected routing protocol is prepared as soon as the sensed signal reaches the particular node or each channel reaches the prescribed density of data, so that, it signals to the CAFA node. The signal will be transmitted instantly using the projected minimum hop routing protocol with the low duty cycle[9,10].

### Algorithm

CAFA -Region Selection Algorithm:

Input: Null

Output: Node List Ns, Route Table Rb

Start

Initialize Node Count Nc.

Generate ACM Message. [Availability of waking node present allotted function for multi functional node]

Broadcast ACM Message.

Initialize Broadcast Timer Ti.

while Bt is running

Receive ACMREP message.

if ACMREP.DNode == True then

Add NodeIf to Node List Nc.

Nc =

Extract the route to reach the node.

Rb =

End

End

Stop.

### Energy harvest based data collection

During the creation of the model, the process of setting up the experimental step is greatly influenced by the energy considerations. Since the transmission power of a wireless radio is proportional to the distance of the sensed target which is very high order in the presence of troublesome season, multi-hop routing will consume less energy than direct transmission. However, minimum/multi-hop routing introduces significant overhead for the topology management and medium access control. Direct transmission will perform effectively if it is a small area and all the nodes are near the sink. Most of the time, sensors are scattered randomly over a section of the large area and sensing becomes unavoidable.

Data aggregation is carried out with the help of the query processing, according to the density of the data from the different agent node by using the functions such as filtering (eliminating duplicates), low, high and medium. Some of these functions are performed regularly in each sensor node, by allowing the sensor nodes to conduct the decrease in network data volume due to duplication. Considerable energy savings are obtained through data

querying process. This technique is used to achieve energy utilization and to remove traffic in the number of routing protocols. In the network model, all the analytical functions are assigned to the more powerful CAFA node and the collection functions to specialized nodes. Data collection is also feasible through signal processing techniques. In that case, it is referred to as data fusion where a node is capable of producing a more accurate signal by reducing the noise and using special techniques such as beam forming to combine the proper and clear signals.

## RESULTS AND DISCUSSION

The proposed CAFA model network approach for data collection is implemented and evaluated for its efficiency, using the different scenarios. The method is evaluated for its accuracy and energy saving in the data collection. The method has produced efficient results in data collection as well as in the other factors like time complexity, latency and energy efficiency with lifetime maximization.

Table: 1. Details of Simulation

Parameter	Value
Simulation Platform	Ns2
Simulation Area	1000×1000 meters
Number of nodes	240
Transmission Range	120 meters

[Table-1] demonstrates the details of the imitation being used to estimate the presentation of the proposed approach.

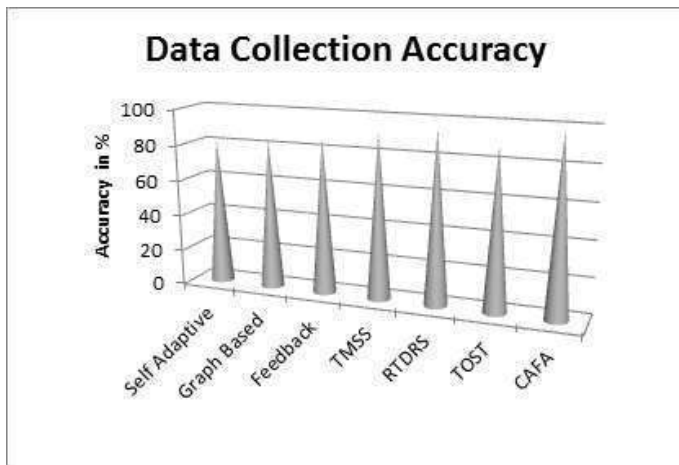


Fig: 1. Evaluation of Rule Generation Accuracy

[Figure- 1] demonstrates the Accuracy of the Data Collection, and it clearly shows that the planned approach achieves data collection with advanced accuracy than the previous approaches.

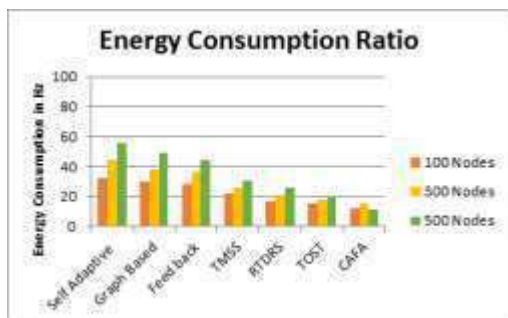




Fig: 2. Comparison of Energy Depletion Occurred

The [Figure- 2] shows that the Energy Depletion Occurred in the data collection is according to the number of nodes available in the network.

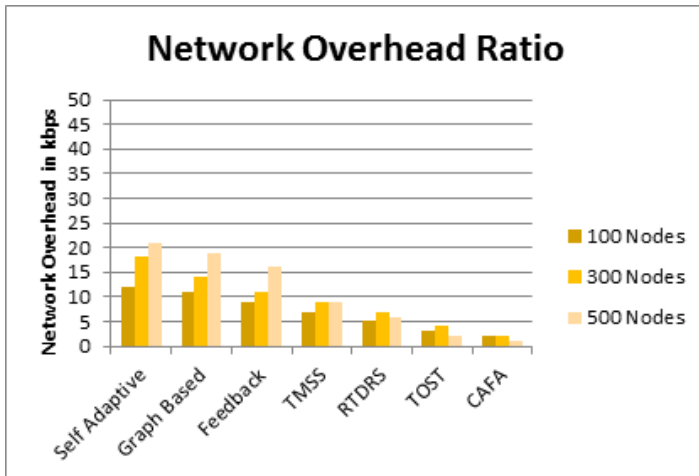


Fig: 3. Comparison of Network Overhead

[Figure- 3] shows the Comparison of the Network Overhead introduced by the different methods in collecting data from the data sensor nodes. The Figure demonstrates that the proposed method has led to less Network Overhead than the other methods.

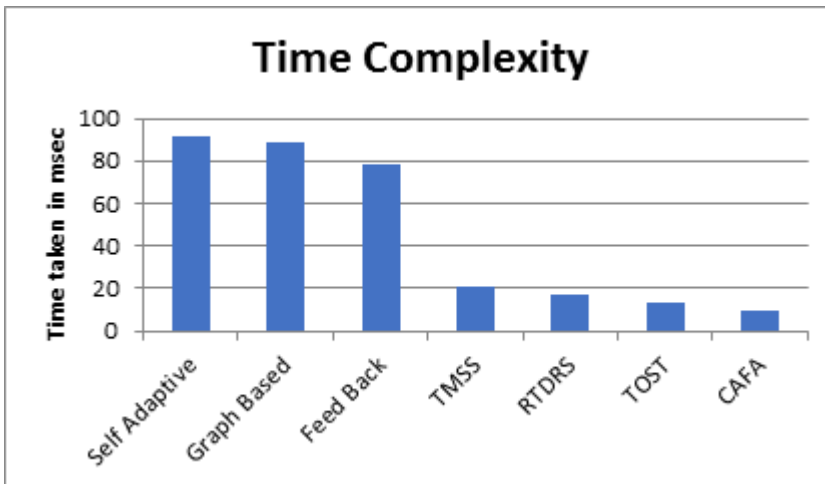


Fig: 4. Time Complexity of Different Methods

[Figure- 4] demonstrates the Time Complexity formed by various procedures in the data gathering where the number of data devices is more than 20, and it shows obviously that the proposed approach forms less time complexity than the supplementary methods.

### CONCLUSION

Constant Factor Analysing Agent node technique is used for data collection in the Wireless Sensor Networks. The method computes the data available under experiential evaluation support measure and for each region. Based on

the above values, the method arrives at two different decisions, by considering the data volumes under a good sensing environment factor and the number of times the region is selected due to the availability of the required density of the data collection support. The proposed method performs an efficient data collection and improves the performance of data collection in the Wireless Sensor Network and reduces energy consumption ratio.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

### REFERENCES

- [1] Andreas R, Christian R. [2015]RoCoCo: Receiver-Initiated Opportunistic Data Collection and Command Multicasting for WSNs, Springer, Wireless Sensor Networks Lecture Notes in Computer Science. 8965:218-233.
- [2] Amir Hossein M, Mohammad Hossein Y, Vahid F. [2015] Total Data Collection Algorithm Based on Estimation Model for Wireless Sensor Network, Wireless Personal Communications. 81(2):745-778.
- [3] Chi-Fu Huang, Wei-Chen Lin. [2015] Many-to-Many Data Collection for Mobile Users in Wireless Sensor Networks, Springer, Advanced Multimedia and Ubiquitous Engineering Lecture Notes in Electrical Engineering. 352:143-148.
- [4] Nidhi G, Sanjeev S, Renu V. [2015] Data collection model for energy-efficient Wireless Sensor Networks, Springer, annals of telecommunications - annales des télécommunications.
- [5] Sivrikaya F, Yener B. [2004] Time synchronization in sensor networks: a survey, Network, IEEE. 18(4):45-50.
- [6] Akyildiz I F, Vuran M C. [2010] Wireless Sensor Networks. John Wiley & Sons.
- [7] Maróti M, Kusy B, Simon G, Lédeczi A. [2004] The flooding time synchronization protocol, in SenSys '04: ACM. 39-49.
- [8] Lenzen C, Sommer P, Wattenhofer R. [2009] Optimal Clock Synchronization in Networks.
- [9] Lu J, Whitehouse K. [2009] Flash flooding: Exploiting the capture effect for rapid flooding in Wireless Sensor Networks, IEEE. 2491-2499.
- [10] Schmid T, Charbiwala Z, Anagnostopoulou Z, Srivastava MB, Dutta P. [2010] A case against routing-integrated time synchronization, ACM. 267-280.

# DEVELOP PEER-TO-PEER NETWORKS PERFORMANCE USING TOP-K QUERY PROCESS OVER P2P LIVE STREAMING

V. Priyanka\*, M. Sai Suchithra, S. Sharmila, M. Narayanan

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha University, Thandalam, Chennai, Tamilnadu, INDIA

## ABSTRACT

In the internet world Peer-to-Peer computing is considered to be a promising technology because of the fact that P2P file sharing plays a brilliant part. Its systems are disseminated systems which aid in the allotment of resources over a large peer population namely file content that is common, software application, computer memory storage and hardware. Cost efficiency and degree of availability of personal computers are two of the main attributes of the P2P file sharing systems. However, the occurrence of free-riding threatens the presentation of P2P file sharing systems. It originates from the denial of peers to quit a few of their resources to the worth of the community. Hence several peer to peer networks lacks the accurate security and in order to shelter these peer to peer networks the top k queries is utilized. Top-k query processing is equivalent to discovering k objects that possess the top grades on the whole. Responding the K top-ranked answers swiftly and competently is the primary aim of these queries.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Distributed Resource, P2P File Sharing, P2P Networks, Top-K Quires.

\*Corresponding author: Email: [priyankavedicherla@gmail.com](mailto:priyankavedicherla@gmail.com); Tel.: +91 97881 37003

## INTRODUCTION

A Peer-to-Peer network is fashioned when two or more computers are linked and contribute to a set of lacking going away during a split server PC. It could be thought as any type of network architecture through swap between systems in an unproblematic distribution of computer resources and services. It operates both as a provider and end user or dispatcher and recipient. P2P categorization are wholesome opposed to assortment, fundamental, decentralized, intended, unformed and assortment. Because the Internet possess certain key assets namely information, bandwidth, and computing, the Peer-to-Peer networks are utilized. This is the way peer-to-peer network operates. Suppose you are sprint a P2P file-sharing; attempt to force away an appeal for the file that is necessary to download. Even the software query of the other PCs are linked to the Internet and function the file-sharing software in order to establish the file in the proper position.

If the software commences to detect a PC that contains the folder essential on its hard drive, then the process of download repeats. Thus by swapping the files, the file-transfer load are disseminated among the computers. Utilizing Top-k queries was my primary and the finest choice although famous technologies like VoD, Social tubes, CDNs exists.

Some people simply download files which further impedes other process and this process is referred to as leeching. Peer-to-Peer has numerous benefits such as:

1. Scalable: It means the assets are given or contributed by clients as well. And as per the need the collective assets would be burgeoning.
2. Reliable: It means the specific type of failure doesn't take place. Geometric distribution
3. Simplicity of Administration: The arrangement of the servers is not much essential in order to place order, replication etc.

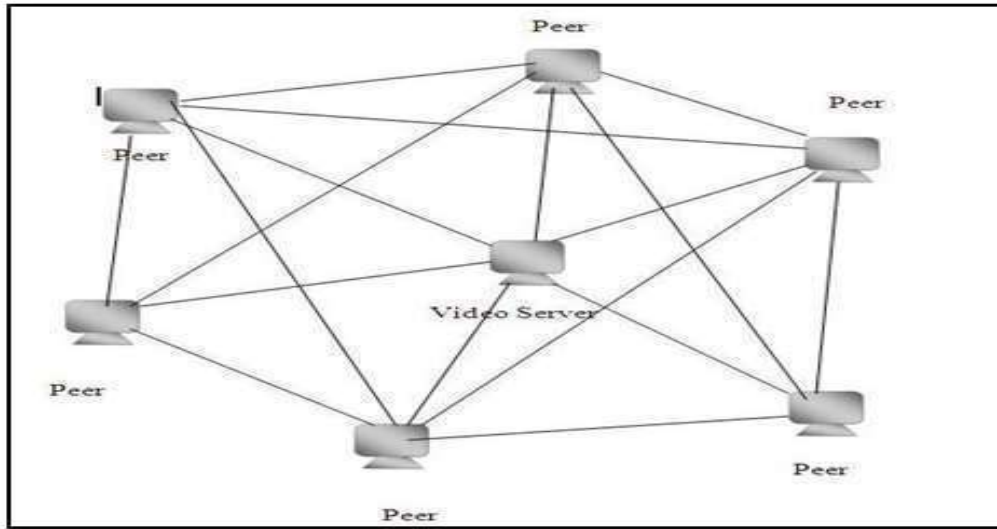


Fig.1: Block diagram of P2P network

[Figure-1] shows that block diagram of the P2P network indicating that all the peers are linked and the data is shared together. In Peer-to-Peer the threats are poisoning, Defection, Adding up of viruses, Denial of Service Filtering, Identity attacks, Spam etc. Also certain safety issues are yet to be addressed by equipments.

The rest of the paper organized the following manner, the Section II focused literature survey of the P2P networks and top-k query processing systems. The Section III gives proposed algorithm the Section IV illustrates Result and implementations and the final section derives the conclusion of the paper.

## RELATED WORK

Ad-hoc is a P2P network that receive a particular concentration payable to the likely applications. It eases the way people access the same data by accessing the particular record from other the data centers. This saves the bandwidth time enabling people in some specific place to contact the internet via wireless equipment's like roof net. Simultaneously, if one user downloads the video or audio the other people can also get the same file. Through this effective way we can realize that the nodes or users share the information with each other. In fact collaborative caching permits quite a few nodes or users to access the same data. This enhances the system performance with lesser bandwidth and time [3].

By utilizing web cache the P2P networks can further be enhanced. This is possible by inter linking the browser cache of each client with the cache of web proxy and thus the cache information is shared. The most important benefit is that hit ratio becomes better and averts unnecessary application. Few client owns its individual browser cache. A set of clients will be linked to an alternative and when the request is processed, the local cache will be initially utilized. In case of the absence of any match the request will then be directed either to other caches or to the original server. Once the request is established it will kept it in the alternative and go back to the browser and store the document in its cache. Replication of document among diverse stages of the cache, misuse of storage space and restricted scalability are the major drawback [10].

The extended of Peer-to-Peer usage for file-sharing systems is unprecedented. Majority of who uses it are constantly incorporating particular file-sharing systems and hence supplementary amount of things were also being made. The perfect amount and probable elevated magnification of Peer-to-Peer traffic may occur the harmful penalty. Few equipment's to be names are static over time or incapable to be changed, bulkiness of the size, streaming media systems like Servers, P2P and multicasting, Software as a service and clients. The key exceptionality of Peer-to-Peer traffic is "germane to caching", "object reputation" Design and evaluation is possible with these distinctiveness. It is possible to improve the algorithms by utilizing diverse techniques. In the

caching mythology, diverse division sizes for special workloads can be performed. P2P traffic can be identified by individual caches. This includes object popularity. The sizes of P2P object can be enlarged to several Giga Bytes [6].

In this generation file sharing, Internet Protocol Television (IPTV), Video on Demand (VoD) using P2P systems is the current killer applications for the internet. Every day the P2P systems produce enormous traffic and due to this challenge universities have a huge budget to be spent. ISP (Internet Service Provider) also encounter similar issue. There also exists negative efforts which cannot be altered. Here caching of P2P traffic also occurs caching which does not require changing the P2P protocols, because this can be bring into effective action transparently from clients.

To set off various Caches the alternate cache is designed for various systems by utilizing algorithms. It enables easier identification of connections belonging to P2P systems. Thus Cache results from many recurrence of a process or utterance and refinements. Even the performance of P2P clients are now speeded up than the other clients. Many research works are undertaken on several extensions for Proxy Cache [5].

There is unprecedented growth in peer to peer networks, many functions such as Bit Torrent, audio conferencing. Most of the systems held in this strategy can be used by any user all around the world. This open internet power is strong point of peer to peer system.

By utilizing closed communities and campus community it can be viewed. The closed community is formed and it is also used by the nation-wide ISP in Europe. The ISP can support bulk bandwidth. This ISP allows the user to use both ADSL and FTTH technology. The ADSL uses the copper cable and FTTH uses the fiber cable. The community which was observed in this network is according to eMule (use multiple ID's to attack the KAD network). To avoid the problems, the eMule has been modified by the users. It was referred as ISP community which is called as closed community. The second community was found in the academy site with rapid Ethernet connection. The campus community user has Direct Connect to the peer to peer application. So that the peer which was running in the host joins community which is called as campus-community [7].

And also many users are using the internet. It has the ability to support large amount of work, can withstand many difficult condition, and has good infrastructure on cost and also it supports many file sharing application. Some networks in P2P which was based on kademila (DHT) like KAD supports million numbers of peers which are connected. One important security which affects peer to peer network is Sybil attack. It is a type of security. And also a malicious user which can gather all information by using fake identities [8]. It monitors the P2P network, polluting the content indexation and also affecting the DHTs. The main target of this Sybil attack is

1. Internet polls – uses multiple IP address to get more votes
2. eMule – use multiple IDs to attack the KAD network.
3. Reputation systems – eBay: create fake accounts and give the feedback to server.

It can be secured using trusted authority that means all nodes should be authenticate with all public keys. The second is resource testing. In that system should have limited IP address, testing the computer power, storage etc. [8].

It was very useful for the reason that algorithms and protocols have been experienced before it is produced. We can make processes and growth in the networks even the simulations are also carried on the same prototype code. This involves either resources to convert low to high range test beds, or it will use the high range services like planet lab so that there will be advantages of practicality of the feral internet and it may to constant too. To avoid this problem they are using the Model Net. It will never move towards or under the outlay of the performance Model Net does not need these experimental facilities like planet lab and one lab. This makes the user to test the P2P application. And it is also suitable for traffic engineering (TE). The main advantage is it can run thousands of unmodified applications. And also social networks are most popular. It provides a greatest share of applications, social connections etc.

The recent rapid development of social network video sharing applications says the evolution is simply communication focused tools to a media portal. Social network needs advanced progression. It depends on content delivery networks (CDNs). And for video content delivery CDNs are very expensive. But the advantage is that

social tube can reduce the work of the server, improves the quality and used in large client application. The present (VoD) technology is not possible to share a particular video in social networks. It results that Social Tube be able to supply a low capture set up delay and small transfer command. To reduce the particular problem they implemented a prototype in planet lab to evaluate the performance of Social Tube. This result from the prototype further confirms the efficiency of Social Tube [1].

The author M.Narayanan describes the tail part of the data segments are not used in the P2P network data communications. So with help of segment table tail part has taken for synchronized with neighboring peers [4].

In current years, P2P networks contains up-and-coming as the best guaranteed approach to tackle problems in the high range powered systems. The co-operative example in these types of networks is that it basically amplifies the check capacity lacking need of particular hold upon or after system communications. Currently Peer-to-Peer networks have accomplished significant outcome to beat the bottom of majority of the applications over Internet. Hence there is no important duty to estimate the particular service. The client has the position for asking any video snip independently at any task of point. We can consider the numerical outcomes of the client performance in VoD service will play the vital roles in gaining accurate understanding of the scalability. Hence we found that various types of video clips are located in different directories on the VoD server [2]

## PROPOSED ALGORITHM

So many peer to peer networks are not secured properly. So to secure these peer to peer networks we are using top k queries to implement this. The [Figure- 2] shows the Top – K process. The Top-k query processing is equal to finding k objects that have the highest overall grades. The base relation is vertically fragmented across the network. Centralized algorithms require several iterations to complete (recall TA) and each iteration introduces additional latency and messaging.

**Table: 1. Proposed Top-K Elements Algorithm**

1. **Algorithm for Finding the top-K Elements**
2. {
3. Sorted access segment
4. }
5. Number of Peer is EQUAL TO 0
6. WHILE Number of Peer is LESS THAN K
7. DO
8. Every Peer  $P_i$  discovers the  $m$  closest local Peer according to the local distance metric  $W_i S_i$
9. {
10. Once the first  $m$  Peer have been created
11. Each peer can minimally locates the subsequent  $k$  Peer in
12. sorted order
13. }
14. END WHILE
15. SET Identity as ID is produced by Peer  $P_i$  from the ID's created in the previous step
16. Number of Peer  $\leftarrow |ID0T...TIDp-1|$
17.  $M \leftarrow M + K$
18. END WHILE
19. { Random access / calculation of segments}
20. Peer  $\leftarrow ID0S...SIDp-1$
21. FOR ID  $\in$  Peer Do
22. FOR all Peer  $j = 0, \dots, p - 2$  DO
23.  $P_j$ : Produced a random  $r_j$  uniformly from the field  $F$
24.  $P_j$ : Calculates the local weighted score  $W_j S_j$  for Peer ID
25.  $P_j$ :  $in_j \leftarrow W_j S_j + r_j \pmod{F}$
26. The local weighted score is returning its function to  $P_j$ .
27.  $P_j$ : Launch  $in_j$  to Peer  $P_{p-1}$
28.  $P_j$ : Send  $-r_j \pmod{F}$  to Peer  $P_{p-2}$
29. END FOR
30.  $P_{p-1}$ : Compute  $spid = P_{p-2} j=0 in_j + w_{p-1} S_{p-1} \pmod{F}$
31.  $P_{p-2}$ : Compute  $sp'id = P_{p-2} j=0 -r_j \pmod{F}$
32. }

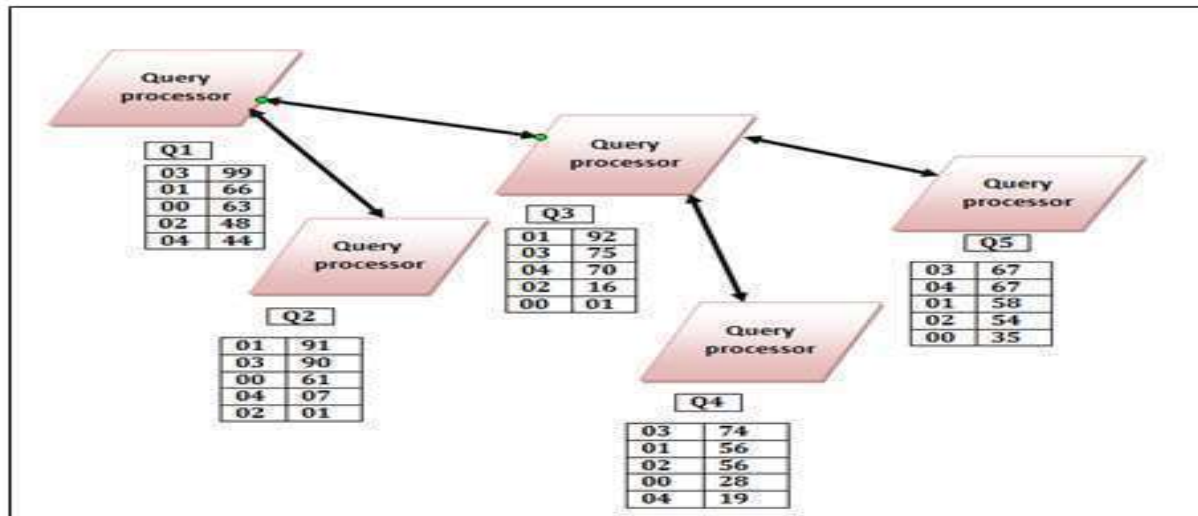


Fig: 2. Top-K query Process

- The main objective of these queries is to return the K highest-ranked answers quickly and efficiently. It is used to
1. Reduce the price metric which is linked or connected with the recovery of all the answer sets. (e.g., disk, network, etc.)
  2. If we maximize the excellence of the answer set, such that the client is not beset by the means of unrelated outcomes.

### RESULT AND IMPLEMENTATION

We implemented in ns2.34, which is discrete event driven simulator tool for learning the dynamic nature of the wireless communication network. The [Table- 2] describes the parameters set by ns2.34 simulator. We calculated the End-to-End delay of videos using trace file information. The results show end-to-end delay of the video packets has less delay.

Table: 2. Implementation Parameters

Parameters	Values
Video Packet size	1048
Mac	802.11
Routing protocol	AODV
Prorogation model	Two-ray ground
Application Protocol	TCP/CBR

Once the parameters are set, the eventual step is setting up the threshold of the traffic. Here threshold set has 250 peers which mean that up to 250 peers the server will provide service considering here that every peer with 40 kbps speeds. After accomplishing the threshold point, 250 peers automatically the Top-K Query Process takes accountability to supply services. Thus, whatever the peers are appealing to the server, the whole appeal is made in line. And then based on any one of the scheduling concept that convey work to the concerned peer, that peer will provide service to the requested peer

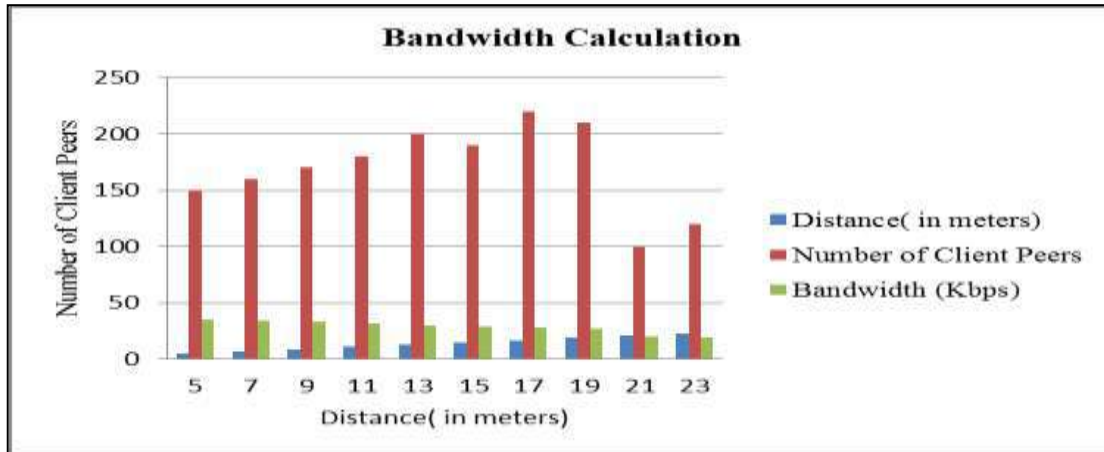


Fig. 3. Bandwidth Calculation

The Top-k query processing is equivalent to discovering k objects that possess the top grades on the entire networks. The [Figure-3] drawn bandwidth calculation of the P2P networks. The x-axis proceeds distance of the peers and y-axis continues the number of clients participates in the Peer-to-Peer Networks.

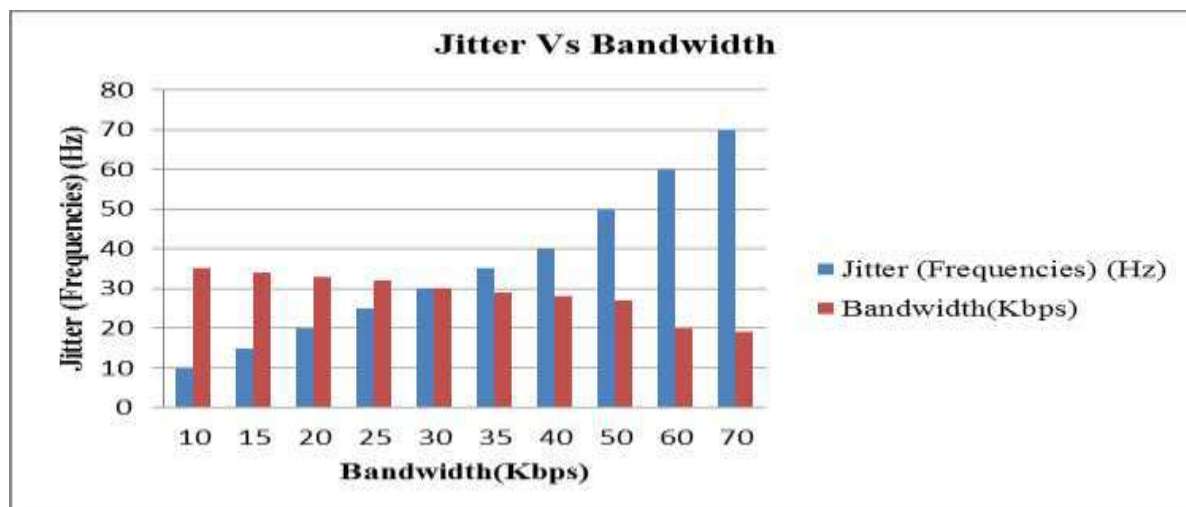


Fig. 4. Bandwidth with Jitter

The P2P networks mainly depends on Jitter, so we have to take the main factor Jitter and to produce better result. The [Figure-4] illustrates Jitter versus Bandwidth. Server provides service to client is employed after the threshold point based on the parameters like Distance of the peers, nearest path of the client/server, Jitter status.

### CONCLUSION

Hence we can bring to a close that one of the most promising technology worldwide is the Peer-to-Peer networks. It is utilized not only for media file sharing but the set-up can initiate a new channel for efficient and competent downloading and sharing of files, data and information. There should be proper security procedures to evade the potential leak of responsive or individual information and other safety breach. Organization administrators must ensure that every demand corresponds with the commercial security guidelines. End users comprising home users should be particular to clear up the possibility of diffusing viruses over the peer-to-peer network. The downside of Top-K Query process can be conquered by introducing new technology and built it as a trusted network.



**CONFLICT OF INTEREST**

The authors declare no conflict of interests.

**ACKNOWLEDGEMENT**

None

**FINANCIAL DISCLOSURE**

None

**REFERENCES**

- [1] Dario R, Paolo V, Matteo S, Federico L. [2013] *Model Net-TE: An emulation tool for the study of P2P and traffic engineering interaction dynamics*, Springer Peer-to-peer Networking and Applications (PPNA). 6(2):194-212.
- [2] Jian-GuangLuo, Qian Z, Tang Yun Y, Shi-Qiang Y. [2009] *A Trace-Driven Approach to Evaluate the Scalability of P2P-Based Video-on-Demand (VoD) Service*, Parallel and Distributed Systems, IEEE Transactions on. 20(1).
- [3] Jing Z, Ping Z, Guohong C. [2008] *On Cooperative Caching in Wireless P2P Networks*, Published in Distributed Computing Systems, 2008. ICDCS '08. The 28th International Conference on Date of Conference. 17-20.
- [4] Narayanan M, Arun C. [2014]. Outliers Based Caching of Data Segment with Synchronization over Video-on Demand using P2P Computing. Research Journal of Applied Sciences, Engineering and Technology. 7(21): 4559-4564.
- [5] Mohamed H. [2011] Senior Member, IEEE, Cheng-Hsin Hsu, Member, IEEE, and Kianoosh Mokhtarian, Student Member, IEEE, *Design and Evaluation of a Proxy Cache for Peer-to-Peer Traffic*, IEEE Trans. Computers. 60(7):964-977.
- [6] Osama S, Mohamed Hefeeda M. [2006] *Modeling and Caching of Peer-to-Peer Traffic*, Network Protocols, 2006. ICNP '06. Proceedings of the 2006 14th IEEE International Conference on Date of Conference.12-15
- [7] Ruben T, Marco M, Maurizio M, Munafò, Sanjay G. Rao. [2012] *Characterization of community based-P2P systems and implications for traffic localization*, © Springer Science + Business Media, LLC.
- [8] Thibault C, Isabelle C, Olivier F, Guillaume D. [2012] *Detection and mitigation of localized attacks in a widely deployed P2P network*, Published online: © Springer Science + Business Media, LLC.
- [9] Rajan C, Shanthy N. [2013] Misbehaving attack mitigation technique for multicast security in mobile ad hoc networks (MANET), Journal of Theoretical and Applied Information Technology. 48(3):1349–1357.
- [10] Xiaodong Z, Artur A, Songqing C. [2004] *Building a Large and Efficient Hybrid Peer-to-Peer Internet Caching System*, IEEE Transactions on Knowledge & Data Engineering. 16(6):754-769. doi:10.1109/TKDE.2004.1.

# ANALYZE AND PREVENT MODERN EMAIL MALWARE PROPAGATION USING SEII MODEL

**S. Sneha\***, **P. Swapna***Dept of Information Technology, M. Kumarasamy college of Engineering, Thalavapalayam, Karur-639113, Tamilnadu, INDIA*

## ABSTRACT

**Aim:** It is necessary to develop a prevention model to avoid potential damages caused by modern email malware. While comparing to earlier versions the modern email malware has two characteristics: reinfection and self-start. In earlier research some models were developed for analyzing the malware propagation still there is needed to improve the accuracy of the model. Further the existing approach uses virtual node concept which increases the computational overhead. To concentrate on these problems, SEII model is proposed to analyze and prevent the modern email malware. Based on the result of the analysis model the impact of parameters in propagation is analyzed and presents the automated email malware detection and control system. Complete inspection and demonstration shows that the proposed model can detect and prevent the modern email malware in effective manner.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

*email malware, reinfection, self-start, SEII model, control system*\* Email: [snekasekaran@gmail.com](mailto:snekasekaran@gmail.com); Tel.: +91 7871473343

## INTRODUCTION

The email malware creates the major security issues for email service users. A computer virus is one of the major forms of malicious information spreading in the Internet. The computer viruses are classified into scanning-based viruses and topological-based viruses. The email malware is based on the topological viruses. Once an email user is infected by email malware, malicious email copies are sent to their friends embedded in email lists. The user will get infected, whenever they open and read the email received from the infected user. The infection processes are spreads quickly and reach a large scale, frequently from one user to adjacent users.

Current research on email malware focuses on propagation dynamics<sup>1, 2, 3, 4, 5</sup> which is used to decrease email malware distribution speed. In email malware, if the user visits a malicious mail again or not the infected user will send the email copies once<sup>1, 2, 3, 6</sup>. The modern email malware is extremely dangerous because of reinfection and self-start. In reinfection, every time healthy or infected recipients open the malicious attached file the modern email malware sends its copy to all users in list. In self-start, each time compromised computers restart or malicious files are visited the malware, and then the malware occupied the memory of the system by spreading its own copy. The existing analytical SII model<sup>10</sup> presented the procedure of malware spreading, but it cannot accurately estimate the spreading of email malware. A SEII systematic model is proposed to describe the email malware propagation. The distribution procedure can be distinguished by a susceptible-exposed-infected-immunized (SEII) process.

In this paper, we define virus transmission rules as follows:

- If susceptible nodes get contact with an infectious node, then it transmits into an exposed.
- If exposed nodes get contact with an infectious node, then that will transmit into an infected.
- All the nodes transmits to an immunized state, when the user immunized.

[13] proposed an analytical model that<sup>13</sup> represented the spreading process by susceptible-infected-susceptible (SIS) process. In this model, susceptible and infected users both can be susceptible again. In<sup>1</sup> [18] presented the SIR model to describe the email malware propagation. In

this model both susceptible and infected users can be recovered and they would receive lifelong immunity. The work of [6]<sup>9</sup> characterized the propagation dynamics of isomorphic malware, such as P2P malware<sup>1</sup>, mobile malware<sup>14, 15</sup> and malware on online social networks<sup>16, 15, 17</sup>. These existing models are not able to effectively present the propagation of modern email malware.

To tackle the problem of reinfection and self-start Sheng Wen<sup>10</sup>, proposed a novel analytical SII model. It cannot perfectly estimate the spreading of modern email malware. This model had some minor deviation between mathematical model and simulations results because of the independent assumption.

### PROBLEM DEFINITION

The malware propagation is based on non-reinfection, refection and self-start. In fact, reinfection alone is not sufficient to describe the propagation of modern email malware because most email malware is having the self-start. Modeling the non-reinfection is simple when compared to reinfection and self-start. Therefore, reinfection and self-start are the two mechanisms used to differentiate the spreading of modern email malware.

Reinfection denotes whenever the infected user opens the malicious mail he will get infected again. The reinfection dominates the non-reinfection in below aspects:

- (i) User will infect again even if he/she infected before.
  - (ii) Whenever the user get infected he/she will send the mail to their neighbor
- Thus, a receiver may continually receive malicious emails from the same compromised user.

Self-start is the behavior of spreading malware whenever the infected computers restart. The sur pass of the self-start is given below:

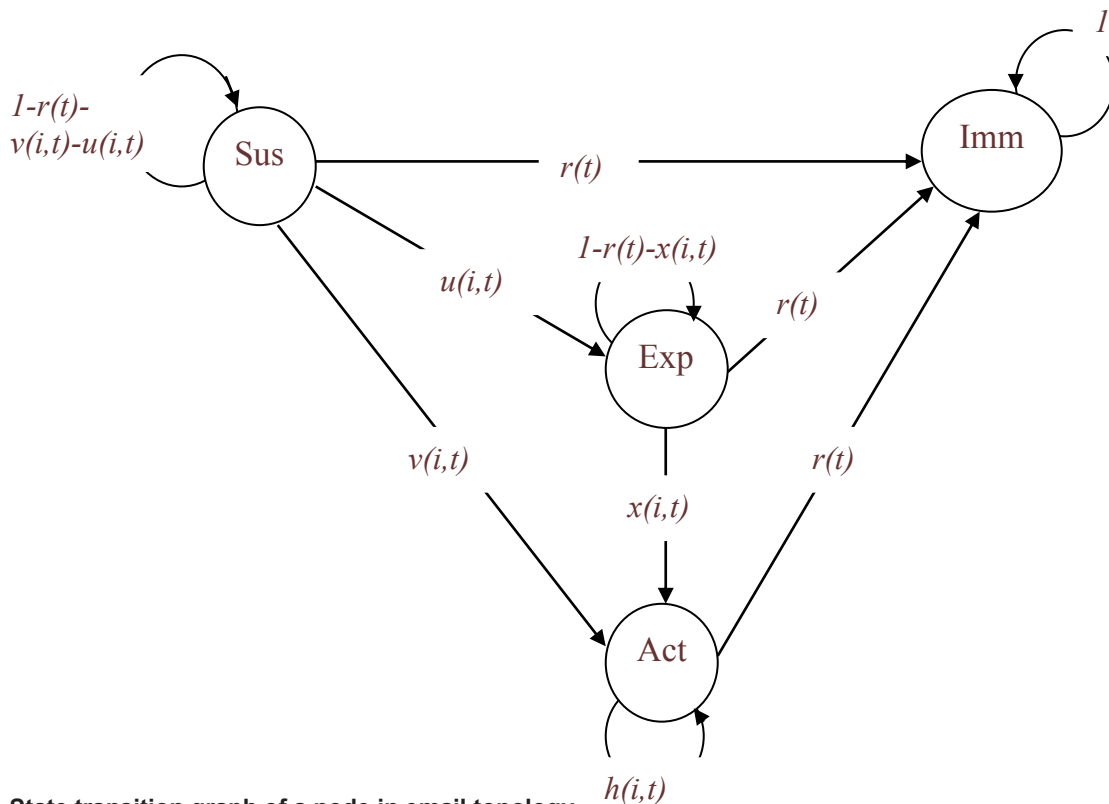


Fig: 1. State transition graph of a node in email topology.

A user has been infected at a particular time. If the user restarts the computer, a malicious email copy is received by a new user because of self-start. The malware can spread much faster in self start, when compare to all other models.

To conquer these complexities, SEII analytical model is proposed which will describe the dissemination of the modern email malware.

## PROPOSED SYSTEM

To overcome the difficulties of previous models, we propose SEII model as shown in [Figure- 1] for modern email malware. SEII model is accurate than SIS and SIR models<sup>18, 10</sup> and SII model<sup>10</sup>. In this model, susceptible, exposed and infected users can be immunized.

## SEII MODEL - INTRODUCTION

The essential elements for the dissemination of modern email malware are topology and node information. A node represents a user in the email network. Let random variable  $X_i(t)$  indicate the state of a node  $i$  at distinct time  $t$ .

$$X_i(t) = \begin{cases} \text{Sus., Susceptible} \\ \text{Hea., Healthy} \begin{cases} \text{Exp., Exposed} \\ \text{Imm., Immunized} \end{cases} \\ \text{Inf., Infected} \end{cases} \quad (1)$$

Initially all the nodes in the network are susceptible. In susceptible state the user have the possibility of getting infected. The susceptible node is transits to active state when the user infected. If the user  $i$  is in the address book of infected user, then the infection possibility of the user  $i$  is higher. Therefore, the user transmits from the susceptible state to the exposed state. Since, the infected user sends out the malware to the user  $i$  when it is compromised, then the user  $i$  transmits from exposed state to active state after the infection of user  $i$ . All the states are transmits to immunized state at final.

Let  $r(t)$  be the probability of immunization.  $H(i, t)$  is the probability of being in the active state.  $V(i, t)$  is the probability of node  $i$  transits from susceptible to active state.  $U(i, t)$  be the probability that node  $i$  transits from susceptible to exposed state.  $X(i, t)$  is the probability that the node  $i$  moved to active state from exposed state. In SEII Model, an  $N$  by  $N$  square matrix with elements  $p_{ij}$  is used to describe a network topology consisting of  $N$  nodes, as in,

$$\begin{pmatrix} P_{11} & \dots & P_{1M} \\ \vdots & P_{ij} & \vdots \\ P_{M1} & \dots & P_{MM} \end{pmatrix} p_{ij} \in [0,1] \quad (2)$$

Where in  $p_{ij}$  stands for the probability, that the user  $j$  visits a malicious email from user  $i$ . If  $p_{ij}$  is equal to zero, mail address of user  $j$  is not present in the contact list of user  $i$ . Therefore, the matrix replicates the topology of an email network. This model assumes the states of neighbouring nodes are dependent.

The infection of malware depends on unsuspecting email users' read-through new emails. This process involves two components in the modeling. First, the flag variable  $open_i(t)$  is introduced.  $Open_i(t)=1$ , if the newly arrived mails are checked by user at time  $t$ , otherwise  $open_i(t)=0$ .

$$P(open_i(t) = 1) = \begin{cases} 0, & \text{otherwise} \\ 1, & t \bmod T_i = 0 \end{cases} \quad (3)$$

$T_i$  – user  $i$ 's email checking period.

Every user has different values of  $T_i$ . User can receive numerous emails at different time but the user checks the mailbox and read the email at one time. Suppose the user  $i$  checks newly received emails at time  $t$ , then the user checks email at time  $t$ . That is user  $i$  receives the new email after the last checking act of her mailbox. Here we initiate a variable  $\tau$  to indicate random time between the user  $i$ 's last email checking time and the current time  $t$ . Then the value of  $\tau$  is,

$$\begin{cases} t - T_i \leq \tau < t, & \text{if } open_i(t) = 1 \\ t - (t \bmod T_i) \leq \tau < t, & \text{if not} \end{cases} \quad (4)$$

An infected user only propagates malware to the adjacent users in topological networks. For each user in email networks, we record and collect every new malicious email from adjacent users at time  $t$ , and at last attain the joint infection probability of each user who checks the malicious emails.

## SPREADING ANALYSIS

Here, the values 0 and 1 are used to substitute the healthy state and the infected state, respectively.  $M$  denotes the nodes in email network topology, the number of infected users at time  $t$  and  $n(t)$ , is computed as in,

$$\begin{aligned} n(t) &= \left[ \sum_{i=1}^M X_i(t) \right] = \sum_{i=1}^M E[X_i(t)] \\ &= \sum_{i=1}^M [(0P(X_i(t) = 0) + (1P(X_i(t) = 1))] = \sum_{i=1}^M P(X_i(t) = 1) \\ &= \sum_{i=1}^M P(X_i(t) = \text{inf}) \quad (5) \end{aligned}$$

The probable number of infected nodes,  $n(t)$ , is equal to the sum of the probability of each node being infected a time  $t$ ,  $P(X_i(t) = \text{inf})$ . As made known in [Figure- 1], susceptible node and an exposed node can be infected and stay in the infected state, and an infected node recovered and stay at the immunized state. The transitions of each state is used to derive the  $P(X_i(t) = \text{inf})$  by difference equations as follows:

$$P(X_i(t) = \text{Inf}) = (1 - r(t))P(X_i(t-1) = \text{Inf}) + v(i, t)P(X_i(t-1) = \text{Sus}) + x(i, t)P(X_i(t-1) = \text{Exp}) \quad (6)$$

To find  $P(X_i(t) = \text{Sus})$  we have,

$$P(X_i(t) = \text{Sus}) = 1 - P(X_i(t) = \text{Inf}) - P(X_i(t) = \text{Exp}) - P(X_i(t) = \text{Imm}) \quad (7)$$

To find  $P(X_i(t) = \text{Exp})$  we have,

$$P(X_i(t) = \text{Exp}) = u(i, t)P(X_i(t) = \text{Sus}) + (1 - r(t))P(X_i(t-1) = \text{Inf}) + (1 - x(i, t))P(X_i(t-1) = \text{Inf}) \quad (8)$$

To find  $P(X_i(t) = \text{Imm})$  we have,

$$P(X_i(t) = \text{Imm}) = P(X_i(t-1) = \text{Imm}) + r(t) \cdot [1 - P(X_i(t-1) = \text{Imm})] \quad (9)$$

Once the values of  $v(i, t)$ ,  $x(i, t)$  and  $r(t)$  are obtained then the value of  $P(X_i(t) = \text{Inf})$  can be calculated by using the iteration of the above equations, (6),(7),(8) and (9).

There are three prerequisites for random user being infected by email malware is given below:

- i) the user should not immunized;
- ii) the user checks for new emails;

iii) the user incautiously visits malicious emails;  
 When the first two prerequisites are satisfied, the  $s(i, t)$  is used to represent the probability of user  $i$  visiting malicious emails from adjacent nodes. Then, the probability for infection  $v(i, t)$  and  $x(i, t)$  can be derived as in,

$$v(i, t) = s(i, t) \cdot P(\text{open}_i(t) = 1)(1 - r(t)) \quad (10)$$

$$x(i, t) = s(i, t) \cdot P(\text{open}_i(t) = 1)(1 - r(t)) (1 - u(i, t)) \quad (11)$$

In SEII model, an user  $i$  visits malicious attachments with  $p_{ji}$  probability, when reading malicious emails from a adjacent user  $j$ .  $N_i$  denotes the set of adjacent nodes of node  $i$  after eradicate the unknown nodes. Then, we can calculate  $s(i, t)$  as in,

$$s(i, t) = \prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(\tau) = Act)] \quad (12)$$

Where in the event  $X_j(\tau) = Act$  means that the node  $j$  is infected and propels a malicious mail copy to adjacent nodes at time.

Let us consider that the variable  $\tau$  may take the different values, the equation (12) is disassembled by not including  $t_1$  from the range of value  $\tau$ . There are two cases: First, the user does not checking new emails at time  $t_1$ . Thus, we have

$$\prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(\tau) = Act)] = \prod_{j \in N_i, \tau \neq t-1} [1 - p_{ji} \cdot P(X_j(\tau) = Act)] \times \prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(t-1) = Act)] \quad (13)$$

$$= (1 - s(i, t-1)) \cdot \prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(t-1) = Act)] \quad (14)$$

Second, the user checks new emails at time  $t_1$ . Thus, the malicious email copies are received at time  $t$  and those are delivered at time  $t_1$  by the infected adjacent users. Here we have,

$$\prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(\tau) = Act)] = \prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(t-1) = Act)] \quad (15)$$

Actually, the difference of equations (14) and (15) are caused by user checks new emails at time  $t_1$ . Then unified expression of (14) and (15) is given below:

$$\prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(\tau) = Act)] = [1 - s(i, t-1) \cdot (1 - P(\text{open}_i(t-1) = 1))] \cdot \prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(t-1) = Act)] \quad (16)$$

Now, the equation (12) becomes,

$$s(i, t) = 1 - [1 - s(i, t-1) \cdot (1 - P(\text{open}_i(t-1) = 1))] \times \prod_{j \in N_i} [1 - p_{ji} \cdot P(X_j(t-1) = Act)] \quad (17)$$

In equation (17), various trials of  $P(X_j(t-1) = Act)$  and  $N_i$  may guide to different distribution performance.

## SIMULATION AND RESULTS

In proposed SEII model the valuation is based on the open analytical model. The spread of the majority email malware is typically impractical to track spreading of malicious mail. In this work, we construct the topology according to the previous investigation of existent email networks. The degree for every node was imitated by the Power-law distribution. The probability of users infected by adjacent nodes ( $p_{ij}$ ), the email checking time ( $T_i$ ) and the event generating period ( $R_i$ ) are determined by human factors. These parameters follow the Gaussian distribution which may provide idealistic values, such as  $p_{ij} < 0$  and  $T_i < 1$ . Here, these values are substituted with their practical range. Thus, if  $p_{ij} < 0$ ,  $T_i < 1$  and  $R_i < 1$ , we let  $p_{ij} = 0$ ,  $T_i = 1$  and  $R_i = 1$ . First, the accuracy for different circulation of  $p_{ij}$  is evaluated. The  $T_i$  and  $R_i$  pursue Gaussian distribution  $N(40, 20^2)$ . As shown in [Figure- 2], the outcomes of SEII model are close to the outcomes of simulations. The SEII model reaches better presentation in correctness.

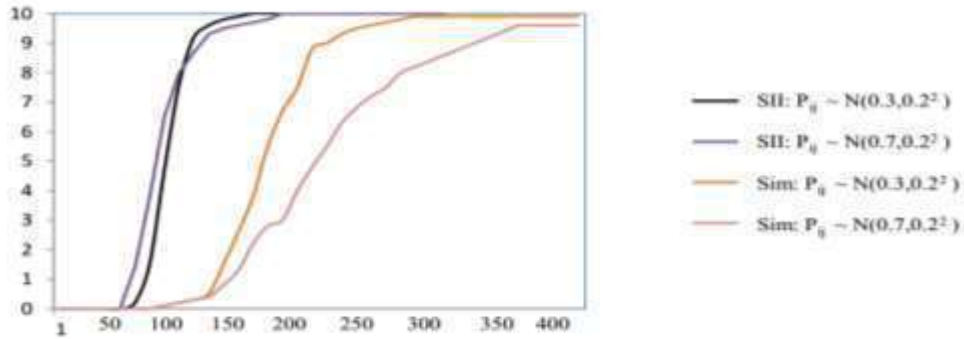


Fig. 2. The accuracy with different distributions

Second, the accuracy for different topologies is evaluated.  $T_i$  and  $R_i$  pursue Gaussian distribution  $N(40, 20^2)$  and the probability of infection  $p_{ij}$  follow  $N(0.5, 0.2^2)$ . As shown in [Figure- 3], the proposed SEII model is efficient in various topologies with various power-law exponents  $\alpha$  and means of degrees  $E(D)$ .

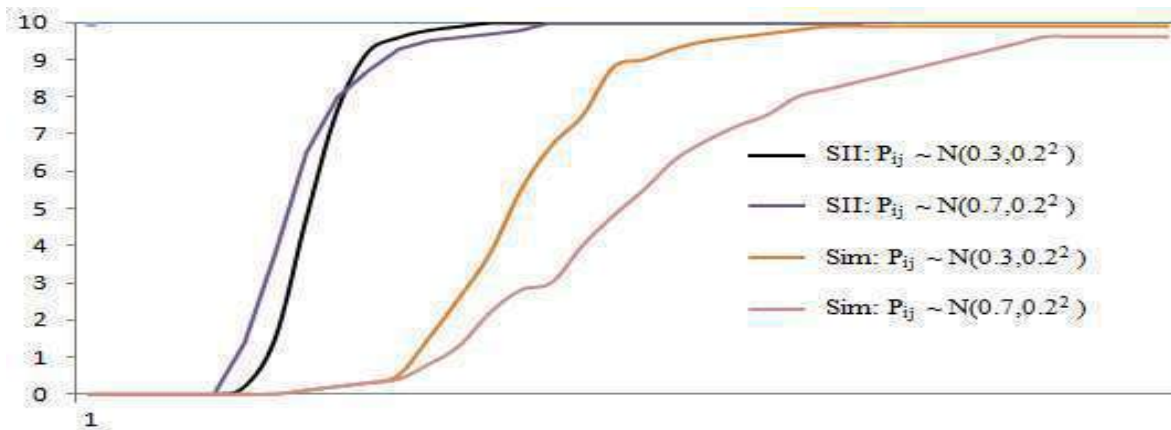


Fig. 3. The accuracy with different distributions of  $T_i$  and  $R_i$

### CONCLUSION

In this work, we proposed a novel SEII model for evaluate the modern email malware propagation. This model is capable to deal with two critical issues, which are reinfection and self-start. By establishing difference equations, the repetitious distribution processes caused by the above mentioned issues are presented. The outcome of the proposed SEII model is closest to the simulations. For the future work, spatial dependence and temporal dynamics problems are taken into account. Novel simulations have to be designed to enclose real system samples, to analyze the behavior of malware against these samples. The aim of this simulation is to avoid system before being infected by real malware.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Fan W, Yeung KH. [2011] Online Social Networks-Paradise of Computer Viruses, *Physica A: Statistical Mechanics and Its Applications*. 390(2):189-197.
- [2] Garetto M, Gong W, Towsley D. [2003] Modeling malware spreading dynamics," in Proc. INFOCOM'03. San Francisco, CA. 3:1869-1879.
- [3] Wen S, Zhou W, Wang Y, Zhou W, Xiang Y. [2012] Locating Defense Positions for Thwarting the Propagation of Topological Worms, *IEEE Comm. Letters*. 16(4):560-563.
- [4] Xiong J. [2004] Act: Attachment Chain Tracing Scheme for Email Virus Detection and Control, Proc. ACM Workshop Rapid Malcode (WORM '04). 11-22.
- [5] Zou CC, Towsley D, Gong W. [2007] Modeling and Simulation Study of the Propagation and Defense of Internet E-Mail Worms, *IEEE Trans. Dependable and Secure Computing*. 4(2):105-118.
- [6] Wen S, Zhou W, Zhang J, Xiang Y, Zhou W, Jia W. [2013] Modeling Propagation Dynamics of Social Network Worms, *IEEE Trans. Parallel and Distributed Systems*. 24(8):1633-1643.
- [7] Calzarossa M, Gelenbe E. [2004] Performance Tools and Applications to Networked Systems: Revised Tutorial Lectures. Springer-Verlag.
- [8] Serazzi G, Zanero S. [2003] Computer Virus Propagation Models," Proc. 11th IEEE/ACM Int'l Conf. Modeling, Analysis and Simulations of Computer and Telecomm. Systems (MASCOTS '03). 1-10.
- [9] Sheng W, Yang X, Weijia J. [2014] Modeling and Analysis on the Propagation Dynamics of Modern Email Malware, *IEEE transactions on dependable and secure computing*. 11:(4)
- [10] Sneha S, Malathi L, Saranya R. [2015] A Survey on Malware Propagation Analysis and Prevention Model, *International Journal of Computer Applications* (0975 – 8887).131(11)
- [11] Zou CC, Towsley D, Gong W. [2007] Modeling and Simulation Study of the Propagation and Defense of Internet E-Mail Worms, *IEEE Trans. Dependable and Secure Computing*. 4(2):105-118.
- [12] Gao C, Liu J, Zhong N. [2001] Network Immunization and Virus Propagation in Email Networks: Experimental Evaluation and Analysis, *Knowledge and Information Systems*. 27:253-279.
- [13] Chen Z, Ji C. [2005] Spatial-Temporal Modeling of Malware Propagation in Networks, *IEEE Trans. Neural Networks*. 16(5):1291-1303.
- [14] Wang Y, Chakrabarti D, Wang C, Faloutsos C. [2003] Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint, Proc. 22nd Int'l Symp. Reliable Distributed Systems (SRDS). 25-34.
- [15] Yanping Z, Tingting S, Shu Z. [2012] A Novel Model to Restrain Email Virus Propagation, *IEEE International Conference on Granular Computing*.
- [16] Ganesh AJ, Massouli L, Towsley DF. [2005] The Effect of Network Topology on the Spread of Epidemics, Proc. IEEE INFOCOM '05:1455-1466.
- [17] Yan G, Eidenbenz S. [2009] Modeling Propagation Dynamics of Bluetooth Worms (Extended Version)," *IEEE Trans. Mobile Computing*. 8(3):353-368.
- [18] Boguna M, Pastor-Satorras R, Vespignani A. [2003] Epidemic Spreading in Complex Networks with Degree Correlations, *Lecture Notes in Physics*. 625:1-23.
- [19] Simarleen K, Arvinder K. [2016] Detection of Malware of Code Clone using String Pattern Back Propagation Neural Network Algorithm. 9(33).



## A SURVEY ON MAC PROTOCOL IN UNDERWATER SENSOR NETWORK

P. Swapna\*, S. Sneha

Department of Information Technology, Kumarasamy College of Engineering, KARUR, TN, INDIA

### ABSTRACT

**Aim:** Underwater sensor network is widely used in many applications such like, environment monitoring, mobile object tracking, navigation applications, etc. In UWSN, MAC protocol plays an important role. Here radio waves are used for communication. We also use the acoustic communication, mainly used in physical layer with longer range but high error probability and node mobility are limited. MAC protocol contains two main nodes they are sensor node and sink node. Sensor nodes are deployed into a connected network to gather sensor data and send the data to the sink node. In this survey paper, we compare the throughput, collision rate, advantage, disadvantage and methodology of each MAC protocol and we also compare the issues and steps to overcome UWSN.

Published on: 2<sup>nd</sup> -December-2016

#### KEY WORDS

MAC protocols, underwater sensor network, wireless sensor network

\* Email: [swapnaengineer@gmail.com](mailto:swapnaengineer@gmail.com) Tel.: 9994234835

### INTRODUCTION

Wireless sensor network is used in wide range. Specifically, in this paper we use WSN in the process of underwater sensor network and its applications. In UWSN it contains many sensor nodes mainly used to collect information and processing the data. These are managed by the sensor node. The sensor node is capable of collecting data from one or more sink nodes. These nodes are mainly used to reduce the power consumption. Here we use MAC protocol as the backbone of the operations. There are many sets of MAC protocol being used in UWSN and in WSN. Some of the MAC protocol are QMAC, Asym-MAC, Z-MAC, etc. Each of the MAC protocol contains different algorithms or steps followed and vary in each set of function and in execution.

In underwater sensor network the detector node which is mainly used for collecting the information through the wireless sensor network. It mainly operates in power unit and its impossible to replace that unit. Detectors are mainly used for minimizing the memory and bandwidth and to increase the range of communication. The main scope of the detectors is to handle the critical section. It also contains another detector called the spatially distributed autonomous used for monitoring the functions. This UWSN is mainly build by nodes which is used to pass the data, monitor the function, intimation of intruder's etc.

MAC protocol plays an important role in UWSN. MAC is also called as medium access control. It acts as medium when a data transaction undertaken. This protocol layer is a communication network protocol and as a sub layer. This layers is also under the layer of OSI. It's the second layer data layer or data link layer. Here we take the IEEE, this is divided into two sectional part they are LLC and MAC. LLC stands for logical link layer and MAC stands for medium access control. This LLC is under IEEE and it contains three sub protocol. IP protocol, IPX protocol these are mainly used to carry the packet or at the period of transmission. Similarly the MAC layer contains two main functions they are data encapsulation and frame initiating

This MAC layer contains more set of protocol such like ALOHA, CSMA, CSMA/CD, CSMA/CA etc. Here we take the ALOHA, this is mainly used in terrestrial broadcasting function and implementations in the satellite system. It has the power to hander more channel. Collisions are expected to occur because the function of this protocol is when the node is ready to send data it transmits and collision occurs. The CSMA protocol is used in

shred topologies. It is used to capture the data before sending a data and if a node s busy to capture the data then it waits. Hence to overcome the waiting process of the node we intake two sections they are CSMA/CD, CSMA/CA.

### MAC PROTOCOL OF WSN

There are more set of MAC protocol used in underwater sensor network. We compare the collision, energy efficient, and their major issues in these compared papers.

Chih-Min [1] major in wireless sensor network and it intake the QMAC for function. This MAC is also called as Quarum MAC. It is used in several application such like environment monitoring and navigation application. QMAC saves the energy and reduce collision. In this the major concept is next hop condition. Using this concept the latency is being reduced. The lifetime of the sensor node are based on the energy efficient protocol. (QMAC) protocol that enables sensor nodes to sleep longer under. To conserve energy, MAC protocol, realizing that sensor nodes have different loads due to their different distances to the sink. To adjust their sleep durations based on their traffic loads the concept of quorum is used. To reduce delays induced by longer sleep durations, we have increased each node's transmission opportunity by enabling group of next-hop nodes to accomplish the packet-relaying job. light loads to reduce transmission latency. The major pros and cons of this paper is QMAC protocol that enables sensor nodes to sleep longer under light loads. QMAC protocol saves more energy and keeps the transmission latency low. To reduce transmission latency. CONS are regular sleep/awake mechanism fails to adjust a sensor node's sleep duration based on its traffic load, thus causing either lower power efficiency or higher latency. Furthermore, sensors may be deployed in hostile environments and may thus unexpectedly fail. Most power-saving protocols do not promptly react to such link breakage, resulting in long transmission delays.

Myounggyu Won [2] he presents the Asym MAC, these are used for several application such like medical and military field. Here the collision occurs at the sender and the energy is consumed. The major issues in this paper is low channel utilization and they are overcome using the concept receiver imitated protocol. This concept provide a performance called combat performance degradation. The lifetime of the sensor node is based on LB-MAC.

Chih-Min Chao [3] he present the MM protocol which states multiple-rendezvous multichannel MAC protocol, undertaken underwater sensor network. The lifetime of the sensor network is based on the distributed multiple rendezvous multichannel MAC protocol and they also provide the channel allocation and overcome the problem of receiver. It reduce the collision and consumes the energy using this concept. This protocol is used in many application some of them are disasters warning and tactical surveillance.

Pei Huang [4] uses the RC-MAC major in wireless sensor network.. It has certain issues like over hearing problem and they use the concept of receiver centric scheduling to overcome this issues. They provide high throughput and reduce the collision rate. The application used in this protocol are target tracking and intrusion detection.

Tiansi Hu [5] major in underwater sensor network. In this paper they use the DSH MAC protocol mainly used for past decade application. The use the concept of decoupled and suppressed handshaking which provide the channel allocation and throughput. It reduces the collision using this concept and the lifetime of the sensor node is based on RTS and CTS MAC protocol. RTS states Request to Send, this is used to send the request when a node is ready to data to another sensor node. At that time this RTS works. CTS states that Clear to Send, when the request RTS is send from one set of node and check the availability. If the nodes are free to transmit the data then it send the acknowledgment called CTS.

Chao Li [6] this paper represents the DTMAC which is mainly used for sparse network and maily use the novel delay tolerant MAC. using this MAC protocol it overcome the long propagation delay and swarm mobility. The energy efficiency is consuming less and it avoid the collision. The lifetime of the sensor node is based on the coupon collection algorithm and RTS/CTS protocol.

Chih-Min Chao [7] Multiple-Rendezvous Multichannel MAC Protocol Design for Underwater Sensor Networks distributed multiple-rendezvous multichannel MAC protocol, MM-MAC is a multiple rendezvous protocol; only one modem is required for each node. Utilizing cyclic quorum systems, nodes running MMMAC are guaranteed to meet their intended receivers, which solves the missing receiver problem. The separation of control and data transmissions also helps reduce the collision probability of data packets. MM-MAC, to reduce collision probability. The major advantage of this paper is Only one modem is needed for each node to solve the missing

receive problem multiple sensor node pairs can complete their channel negotiations on different channels simultaneously. Data packets will not be collided by control packets. The disadvantage is Fail to function effectively in a multi hop network consisting of more sensor nodes with heavier traffic loads. In underwater sensor networks (UWSNs) include Lower transmission rate, longer delay time, and higher power consumption.

Maaz M. Mohiuddin, [8] this paper represent the concept of clock driven and event driven. Here the clock driven is based on periodic scheduling and they mainly used under internal clock function. The data packets also execute in periodic function. Then we take event driven, these are based on the external agent. The major issues are timely delivery of packets. These issues are overcome by the using the EEDF MAC. There are several application used under this MAC protocol they are environmental parameters. The lifetime of the sensor node is based in the cluster size and energy efficient. In this paper there are five phases they are network initialization phase, schedule broadcasting phase, data transmission phase, synchronization phase and control phase. The first phase is taken over at the first time of the set up and event driven transmit the data from one to another using the CSMA/CA . Here now second phase start working , it has the combination of the event driven and the clock driven are taken and they are schedule the sink node data collected by the sensor node. They form these result as a table and transmit to the third phase. The third phase is the data transmission phase and they transmit the data which is having the high priority and transmit the low priority data in later. Then fourth phase is synchronization phase here they has the topology network and they collect the data of the failure nodes and consolidate the data to one document and transmit them. Here they create a new set of document using the old and failure data.

Lei Tang[9] et al present the paper in EM MAC , also called as energy efficient MAC protocol. These are suitable for wireless sensor network. Their major issue is they doesn't relay on the dedicated control channel and hence we use the reduce duty cycle function which provide the multichannel redezeous. They reduce the collision rate and application used under this MAC protocol is zigbee channel. The lifetime of the sensor node is based on the O-MAC. This O-MAC contains wake up time section for receiver and generate the slot number with frames. They are based on the Pseudo-random Staggered on scheme

Geethanjali and pravin [10] describe the protocol for energy and its efficiency for this we include the multi-channel MAC. The main mechanishm of this paper is multichannel. They avoid collision and consume the energy. The major issues are distributed channel assignment and efficient cross channel communication. This MAC protocol overcome this issues using the multiple channel and sleep allocation. There are many application used under this MAC function some of them are industrial control, monitoring, security and military intelligence. The lifetime of the sensor node is based on the MAMAC, this utilize the quorum system and also deal with the mutual exclusion problem.

Here we have taken MAC protocol which plays major role in Wireless sensor network and underwater sensor network. Each of the MAC protocol avoids the collision and reduces the collision using the concept of the different set of MAC protocol

### LIFETIME OF MAC IN WSN

Here we combine all set of MAC protocol papers based on wireless and underwater sensor network each of the sensor nodes increases their lifetime by certain MAC. each paper takes different algorithm , different MAC protocol , different function for improving the strength of the sensor nodes.

Lifetime of the sensor node should be improved and using this latency then power will be consumed in low rate. We can achieve the long battery life. Chih-Min [1] et al represent the paper under QMAC, the lifetime of this MAC is improved by using the energy efficient protocol. This protocol overcome the problem like idle listening and conserves energy. They are maintained by this MAC function using QMAC\_LR it increase the power by saving and it overcome the problem depletion.

Tiansi Hu [5] this paper's sensor node lifetime is based on the RTS/CTS MAC protocol. This protocol is mainly used for the handshaking process, in this they reserve the channel of communication. When the sender is ready to send the data then it request for the near by neighbour nod by RTS function. These nearby node check for the availability, if any data is working under this node then it avoid this RTS request to overcome the collision. These process are mainly taken over by the two control periods they are NOTE and GRANT. The NOTE packet is used to inform the number of buffered packet and it also intimates the transmission intension of the sender. Similarly

the GRANT packet is used to inform the readiness of the data. these control packet is based on decoupling and data transmission /handshaking process. This is the interior process of RTS/CTS function. Using this the lifetime of the sensor node is increased.

Table -1

S.NO	MAC LAYERS	MAJOR IN	COLLISION	ENERGYEFFICIENCY	LIFETIME	ISSUES	APPLICATION	PROVIDE	CONCEPT
1	QMAC	WSN	Expected	Save	Energy Efficient Protocol	Node Delay Pending Packet	environment monitoring, navigation application	reduces latency	Next Hop
2	Asym-MAC	WSN	At Sender	Consume	LB-MAC	Low channel utilization	Medical,military	combat performance degradation	Receiver Initiated Protocol
3	MM-MAC	UWSN	Reduces	Consume	distributed multiple-rendezvous multichannel MAC	channel assignment & transmission scheduling	tactical surveillance, disaster warning	channel allocation & to solve the missing receiver problem	Cyclic Quorum System
4	RC-MAC	WSN	Reduces	Improve	novel receiver-centric MAC protocol	overhearing problem	Intrusion detection, Target tracking	High Throughput	receiver-centric scheduling
5	DSH-MAC	UW-ASN	Reduce	More	RTS/CTS MAC protocol	Long Probagation	Past decade	Channel utilization & throughput	Decoupled & suppressed handshaking
6	DTMAC	UW-WSN	Avoid	Consume less	coupon collection algorithm & RTS/CTS protocol	Swarm mobility,long probagation delay	Sparese n/w	Optimimal through put	Novel delay tolerant MAC
7	ZMAC	WSN	Reduces	Good	B-MAC	slot assignment failures & time-varying channel conditions	Real Time	high channel utilization and low-latency	time slot assignment algorithm.
8	EEDF-MAC	WSN	Avoid	Save	cluster size and energy efficient	timely delivery of packets	environmental parameters	energy efficiency & improved latency performance	clock-driven and event-driven
9	EM-MAC	WSN	Reduces	High	O-MAC	Doesn't rely on dedicated Control channel	ZigBee channel	multichannel rendezvous	Reduce Duty Cycle
10	Multi channel MAC	WSN	Avoid	Consume	MAMAC	Distributed channel assignment & efficient cross channel Communication	Industrial control and monitoring, Security and military intelligence	multiple channels & sleeping mechanism	Multichannel
11	RPMAC	WSN	Avoid	Improve	novel receiver-pivotal MAC	Adapt to Low Duty Cycle	surveillance, intrusion detection and target tracking	integrates duty cycling and receiver centric scheduling	Multichannel
12	MC-LMAC	WSN	Avoid	Save		Scheduled access	intruder detection structural health monitoring	higher throughput	single channel LMAC protocol
13	Multihop Fair Access-MAC	WSN & UWSN	Reduce	Reduce	TDMA scheduling algorithms	propagation delay	moored oceanographic	composite throughput	single-channel and half-duplex radios & grid networks
14	X-MAC	WSN	mitigate	Save	awake state and a sleep	long sleep time	traffic loads	energy consumption,	asynchronous duty-cycled

					state			latency, and throughput	MAC
15	BiC-MAC	UW-AN	depends on the distribution	Save	time-slotting	channel utilization	terrestrial-based models	novel analytical framework	bidirectional concurrent data transmissions

### COMPARISON OF MAC PROTOCOLS

Shuguo Zhuo [16] et al is the process of I Queue MAC. In this they balance the traffic load by adaptive duty cycle and increase the through put and energy efficient. There are five key features they are without any overhead the load information will be accurate , allocation of TDMA here the node takes the data from the sender and exchange the simple node to the time slots using this the throughput is increased. Synchronization of the nodes which are of neighbouring nodes using LPL condition, shortening the channel from a router to another router transmission and finally multi-channel gains the acknowledgement. The lifetime of the sensor node is based on the sleep and active

### CONCLUSION

In this paper we have taken a comparative study about the MAC protocols major in underwater sensor network, wireless sensor network and acoustic sensor network. In this, there are many different MAC protocol with different mechanism and different algorithm and different function. Here we compare the efficiency, throughput and collision rate. Using this comparison we improve the old scheme with new concept. This new concept improve the functionalities, throughput rate and avoid collision. It also reduce issues taken over in old scheme. In each mechanism there will be new challenges that cannot be avoided, and some can be avoided with new concept. [Table-1] describe about the MAC protocol and their performance. Using this we can improve the functionalities in better way.

[Table-1] defines the different protocols the first paper Chih-Min Chao [1] et al, he represent the major issues in the paper is node delay pending packet. This issue is rectified using the MAC protocol called QMAC. QMAC is quorum and provide a next hop concept, this concept is used to avoid the issue node delay pending rate because the next hop which used to have cyclic quorum in the time slot so its harder to leave a data without any functionalities.

Tiansi Hu [5] he represents the issue long propagation. This drawback is rectified by DSH-MAC ie. Decoupled and supressed handshaking process. Its followed by RTS/CTS MAC protocol (request to send / clear to send). Here two sets of mode are taken sleep and awake. The nodes will be awake when a data is ready to transmit and reaming time the node will go to sleep mode. When the node is in sleep mode then it can't transmit any set of data so, at this part of section the node send RTS is send to the neighbour by data. Then the node awake for the time of transmitting and after that it move on to the sleep mode.

Here we compared two papers in theory and such like there are more set of papers. The large scale of localization system which affects the speed, communication cost, coverage and improve the performance of MAC protocol.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

### REFERENCES

- [1] Chih-Min C, Yi-Wei L. [2010] A Quorum-Based Energy-Saving MAC Protocol Design for Wireless Sensor Networks" IEEE transactions on vehicular technology. 59(2).
- [2] Myounggyu W, Taejoon P, Sang H. [2014] Son Asym-MAC: A MAC Protocol for Low-Power Duty Cycled Wireless Sensor Networks with Asymmetric Links, IEEE communications letters. 18(5).

- [3] Chih-Min C, Yao-Zong W, Ming-Wei L. [2013] Multiple-Rendezvous Multichannel MAC Protocol Design for Underwater Sensor Networks, *IEEE transactions on systems, man, and cybernetics: systems*. 43(1).
- [4] Pei H, Chen W, Li X. [2015] RC-MAC: A Receiver-Centric MAC Protocol for Event-Driven Wireless Sensor Networks, *IEEE transactions on computers*. 64(4).
- [5] Tiansi H, Yunsi F. [2013] DSH-MAC: Medium Access Control Based on Decoupled and Suppressed Handshaking for Long-delay Underwater Acoustic Sensor Networks, *IEEE conference on local computer networks*.
- [6] Chao Li, YongJun Xu, ChaoNong Xu, ZhuLin An, BoYu Diao, and XiaoWei Li. [2015] DTMAC: A Delay Tolerant MAC Protocol for Underwater Wireless Sensor Networks, *IEEE Sensors Journal*, DOI 10.1109/JSEN.2462740.
- [7] Injong R, Ajit W, Mahesh A, Jeongki M. [2005] ZMAC: a Hybrid MAC for Wireless Sensor Networks, *ACM 159593054X/05/0011*.
- [8] Maaz M, Mohiuddin, Adithyan I, Rajalakshmi P. [2013] EEDF-MAC: An Energy Efficient MAC Protocol for Wireless Sensor Networks, *International Conference on Advances in Computing, Communications and Informatics*.
- [9] Lei T, Yanjun S, Omer G, David B. [2011] Johnson EM-MAC: A Dynamic Multichannel Energy-Efficient MAC Protocol for Wireless Sensor Networks, *ACM 978-1-4503-0722-2.16-19*.
- [10] Geethanjali S, Pravin reold A. [2013] Multi-channel mac protocol for energy saving in wireless sensor networks, *International Journal of Smart Sensors and Ad Hoc Networks (IJSSAN)*. 3(1):2248-9738,
- [11] kancherla V, Ramesh babu P, Nirupama P. [2014] RPMAC: A Novel Receiver Pivotal MAC Protocol Event Driven Wireless Sensor Networks, *International Journal of Advance Research in Computer Science and Management Studies*. 2(11).
- [12] zlem Durmaz Incel O, Pierre J, Sape M. MC-LMAC: A Multi-Channel MAC Protocol for Wireless Sensor Networks.
- [13] Yang X. [2012] Senior Member, IEEE, Miao Peng, John Gibson, Geoffrey G. Xie, Ding-Zhu Du, and Athanasios V. Vasilakos, Senior, Tight Performance Bounds of Multihop Fair Access for MAC Protocols in Wireless Sensor Networks and Underwater Sensor Networks, *IEEE Transactions on mobile computing*. 11(10).
- [14] Mohan R, Rajan C, Shanthi N. [2012] A Stable Mobility Model Evaluation Strategy for MANET Routing Protocols. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2:58-65.
- [15] Michael B, Gary V, Yee, Eric A, Richard H. [2006] X-MAC: A Short Preamble MAC Protocol for Duty-Cycled Wireless Sensor Networks, *ACM 1-59593-343-3/06/0011*, 1-3.
- [16] Hai-Heng N, Wee-Seng S, Mehul M. [2015] Saturation Throughput Analysis of the Slotted BiC-MAC Protocol for Underwater Acoustic Networks, *IEEE Transactions on wireless communications*. 14(7).
- [17] Shuguo Z, Zhi W, Ye-Qiong S, Zhibo W, Luis A. iQueue-MAC: A Traffic Adaptive duty-cycled MAC Protocol With Dynamic Slot Allocation.

# A SURVEY ON EFFICIENT CRYPTOGRAPHIC APPROACH FOR DATA SECURITY IN WIRELESS SENSOR NETWORKS

S. Ilakkiya\*, M. Mailsamy, J. Gladson Maria Britto

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for women, Namakkal, Tamilnadu, INDIA

## ABSTRACT

*Aim: A new symmetric encryption standard algorithm which is the amalgamation of two different encryption algorithms proposed by Nath et. Al namely TTJSA and DJSA algorithms in randomized method. The algorithm is named as Modern Encryption Standard version – I algorithm. The idea of modern encryption standard is to make a symmetric key cryptographic method which should be unbreakable. The MES version –I algorithm is effective against frequency analysis and spectral analysis. Further improvements can be bit level encryption can be performed on the text files after dividing the plaintext into two text files.*

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Keymanagement, WSN,  
random pre distribution

\*Corresponding author: Email: [ilakskiya13081994@gmail.com](mailto:ilakskiya13081994@gmail.com); Tel.: +91 9944485920

## INTRODUCTION

In modern digital communication era, allocation of information is collective significantly. The information being diffused is exposed to innumerable attacks. Therefore, the information security is solitary of the most challenging aspects of communication in any current network. This also smears to WSNs, specially those used in applications that monitor sensitive information (e.g., health care applications). These networks are quickly gaining popularity that they are potentially low cost solutions to a variety of real world challenges and are expected to play an essential role in the upcoming age of pervasive computing. However, the highly constrained nature of sensors imposes a difficult challenge: their reduced availability of memory, processing power and energy delays the deployment of many modern cryptographic algorithms considered safe.

### Cryptography

The encryption-decryption techniques devised for the traditional wired networks are not viable to be applied directly for wireless sensor networks. WSNs consist of tiny sensors which really ache from the deficiency of battery power, processing and memory. Any encryption scheme applying on WSNs require transmission of extra bits, hence extra processing, memory and battery power which are very central resources for the sensor's durability. Applying the security mechanisms such as encryption could also increase delay, jitter and packet loss in wireless sensor networks [3]. There are some key queries arise when applying encryption schemes to WSNs like, how the keys are generated or dispersed. There is an important issue how the keys could be changed time to time for encryption as there is insignificant (or no) interaction for the sensors. There are other many issues how keys are revoked, assigned to a new sensor added to the network or rehabilitated for ensuring robust security for the network. There could not be an efficient solution for adopting of pre-loaded keys or embedded keys.

## Data security

For this reason, the choice of the most memory, processing and energy efficient security solutions is of spirited importance in WSNs. To date, several dramatists have developed wide studies comparing different encryption algorithms. WSNs can be seen as a special type of ad-hoc network poised by a large number of little, low-cost and highly resource forced sensor nodes, known as spots. The sensors are spread in the area of interest, and can then meet and process data from the environment (e.g., mechanical, thermal, biological, chemical, and optical readings). They have applications in a variety of fields such as environment monitoring which involves checking look, dirt and marine, state based maintenance, habitat monitoring (determining the plant and animal species population and behaviour), seismic detection, military following, inventory chasing, smart spaces and rally sensing material in hostile locations, medical and home security to machine diagnosis, chemical/biological detection etc[6]. Spots are normally battery-powered, which has moved considerable research efforts on the development of energy mindful protocols, such as data link layer protocols. In general, one of the main goals driving the design of these systems is to improve network communications in order to save energy, and thus extend the network's lifetime. On the other hand, security is regularly very forlornly considered at the very last step in the design of WSNs. Actually, most WSN organisations do not even consider security between their requirements because the effecting and energy outlays it adds to the system is seen as an adverse "extra cost" in such forced environments. However, in WSN-based applications that monitor sensitive information, it is vital to avoid eavesdropping, which is typically obtained by means of encryption systems (e.g., symmetric ciphers). Even when the evidence acquired is not confidential, it is still necessary to ensure data integrity and reality by means of message authentication mechanisms, since the approval of invalid data (generated either by natural causes or with malicious purposes) could lead to mistaken actions and severe values.

## ARCHITECTURE

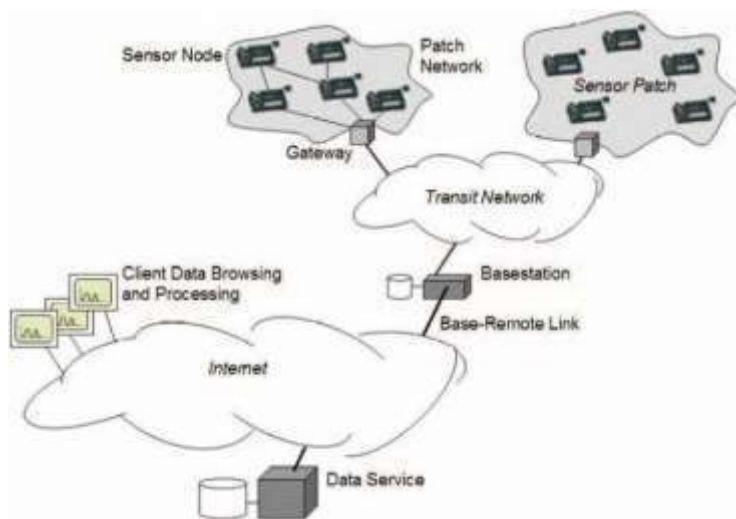


Fig: 1. Wireless sensor network

## LITERATURE REVIEW

Key management deals with the secure generation, distribution, and Storage of keys. It plays a vital role in computer security today as practical attacks on public-key systems are typically aimed at key management as disparate to the cryptographic algorithms themselves.

In[1] Authors Abtin Keshavarzian, Elif Uysal-Biyikoglu in the paper "Energy-efficient Link Assessment in Wireless Sensor Networks" "For energy on strained stationary wireless networks of sensom, selection of links with high quality rate helps to guarantee consistent long-term operation. During the implementation of a protocol aiming industrial applications of such systems, it was found that it is useful to acquire exact information about the availability and quality of the RF communication links prior to the network topology formation. "Link



assessment” as part of the initialization process, undertakes this task by assessing a enough amount of packets exchanged between neighboring nodes. It introduces and analyze two different approaches to link valuation: The first attitude is a random nondeterministic scheme that permits for a probabilistic guarantee of collision-free packet exchange. An alternative method is described which employs ‘ronstant-weight codes’ and provides a determinklic guarantee 01 success [2]. In particular, a speciel class of constant-weight codes, known as optical orthogonal codes, are considered. Since, these codes are regularly permutable, they make the link assessment process simpler, and therefore they are preferred over other codes. And evaluate the performance of these methods based on their energy consumption, time duration, and implementation complexity.

In[2]Authors Theodoros Salonidis<sup>1</sup>, Pravin Bhagwat<sup>2</sup>, Leandros Tassioulas<sup>1</sup>, and Richard LaMaire<sup>3</sup>in the paper” Distributed Topology Construction of Bluetooth Personal Area Network” In recent years, wireless ad hoc networks have been a emergent area of research. While there has been considerable research on the topic of routing in such networks, the topic of topology creation has not received due devotion. This is since almost all ad hoc networks to date have been built on top of a single channel, broadcast based wireless media, such as 802.11 or IR LANs. For such networks the reserve relationship between the nodes obliquely (and uniquely) determines the topology of the ad hoc network. Bluetooth is a promising new wireless technology, which enables portable devices to form short-range wireless ad hoc networks and is built on a frequency hopping physical layer. This fact implies that hosts are not able to communicate unless they have previously discovered each other by matching their frequency leaping patterns. Thus, even if all nodes are within direct communication range of each other, only those nodes which are synchronized with the mast can get the transmission. To provision any-to-any communication, nodes must be synchronized so that the pairs of nodes (which can communicate with each other) together form a connected graph. Using Bluetooth as an example, this rag first provides deeper insights into the issue to link establishment in frequency hopping wireless systems [6]. It then announces the Bluetooth Topology Costruction Protocol (BTCP), an asynchronous distributed protocol for constructing scatternets which starts with nodes that have no knowledge of their surroundings and sacks with the formation of a connected network satisfying all connectivity constraints posed by the Bluetooth technology [6]. To the best of our knowledge, the work presented in this paper is the first attempt at building Bluetooth scatternets using distributed logic and is quite “practical” in the sense that it can be implemented using the communication primitives offered by the Bluetooth 1.0 specifications.

In[3]Authors Elyes Ben Hamida, Guillaume Chelius in the paper”Revisiting Neighbor Discovery with Interferences Consideration” In wireless multi-hop networks, hello protocols for neighbor finding are a basic overhaul existing by the networking hoard. However, their study usually rely on quite simplistic models which do not take into version problems ensuing from low level layers, such as the physical layer. One of the individualities of radio communications is the presence of interferences which decrease the size of the medium. A random hello protocol inspired by aloha and study the impact of the interferences on the neighbordiscovery process[7]. As expected, and prove that, in average and in the presence of interferences, a node discovers only a subset of its neighbours and analytical model to compute the average number of nodes that a given node may expect to discover in its neighborhood. Finally,can present a hello protocol with sleep periods and show how to optimize this protocol using our hybrid model. A real scenario stemming from the CAPNET project is then analyzed and studied.

In[4]Authors Sudarshan Vasudevan, Jim Kurose, Don Towsley presented in the paper”On Neighbor Discovery in Wireless Networks With Directional Antennas” The problem of neighbour discovery in static wireless ad hoc networks with directional antennas. Then propose several probabilistic processes in which nodes perform random, independent shows to discover their one-hop neighbors. Our neighbor sighting algorithms are confidential into two sets, viz. Direct- Discovery Procedures in which nodes discover their neighbors only upon receiving a transmission from their neighbors and Gossip-Based Algorithms in which nodes gossip about their neighbors' location evidence to allow faster discovery. consider the operation of these algorithms in a fit, synchronous system and mathematically derive their optimal parameter settings. How to cover these algorithms for an asynchronous system and describe their optimal design. Analysis and simulation of the algorithms show that nodes discover their neighbors much sooner using conversation-created processes than using through-discovery algorithms. Furthermore, the routine of gossip-based algorithms is insensitive to an surge in node density. The efficiency of a neighbor discovery algorithm also depends on the choice of antenna beamwidth. How the choice of beamwidth effects the performance of the finding process and provide insights into how nodes can construct their beamwidths.

In [5] Authors Jason Hill and David Culler {jhill, culler}@cs.berkeley.edu presented in the paper “A wireless embedded sensor architecture for system-level optimization” Power consumption and abilities of the radio statement layer are the prevailing factors in total system concert. It presents a wireless sensor node architecture to achieve high communication bandwidth with the bounce to efficiently implement novel communication protocols. The architecture is instantiated in an operational design using viable microcontroller and radio technology. Its ability to adjust system act by using unusual protocols is by four case studies involving power management, synchronization, localization, and wake-up.

In[6] Authors N. B. Anuar, M. L. M. Kiah, V. A. Rohani, D. Petkovi presented in the paper “Key management protocol with end-to-end data security and key revocation for a multi-BS wireless sensor network”, A significant wireless device network with several base stations (BS), a key management protocol is planned in it. For confidently conveying data between a node and a base station or two nodes, an end-to-end data security method is accepted by this protocol. Further using a distributed key reversal scheme to efficiently remove conceded nodes then forms our key management protocol celled multi-BS key management protocol (MKMP).

In [7] Authors W. Bechkit, Y. Challal, A. Bouabdallah, and V. Tarokh presented in the paper “BHNFDIA Energy Efficient Elimination of Black Hole and False Data Injection Attacks in Wireless Sensor Networks “The trustworthy containers will be accelerated and malicious packets will be unwanted nearly. The proposed scheme can exclude false data injection by outside malicious nodes and Black hole attack by compromised insider nodes. Replication results show that the scheme can successfully identify and eliminate 100% black hole nodes. Malicious packets are immediately impassive with 100% filtering efficiency. The pattern ensures more than 99% packet delivery with better network traffic.

In [8] Authors D. H. Yum and P. J. Lee presented in the paper “Hands-On Experiences in Deploying Cost-Effective Ambient-Assisted Living Systems “The prototype is built upon inexpensive, off-the-shelf hardware (e.g., various sensors, Arduino microcontrollers, ZigBee-attuned wireless communication modules) and license-free software, there by warranting low system deployment costs. The network embraces nodes placed in a house’s main rooms or mounted on furniture, one wearable node, one actuator node and a centralized processing element (coordinator). Upon detecting significant abnormalities from the conventional movement patterns of individuals and/or sudden falls, the system issues automated alarms which may be forwarded to legal care patrons via a variety of statement channels.

In [9] Authors F. Gandino, B. Montrucchio, and M. Rebaudengo presented in the paper “Information Assurance Based on Cryptographic Checksum with Clustering Security Management Protocol “To swear information security beside security attacks and particularly node seizing attacks and propose a cluster security management protocol, called Cryptographic Checksum Clustering Security Management (C3SM), to offer an competent reorganised security management for hierarchal networks. In C3SM, every cluster selects dynamically and alternately a node as a cluster security manager (CSM) which issues a episodic shared secrete key for all nodes in the cluster.

In [10] Authors W. Du, J. Deng, Y. S. Han, P. K. Varshney, J. Katz, and A. Khalili presented in the paper “Hybrid Intelligent Routing in Wireless Mesh Networks: Soft Computing Based Approaches “Wireless Lattice Networks (WMNs) are the evolutionary identity-forming multi-hop wireless networks to ability last mile access. Due to the rise of stochastically varying network environments, steering in WMNs is disapprovingly affected. This Integrated Link Cost (ILC) is added for each link based upon throughput, delay, jitter of the link and remaining energy of the node and is used to calculate shortest path between a given source-terminal node pair.

## COMPARISION OF DIFFERENT TECHNIQUE

By Comparing these keys to rally the protocols for proficiency for energy consumption can be highly accomplished. And expense of an increased node density and network latency.

Table I

S.NO	TITLE OF THE PAPER	KEYS/PROTOCOLS	SECURITY	DEMERITS
	Energy efficient	RandomKey Distribution	Energy Efficient	Require Large Recall

[1]	Link Assessment in Wireless Sensor Networks			Space To Store The Ring.
[2]	Distributed Topology Construction of Bluetooth Personal Area Network	Bluetooth Topology construction protocol	Short-range radio link between movable devices	The Topology certainly affects its operation and performance
[3]	Revisiting Neighbor Discovery with Interferences Consideration	Hello protocol	A Symmetric link neighbor will include your id in its neighbor list.	Larger Hello message size in dense networks.
[4]	On Neighbor Discovery in Wireless Networks With Directional Antennas	Neighbor discovery Algorithm	Efficient in tracking the nodes with in a node's communication range.	Static and dynamic multihop Sensor Networks. Conventional node discovery techniques were found to be inadequate and they give less significance to the QoS parameters.
[5]	A wireless embedded sensor architecture for system-level optimization	Multi BS Management protocol	Additional load on router equipment	It may affect the stability of the other protocols.
[6]	Key management protocol with end-to-end data security and key revocation for a multi-BS wireless sensor network	Key Management Protocol	It deals with the secure generation, distribution and storage of keys.	Difficult to check Routing information.
[7]	BHNFDA Energy Efficient Elimination of Black Hole and False Data Injection Attacks in Wireless Sensor Networks	Black hole and false data injection	Simple form of Selective forwarding attack, where a malicious node may drop all the packet passing through it without forwarding to the sink node.	Malicious Node Communicates the Destination Node with false route information.
[8]	Hands-On Experiences in Deploying Cost Effective Ambient Assisted Living Systems	Plain Global Key Scheme	Modification	The System Administrator stores this information in the memory of the nodes.
[9]	Information Assurance Based on Cryptographic Checksum with Clustering Security Management Protocol	cluster security management protocol	Increase the Energy Consumption of Sensor Nodes, Strong Resilience Against Node Capture With Lower Key Storage.	An Enemy That knows the master key and the identification number of the cluster head can extract the cluster key and easily Attack the cluster.

[10]	Hybrid Intelligent Routing in Wireless Mesh Networks: Soft Computing Based Approaches	Transistor Key	Fabrication, modification	False information to the neighboring nodes, packet loss, selective forwarding attack
------	---	----------------	---------------------------	--

## SUSCEPTIBLE SOLUTION

The main benefit of q-s-composite is represented by an efficient memory management, which allows to store a larger quantity of keys and consequently it can improve the resilience of the protocol. This result is reached by means of a new key generation mechanism and by limiting the quantity of starting keys per link.

## CONCLUSION

The potential drawbacks of the proposed scheme have been analyzed and an in-depth analysis has shown that their effects are overcome by the security improvements. A comparison with state-of-the-art schemes shows that the proposed approach represents the best solution for large mobile WSNs, and that it is also the best solution for static WSNs, if the nodes can be compromised during the initialization phase.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Seo SH, Won J, Sultana S, Bertino E. [2015] Effective key management in dynamic wireless sensor networks, *Information Forensics and Security, IEEE Transactions on.* 10(2):371–383.
- [2] Gisbert J, Palau C, Uriarte M, et al. [2014] Integrated system for control and monitoring industrial wireless networks for labor risk prevention, *Journal of Network and Computer Applications.* 39:233–252.
- [3] Hackmann G, Guo W, Yan G, et al. [2014] Cyber physical co design of distributed structural health monitoring with wireless sensor networks, *Parallel and Distributed Systems, IEEE Transactions on.* 25(1):63–72.
- [4] Rezaee AA, Yaghmaee MH, Rahmani M, et al. [2014] Hoca: Healthcare aware optimized congestion avoidance and control protocol for wireless sensor networks, *Journal of Network and Computer Applications.* 37:216–228.
- [5] Moosavi H, Bui F. [2014] A game-theoretic framework for robust optimal intrusion detection in wireless sensor networks, *Information Forensics and Security, IEEE Transactions on.* 9(9):1367–1379.
- [6] Shamshirband S, Anuar NB, et al. [2014] Co-fais: Cooperative fuzzy artificial immune system for detecting intrusion in wireless sensor networks, *Journal of Network and Computer Applications.* 42(0):102–117.
- [7] Bechkit W, Challal Y, Bouabdallah A, Tarokh V. [2013] A highly scalable key pre-distribution scheme for wireless sensor networks, *Wireless Communications, IEEE Transactions on.* 12(2):948–959.
- [8] Dai H, Min Zhu Z, Gu XF. [2013] Multi-target indoor localization and tracking on video monitoring system in a wireless sensor network, *Journal of Network and Computer Applications.* 36(1):228–234.
- [9] Das AK. [2012] Improving identity-based random key establishment scheme for large-scale hierarchical wireless sensor networks. *IJ Network Security.* 14(1):1–21.
- [10] Yum DH, Lee PJ. [2012] Exact formulae for resilience in random key pre distribution schemes, *Wireless Communications, IEEE Transactions on.* 11(5):1638–1642.
- [11] Gandino F, Montrucchio B, Rebaudengo M. [2014] Key management for static wireless sensor networks with node adding, *Industrial Informatics, IEEE Transactions on.* 10(2):1133–1143.
- [12] Blom R. [1985] An optimal class of symmetric key generation systems, in *EUROCRYPT 84 workshop on Advances in cryptology: theory and application of*

cryptographic techniques. New York, NY, USA:Springer-Verlag. 335–338.

- [13] Du W, Deng J, Han YS, Varshney PK, Katz J, Khalili A. [2005] A pairwise key predistribution scheme for wireless sensor networks, *ACM Trans. Inf. Syst. Secur.* 8(2):228–258.
- [14] Xiao Y, Rayi VK, Sun B, Du X, Hu F, Galloway M. [2007] A survey of key management schemes in wireless sensor networks, *Computer Communications*. 30:11-12,2314–2341.
- [15] Madhusudhanan B, Chitra S, Rajan C. [2015] Mobility Based Key Management Technique for Multicast Security in Mobile Ad Hoc Networks, *The Scientific World Journal*, Hindawi Publishing Corporation.
- [16] Bechkit W, Challal Y, Bouabdallah A, Tarokh V. [2013] A highly scalable key pre-distribution scheme for wireless sensor networks, *Wireless Communications, IEEE Transactions on*. 12(2):948–959.
- [17] Das AK. [2012] A random key establishment scheme for multi-phase deployment in large-scale distributed sensor networks, *International Journal of Information Security*. 11(3):189–211.
- [18] Zhu S, et al. [2006] Leap+: Efficient security mechanisms for large scale distributed sensor networks, *ACM Transactions on Sensor Networks*. 2(4):500–528.
- [19] Gandino F, Montrucchio B, Rebaudengo M. [2014] Key management for static wireless sensor networks with node adding, *Industrial Informatics, IEEE Transactions on*. 10(2):1133–1143.

# EXTRACTION AND DETECTION OF BRAIN TUMOR FROM MAGNETIC RESONANCE IMAGES - A SURVEY

M. Karthika\*, N. Mohana Priya, B. Kalaavathi

Dept of Computer Science and Engineering, Vivekanandha College of Engineering for Women, K.S.R Institute for Engineering and Technology, Chennai, Tamilnadu, INDIA

## ABSTRACT

*Aim: Tumor is a rapid uncontrolled growth of cells; it can be three varieties malignant, benign or pre - malignant. When a tumor classified as malignant then the tumor leads to cancer. Tumor of brain is a collection or mass of abnormal cells in your brain. MRI (Magnetic Resonance Imaging) system is used for detection particularly in medical field and visualization of details in the internal structure of the body. It can detect the differences in the body tissues which is the best technique as compared to CT (Computed Tomography), Endoscopy, PET (Positron Emission Topography), Ultrasonic imaging system. This paper presents survey of various techniques applied on brain tumor extraction & detection from magnetic resonance image (MRI). It has also calculated the parameters of accuracy, sensitivity, and specificity. From the comparative analysis results, Region growing method is best suited for brain tumor extraction and fuzzy C-means can be used for efficient detection of brain tumor from MRI images. This paper is also used to show the relevant feature extraction technique that improves the classification accuracy rate.*

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Medical images, brain tumor, Feature extraction, detection, MRI image, region growing, fuzzy C-means.

\*Corresponding author: Email: [karthikasweet38@gmail.com](mailto:karthikasweet38@gmail.com); Tel.: +91 9524932257

## INTRODUCTION

Medical imaging is process of creating the visual representations of the interior of a body for clinical analysis and medical intervention, as well as visual representation of the function of some organs or tissues. Medical Imaging is used to study the structure and pathological condition of human organ. Image Processing is a method to develop raw images received from sensors for various applications. Image Processing consist of two types are Analog Image Processing refers to the alteration of image through electrical means. Digital Image Processing refers to processing of a two-dimensional picture by a digital computer. Brain tumor detection and extraction is one of the most challenging and time consuming task in medical image processing.

## IMAGE EXTRACTION

Feature extraction describes the relevant shape information contained in a pattern. It is used to create new features as a function of existing features. It has two functions are linear functions (PCA, ICA) and Non linear functions (hidden units in a neural network). It starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization.

## IMAGE DETECTION

Feature detection is a low-level image processing operation. It is usually performed as the first operation on an image and examines every pixel. If this is part of a larger algorithm, then the algorithm will typically only examine the image in the region of the features. It may also provide complementary attributes, such as the edge orientation, edge detection, the polarity and the strength of the blob in blob detection.

## EXISTING SYSTEM

In this work, the detailed study on brain tumor extraction and detection has been reviewed and analyzed. Some paper [2],[3],[4] based on different technologies has been discussed about extraction of brain tumor. Some paper [1],[5],[6],[7],[8],[9],[10] based on different technologies has been discussed about detection of brain tumor. The detailed description of various techniques was given below:

In [1] Elisee Ilunga-Mbuyamba, Juan Gabriel Avina-Cervantes, Dirk Lindner, Jesus Guerrero-Turrubiates, Claire Chalopin was proposed "Automatic Brain Tumor Tissue Detection based on Hierarchical Centroid Shape Descriptor in T1-weighted MR images" presents a novel scheme which uses a two method, the k-means method and the Hierarchical Centroid Shape Descriptor (HCSD). The clustering stage is applied to discriminate structures based on pixel intensity while the HCSD allow to select only having a specific shape. The automatic tumor detection can be achieved by using some features like texture, shape, intensity and symmetry.

Hierarchical Centroid Shape Descriptor is a binary shape descriptor with the centroid coordinates extracted from a binary image and it is based on the k-tree technique decomposition. The two methods for brain tumor tissue detection were introduced.

This method combines the k-means clustering algorithm followed by the use of a shape descriptor based on features called Hierarchical centroids. On the first step, the k-means algorithm group image pixels in k clusters, then the image is binarized by using a threshold value equal to k. The tumor structures are found in remained binary elements but they are often surrounded by healthy structures. The second step method is used to discard other tissues in order to detect only those corresponding to the tumor.

The Fuzzy C means algorithm is used for classifying brain tumor images and it works well for tumor detection. Like the mean-shift this method has a high computation complexity; however it is suitable when the number of clusters are unknown a priori. In the brain symmetry was used for tumor segmentation and detection by using the texture and intensity. Another automatic method for tumor detection based on the brain symmetry is introduced. The application of this kind of feature is limited on axial and coronal planes because there is no symmetric structures in the sagittal plane.

As applied in various works, the thresholding technique and the morphological operations such as erosion, dilation, closing and opening are used in a pre-processing step for skull removing. To overcome the ineffectiveness of algorithms to automatically detect and localize tumor in human brain, to use the k-means clustering method followed by the selection of the shape that can better describe the tumor.

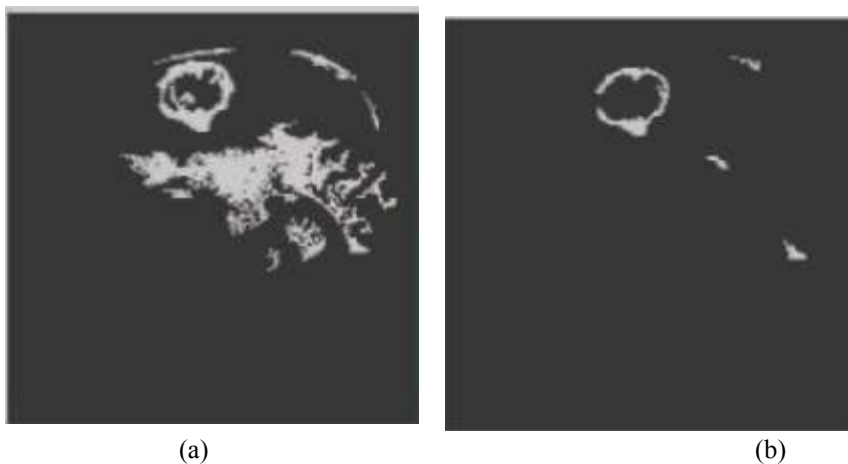


Fig: 2.1. (a) Otsu method and (b) k-means method

k-means is one of the most popular clustering algorithm. HCSD & K-Means methods is robust in detecting brain tumor tissue even this kind of data. Otsu and K-Means results are very similar because the segmented structures are often identical, but sometime they have different shape.

In [2] Arashdeep Kaur was proposed "An Automatic Brain Tumor Extraction System using Different Segmentation Methods" presents some algorithms for brain tumor extraction namely Otsu, K-means, Fuzzy-c-Means and thresholding. In these four techniques have been used to identify region of interest i.e. tumor. Morphological Analysis forms an integral part of Image processing. Binary morphology can be used to extract objects from a binary image.

**a) K-means**

K-means is a clustering technique which aims to partition a set of observations so as to minimize the within cluster sum of squares. The evaluating function for an image a (m, n) is given as:

$$C(i) = \text{Arg min} \sum_{x,y} |mxy^2 - nxy^2| x^2 \tag{1}$$

Where i is the number of clusters.

**b) Otsu's Method**

Otsu's Method divides the image into two classes of regions namely foreground and background. The background and foreground regions are selected using the following weighted class variance:

$$\sigma^2 = W_1\sigma_1^2 + W_2\sigma_2^2 \tag{2}$$

Where W1 and W2 are class variance for background and foreground region respectively.

**c) Fuzzy-c-means**

Fuzzy-c-means is also a clustering technique based on fuzzy logic to segment the image. Fuzzy clustering algorithms determine an optimal partition of a given data set as follows:

$$A = \{x_j \mid j=1 \dots n\} \tag{3}$$

Where, A is the partitioned image into c clusters by minimizing the following objective function:

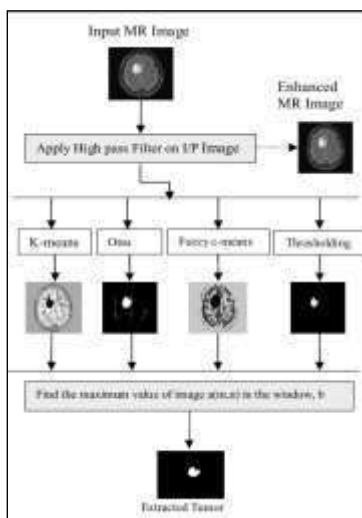
$$f = \sum_{m=1}^c \sum_{j=1}^n u_{mj} d_{mj} \tag{4}$$

Where, m and n are the rows and columns of the image.

**d) Thresholding**

It is a process of creating a black-and white image of a grayscale image consisting of setting exactly those pixels to white whose value is given threshold, setting the other pixels to black.

$$a(m,n) = f(x) = \begin{cases} 1, & \text{object if } a(m,n) \geq 0 \\ 0, & \text{background otherwise} \end{cases} \tag{5}$$

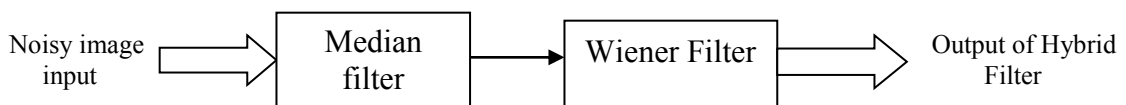




**Fig: 2.2.Flowchart for brain tumor extraction**

This paper presents accuracy and computation time of tumor extracted using four different segmentation techniques. Computational time is defined as the time taken for the algorithm to execute and extract the tumor.

In [3] Rana Banik, Md. Rabiul Hasan, Md. Saif Iftekhar was proposed “**Automatic Detection, Extraction and Mapping of Brain Tumor from MRI Scanned Images using Frequency Emphasis Homomorphic and Cascaded Hybrid Filtering Techniques**”, presents these techniques to detect brain tumor from MRI scanned images. Brain tumor regions are detected, extracted and mapped by Frequency Emphasis Homomorphic and Cascaded Hybrid Filtering Techniques. The hybrid filter is a combination of wiener filter and median filter. The salt and pepper noise, Gaussian noise, impulse noise, Rayleigh noise are the type of noises are produced during transmission. The methods now present to detect brain tumors are generally three types: atlas-based methods , feature-based methods and symmentary-property-based methods. High pass filter also reduces frequency domain Gaussian noises.


**Fig: 3. Block diagram of hybrid filter**

This method is used to difference of the intensity perfectly by using Frequency Emphasis in Homomorphic Filtering. The final results are mapped with edge detected image of the real time patient by mathematical logic operations.

In [4] Jitendra Singh Sengar and Priyanka Chanderiya was proposed “ **Review: A Survey on Brain Tumor Extraction from MRI**” presents the brain image are segmented & artificially colored to represent original data through modified fuzzy c-means algorithm, SCNNA(supervised computational neural network approach), KBT(knowledge based technique), FBTD(fractal-based brain tumor detection), ASBT(automatic segmentation of brain tumor). Segmentation applied for Gray Matter (GM), White Matter (WM), and Cerebra Spinal Fluid (CSF) and tumor region extraction of brain images.

In [5] Ishmam Zabir, Sudip Paul, Abu Rayhan, Tanmoy Sarker, Shaikh Anowarul Fattah, and Celia Shahnaz was proposed ” **Automatic Brain Tumor Detection and Segmentation from Multi-Modal MRI Images Based on Region Growing and Level Set Evolution**” They presents a type of tumor named as GLIOMAS. It has high frequent and survival rate of patients from both high grade & low grade glioma is approximately less than two or three years. The drawback of using only region-based models is that they fail to detect exact boundary of tumor because of tumor intensity variations from center to boundary tumor regions. It provide outcome of the region growing approach automatically equally the initial contour and the final decision is made iteratively. The contour based iterative Distance regularized level set evolution (DRLSE) method is thus aided by region growing approach and improves the segmentation performance than both the region growing and DRLS methods. The accuracy value calculated in detecting tumor containing slices & no tumor from the normal slices are also calculated taking 26 tumor containing slices and 14 normal slices. Level set function is defined as a partial differential equation-

$$\frac{\partial \Phi}{\partial t} = F |\nabla \Phi| \quad (6)$$

where F is the speed function that controls the motion of the contour. However, in DRLSE level set function is formulated in terms of distance regularized term & an external energy term.

In [6] William Thomas H M,S C Prasanna Kumar was proposed ” **Detection of a brain tumor using segmentation and morphological operators from MRI scan with FPGA**” work uses K-Means clustering where the detected tumor shows some abnormality is rectified by the morphological operators are used with basic image processing techniques to meet the goal of separating the tumor cells from the normal cells. The Threshold and Watershed segmentation is very simple and popular but using morphological operators on applying to the output image of other two provided a better detection of tumor. It has used to detect the tumor using Image Segmentation

approach for the detection of damaged cells of brain but detect particularly the abnormal cells of human brain is not an abstract rather it is possible by using combination of thresholding and watershed segmentation along with applying the morphological operators we get the output of the MRI image which is possible for doctors to detect accurately where the tumor is located.

In [7] Rasel Ahmmed, Md. Faisal Hossain was proposed” **Tumor Detection in Brain MRI Image Using Template based K-means and Fuzzy C-means Clustering Algorithm**”, presents a combination of TKFCM(template based K-means and modified Fuzzy C-means) clustering algorithm that reduces operators and equipment error. In TKFCM, the small deviation of gray level intensity of normal and abnormal tissue is detected and the performances of TKFCM method is analyzed over neural network provide a better regression and least error.

The performance parameters show comparable results which are effective in detecting tumor in multiple intensity based on brain MRI image. The main drawback of thresholding is cannot be applicable for multiple channel images. In addition, it does not provide spatial characteristics, which causes it to be sensitive to noise as well as inhomogeneity intensity. On the other hand, foundation of restricting threshold is also used methods such as classifier, ANN, clustering etc. Based on some predefined criteria i.e. intensity information and/or edges, the connected region of an image is extracted in region growing.

In [8] Praveen G.B, Anita Agrawal was proposed” **Hybrid Approach for Brain Tumor Detection and Classification in Magnetic Resonance Images**” presents the brain tumor are detected and classified in magnetic resonance images has been proposed using hybrid approach. Segmentation methods includes a various approaches based on classification using extracted features, level set methods, Markov random field (MRF) methods, fuzzy c-means (FCM), k-nearest neighbor (KNN) and region growing methods. Level Sets method requires initial curves identification. Advantage and drawbacks of hybrid approach is a combination of region based and texture based methods for brain tumor detection and classification. Fast bounding box algorithm is used as region based method for tumor segmentation. This methodology is more efficient than the existing methods and segmentation accuracy is 96.63%.

In [9] Anupurba Nandi was proposed” **Detection of human brain tumor using MRI image segmentation and morphological operators**” presents clustering is suitable for biomedical image segmentation uses unsupervised learning. This paper work uses K-Means clustering used for detected tumor shows some abnormality is rectified by the use of morphological operators with image processing techniques to meet the goal of separating the tumor cells from the normal cells. In [8] presents preprocess the two-dimensional magnetic resonance images of brain and subsequently detect the tumor using Image Segmentation approach.

In [10] William Thomas H M ,S C Prasanna Kumar was proposed ” **A review of segmentation and edge detection methods for real time image processing used to detect brain tumor**” presents surveyed based approaches to extract the tumor from set of brain images. Some methods are thresholding, clustering, level set method, region growing, morphological based segmentation, graph based detection, histogram thresholding. The various types of edge detection is Robert edge detection, Prewitt edge detection, Sobel edge detection.

## COMPARATIVE ANALYSIS OF EXISTING SYSTEM

This section describes a comparison of the brain tumor extraction and detection of MRI and processes using rate of parameters for several techniques. To measure the accuracy, sensitivity, and specificity show that below table.

**Table: I. Comparison of tumor detection and extraction techniques**

Tumor Extraction				
S.no	Algorithms	Sensitivity (%)	Specificity (%)	Accuracy (%)
1.	Thresholding	84	80	83.3
2.	Region Growing	88.46	75	86.7
3.	Second order + ANN	91.42	90.1	92.22
Tumor Detection				
4.	Texture Combined + ANN	95.4	96.1	97.22
5.	Fuzzy C-Mean	96	93.3	86.6

6.	K-Mean	80	93.12	83.3
7.	TKFCM	96.67	100	97.1

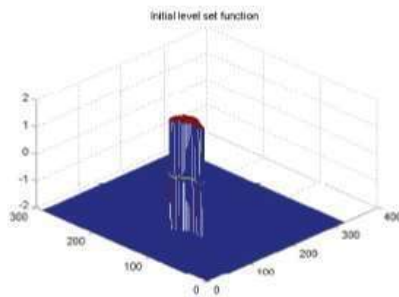
## PROBLEM DEFINITION

We have described that survey of various techniques for extract and detect the brain tumor. Tumor is one of the problem for uncontrolled patient. The doctors rectify the tumor of brain. In various techniques, doesn't have accurate values. A tumor does not mean cancer – tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). If tumor is cancer, possible treatments are chemotherapy, radiation, surgery, Targeted cancer therapy.

Brain tumour was detected and clear that tumor of patient. Then by using some algorithms in mathematical operations. Thus this solution is one of the problem of brain tumor, Region Growing method is best to extract the particular boundary regions of tumor and Fuzzy-c-means is also a clustering technique based on fuzzy logic to detect the tumor in brain image. Fuzzy C-means method is used to detect the tumor with accuracy.

## PROPOSED SYSTEM

Thus the paper proposed that region growing methods to extract the tumor of brain. The initial contour of the seed point is to be selected to the iterative level set method. Thus the need of selecting the initial region of interest is removed. Region grow areas are automatically detected as the initial Level Set Functions. The proposed method is capable of improving the accuracy of overall extraction and detection performance of tumor for different parameters publicly available database. T2 weighted and flair modalities of MRI images are used in parallel to find the tumor Extraction and Detection.



**Fig: 3.1.Initial level set function obtained after applying**

region growing approach in T2 image.

## MODULES

Our project, brain tumor has been extracted and detected is made up of three modules. They are:

- Region of Interest (ROI)
- Feature Extraction
- Fuzzy C- means algorithm.

## MODULE DESCRIPTION

### Region of Interest (ROI)

The segmented area obtained from the region-growing approach is automatically selected as the initial contour to the iterative distance regularized level set evolution method thus removing the need of selecting the initial region of interest by the user. As diagnosis tumor is a complicated and sensitive task ; accuracy and reliability are always assigned much importance. Hence, an elaborated methodology that highlights new views for extracting more robust image extraction technique is much sought.

### Feature Extraction

Feature Extraction is extract the boundary of regions. It is used to create new features as a function of existing features. In this proposed method, we use region growing method for extracting the brain tumor. The initial seed point starts to grow by searching for neighborhood pixels or regions with similar properties that create connected regions. In this method, a tolerance level of 12 is selected. Each pixel around the boundary of growing region having less than 12 pixel intensity value is included in that region. However, this region may not be the actual segmentation of tumor, because of irregular pixel intensity in tumor region. In this method, the performance of the conventional region growing approach is improved by the re-estimation of proper seed point. It is mainly used in this report, for extraction of MRI brain tumor image.

### Fuzzy C- means algorithm

The Fuzzy C means algorithm is used for clustering brain tumor images and it works well for tumor detection. Fuzzy C means uses fuzzy logic values to each pixel. Tumor is detected using fuzzy c-means has the highest accuracy is 90.5% amongst all the other segmentation. The accuracy achieved is better than the methods or algorithms already existing in the literature. Also, when proposed algorithm is compared with the manual tumor extraction, provides more accuracy.

Region based methods mainly rely on the assumption that the neighboring pixels within one region have similar value. It has to extracted and fuzzy C means algorithm has to detected. The radiologist found that tumor with accuracy and rectify to the controllable patient.

### CONCLUSION

This paper presents survey of various extraction and detection techniques for real time MRI brain image has been accomplished. As the diagnosis tumor is a complicated, sensitive task, accuracy and reliability are always assigned much importance. It is used to doctor rectify the brain tumor easily. This paper presents survey has been made on the different types of methods and algorithms used for tumor extraction and detection. It was also found out that region growing for extraction and Fuzzy C-Means for detection gives better performance. It gives more accuracy value compared to the various techniques used to extract and detect the tumor in human brain using MRI images.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

### REFERENCES

- [1] Amir A, Marwa I, Ahmed S, Fahmi K, Matthew N, et al. [2015] Infant Brain Extraction in T1-weighted MR Images using BET and Refinement using LCDG and MGRF Models, IEEE Journal of Biomedical and Health Informatics.
- [2] Elisee Ilunga M, Dirk L, Jesus Guerrero-Turrubiates, Claire C, et al. [2016] Automatic Brain Tumor Tissue Detection based on Hierarchical Centroid Shape Descriptor in T1-weighted MR images IEEE International Conference On Electronics, Communications And Computers (CONIELECOMP). 62-67.
- [3] Arashdeep K. [2016] An Automatic Brain Tumor Extraction System using Different Segmentation Methods, Second International Conference on Computational Intelligence & Communication Technology. 187-191.
- [4] Rana B, Rabiul Hasan MD, Saif Iftekhar MD. [2015] Automatic Detection, Extraction and Mapping of Brain Tumor from MRI Scanned Images using Frequency Emphasis Homomorphic and Cascaded Hybrid Filtering Techniques, 2nd International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT). 21-23.
- [5] Jitendra Singh S, Priyanka C. [2016] Review: A Survey on Brain Tumor Extraction from MRI, International Journal of Signal Processing, Image Processing and Pattern Recognition. 9(6):171-176.
- [6] Ishmam Z, Sudip P, Abu Rayhan MD, Tanmoy S. [2015] Automatic Brain Tumor Detection and Segmentation from Multi-Modal MRI Images Based on Region Growing and Level Set Evolution, IEEE International WIE Conference

- on Electrical and Computer Engineering (WIECON-ECE). 503-506, 19-20.
- [7] Anupurba N. [2015] Detection of human brain tumour using MRI image segmentation and morphological operators, IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS). 55-60.
- [8] Rasel A, Foisal Hossain MD. [2016] Tumor Detection in Brain MRI Image Using Template based K-means and Fuzzy C-means Clustering Algorithm. International Conference on Computer Communication and Informatics (ICCCI -2016). 07-09.
- [9] Praveen GB, Anita A. [2015] Hybrid Approach for Brain Tumor Detection and Classification in Magnetic Resonance Images, International Conference on Communication, Control and Intelligent Systems (CCIS).
- [10] William Thomas HM, Prasanna Kumar SC. [2015] Detection of human brain tumor using MRI image segmentation and morphological operators International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). 728-731.
- [11] William Thomas HM, Prasanna Kumar SC. [2015] A review of segmentation and edge detection methods for real time image processing used to detect brain tumour IEEE International Conference on Computational Intelligence and Computing Research.
- [12] Suganya A, Mohana Priya N, Kalaavathi B. [2015] Lung Nodule classification techniques for Low Dose Computed Tomography(LDCT) Scan Images as Survey, International Journal of Computer Applications(09752-8887). (131):14
- [13] Devika KK, Mohanapriya N, Kalaavathi B. Survey on Medical Image Segmentation using Region based Watershed Algorithm proceeding of International Journal of Future Innovative Science and Engineering Research. 2(1):22-31,2454-1966.
- [14] Mohanapriya N, Kalaavathi B. [2014] Image Enhancement Using Multilevel Contrast Stretching and Noise Smoothing Technique for CT Images, International Journal of Scientific & Engineering Research. 5(5):2229-5518.
- [15] Mohanapriya N, Kalaavathi B. [2015] Medical Image Enhancement Using Adaptive Wiener Filter And Contrast Stretching Techniques, Australian Journal of Basic and Applied Sciences. 9(10):156-160.

## IMPROVING THE ACCURACY OF IR TECHNIQUES USING TRACEABILITY

MB. Suseela, PP. Devi

Department Of Computer Science and Engineering, Vel Tech Multi tech Dr. RR and Dr.SR Engineering College, INDIA

### ABSTRACT

Traceability links between the requirements of a system and its source code are helpful in reducing system comprehension effort. During software maintenance and evolution, requirement traceability links become obsolete because developers do not/cannot devote effort to updating them. Traceability is a sub discipline of requirements management within software development and systems engineering. Consequently, the literature has proposed methods, techniques, and tools to recover these traceability links mechanically or semi automatically. In proposed system the information retrieval technique can automatically recover the traceability links between specified requirements and source code. However, IR techniques lack accuracy. In this paper, we show that mining software repositories and combining mined results with IR techniques can improve the accuracy of IR techniques and we propose a dynamic specification mining technique which used to find the traceability links and also will detect the bugs in the source code.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Traceability links, Requirement traceability, Information retrieval, dynamic specification mining

\*Corresponding author: Email: [suseelambveltech@gmail.com](mailto:suseelambveltech@gmail.com)

### INTRODUCTION

Traceability is the ability to verify the old process, location, or application of an item by means of documented recorded identification. [1] It is one of important criteria in software engineering quality factors in which it describes and follows the life of a software in term of modelling the relations between software artifacts explicitly. Requirements traceability text file the life of a requirement, tracks every change and links its relationships with other items within a project [2]. Traditionally, a manual traceability matrix connects high-level requirements with design, test plan and test cases. Traceability links have been used to connect requirements gathered resulting from the initial requirements. However, traceability links can be created between any items, be it an artifact, folder or module, in any version [2]. Traceability links can be used for various additional features in a project's development. Trustrace consists of three parts [5]:

1. Histrace stores all the required links between requirements and software repositories it also mines software repositories to create links.
  2. Trumo compares the duality of the recovered links provided by the experts and also combines the requirements traceability links from that trumo will ranks/discards.
  3. DynWing analyses each expert's similarity value for each links and assign values also for each links
- We thus show [5] that our trustbased approach indeed improves precision and recall and also that CVS/SVN change logs are useful in the traceability recovery process.

An dynamic specification mining technique is used to recover the requirement traceability and also to detect the bugs in the source code. To control the fault in source code those fault report are been thrown to the developer mail id for future references.

The remaining paper has been followed as. Relatedfocus on the literature review, focus on implementation, focus on the modules, focus on the result and finally focus on conclusion.

## LITERATURE REVIEW

### Trustrace: Trust based traceability

Trustrace uses software repositories e.g. bug tracking system, as experts to trust more or less some base line links recovered by an IR technique.

#### Definition

In trustrace, we symbolize a traceability links as a triple {source document, target document, and similarity}. In [\[Figure- 1\]](#) let  $R(r_1, \dots, r_n)$  be set of requirements or high level document,  $C(c_1, \dots, c_m)$  be a set of implementing classes. Let  $T(t_1, \dots, t_p)$  be collections of set where each set  $T_i$  is a set of homogeneous piece of information. [\[4\],\[5\]](#)

Then let us assume that for each  $T_i \in T$  possible to define function  $\delta_{T_i}$  consequently, we define R2C as the set of traceability links rebound between R and C and, for each set  $T_i \in T$ , a set  $R2CT_i$  generated by each piece of information  $t_k$ : [\[4\], \[5\]](#)  $R2CT_i(r_j, t_k) = \{(r_j, c_{ss}, \sigma_{j,k}) | C_s \in \delta_{T_i}(t_k) \& t_k \in T_i\}$

#### Model

In this trust race use the following equation[\[4\]](#). Trustrace uses the set of candidate links  $l_r(r_j, c_s, \sigma_{j,s})$  with  $j \in [1, \dots, N]$  and  $s \in [1, \dots, I]$  also uses the sets of candidate links  $l_{rt}=(r_j, t_k, \sigma'_{j,k})$   $j \in [1, \dots, N]$  and  $k \in [1, \dots, N_i]$  generated from set of some other pieces of information  $T_i$  and for requirement  $r_j \in R$ . Indeed, for each set  $T_i \in T$ , Trustrace builds a trustable links  $T_{ri}$  as follows  $T_{ri} = \{(r_j, c_s, \sigma_{j,s}) | \exists t_k \in T_i : (r_j, c_s) \in \alpha(R2CT_{i,r_j, t_k}) \& (r_j, c_s) \in \alpha(R2C)\}$

Finally Trustrace merges the trust levels of each  $T_{ri}$  into global level of trustworthiness:[\[5\]](#)

### Histrace

Histrace produce links between the set of requirements, R and the source code C, using the software repositories  $T_i$ . Histrace consider the requirements, textual descriptions, CVS/SVN commit messages, bug reports, and classes as separate documents. It also uses the information from two experts. A bug report can also be produced to improve the accuracy of the IR techniques. Histrace mines the software repositories.

#### Document Pre processing

Depending on the input information source we perform specific pre-processing steps to remove irrelevant details, split identifiers, and, finally, normalize the text using stop words removal and stemming.

#### Requirements and source code

Histrace first processes source code files to extract all the identifiers and comments in each class [\[4\] \[5\]](#).

Histrace then performs the Following steps to normalize source code documents and requirements: (i) compute all upper case letters into lower case and remove punctuation: (ii) remove all stop words.

#### CVS/SVN commits

To build  $T_o$ , histrace extracts CVS/SVN commits and discards those that (i) are tagged as “delete” (ii) does not concern source code (iii) have messages of length shorter or equal to two words [\[4\] \[5\]](#).

#### Bug reports

To create histrace bug, histrace extracts all the bug reports from a bug tracking system. Usually, bug reports do not contain declared information about the source code files that developers updated to fix a bug [\[5\]](#). Histrace use regular expression, i.e, a simple text machine approach but with reasonable results to link CVS/SVN commit messages to bug reports.

### Trumo

Trumo assumes that different experts know useful information to discard or rerank the traceability links between two documents, e.g., requirements and source code classes. By the definitions, section 2.1.1  $r_j$  is a requirement with  $r_j \in R$ , and class  $cs \in \delta_{T_i}(t_k)$  [5], [6]

### Dynwing

Dynamic weighting technique automatically decides the weights. Choosing a right weight per link is a problem that we articulate as a more problem. Essentially we have different experts, i.e, CVS/SVN commits, bug reports, and others, to trust a link. By maximizing the similarity value  $\psi_{r_j,cs}(Tr)$ . Dynwing automatically identifies the experts that are most trust worthy and those that are less trustworthy. [5]

$$\lambda_1(r_j,cs) \dots \max \dots \lambda_{p+1}(r_j,cs) \{ \psi_{r_j,cs}(Tr) \}$$

With the following constraints: [5]

$$0 \leq \lambda_i(r_j,cs) \leq 1, i=1, \dots, P+1 \quad \lambda_1(r_j,cs) + \lambda_2(r_j,cs) + \dots + \lambda_{p+1}(r_j,cs) = \lambda_{k1}(r_j,cs) \geq \lambda_{k2}(r_j,cs) \geq \dots \geq \lambda_{kp+1}(r_j,cs).$$

Therefore, developer may further constraint [5] by imposing

$$\lambda_{commits}(r_j,cs) \geq \lambda_{bugs}(r_j,cs) > 0$$

To mechanically decide the weights  $\lambda_i(r_j,cs)$  for each expert.

### Topic modeling

Topic modeling is a widely-used machine learning technique for automatically [3] inferring semantic topics from a text corpus.

### Latent dirichlet allocation

In LSI (latent semantic indexing), each document in the corpus is represented as a word count vector of length  $W$ , where  $W$  is the number of words in the principal vocabulary. When the vectors of all  $D$  documents are placed side by side, one obtain a  $W \times D$  matrix of counts.

It is possible to learn an LDA model in near real time on a moderately-sized set of documents. The algorithm consists of iteratively performing variational updates in a systematic scan over tokens [3],

$$q(z_{nd=t}) \propto N_{wd=nd} + \beta \cdot N_{t=nd} + \gamma_{nd} \quad [3]$$

$$W \beta(N_{td})$$

$N_{wd=nd}$  and  $N_{td=nd}$  are expected counts derived from  $q(z)$ .

### Current support for requirements traceability

It has been noted that most tools do not cover RT. [6] They differ mainly in cosmetics, and in time, effort, and manual intervention they require to achieve RT.

### Basic techniques

Various techniques used for providing RT, are as follows: cross referencing schemes, key phrase dependencies, templates, RT matrices, matrix sequences, hypertext, consolidation documents, assumption-based truth maintenance networks and constraint networks.

Moreover, some form of RT can result from using certain languages, models, and methods for development.

### Automated support

More commercial tools and research products support RT, we high spot some representative examples: they are as follows [6]

**General- Purpose tools:** These incorporate; hypertext editors, word processors, spread sheets, database systems and even more. They can be hand-configured to allow previously manual and paper-based RT tasks to be carried out on-line.



**Special-Purpose Tools:** foundation dedicated activities related to RE and some achieve restricted RT.

**Workbenches:** It limits a collection of the above to support coherent sets of activities. Less restricted RT can be achieved.

**Environment:** RT can be acclaimed by the basis of integration; a common language, a common structure, a common method or a specialized RT tool or repository structure.

## IMPLEMENTATION

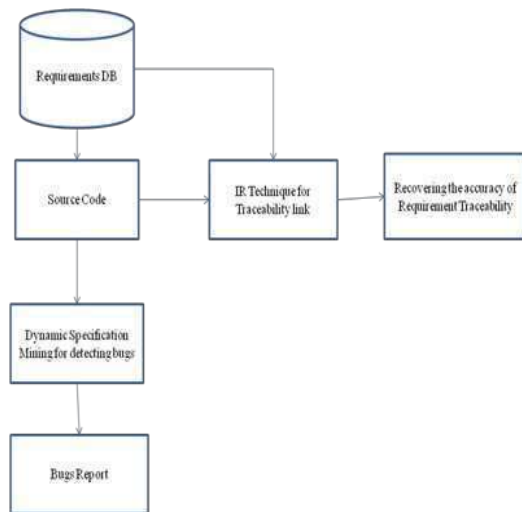


Fig: 1. Work flow diagram

Requirements are stored in a database and then a link is created between the source code and the specified requirements. Then a bugs can be detected from the source code and that bugs are been thrown to the developer mail id.

### Requirement traceability

The requirements and bug report from the user is given a link and that link is given to the source code if any requirement is missing by doing the traceability method the missed requirements will be specified.

### Bug detection

The fault that are been present in the code will be detected and the misspell synax word can be found in this step.

### Recovery process

In this process the words that are misspelled will be recovered automatically by clicking the recover process.

### Accuracy of traceability links

The accuracy of the traceability links between requirements and source code recovered by an IR technique are improved by

1. Mining software repositories
2. Trust model
3. Weighting the experts' opinions

### SMTP mail server

Simple mail transfer protocol is used to send the error report to the developer mail id.

### Expected Outcome



Fig. 2. Code Correction Report

The source code fault will be detected in this process and that report will be send to the developer mail id.

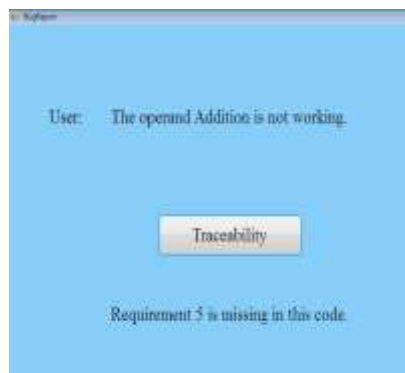


Fig. 3. Requirement Traceability

The requirement missing will be specified in this screen shot.

## CONCLUSION

Requirement tracing is inevitable. The concepts of requirements tracing is quite simple: to follow relationships or links. Requirement traceability concern to the ability to describe and follow the life of requirements in both forwards and backwards direction. In this paper, detailed explanation given about various information retrieval techniques. First we proposed Trust race, inspired by web trust models to improve precisions and recall of traceability links: trust race use any traceability recovery approach as a basis on which it applies various expert's opinion. Second we proposed Histrace an expert supporting the identification of traceability links between requirements and source code. Third we proposed Trumo, inspired by web trust models improves the precision and recall values of some baseline traceability links. Fourth we proposed Dynwing combines and assigns weights to Histrace experts' opinion using a dynamic weighting technique. Fifth we proposed Topic modelling is used for automatic machine learning technique.

## FUTURE ENCHANCEMENT

In future work, we plan more experiments with other combinations of IR-based approaches to future improve the precisions and recall values. We will also perform more experiments on heterogeneous software arti facts to measure the usefulness of these other arti facts for a traceability recovery process. In particular, we are currently

building new traceability approaches using BCR techniques (binary class relationships) this will improve the precision and recall values of the traceability links.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

### REFERENCES

- [1] Nuv Adila A, Rodziah A. [2012] Traceability Method for Software Engineering Documentation IJSC international journal of computer science issues. 9(2):2.
- [2] Ale Ksandra K, Kalman G, Andy S, Ralf S. [2011] Traceability Link Evolution with Version Control in: IEvolution are software- und system entwickleny- method und erfahurgen workshop I Am Rahmen der Konferenz SE2011.
- [3] Asuncion H, Asuncion A, Taylor R. [2010] software Traceability with Topic modeling proc.32ndACM/IEEE int'l conf. Software eng. 1:95-104.
- [4] Ali N, Gueheneuc YG, Antoniol G. [2011] Trust-Based Requirements traceability, Proc. 19th IEEE Int'l Conf. Program Comrehension, S.E. Sim and F.Ricca, eds. 111-120.
- [5] Nasir A. [2013] Student Member, IEE, Yann-Gael Gueheneuc, Senior Member, IEEE, and Giuliano Antoniol, Member, IEEE, Tustrace: Mining Software Repositories to Improve the Accuracy of Requirement Traceability Links, of Software Engineering. 39(5).
- [6] Gotel OCZ, Finkelstein CW. [1994] An Analysis of the Requirements traceability Problem, Proc. First Int'l Conf. Requirements Eng. A. Marcus and J.I. Maletic. 94-101.
- [7] [2003] Recovering Documentation-to source-Code Traceability Links Using Latent Semantic Indexing, proc. 25th Int'l Conf. software Eng. 125-135.
- [8] McKnight DH, Choudhury V, Kacmar CK. [2002] The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: A Trust Building Model, The J. Strategic Information Systems. 11(¼):297-323.
- [9] Palmer JW, Bailey JP, Faraj S. [2000] The Role of Intermediaries in the Development of Trust on the WWW: The Use and Prominence of Trusted Third Parties and Privacy Statements, J. Computer-Mediated Comm. 5(3).

HYBRID ARTIFICIAL IMMUNE SYSTEMS FOR CLASSIFICATION ON MRI  
BRAIN IMAGES

S. Valarmathy, N. Suthanthira Vanitha

<sup>1</sup>Associate Professor / ECE, V.M.K.V Engineering College<sup>2</sup>Professor / Head / EEE, Knowledge institute of technology, Salem, TN, INDIA

## ABSTRACT

*Aim: Dementia is a common neurodegenerative disease which propagates itself through minute symptoms and develops into a form of severe brain damage. Magnetic Resonance Imaging (MRIs) are currently the best medical imaging tools that permit cross-sectional views of the human body with excellent tissue contrasts. MRIs play a significant part in the appraisal of pathological features of brains and are effective in diagnosis of dementia. In this work, the brain MRI are classified as dementia and non-dementia. Features are extracted using Discrete Wavelet Transform (DWT) and feature selection is via proposed hybrid Artificial Immune System (AIS). Genetic Algorithms (GAs) are combined with AIS to optimize the feature subset selection. Naïve Bayes, C4.5 and K nearest neighbour then classifies the selected features as dementia or non-dementia. Experimental results show that the proposed method is effective in improving the efficiency of the classifiers*

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Magnetic Resonance Imaging (MRI), Dementia, Feature Selection, Artificial Immune System (AIS), Naïve Bayes, KNN, C4.5, Radial basis function (RBF).

\*Corresponding author: Email: [valarmathysr@gmail.com](mailto:valarmathysr@gmail.com);

## INTRODUCTION

Alzheimer's Disease (AD), an old age-related illness, characterized by progressively declining cognitive facilities. It is widely accepted that this is a disease that affects only the elderly, i.e. people who are sixty-five, and is a predictable onset of dementia. However, a less-known fact about AD's and more importantly about Dementia is that this disease is not restricted to the elderly, and there are several variations of the illness [1]. Not all the variants of Dementia are age based, and vascular or multi-infarcted illnesses are an example of irreversible dementia. Apart from this, Alzheimer's is a kind of disease that progressively declined mental capacities, like cognition, memory and learning capabilities. AD is accepted as a predominant old age illness and cannot be associated with vascular dementia, as the latter could be an onset of motor disabilities and cognitive malfunctions. However, the presence of AD is quite subtle as the disease does not immediately present itself. It propagates itself through minute symptoms which can develop into a form of severe brain damage. It is, therefore, essential to recognize the onset of AD before it gradually takes over as a serious illness. Some common methods used for early diagnosis are Magnetic Resonance Imagers (MRI) and Computed tomographers (CT). These recommended techniques follow prescribed guidelines to capture the illness at any stage of its growth. It must be noted that only a rare percentage of dementia patients can be completely treated, and individuals with Dementia Lewy bodies (DLB), Front Temporal Dementia (FTD) and Vascular Dementia (VaD) are harder to treat [2].

The brain is the major organ which is affected by AD and it is through the loss of healthy cognitive functioning that Alzheimers can eventually result in loss of tissues and large-scale cerebral nerve cell damage. The operation of the brain, with age, deteriorates to the point where the slow loss of cognitive and intellectual functions leads to dementia. Here, it is observed that the brain can lose five to ten percent of its total volume during the transition to old age. Some of the leading causes of dementia, apart from Alzheimers, are caused due to organ failures, drug toxicity, and other similar reasons [3].

Recently, a new field of research is slowly gaining popularity, due to its likeness to the study of vertebrate immune systems. The study of this immune system could propagate a new terrain of study based on a bundant theories to generate resources for computer-based solutions. This growing field has is called the Artificial Immune System (AIS), and it utilizes notions from immunology to help build appropriate models which can perform tasks in engineering applications. Although this field might lack detailed resources for remote sensing, there is scope for

developing this field within the paradigm of AIS study [4].

CT generated information provides relevant information for positron emission tomography (PET) and holistic care of the patients. There are numerous protocols undertaken by using PET/CT scans, commonly used for low -dosage attenuation corrections and anatomic localizations, in the situation where patients have an existent MR of their skull. If in the case where the MR IS contraindicated, a portion of the CT exam is an optimized performance with standard imaging parameters. It is important to review carefully CT images to minimize radiation dose to the lens [5].

An inherent initial stage in mining of huge databases is to calculate feature selections for classification of any pattern problem. Computational burdens are vastly decreased, by the reduction of data dimensionality. Feature selections algorithms constructed based on the following categories: exponential, randomized, and sequential, and the task of these algorithms is to search for best feature subsets which will reduce possible feature space with fewer losses in classification accuracy. The ideal goal here is to lower the number of features which are being analyzed, without the sacrificing of class discrimination, and eventually lead to classification accuracy [6].

This paper suggests a machine learning approach to help classify dementia and MRI medical images from each other, based on Discrete Wavelet Transform (DWT) to analyze features. It is proposed to use Artificial Immune System (AIS) as a feature selection methodology. The chosen features are categorized using Naïve Bayes, CART, C4.5 and K-Nearest Neighbour techniques. To evaluate the techniques, MRI image samples are acquired from Open Access Series of Imaging Studies (OASIS) dataset. The maining paper is arranged thus: Section 2 reviews the essential literature of the above problem. Section 3 & 4 provides a detailed methodology, and the experimental results, which is concluded in Section 5.

## LITERATURE SURVEY

One of the methods to segment MRI brain images is presented by Adhikari et al., [7] who considers utilizing the intensity of non-uniformity (INU) and its spatial knowledge, with the help of a fuzzy C-mean clustering algorithm. The non-uniformity of brain images procured by MRI Scanners are repaired by Gaussian surface fusion cells. A single Gaussian surface calculated uniformly over various corresponding surfaces by investigating its nucleus in the middle of a mass of other similar homogeneous region. Another aspect is the intensity of the inhomogeneity (IIH) where a repaired image is divided based on FCM algorithm probabilities, considering the spatial features of image pixels. The above method has proved to perform well, after utilizing 3D synthetic phantoms as well as actual patient MRIs of brains to conduct the above experiments.

Al-Badarnah [8] et al. proposed an automated classification model, to be utilized for MRI image tumor classifications. The results of this classification depicted how NNs as well as KNN models affect tumor classifications. The achievements of the experiments showed full-fledged accuracies in ranking, using the KNN and 98.92% through NNs. Similarly, Jeena and Kumar [9] illustrated a comparative analysis for diagnosing strokes on CT and MRI images. Their proposed study proposed how to use digital images as a processing tools for identifying hemorrhaging in the cranium. They were able to perform these segmentation tasks using Gabor filtering as seeded region growth algorithms. Finally, the end product of this method was depicted using MRI brain images and CT scans, showing various levels of infarcts.

Extensive efforts to study the above experiment was also carried out by Hamdaoui et al., [10] which added a specific controlled unit that routine every task under various blocks of architecture. Hence, the newly obtained MRI images adhere to the synchronous art of image segmentation under the Particle Swarm Optimization method (PSO) which allows the researchers to cut short during the execution period, thus narrowly reducing the procedural search threshold to an optimal. Therefore, the performance of synchronous hardware infrastructure is judged and verified based on a collection of medical MRIs. Another researcher Pinto [11] and his fellow colleagues, popularly known for introducing a fuzzy automatic segmentation algorithm, bases his work on a k-fold cross-validation approach under the Random Decision Forest Mechanism. The following features which were extracted, complemented itself with other intensity based appearances and context-based features. The researcher applied morphological filters for dealing with errors of misclassifications during the post-processing phase. Therefore, the above method achieved highly satisfactory results and was able to detect a tumor and segment the possible variations of tumorous tissues in the glioma,

A new approach was proposed by Rajini and Bhavani [12], who automated the diagnosis system based on MRI classification. Their proposed idea aggregates two different stages divided based on feature extraction and classification. The authors had procured the features related to MRI images in the first step and using a discrete wavelet transformation (DWT) they extracted which features of MRIs had become diminished. By adopting this principle component analysis (PCA), the derived features will be conditioned to figure neural networks based a binary classifier. Therefore, this will automatically conclude if the image procured are of a healthy brain or a pathologically suffering lesioned brain. Ragini and Bhavani's research proved to have successful conclusions and can help further the MRI classifications. However, these are not the only successful studies available, as other successful stories are available from researchers like Rehman, Saraswathi, and Jang [13]. A Discrete Binary Particle Swarm Optimization (DBPSO) method for MRI feature selection was popularized by Rehman, who utilized it to classify the functioning of normal and abnormal brains, under the purview of Support Vector Machines and K-Nearest Neighbor experiments. The results of his experiment provided high accuracy possibilities and reduced other features which could hinder with performance.

However, a new technique for the detection of start of AD through MRIs was produced by Saraswathi et al., [14] who structured this classifying model into three groups: normal, extremely mild AD as well as moderate AD. The mechanized algorithm study she utilized was commonly known as the Extreme Learning Machine (ELM), which helped in optimizing performance by modifying the PSO as well as GA. Also for the study, a Voxel-Based Morphometry (VBM) method was carried out to extract features in the MRI developed images, and GA was adopted to diminish highly dimensioned features that needed classification.

Finally, Jang et al. [15] suggested a method to extract the cortex of inter-brain subjects, based on co-segmentation method. This method intends to divide binary images together. Jang emphasized the use of Markov Random Field (MRF) as a structure for creating basic functions and utilizing an optimal graph-cut algorithm for identifying identical voxel pairs, using a transformation matrix computed through the matching of 3D SIFT attributes. However to conduct experiments, the author suggests the use of pre-segmented cortex images and copies of segmented brain images, used as references.

## METHOD

Since it is not an easy task to determine shapes utilizing feature extractions can be quite useful. The process of extracting features that are necessary for diagnosing the pattern of dementia. This is defined as an extraction process, containing specific characteristic attributes which help in the generation of collection of useful descriptors for images. The process is utilized to discover tissue sets which can precisely differentiate between dementia and non-dementia. However, the process of determining significantly extracted features is tricky as dementia cannot be assessed through a single feature. The various techniques used in feature extraction, feature selection, classification and the proposed technique is detailed in this section.

### Dataset

Datasets comprising 436 neurological MRIs were made accessible by Open Access Series of Imaging Studies (OASIS) project. The ages of the subjects are between 18 and 96, with 100 individuals having clinical diagnoses from mild to moderate Alzheimer's disease. All scans comprise of 176x208x176 voxels that were pre-processed to remove the skulls, retaining merely brain matter in the images. All scans are linked to further data regarding the subjects, such as ages, sex, education levels, socio-economic status, intra-cranial volumes as well as normalized brain volumes and two metrics of dementia, which are clinical dementia ratings (CDR) as well as mini-mental state exams (MMSE).

Apart from pre-processing for removal of skulls scans are additionally processed under OASIS through k-means approach for binning pixel intensity as one among four colours that relate to the backgrounds (0), cerebrospinal fluids (1), grey matter (2) as well as white matter (3) [16].

### Discrete Wavelet Transform (DWT)

Wavelet transform permits time-frequency localizations. Wavelets refer to small waves such that wavelet analyses refer to analyses of signals of short duration, finite energy function. It transforms signals which are being

investigated to another abstraction that yields signals in better forms. In a mathematical manner, wavelets are given by:

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right)$$

Wherein  $b$  refers to position variable  $a$  refers to scaling variable for given scaling variable  $a$ , translates wavelets through variation of variable  $b$ . Wavelet transforms are given

$$w(a,b) = \int_t f(t) \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right)$$

Accordingly, for all  $(a, b)$ , wavelet transform coefficients, denoting how much scaled wavelets is identical functions at position,  $t = b/a$ . Scales as well as positions are altered in a smooth manner, and later transforms are known as continuous wavelet transforms. When scales as well as positions are altered in a discrete manner, transforms are known as discrete wavelet transforms [17].

### Feature selection

The removal of unnecessary features via the application of feature selection immediately impacts the speed of the classification process. Typically, a basic feature selection process has four relevant steps followed procedurally, namely:

- 1) Generating subsets
- 2) Evaluating subsets
- 3) Terminating criteria
- 4) Validating results

Apart from these basic steps, the above method can be represented by the following models. The first type of representation is known as the “Filter Model,” characterized by generic selection of data to access which features can provide an optimal subset(s), without applying learning algorithm, based on methods of correlation, entropy, mutual information, etc. The second model illustrated in this study is called the “Wrapper Model,” which adopts a learning algorithm to investigate features, and benefit its performance, based on the given algorithm. The space provided for the feature can be vast and complex, and therefore, it is essential to find the most efficient features through the selection and extraction process [18].

### Proposed hybrid a is feature selection

In the proposed hybrid AIS feature selection, the Genetic Algorithm is incorporated with AIS. The AIS algorithm is summoned to assist GA in influencing the feasibility of units in populations. But rather than merely implanting an alternative GA to the primary search, the study can utilize a basic procedure which can be stimulated through the clonal selection principle. This relationship between AIS and GA is focused on a search outer loop among the current population for optimization constraints and is split into possible antigens and non-feasible antibodies after being investigated for any constraint. If it is found that there is no certainty for feasible individuals, then it is proposed to shift two infeasible individuals to antigen population and thus quantities of duplicates of better infeasible units is adjustable and may be proposed as the potential users.

In AIS a loop situated within alongside antibodies which are cloned and mutated. Following this, the distances are computed between antibodies and antigens and whichever bodies have a higher affinity distance (smaller sum of distances), are chosen to define new antibodies situated near the feasible regions. Thus, the cyclic procedure of the AIS system can be repetitive and some population of antibodies which survive, are passed through a pre-calculated GA system. Finally, this procedure is ended with a selection operation, to apply any recombinations and mutation operators among the parents chosen to reproduce a young population and thus to finish the outer loop of the GA.

Therefore, the GA selection procedure constitutes of binary tournaments in which the chosen candidates and its opponents are selected at random. The ground rules of this tournament are:

- (i) The preference of feasible individuals over infeasible ones;
  - (ii) Healthier candidates will be chosen from a pair of infeasible candidates;
  - (iii) But it should be noted that individuals with lesser constraints are chosen between a pair of infeasible individuals, as this helps in computing the affinity factors from a genotypical distance, after adopting a Hamming Distance standard [23].
- The feature set is encoded in binary form, where

### K-Nearest neighbour classifier

KNNs are non-parametric or simple machine learning methods of classification. KNNs are among the most simple classification methods. Classifications are carried out through determination of k nearest training vectors as per adequate distance measure [19]. Vector X is designated to the class to which most KNNs are a part of. KNN models are grounded in distance functions as well as voting functions in KNNs, the measure utilized is Euclidian distances. KNN classifiers are traditional non-parametric monitored classifiers which are supposed to provide excellent performances for best values of k. Like other monitored learning methods, KNN algorithms comprise training phases as well as testing phases. During the first one, data points are provided in n-dimensional spaces.

Training data points possess labels connected to them which assign their classes. During testing, unlabelled data are specified while algorithms generate lists of k nearest data points to unlabelled points. Algorithms later return classes of most of the list. Needed: distance functions on samples. Models are labelled training data (a1; c1); (aN; cN). Classification of fresh samples occurs thus:

Let (aj1 ; cj1 ) ; : : : ; (ajK; cjK) be K training samples whose features are nearest to a. Label a with class label which occurs mostly amongst cj1..... cjK., may provide greater weights to near neighbours weighted votes for labels.

$$v(c) = \sum_{x=1, c_{ji}=c}^k \frac{1}{d(a_{ji}, a)}$$

Weighted votings are computed with c which have greatest v(c) values.

Algorithm of KNN:

1. Adequate distance measures are determined
2. During training stage: Stashes entire training data sets P as pairs (as per chosen attributes) P = (yi, ci), i=1. Wherein yi refers to training patterns in training datasets, ci refers to respective classes while n refers to quantity of training patterns.
3. During testing stage: Calculates distances between fresh features vectors as well as all stashed attributes, which is the training data.
4. KNN are selected and demanded to vote for classes of fresh samples. Accurate classifications provided in testing phases are utilized for measuring precision of the system. If it is not adequate, k values may be tuned till appropriate levels of precisions are attained [20].

### Naïve bayes classifier

These are grounded in Bayesian theories and are simple yet effective probability classification techniques on the basis of monitored classification methods. For all class values, it predicts if specified sample is part of that class. Attribute items in a class are presumed as independent of others known as class conditional independences. These classifiers require limited quantity of training sets for estimation of variables for classifications. The classifier is denoted by Naïve Bayes (NB)

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Wherein P (A) refers to the marginal probability of A, P (A|B) refers to the conditional probability of A, given B known as posterior probability, P (B|A) refers to the conditional probability of B given A while P(B) is the marginal probability of B that performs as normalizing constant. It gives the method through which conditional probabilities of occurrence A given B may be connected to the conditional probabilities of B given A, mathematically. Probability values of final classes dominate over the rest [21].

### Radial basis function (RBF) networks



RBF networks perform identical function mapping with multi-layer NNs, though their structures as well as functions are distinct. RBFs are local networks which are trained in a monitored fashion. RBF carries out local mappings which imply that merely inputs close to receptive fields yield activation. Input layers of networks are sets of  $n$  components that accept components of  $n$ -dimensional input features vectors.  $n$  components of input vector  $x$  are inputs to hidden functions, outputs of hidden functions that are multiplied by weighting factors  $w(i, j)$ , are inputs to output layers of networks  $y(x)$ . For all RBF units  $k$ ,  $k = 1, 2, 3, \dots, l$  centres are chosen as mean values of sample patterns which are a part of class  $k$ , that is

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_k^i, \quad k=1, 2, 3, \dots, m$$

Wherein  $x_k^i$  refers to eigen vectors of  $i$ th image in class  $k$ , while  $N_k$  refers to the overall quantity of trained images in class  $k$ .

As RBF NNs are a set of NNs, activation functions of hidden components are defined using distances between input vectors as well as prototype vectors. Generally, activation functions of RBF components (hidden layer component) are chosen as Gaussian functions with mean vectors  $\mu_i$  as well as variance vectors  $\sigma_i$  thus

$$h_i(x) = \exp\left[-\frac{\|x - \mu_i\|^2}{\sigma_i^2}\right]$$

It is to be noted that  $x$  refers to  $n$ -dimensional input features vector while  $\mu_i$  refers to  $n$ -dimensional vectors known as centre of RBF units,  $\sigma_i$  refers to widths of  $i$ th RBF component and  $l$  refers to number of RBF components. Responses of  $j$ th output units for inputs  $x$  are as follows:

$$y_j(x) = \sum_{i=1}^l h_i(x)w(i, j)$$

Wherein  $w(i, j)$  refers to connection weights of  $i$ th RBF component to  $j$ th output nodes [22].

### Algorithm

Specific algorithms were adopted by researchers, to compute the data received from the above tournaments. One such algorithm, known as the C4.5, created by Quinlan, as a basic software extension of the ID3 algorithms, was used to handle problems which have not been accessed by the ID3 algorithm. Ignoring the overriding of data, shortened pruning of errors and post-pruning issues, the mishandling of continuous data through missing variable attribute, these are the common problems faced when adopting the ID3 algorithm. Therefore, the C4.5 classification algorithm benefits users, as it reinforces tree splitting via entropy and information gain. When researchers adopted this method in their training phase as a test, the algorithm was useful in obtaining the rule set and through this testing phase, it was possible to classify the rules to the pre-processed data [24].

## RESULTS AND DISCUSSION

Experimental evaluations are carried out through 436 neurological MRI scans from the OASIS project. The image is segmented using Multi-scale Image Segmentation. Features are extracted using DWT. The features are reduced using proposed AIS feature selection and classified using Naïve Bayes, RBF Classifier, C4.5 and KNN.

**Table 1. Classification Accuracy**

Classifiers	GA	AIS	GA-AIS
KNN	92.66	94.95	96.79
Naïve Bayes	93.12	95.87	97.94

C4.5	95.41	98.17	98.62
RBF Classifier	97.02	98.85	99.08

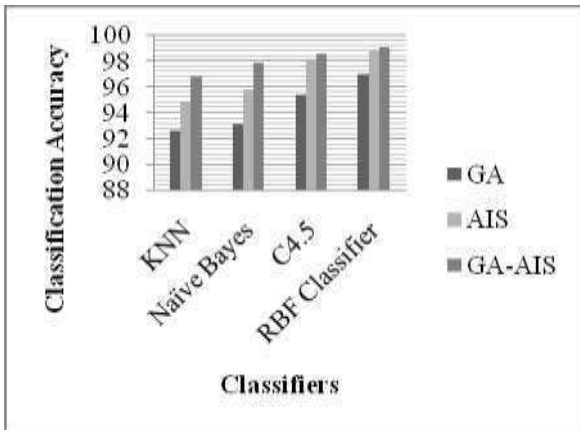


Fig. 1. Classification Accuracies

Observing [Figure- 1], it is seen that proposed AIS technique increased classification accuracies for all the classifiers. RBF classifiers with the proposed hybrid GA-AIS achieve the best accuracy of 99.08%, and it improves accuracy by 2.1% compared to GA feature selection and by 0.23% when compared to AIS feature selection.

Table: 2. Sensitivity for Normal

Classifiers	GA	AIS	GA-AIS
KNN	0.9547	0.9756	0.9826
Naïve Bayes	0.9582	0.9756	0.9826
C4.5	0.9756	0.9895	0.993
RBF Classifier	0.9791	0.993	0.993

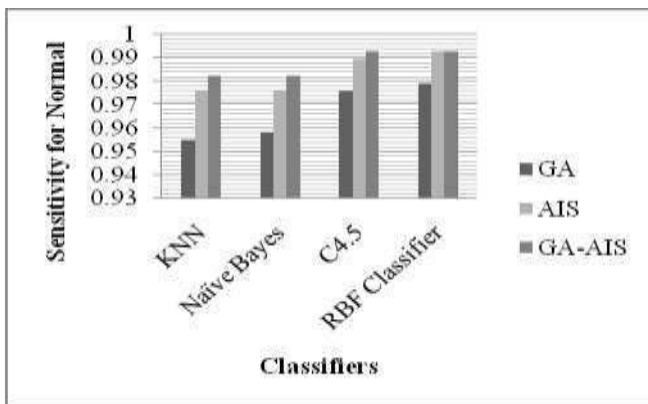


Fig. 2. Sensitivity for Normal

From the [Figure- 2], it is seen that proposed AIS technique improved sensitivities for Normal by 0.71%, 0.71%, and 0.35% for GA-AIS when compared with AIS for KNN, Naïve Bayes, C4.5 classifiers.

Table: 3. Sensitivity for Dementia

Classifiers	GA	AIS	GA-AIS
KNN	0.87	0.89	0.94
Naïve Bayes	0.88	0.93	0.97
C4.5	0.91	0.97	0.97
RBF Classifier	0.95	0.98	0.99

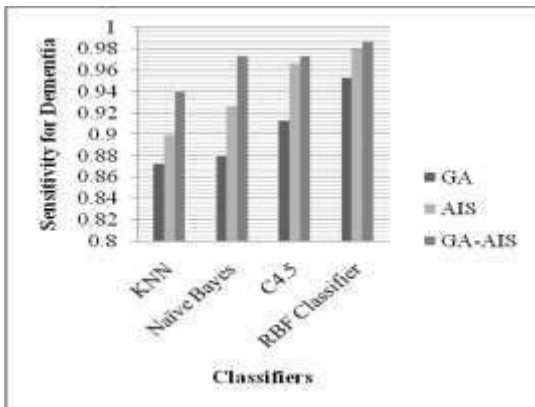


Fig: 3. Sensitivity for Dementia

From the [Figure -3], it is seen that presented GA-AIS technique improved Sensitivity for Dementia by 3.03%, 5.21%, 5.7% and 2.78% for GA feature selection compared with KNN, Naïve Bayes, C4.5, and RBF classifiers.

Table: 4. Specificity for Normal

Classifiers	GA	AIS	GA-AIS
KNN	0.87	0.89	0.94
Naïve Bayes	0.88	0.93	0.97
C4.5	0.91	0.97	0.97
RBF Classifier	0.95	0.98	0.99

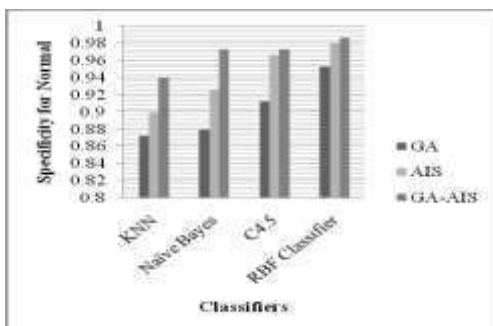


Fig: 4. Specificity for Normal

From the [Figure- 4], it is seen that presented AIS technique improved Specificity for Normal by 4.38%, 4.95%, 0.7% and 0.68% for GA-AIS when compared with KNN, Naïve Bayes, C4.5, and RBF classifiers.

Table: 5. Specificity for Dementia

Classifiers	GA	AIS	GA-AIS
KNN	0.95	0.98	0.98
Naïve Bayes	0.96	0.98	0.98
C4.5	0.98	0.99	0.99
RBF Classifier	0.98	0.99	0.99

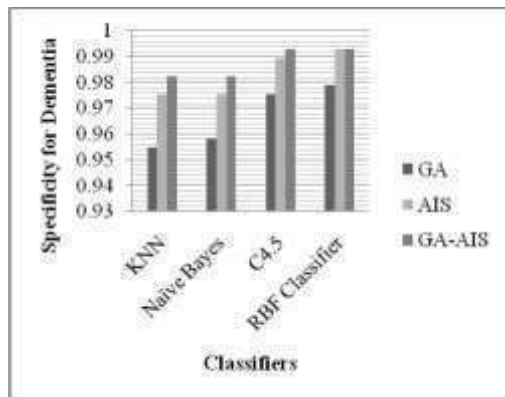


Fig: 5. Specificity for Dementia

From the [Figure- 5], it is seen that presented AIS approach improved Specificity for Dementia by 2.17%, 1.8%, 1.41% and 1.41% for GA when compared with KNN, Naïve Bayes, C4.5, and RBF classifiers.

### CONCLUSION

Neuroimaging is an important tool in the diagnostic work up of dementia. This paper presents an automatic MRI medical images classification process for classifying dementia. A feature selection based on AIS is proposed. These extracted features are classified with Naïve Bayes, C4.5 and K Nearest Neighbour. Comparative study with several alternate algorithms was carried out and AIS-GA yielded best outcomes in all categories. Results demonstrate that KNN with the proposed feature selection method achieves the best results and the proposed AIS feature selection improves the efficacy of the classifiers. Additional work is required to improve classification accuracy further through supervised learning.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

None

### REFERENCES

- [1] Shanthi V, Singh DJ. [2011] Estimation of Hippocampus Volume from MRI Using ImageJ for Alzheimer’s Diagnosis.
- [2] Ciblis AS, Butler ML, Bokde AL, Mullins PG, O’Neill D, McNulty JP. [2015] Neuroimaging referral for dementia diagnosis: The specialist’s perspective in Ireland. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*. 1(1): 41-47.
- [3] Desai KD, Parmar S. [2012] Effective early detection of Alzheimer’s and Dementia disease using Brain MRI Scan Images. *International Journal of Emerging Technology and Advanced Engineering*. 2(4).

- [4] Cao Q, Guo Z, Yang Y. [2015] An improved back propagation neural network approach to the remote sensing land use and land cover classification. In *Computer Science and Applications: Proceedings of the 2014 Asia-Pacific Conference on Computer Science and Applications (CSAC 2014)*, Shanghai, China, 27-28 December 2014. CRC Press.369.
- [5] Frey KA, Lodge MA, Meltzer CC, Peller PJ, Wong T Z, Hess CP, Subramaniam RM. [2016] ACR–ASNR Practice Parameter for Brain PET/CT Imaging Dementia. *Clinical nuclear medicine*. 41(2):118-125.
- [6] Chung D, Yun K, Jeong J. [2015] Decoding covert motivations of free riding and cooperation from multi-feature pattern analysis of EEG signals. *Social cognitive and affective neuroscience*, nsv006.
- [7] Adhikari SK, Sing JK, Basu DK, Nasipuri M, Saha P K. [2012] Segmentation of MRI brain images by incorporating intensity inhomogeneity and spatial information using probabilistic fuzzy c-means clustering algorithm. In *Communications, Devices and Intelligent Systems (CODIS), 2012 International Conference on IEEE*. 129-132.
- [8] Al-Badarneh A, Najadat H, Alraziqi AM. [2012] A classifier to detect tumor disease in MRI brain images. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on IEEE*. 784-787.
- [9] Jeena RS, Kumar S. [2013] A Comparative analysis of MRI and CT brain images for stroke diagnosis. In *Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy (AICERA/ICMiCR), 2013 Annual International Conference on IEEE*. 1-5.
- [10] Hamdaoui F, Sakly A, Mtibaa A. [2015] Real-time synchronous hardware architecture for MRI images segmentation based on PSO. In *Systems and Control (ICSC), 4th International Conference on IEEE*. 498-503.
- [11] Pinto A, Pereira S, Dinis H, Silva CA, Rasteiro DM. [2015] Random decision forests for automatic brain tumor segmentation on multi-modal MRI images. In *Bioengineering (ENBENG), IEEE 4th Portuguese Meeting on IEEE*. 1-5.
- [12] Rajini NH, Bhavani R. [2011] Classification of MRI brain images using k-nearest neighbor and artificial neural network. In *Recent Trends in Information Technology (ICRTIT), International Conference on IEEE*. 563-568.
- [13] Rehman AU, Khanum A, Shaukat A. [2013] Hybrid Feature Selection and Tumor Identification in Brain MRI Using Swarm Intelligence. In *Frontiers of Information Technology (FIT), 11th International Conference on IEEE*. 49-54
- [14] Saraswathi S, Mahanand BS, Kloczkowski A, Suresh S, Sundararajan N. [2013] Detection of onset of Alzheimer's disease from MRI images using a GA-ELM-PSO classifier. In *Computational Intelligence in Medical Imaging (CIMI), IEEE Fourth International Workshop on IEEE*. 42-48.
- [15] Jang J, Kim HW, Kim YS. [2014] Co-segmentation of inter-subject brain magnetic resonance images. In *Ubiquitous Robots and Ambient Intelligence (URAI), 11th International Conference on IEEE*. 80-84.
- [16] Miller V, Erlien S, Piersol J. [2012] Identifying dementia in MRI scans using machine learning. Available at <http://cs229.stanford.edu/proj/ErlienMillerPiersolIdentifyingDementiaInMRIScansUsingMachineLearning.pdf>.
- [17] Mishra HOS, Bhatnagar S. [2014] MRI and CT image fusion based on wavelet transform. *International Journal of Information and Computation Technology*. ISSN. 0974-2239.
- [18] Koundal D, Gupta S, Singh S. [2012] Computer-aided diagnosis of thyroid nodule: a review. *International Journal of Computer Science and Engineering Survey*. 3(4):67.
- [19] Takate VS, Vikhe PS. [2012] Classification of MRI Brain Images using K-NN and k-means. *International Journal on Advanced Computer Theory and Engineering (IJACTE)*. 1(2).
- [20] Revathy M. [2012] Image classification with application to MRI brain using 2nd order moment based algorithm. *International Journal of Engineering Research and Applications (IJERA)*. 2(3):1821-1824.
- [21] Karande SM, Jayapal PC, Kale NA. [2014] A New Approach for Brain Tumor Detection and Area Calculation Using Median Filter, K-Means, SVM and Naïve Bayes Classifier.
- [22] LP, Verma VK. [2012] Classification of MRI Brain Images Using Neural Network. 2(5):751-756.
- [23] van Rijn S, Emmerich M, Reehuis E, Back T. [2015] Optimizing highly constrained truck loadings using a self-adaptive genetic algorithm. In *Evolutionary Computation (CEC), IEEE Congress on IEEE*. (4):227-23
- [24] Rajesh K, Anand S. [2012] Analysis of SEER dataset for breast cancer diagnosis using C4. 5 classification algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*. 1(2):2278-1021.

# DESIGN OF A NOVEL ARRAY MULTIPLIER USING ADIABATIC LOGIC IN 32NM CMOS TECHNOLOGY

Suresh Kumar Pittala, A. Jhansi Rani

Research Scholar, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, Andhra Pradesh, INDIA  
Department of ece, Velagapudi ramakrishna siddhartha engineering college, Vijayawada, INDIA

## ABSTRACT

*Aim: The paper presents a new adiabatic multiplier circuit based on Complementary Energy Path Adiabatic Logic (CEPAL). The proposed multiplier consumes lesser power when compared to the conventional CMOS multiplier. The proposed adiabatic array multiplier performs 8 bit multiplication. The proposed adiabatic multiplier is also designed with leakage reduction technique the performance of which is better when compared to the CMOS multiplier. The operating speed of the complementary metal oxide semiconductor is increased. This paper presents the implementation of adiabatic CEPAL multiplier using CMOS. The measurement results of the adiabatic CMOS Multiplier demonstrates a reduction in power and reduction in energy. The operating frequency is in GHz range. These results shows that the proposed circuit can be used in high speed application. The proposed adiabatic circuits are designed in HSPICE using predictive technology models (PTM) in 32nm CMOS Technology. The experimental results for the proposed adiabatic designs demonstrate their effectiveness with energy consumption and with power optimization.*

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

CMOS, adiabatic logic, Adder, NAND gate, shorted gate, energy efficient, power optimization.

\*Corresponding author: Email: [dr.sureshkumarpittala@gmail.com](mailto:dr.sureshkumarpittala@gmail.com)

## INTRODUCTION

In recent times, mobile devices are emerging in faster rate replacing the existing devices which are bulky and consume more power. The mobile devices have high device density and is computationally faster. But increase in device density increases the power consumption. Circuits fabricated in CMOS technology below 65nm have less control over temperature and power wastage happens due to leakage current. Even though full swing voltage mode CMOS logic styles have been extremely successful they suffer from the lower limit of power dissipation on  $CLV_{dd}/2$  during switching instants. So new methodologies and technologies are evolved to solve the problems occurring due to undesired power dissipation. Adiabatic circuits have power dissipation lower than the limit level of CMOS. But the speed of operation reduces. The total energy is reduced through energy recycling which is the basic concept behind adiabatic. Power supplies are pulse shaped and to be designed separately to meet the required load. In literature the adiabatic circuits are proved to have the best performance in power consumption. For computing units in processor core, faster and low power processing elements (PE) are required. Circuits like inverters, adders, multipliers, shifters and latch for the basic building blocks of a microprocessor or digital signal processor. Several adiabatic logics are proposed in literature like ECRL [1], differential logic [2], Dual rail [3], pass transistor [4] etc. In the literature chanda et al [5] proposed a multiplier architecture based on Urdhva Tiryakbhyam in CMOS adiabatic logic. A  $N \times N$  vedic multiplier is designed using sum block, carry block and NAND-AND adiabatic block. For the implementation energy efficient adiabatic logic (EEAL) is used. The work is carried out in 180nm technology. The power can be further modified in this work by implementing in a 90nm technology or utilizing a different adiabatic logic instead of EEAL.

Cancio Monteiro et al [6] proposed a charge-sharing symmetric adiabatic logic (CSSAL) multiplier. The paper reports the NAND/NOR logic and their implementation in multiplier. The advantage of the proposed circuit is nodes internal to the circuit is maintained at same charge for the different combinations of input. The disadvantage is the less successful of the method at higher frequencies so used in cryptography applications.

In literature [7] a two phase clocked adiabatic static CMOS based Baugh wooley and Wallace tree multiplier is designed. The paper reports that the Wallace tree Multiplier shows less power consumption about 62.66%



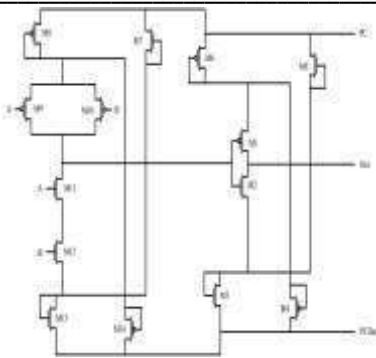


Fig: 2. CEPAL AND gate structure

### Processing elements in DSP architectures

In DSP applications design of Filters is the critical block which deals with lot of computational units. The filter structure differs from application to algorithm specific. The adders, multipliers, shift registers, inverters, buffers are the different components of the filter structure. The multiplier is the most critical unit and requires an optimized structure based on partial product reduction, multiplier and options etc. In multipliers the partial products are first generated and finally combined to obtain the final product. In this work the multiplier is optimized in the adder block where the number of transistors are reduced by 80% compared to the conventional adders. For partial product generators the array multiplier and vedic multiplier are implemented using adiabatic logics. The adder designed in this work is based on CEPAL adiabatic logic. As explained earlier, the adder design is a prime concern for the implementation of an efficient multiplier. The proposed adder is more suitable for efficient implementation of large operand adders. However, these designs are very efficient in terms of power consumption but special power clock generator circuits are required which is not the scope of this paper.

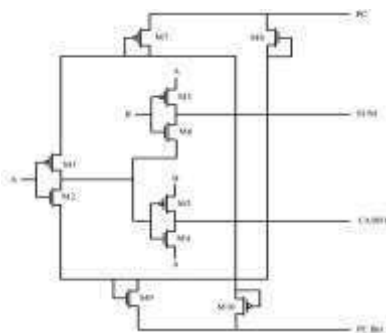


Fig: 3. CEPAL Adder structure

### Multiplier

Easy to design and smaller in size features corresponds to the array multiplier. As the name indicates it uses parallelogram technique in its operation. The partial products are generated at the same time. The combining adder receives the partial products as inputs and produces the output. The multiplier is well suitable when performing the matrix multiplication. The structure is regular such that the vertical and horizontal delays are same. The critical path delays in the terms of full adder and gate have same value. Implementation of DSP algorithms with pipelining can be made easy using array multipliers. Array multiplier is shown in [Figure- 4]



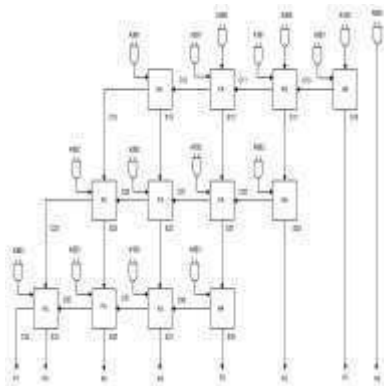


Fig: 4. Array multiplier using conventional adders

### PROPOSED CMOS CEPAL MULTIPLIER

CMOS based adiabatic CEPAL processing element implementation using half adder, full adder and multiplier is proposed. The proposed array multiplier block diagram is shown in [Figure- 5]. The implementation of the proposed circuit for multilier is shown in [Figure- 6].

The basic advantage of 6T half adders is smaller area and lower power utilization. It becomes more not easy and even obsolete to keep full output voltage swing operation as the design with fewer transistor count and lower power utilization are pursued. In pass transistor logic the output voltage swing may be degraded due to the threshold voltage defeat problem. The reduction in voltage swing leads to lower power consumption but may also lead to slow switching in the case of cascaded operation such as ripple carry adder. A low VDD operation the corrupted output may even cause break down of circuit. The smallest voltage that 10 T adder can work at 0.7V. The excessive power dissipation and long delay are attributed to the threshold voltage drop problem and the poor driving capability of some internal nodes at input combinations that create non full-swing transitions. The elimination of the path to the ground reduces the total power use by reducing the short circuit power. The combination of low power and low transistor add up makes the SERF adder circuit a viable option for low power design.

In this paper, we will propose a novel full adder design featuring complementary and level restoring carry logic (CLRCL).The goal is to reduce the circuit complexity and to achieve faster cascaded operation. The strategy is to avoid multiple threshold voltage losses in carry chain by proper level restoring. We first rewrite the full adder and Boolean functions as

$$\begin{aligned} \text{Sum} &= (A \oplus C_{in}) \cdot \overline{C_{out}} + (A \odot C_{in}) \cdot B \\ C_{out} &= (A \oplus C_{in}) \cdot B + (A \odot C_{in}) \cdot A. \end{aligned}$$

“Urdhva Tiryagbhyam (UT)” sutra based multiplier. The multiplication algorithm based on the vedic sutra Urdhva Tiryagbhyam is faster and consumes low power. The sutra performs vertically and crosswise operation between the different digits of the multiplier and multiplicand. The partial product and sum are calculated in a single iteration step. The vedic based multiplier is implemented to compare the performance of the proposed circuit.

Table: 1. Power and energy analysis of Half adder

Circuit	Power clock= 1GHZ				
	P <sub>avg</sub> (W)	I <sub>avg</sub> (A)	VDD(V)	(T <sub>stop</sub> -T <sub>start</sub> )(s)	E(J)
Conventional	2.29E-05	1.45E-05	1	9.99E-07	1.45E-11
CEPAL conventional HA	1.23E-05	6.72E-06	1	9.99E-07	6.72E-12
CEPAL proposed HA	7.85E-05	5.81E-07	1	9.99E-07	5.80E-13

**Table: 2. Power and energy analysis of Half adder**

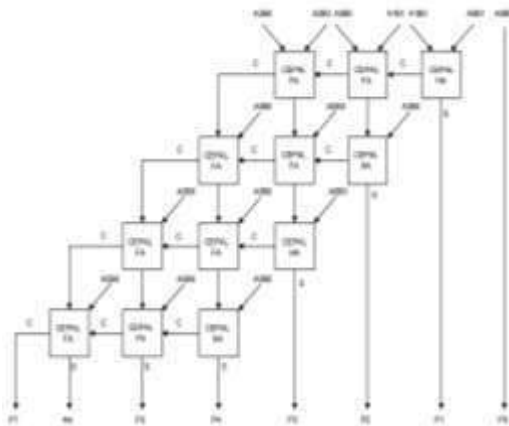
Circuit	Power clock= 1GHZ				
	P <sub>avg</sub> (W)	I <sub>avg</sub> (A)	VDD(V)	(T <sub>stop</sub> -T <sub>start</sub> ) (s)	E(J)
Conventional	8.12E-05	1.77E-05	1	9.99E-07	1.76E-11
CEPAL conventional FA	8.70E-05	2.49E-06	1	9.99E-07	2.48E-12
CEPAL proposed FA	7.93E-05	5.04E-07	1	9.99E-07	1.04E-13

**Table: 3. Power and energy analysis of Half adder**

Circuit	Power clock= 1GHZ				
	P <sub>avg</sub> (W)	I <sub>avg</sub> (A)	VDD (V)	(T <sub>stop</sub> -T <sub>start</sub> ) (s)	E(J)
CEPAL proposed 4X4 VEDIC MULTIPLIER	7.64E-04	3.04E-05	1	9.99E-07	3.04E-11
CEPAL proposed 4X4 ARRAY MULTIPLIER	6.44E-04	1.60E-05	1	9.99E-07	1.60E-11

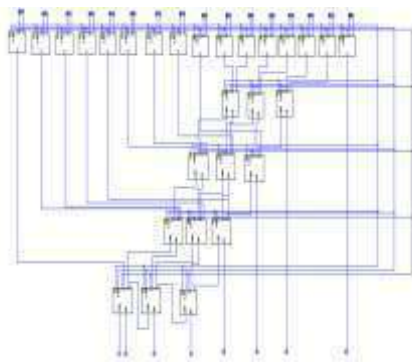
For the implementation predictive technology model for 32nm is used. The supply voltage is 1V. The number of transistors is 14, 10 and 76 for full adder, half adder and multiplier respectively. XOR and AND gate having 14 transistors and 6 transistors respectively are used for the design of adder. The array structure for 8 bit results is shown. The circuit is faster in the operation and consumes lesser power and energy. The power and energy analysis for the conventional and proposed half adder is shown in [Table- 1].

From [Table- 2] The results show that the power dissipation in conventional FA is more compared to the proposed. The proposed multiplier dominates and outperforms the vedic multiplier design. The vedic multiplier is faster when implemented in an non adiabatic way but still suffers from the number of transistors used. But when comes to adiabatic logic, the CMOS CEPAL based multiplier performance is better. [Table- 3] shows the comparison table for the two multipliers. The circuit runs with a power clock of 1 GHz. The power consumption is reduced by 15 % and energy consumption by 47%.



**Fig: 5. Proposed Array multiplier using CEPAL adiabatic logic**

www.iioab.org THE IIOAB JOURNAL www.iioab.webs.com



**Fig. 6. Proposed Array multiplier implementation using CEPAL adiabatic logic**

## CONCLUSION

A new adiabatic multiplier circuit based on Complementary Energy Path Adiabatic Logic (CEPAL) is proposed in this paper. Compared to the conventional methods the proposed circuit is advantages. The proposed multiplier consumes less power and energy when compared to the conventional multiplier. The proposed adiabatic array multiplier performs 8 bit multiplication and is designed with leakage reduction technique. The measurement results of the half adder, full adder and adiabatic CMOS Multiplier demonstrates a reduction in power and reduction in energy. The power clock frequency is set to 1GHz. The implementation were carried out using HSPICE tool with predictive technology models (PTM) in 32nm CMOS Technology.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Chun-Keung L, Philip CH, Chan. [1999] An Adiabatic Differential Logic for Low-Power Digital Systems, IEEE Transactions on Circuits and Systems II:Analog and Digital Signal Processing. 46(9):1245-1250.
- [2] Matthew M. [2014] Synthesis of Dual-Rail Adiabatic Logic for Low, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 33(7):975-988.
- [3] Liu F, Lau KT. [1998] Pass-transistor adiabatic logic with NMOS pull-down configuration, Electronics Letters. 34(8):739-741.
- [4] Chand M, Banerjee S, Saha D, Jain S. [2013] Novel transistor level realization of ultra low power high-speed adiabatic Vedic multiplier, International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 22-23. Kottayam,India. 801-806.
- [5] Cancio M, Yasuhiro T, Toshikazu S. [2013] Robust secure charge-sharing symmetric adiabatic logic against side-channel attacks, 36th International Conference on Telecommunications and Signal Processing (TSP), 2-4. Rome, Italy. 732-736.
- [6] Vishal Shankarrao M, Anchu T, Vigneswaran T. [2015] Design of Baugh Wooley and Wallace tree multiplier using two phase clocked adiabatic static CMOS logic, 2015 International Conference Industrial Instrumentation and Control (ICIC), 28-30. Pune, India. 1178-1183.
- [7] Hardik S, Tanay M, Modi, Kanchana Bhaaskaran VS. [2014] Low power vedic multiplier using energy recovery logic, 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 24-27. New Delhi,India. 640-644.
- [8] Shashank S, Trailokya Nath S. [2015] Design of vedic multiplier using adiabatic logic, International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 25-27. Noida, India. 438-441.
- [9] Sarita U, Saumya P, Garima B, Jasdeep K. [2015] 4X4 Bit Multiplier using Adiabatic 2XOR and sleep mode transistor logic, 2015 International Conference on Signal Processing, Computing and Control (ISPCC), 24-26. Wanknaghat, India. 262-265.
- [10] Kazunari K, Yanagido, Yasuhiro T, Toshikazu S. [2014] Skew tolerance analysis and layout design of 4 $\times$ 4 multiplier using two phase clocking subthreshold adiabatic logic, 2014 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), 17-20 Ishigaki, Japan. 495-498.

# A NOVEL ALGORITHM FOR THE ARRHYTHMIA DIAGNOSIS IN FETAL MONITORING SYSTEM

Radha Abburi<sup>1</sup>, A.S.Chandrasekhara Sastry<sup>2</sup>

<sup>1</sup> Department of ECE, BVRIT Hyderabad College of Engineering for Women, Bachupally, INDIA

<sup>2</sup> Department Of ECE, KL University, Vaddeswaram, Vijayawada, Andhra Pradesh-522502, INDIA

## ABSTRACT

The fetal ECG signal is analyzed using wavelet transform for extraction of Fetal Electrocardiogram (ECG) signals to diagnose. The paper presents the various wavelet transform methods available for denoising and delineating the abdominal electrocardiogram(AECG).The AECG signal has been denoised and delineated using a proposed methods which combines the features of adaptive thresholding with wavelet transform and SVM classifier. Both the simulated version and real-time version of the abdominal electrocardiogram signal was used for performing extraction and classification operation. Extraction of fetal electrocardiogram signal was done using various wavelet transforms such as Daubechies, Biorthogonal and Symlet wavelets. While experimenting with various wavelets it was evident that the Daubechies wavelet proved more efficient in the removal of noise for extraction. The extracted signal was slope threshold to find the maximum point (R peak) in the fetal electrocardiogram signal. Hence the abdominal electrocardiogram was delineated to get the individual parameters which can then be utilized for analysis. The performance of the hybrid method is been compared with other methods. The performance of the classifier was evaluated in terms of training performance and classification accuracies. The results are to be compared to find a better method for ECG classification.

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

Fetal ECG, FECCG extraction, Wavelet Transform, SVM classifier

\*Corresponding author: Email: [radhasundi@gmail.com](mailto:radhasundi@gmail.com)

## INTRODUCTION

The objective of the work is to formulate and implement a new method for the extraction of Fetal ECG features from the abdominal ECG signal using Hybrid signal processing technique. Extraction of fetal ECG features and classification during first trimester of pregnancy will help the physician to know the well being of the fetal. Unwanted Noise signals significantly distort fetal ECG recordings, and consequently the presence of noises is troublesome in extracting the features in ECG signal. Hence, devising efficient methods for successful removal of noises and extraction of fetal ECG from ECG recordings have been still a major challenge. The performance of the algorithms employed previously for extraction of adult ECG signal and classification will not be efficient enough to do the same for the Fetal ECG. Therefore, an efficient algorithm to extract and classify fetal ECG is the objective of this work. In the method of recording, the fetal ECG signals have a very low power relative to that of the maternal ECG. In addition, there will be several sources of interference, which include intrinsic noise from a recorder, noise from electrode-skin contact, baseline drift (DC shift), 50/60 Hz noise etc.

The situation is far worse during the uterine contractions of the mother. During these contractions, the ECG recordings will be corrupted by other electrophysiological signals called uterine electromyogram (EMG) or electrohysterogram (EHG), which are due to the uterine muscle rather than due to the heart. The response of the fetal heart to the uterine contractions is an important indicator of the fetal health. As such a need arises to effectively monitoring the fetal ECG during the uterine contractions. But monitoring the fetal ECG during these contractions is a difficult task because of very poor SNR. The three main characteristics that need to be obtained from the fetal ECG extraction for useful diagnosis includes the Fetal heart rate (FHR), Amplitude( P, Q, R, S, T, QRS) of the different waves and Duration of the waves (S-to-T, R-R Interval,). In literature several methods are been proposed for the extraction of fetal ECG signal. Ruben Martin Clemente et al [1] proposed a simple algorithm based on independent component analysis with simple procedure. The method reduces the dimension and has computational simplicity for extraction. The modification is done based on reduction in dimension and simple post processing stage. But the convergence behavior is always a problem when the algorithm is implemented using reconfigurable architectures. An autonomous algorithm to locate the QRS complex in the fetal ECG signal based on subspace decomposition is presented in [2]. The method is iterative so computational time increases. A novel merging

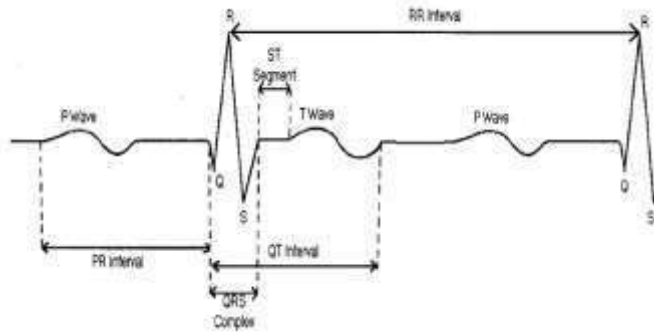
technique is used to detect the fetal ECG R-peaks which requires proper phase alignment which is difficult in long time data. J. L. Camargo et al [3] in their paper presented a multidimensional ICA method to detect the fetal ECG signal. The paper also presents the difficulty in analyzing the delay in the transit time of the signal from maternal heart to mother's abdomen for which the method fails its applicability. To support in clinical implementation for classification of fetal heart rate Shishir Das et al [4] proposed generative models instead of scalar features for the process. The method depends on feature sequences of local patterns. The patterns along with umbilical cord pH values classifies known category with good performance. Liang Han et al [5] proposed a V-support vector regression method to extract the fetal electrocardiogram from the multichannel abdominal ECG.

The nonlinear dependency of maternal ECG on abdominal ECG is utilized for the estimation of V-SVR. The estimated maternal component is subtracted from the abdominal ECG to get the fetal ECG. The disadvantage is phase difference will lead to complete erroneous extraction. The requirement of new methods for classification of fetal ECG signal is getting increased day by day. One such work is proposed earlier in literature where the authors [6] present a neural network based extraction algorithm. Different learning constants were utilized in the project. Few works [7] are done towards giving solutions to the age old problem of electrode numbers. When most of the works are based on multiple channel electrodes the work is based on extraction of fetal ECG from single lead channel. Heart rate variability analysis is been done based on spectral analysis technique. A similar single lead extraction method using extended nonlinear Bayesian filtering framework is proposed in [8]. The extended state Kalman filtering based proposed work uses the dynamic model to discriminate the information in the simulated data. An hybrid method using wavelet transform and adaptive filtering is used in the paper [9] for the fetal ECG extraction. The obtained scale coefficients in wavelet transform were processed using LMS and spatial selective filter. Proper value for threshold determines the efficiency of the system. In [10] George Georgoulas et al have proposed a support vector machine methodology to classify and predict the metabolic acidosis on newborns. The paper presents feature extraction methods through time and frequency domains and classification is done using support vector machines (SVMs). Krupa et al [11] proposed a new method to feature extract and classify the fetal ECG using empirical mode decomposition and support vector machine. The classification is done using the standard deviations of the EMD components. The accuracy of the findings is 86% and the geometric mean of sensitivity and specificity was 94.8%. An adaptive neuro-fuzzy inference system (ANFIS) based fetal diagnosis algorithm is proposed in [12]. Fetal heart rate and uterine contraction data are used for the classification of the normal and the pathologic state. The accuracy of the method is reported as 97.2. Hasan Ocak proposed a SVM – genetic algorithm based fetal ECG arrhythmia diagnosis algorithm [13]. The SVM is constructed using the features extracted from normal and pathological fetal heart rate and uterine contraction. The accuracy reported in the paper is 99.3% and is shown as a better method when compared to the existing neural network based algorithm like ANFIS and ANN. Many methods proposed and executed in the literature classify only few classes [14]. For multiclass training and classification Ben Fei and Jinbai Liu proposed a Binary tree of SVM (BTS) algorithm. The reported BTS training speed is higher due to Log complexity. Similarly a multi-class classification algorithm based on support vector machines for decision tree architecture is proposed in [15]. The hierarchy of binary decision subtasks was determined using clustering algorithm. The method is reported to be faster due to its Log complexity than the widely used multi-class SVM methods like “one-against-one” and “one-against-all”. In this paper the ECG signal recorded by the standard twelve lead system and those obtained by simulating it in MATLAB software is used.

## LITERATURE SURVEY

### a) ECG waves and their intervals

The ECG signal consists of five waves each with definite time period and amplitude. The ECG signal features are similar for the fetal and maternal but the duration and amplitude differs. The P wave representing the spread of electrical impulse through the atrial musculature (activation or depolarization). But in the fetal ECG it's difficult to extract the P wave since the wave is been contaminated by the noises. Its duration is not more than 0.11 seconds and amplitude of not more than 10% that of R-Wave for maternal and for the fetal it will be half and one-tenth in the maternal duration and amplitude respectively. The QRS complex duration is from 0.08s to 0.12s. The S-T segment follows the QRS complex. The T wave is slightly rounded and slightly asymmetrical. The heart rate is calculated based on the Interval between R wave and the next R wave. RR intervals are irregular for sinus node disease and ventricular arrhythmias.



**Fig. 1: Basic ECG waveform with the characteristics waves**

ECG signal characteristics can be studied using several techniques such as Pan and Tompkins algorithm, Kalman filter, Extended Kalman filter, Wavelet transforms. The basic principles used here for denoising is Wavelet transform. This method has the advantage of preserving both time and frequency information in the signal

### b) Wavelet transform

The mother wavelet similar to the feature of the fetal ECG signal wave is allowed for simultaneous time and frequency analysis by the way of a time-frequency localization of the signal. Like conventional Wavelet systems dilating and translating the ECG mother wavelet  $\psi(t)$  is given by

$$\Psi_{a,b}(t) = |a|^{-1/2} \Psi\left(\frac{t-b}{a}\right) \quad (1)$$

Where the scaling factor and translation factor are real ( $a \neq 0$ ). To analyze the low frequency components the mother wavelet is stretched by a large value of 'a'. Since the high frequency components is less important or not required for clinical investigation too high value of "a" is avoided.

**Thresholding and Denoising Scheme:** The wavelet transform decorrelates the fetal ECG signal information into a small number of coefficients which are compared with a threshold. The decomposition level and largest transform co-efficient determines the threshold value (th). In this work the threshold is set adaptively.

**Thresholding Method:** The selected detail coefficient is used to perform the detection of fetal R-wave. For this purpose a threshold limit is set to remove the noises or unwanted peaks in the signal. There are several thresholding methods known. Here we apply both soft and hard thresholding in which the samples below a predetermined threshold are set to zero. The threshold is selected as 12% of maximum value In some researches the detail coefficient has been chosen to detect R wave based on Energy, frequency and correlation Analysis

### c) The feature extraction using wavelet transform

In the preprocessing algorithm, the wavelet transform is used to remove the unwanted noises in the signal. The wavelet decomposes the components, on thresholding the noise components and taking inverse wavelet transform the information of the fetal ECG signal will be retrieved. The noise components are occurring in the finest scales which can be removed. Further, by employing the thresholding mechanism the slope of the R wave is identified to detect its onset. The following equation of slope thresholding is

$$Z = (2 * c) / 16$$

Where , c is maximum value of the slope

Z is slope thresholding value.

Then, the QRS complex is extracted by employing the method of forward searching and reverses searching ten samples with respect to the identified peak point in the denoised signal. Then by employing same mechanism, other waves such as the P wave and T wave were also extracted separately from the denoised signal. The individual delineated signals (P wave, QRS complex, T wave) are then separately correlated with each of the various diseased ECG signals to obtain the maximum correlation value. Based on the correlation values the arrhythmia present in the test signal can be identified. Then the results can be compared with the physician's annotations for accuracy.

#### d) SVM classifier

In pattern recognition problem selection of most suitable subset of the extracted features is very important. Using fewer features better generalization can be done. Various dimensionality reduction technique like principal component analysis use linear combination of original features by which the relevant information are preserved, The computed Eigen values has uncorrelated features which improves the classification operation. Once the dimensionality is reduced the feature vectors are allotted suitable labels. Based on the feature vectors and the kernel functions choice the SVM decision surfaces are constructed. The most important kernels used are polynomial learning machines, the radial basis function (RBF) networks and the two-layer perceptrons. For the classification of our work RBF kernels whose width is specified depending on the data in priori and is common for all the kernels and Polynomial kernels of degree is used. Using same penalty parameters for different classes will reduce the specificity. The two penalty parameters ratio is set to inverse of the corresponding cardinalities of the classes. In this research work, we experimented with various configurations of the learning machines varying the width for the RBF kernels and the degree for the polynomial kernels. For each configuration of the kernel we tested different values for the parameters.

The New Tree based SVM classifier depending on recursive operation is performed with less classifier iteration / groups. In tree based classifier binary classifiers are used in every nodes of the tree and the branch contains the decision output. Recursively each node finalizes with one class samples by employing the probabilistic outputs to measure the similarity between samples are classes used for training.

### PROPOSED METHOD

A SVM-Wavelet based Extractor-classifier is presented as a diagnostic tool to aid physicians in the classification of fetal heart diseases. The proposed methodology using a strategy of hybrid approach of Wavelet transform and Support Vector Classifier system. In this work two intelligent approaches are composed, it will achieve good reasoning in quality and quantity. In other words we have Wavelet based extraction and Machine learning calculation. The feature vectors extracted using wavelet transform were applied as the input to an SVM classifier. The experimental procedure involved in the methodology is as follows:

#### **Data Collection:**

The source of the ECG is obtained from MIT-BIH and EDF Arrhythmia Database (MITDB). From this database variety of data including normal and abnormal cases extracted.

#### **Preprocessing:**

The obtained ECG signal is preprocessed by wavelet transform for the removal of noise components to enhance the quality of ECG signals and help us to detect significant signal events.

#### **Feature extraction and selection:**

Extraction of salient features from the ECG to allow detailed waveform analysis. The features, which represent the classification information contained in the signal are used as inputs to the classifier.

#### **Classifier design:**

Designing an intelligent system to automatically classify the shape of the ECG waveform and interpret shape changes by using an appropriate classifier model. Then the training algorithm is used to train and test the input signal and classify them into different categories.

#### **Optimization/Diagnostic decision:**

To validate the findings and show agreement between the system and human experts. The performance of the classifier is evaluated in terms of training performance and classification accuracies. All the necessary algorithms are implemented in MATLAB.

#### **Database:**

The proposed algorithm is tested with various databases and simulated database. The database used throughout the work was from MIT-BIH Physionet database, EDF database and DaISy (Database for the Identification of Systems). The database contains cutaneous potential recordings of pregnant woman (8 channels), Sampling of 10s, The channels 1-5 is abdominal and 6,7,8 thoracic. The simulated ECG signals using dynamic model are used in this work. The simulated database is generated using dynamic models with Gaussian equation and simple MATLAB functions. The synthetic ECG signals for maternal and fetal is generated for various noise levels.

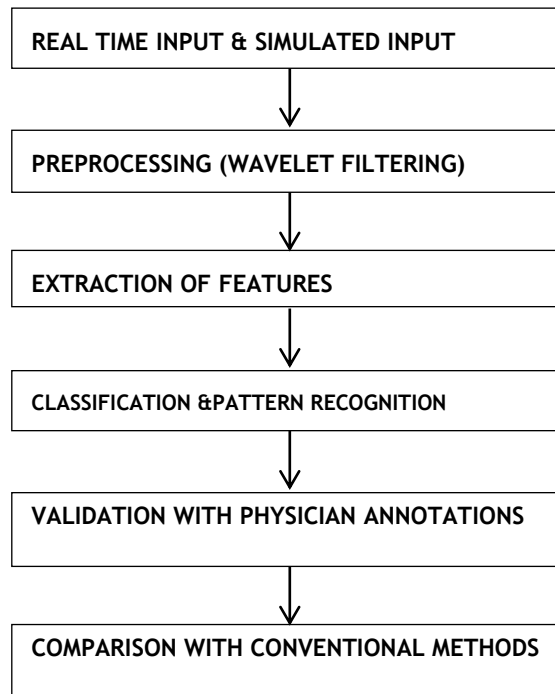


Fig. 2: Block diagram of proposed method

## RESULTS AND INFERENCE

The wavelet transform methods were implemented for preprocessing stage with adaptive thresholding using different wavelets like biorthogonal wavelets, daubechies, coiflets [Figure 3-7] and symlets and tested with simulated and real time signals. For Example Considering bi-orthogonal [Figure- 4] and compactly supported wavelet families (bior1.3, bior2.6, bior3.5, bior5.5, bior3.7) for the 3-level and 5-level decompositions with discrete wavelet transform, the performances are very close to each other and they generally give better results for soft thresholding than hard thresholding denoising rule. Other orthogonal wavelets like Daubachies Db1, Db2, Db3, Db8 [Figure- 6] are applied. Uniformly distributed white noise is added to the ECG signal. From the Table-1 it is seen that 5-level decomposition gives better denoising. The visual inspection of the denoised signal for the bior2.6 and bior5.5 is better than the rest of the biorthogonal set of wavelets and wavelet packet analysis. The computed signal-to-noise ratios are approximately 10.5 dB for the used biorthogonal wavelets family. 1 thoracic and 3 abdominal signal are used with different noise levels imitating the placement of electrodes at different location of the maternal's abdomen. From the application of wavelet denoising techniques all wavelets removes the noise at lower energy levels while failing to remove at higher amplitudes. The shape of the signal/frequency components are altered when high noise components present.

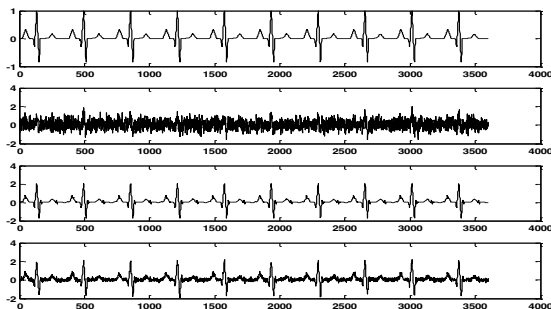


Fig. 3: Multichannel input channel 1, 2,3 and 4 given to the algorithm, thorax containing maternal signal abdominal signal containing maternal+fetal+noise of various levels



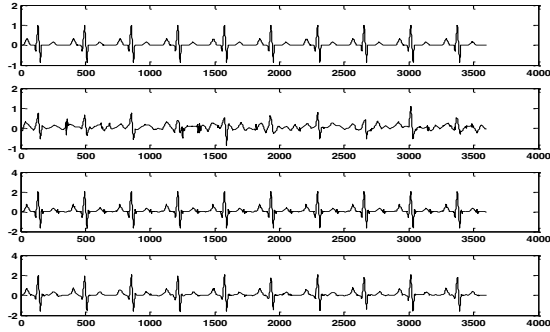


Fig.4: Multichannel output denoised by Biorthogonal wavelet BIOR 2,6 decomposition level 5

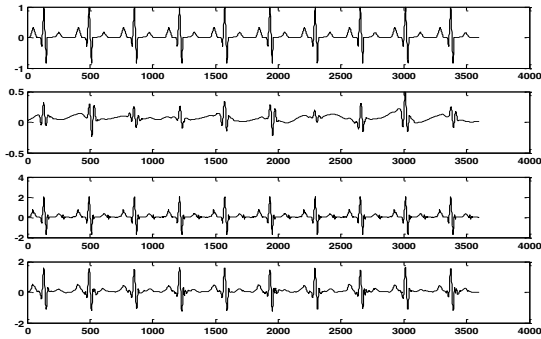


Fig.5: Multichannel output denoised by coiflet 5 decomposition level 7

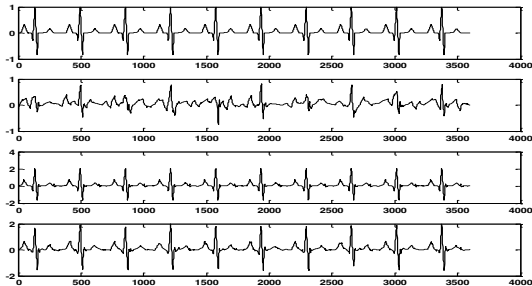


Fig. 6: Multichannel output denoised by Daubechies 5 decomposition level 7

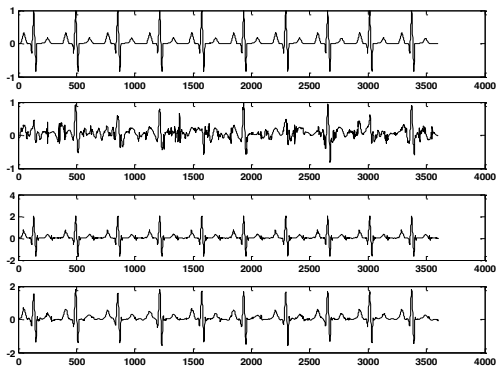


Fig.7: Multichannel output denoised by rbio 3,5 decomposition level 7

**Table 1. PSNR and MSE comparison of few wavelets which is used in the extraction and denoising of ECG signals**

Wavelet	Coif1	Sym2	Db1	Db2	Bior 5,5	Rbio 3,5
PSNR	68.8789	68.6466	68.9327	68.6466	68.8974	69.0033
MSME	0.0084	0.0089	0.0083	0.0089	0.0084	0.0082

The extraction of fetal ECG is done and the patterns are classified using SVM classifier. The performance of the classifier is obtained and our classifier gives better performance and accuracy. The SVM model for classification of fetal Electrocardiogram (ECG) signals into one of the few known categories is done and diagnostic decision is to be arrived at a regarding the condition of the fetus.

**Table 2. Performance of the classifier**

Classifier	Positive Predictive Value	Negative Predictive Value
Neural Network	100	100
SVM	100	100
KNN	100	100
Naive Bayes	89.21	93.7
Daubechies	94	93
Symlet	95	94
PSO-based	96	95

Features like Heart Beat and amplitude is extracted. Using derivative method the features are extracted. m=maternal f=fetal.

Classification of basic features using different classifiers like Neural Networks, K-Nearest Neighbor and Naïve bayes algorithm are investigated and compared with the SVM classifier. The extracted QRS amplitude and Heart rate is given to the different classifier and trained .For testing different data set is given and checked .The performance is given in **Table- 2**.

**CONFLICT OF INTEREST**

The authors declare no conflict of interests.

**ACKNOWLEDGEMENT**

None

**FINANCIAL DISCLOSURE**

None.

**REFERENCES**

- [1] Ruben Martin Clemente, Jose Luis Camargo Olivares, Susana Hornillo Mellado, Mar Elena, Isabel Roman. [2011] Fast Technique for Noninvasive Fetal ECG Extraction, *IEEE Transactions On Biomedical Engineering*, 58(2):227-230
- [2] Masoumeh Haghpanahi, David A Borkholder. 2011] Fetal ECG Extraction From Abdominal Recordings using Array Signal Processing, *IEEE Signal Processing Letters*, 18(3): 173-176.
- [3] J. L. Camargo Olivares, R Martin Clemente, S Hornillo Mellado, M M Elena, and I Roman. [2011] The Maternal Abdominal ECG as Input to MICA in the Fetal ECG Extraction Problem , *IEEE Signal Processing Letters*, 18( 3): 2161-164.
- [4] Shishir Dash, J Gerald Quirk, Petar M, Djuri C. [2014] Fetal Heart Rate Classification Using Generative Models, *IEEE Transactions On Biomedical Engineering*, 61(11): 2796-2805.
- [5] Liang Han, Xiu Juan Pu , Xiao Jun Chen.[2015] Method of fetal electrocardiogram extraction based on n-support vector regression", *IET Signal Processing*., 9(5):. 430-439.
- [6] MA Hasan, MI Ibrahimy, MBI Reaz.[2009] Fetal ECG Extraction from Maternal Abdominal ECG Using Neural Network, *J Software Engineering & Applications*, 2(5): 330-334.
- [7] Gizeaddis Lamesgin, Yonas Kassaw, and Dawit Assefa, "Extraction of Fetal ECG from Abdominal ECG and Heart Rate

Variability Analysis”, , *Advances in Intelligent Systems and Computing*, 334 : 65-76.

- [8] Mohammad Niknazar, Bertrand Rivet, Christian Jutten.[2013] Fetal ECG Extraction by Extended State Kalman Filtering Based on Single-Channel Recordings”, *IEEE Transactions On Biomedical Engineering*, 60( 5): 1345-1352.
- [9] Shuicai Wu A, Yanni Shen A, Zhuhuang Zhou, Lan Lin, Yanjun Zeng, Xiaofeng Gao, Research of fetal ECG extraction using wavelet analysis and adaptive filtering” *Comput Biol Med.* 2013 Oct;43(10):1622-1627.
- [10] George Georgoulas, Chrysostomos D. Stylios, Peter P. Groumpos.[2006] Predicting the Risk of Metabolic Acidosis for Newborns Based on Fetal Heart Rate Signal Classification Using Support Vector Machines, *IEEE transactions on biomedical engineering*, 53(5):.875-884.
- [11] Niranjana Krupa, Mohd Ali MA, Edmond Zahedi, Shuhaila Ahmed, Fauziah M Hassan.[ 2011] Antepartum fetal heart rate feature extraction and classification using empirical mode decomposition and support vector machine, *BioMedical Engineering OnLine*, 10:6
- [12] Hasan Ocak, Huseyin Metin Ertunc.[2013] Prediction of fetal state from the cardiotocogram recordings using adaptive neuro-fuzzy inference systems”, *Springer Neural Computing and Applications*,23( 6):1583-1589.
- [13] Hasan Ocak. .[2013] A Medical Decision Support System Based on Support Vector Machines and the Genetic Algorithm for the Evaluation of Fetal Well-Being, *Springer Journal of Medical Systems*,37:1-7.
- [14] Ben Fei, Jinbai Liu.[2006 ] Binary Tree of SVM: A New Fast Multiclass Training and Classification Algorithm, *IEEE Transactions On Neural Networks*, 17(3): 696-704.
- [15] Gjorgji Madzarov, Dejan Gjorgjevikj, “Multi-class classification using support vector machines in decision tree architecture” 2009 *IEEE Proceedings of EUROCON*, St.-Petersburg, pp.285-295.

# VOIP PERFORMANCE ENHANCEMENT THROUGH SPIT DETECTION AND BLOCKING

Saira Banu\*, K.M. Mehata

\*Assistant Professor (Senior Grade), B.S.Abdur Rahman University, Chennai. INDIA  
Dean(SCISM), B.S.Abdur Rahman University, Chennai, INDIA

## ABSTRACT

SPIT will burst as a major threat in the coming future because of the exponential growth of the VOIP users. spam callers like advertiser, telemarketers, prank callers make use of this VOIP for generating the bulk unwanted calls and messages. SPIT is hard to perceive than the mail spam. Injection of spam caller consumes more bandwidth and congest the network This paper proposes a pre acceptance method to detect the Voice spam based on the call intervals between the call and the online network reputation to identify the legitimate and non-legitimate caller. Realistic simulation results prove that our approaches are effective in discriminating the spammer from legitimate user.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

VoIP, SPIT, CDR, Call interval, Pre acceptance.

\*Corresponding author: Email: [saira.atham@gmail.com](mailto:saira.atham@gmail.com); Tel.: +91 9444420675

## INTRODUCTION

VoIP is used for data transition on IP based networks. With the growth of broadband connectivity, Voice over Internet Telephony is widely used for voice communication. The VoIP technique converts the analog signal to digital signal by using specific codecs. VoIP can support different types of data such as image, voice, video, and fax. The different services provided by VoIP are remote conference, call barring and call forwarding.

The important threats in VoIP network is known as SPIT (Spam over Internet Telephony). Advertiser and prank callers are using VoIP for sending unsolicited bulk calls like spam in emails. The spam detection in VoIP is more difficult than the detection of email spam because, the callee is directly connected by the incoming call. The Spam call makes irritation to VoIP users. For example a spam email that arrives to the inbox at 3a.m. will not disturb the particular user. But the spam call at 3a.m. will be irritated for the user. SPIT is the cost effectiveness for the spam callers. In VoIP network SPIT is most popular because of cheap hardware cost, low call fee, and there is no boundary for international calls [13]. True caller is an android app that identifies the caller and it also identifies the spammer if the number is already in the spam list. This spam list is manages by getting feedback from the caller.

In this paper we use the call intervals between calls and the online shopping pattern of the caller is taken for identifying the spam caller. The online reputation is based on the online shopping pattern of the caller. This method is used to find the spam callers without analyzing the data and without getting feedback from the user.

The CDR (Call Detail Record) contains the information of the call duration, call rate of the VoIP users. The details of call rate and call duration is used to generate the social network graph. The social network graph is used to find direct trust value. Direct trust between caller and the callee is the combination of the amount of time both users engaged in talking and the number of reciprocal calls between them, and number of unique callees of the caller [1]. This direct trust value is taken as the input for finding global reputation score. The global reputation value will be compared with the threshold value to identify the behavior of user. If the global reputation value is lesser than the threshold value then the caller is identified as the non legitimate caller, else he/she is identifying as legitimate caller. After finding the spam call, the information about the spam call will shared with other VoIP users to know whether they have interested to attend those calls or not interest. If they are not interest, in future those spam call will not establish through SIP (Session Initiation Protocol).



**Fig: 1.True caller app for identifying the spam call**

The major works are:

- Getting the caller-callee information from CDR.
- The call detail record has the complete information about the starting time, ending time, call duration, call type and the call routing information of each call.
- The call intervals between the calls can be computed using the starting time of the calls.
- The details of the incoming and outgoing sms are recorded in the call detail record.
- One Time Password(OTP) is a type of sms used for authentication purpose during online shopping, online money transaction etc.
- This OTP registered in the CDR is used for calculating the online network reputation of the caller.
- The call intervals between calls and the online network reputation together used to differentiate the spam callers from legitimate callers.
- This information is used to block the spam calls before the call is setup using the session initiation protocol.
- This scenario is simulated in a WiMax environment and the bandwidth consumed by the spammer for a single day is calculated.
- Thus SPIT detection and blocking will enhance the performance of the VOIP.

## **BIGDATA**

### **A. IP Telephony**

IP telephony is a technology, which compresses the telephone voice signal into data packets for transmission over the Internet. Protocols used in carrying the voice signals over the IP networks are referred to as Voice over IP (VoIP).

IP telephony moves away from the traditional circuit switched voice networks like Public Switched Telephone Networks (PSTN's) to a packet switched one where IP packets containing voice data are sent over the network.

The advantages of IP telephony over traditional telephony are lower infrastructure costs and lower costs per call (or even free calls). IP telephony comprises, independent of the protocols used, a multimedia plane and a signalling plane. The signalling plane is used for transporting the signalling information, during call setup. The media transport plane is used to carry voice data packets between IP telephony components [14].

### **B. VoIP (Voice over Internet Telephony)**

Voice over IP is a technology for transmitting voice packets on the existing IP network between two communicating parties which are connected to the Internet. Unlike PSTN networks, an IP network is packet

switched. In PSTN networks, when the calling party calls the called party, there exist a physical between the two parties. Then the parties can communicate with each other, and the circuit is reserved until they finish their communication.

In an IP network, all communication is carried out using IP packets. When a calling party communicates with a called party, the analog signals converted into digital signal, encoded, and it is packed into an IP packet at the transmitting end and converted back to analog signals at the receiving end.

### C. SIP (Session Initiation Protocol)

SIP, published as RFC 3261, is an application-layer protocol [14]. It can be used for creating, modifying and terminating sessions with one or more participants. The protocol which specifies a set of signalling messages for connection establishment and connection termination. It is used with other transport protocols like RTP (Real-time Transport Protocol), RTCP (Real time Control Protocol) for enabling voice-communication services between two parties. Requests are generated by the client and sent to the server. The server will process the requests and then sends back a response to the client.

SIP makes minimal assumptions about the transport protocol and this protocol provides reliability and it does not depend on TCP for reliability. Session Initiation Protocol depends on the Session Description Protocol to carry out the negotiation for codec identification. SIP supports for session descriptions that allow the participants to agree on a set of compatible media types. The services provided by SIP include:

- ❖ Location of user: Determination of the end system used for communication.
- ❖ Capabilities of user: Determination of the media and parameters to be used
- ❖ Call handling: Used for the transfer and termination of calls
- ❖ Call Setup: Establishing and ringing call parameters at both the calling and the called party

## FREQUENT SUBGRAPH MINING

The caller generates a call request to a callee through the SIP proxy server. The server checks for the registered user with the domain SIP register. After identifying the caller as a registered user the proxy server uses the SIP location server. If the callee is in the same domain as the calling party, then the proxy server forwards the call directly to the callee's end device.

The SIP messages are used for communicating between the client and the SIP server shown in **Figure- 2** are discussed below:

- ❖ INVITE used for inviting a user to a call
- ❖ BYE message used for terminating a connection between the two end points
- ❖ ACK is used for reliable exchange of invitation messages
- ❖ OPTIONS message used for getting information about the capabilities of call
- ❖ REGISTER gives the information about the location of a user to the SPIT registration server
- ❖ CANCEL message is used to terminating the sessions

## LITERATURE SURVEY

Spam over Internet Telephony is a major problem for VoIP users. It affects the private life of VoIP customers and their correspondents.

Ricardo Morla et al, [1] proposed novel content independent, non-intrusive approaches based on caller trust and reputation to block spam callers in a VoIP network. This approach is based on interaction rate, call duration and caller out-degree distribution. It is used to establish a trust network between VoIP users and computes the global reputation of a caller across the network. Gamal A. Ebrahim [2] followed to reduce VoIP Spam by ranking VoIP callers based on a set of parameters. The parameters are caller's reputation, the feedbacks (if any) collected from the called party, and whether the callee responds the call or ignores the call. In Addition, a set of dummy directory numbers is introduced in the callee's domain and these numbers work as traps for certain kinds of VoIP spammers who try to guess the directory numbers in the domain of victim by dialling random directory number. Based on this the spam callers are identified.

Dirk Lentzen et al, [3] designed a system that combines with the advantages of signalling-based SPIT

prevention and audio content-based SPIT detection. This is achieved by calculating spectral audio fingerprints and detecting SPIT calls with identical or similar voice data. The result of the fingerprint comparison is used to generate black list entries. This system integrated into existing VoIP environments and SPIT filter systems. Kentaroh Toyoda et al, [4] proposed a multi-feature call pattern analysis with unsupervised Random Forests Classifier. It is one of the classification algorithms.

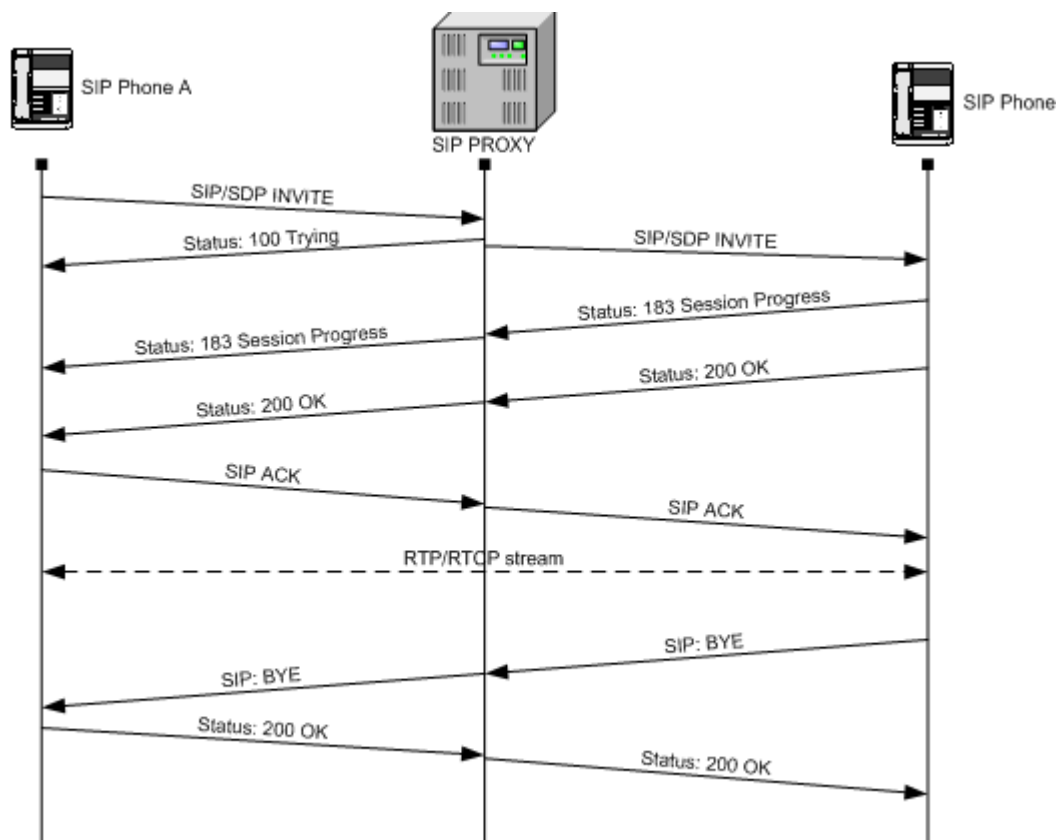


Fig: 2 . Basic SIP Messages

There are two features in Random Forests Classifier. That is BDR (Bi-Direction Ratio) and IOR (Incomings/Outgoings Ratio). BDR indicates the ratio of the intersection of Incomings and Outgoings to the number of Outgoings and IOR represents the ratio of Incomings to the union of Incomings and Outgoings. Based on this classifier the spammer is detected.

Yan Bai et al, [5] followed a user-behavior-aware anti-SPIT technique. It is implemented at the router level for detecting and filtering SPIT. Based upon this rationale technique voice spammers behave significantly different from legitimate callers because of their revenue-driven motivations. This technique defines and combines three features developed from user behavior analyses. This approach is applicable for detecting and filtering both machine-initiated and human-initiated spam calls. He Guang-Yu et al, [6] a detection and prevention method based on feedback judgment, the SPIT problem can be resolved. The improved inference algorithm embodies the characteristic of SPIT behavior and also it's reflects the weight relation of factors which influence the result. The incremental learning algorithm gives the Real-time trust and reputation. The algorithm integrates the trust with the reputation and it makes a comprehensive evaluation of the SPIT.

## PROPOSED APPROACH

Call intervals method in VoIP users is a mechanism for detecting the SPIT (Spam over Internet Telephony) callers. The existing approaches are based on content analysis, user involvement, list based (white, black, gray) and reputation based. But Call interval method does not depends on above list. Call interval is the observation of the difference in time intervals between the consecutive calls of the caller. If the mean call intervals is lesser than the threshold value and if the online network reputation is also lesser than the minimum value then the caller is fully decided as a spam caller. The Wimax environment with the injected Spam caller is simulated for 9 hours and the consumed bandwidth is calculated. The performance of the WiMax scenario is found to degrade when simulated with the spam caller.

Caller reputation method is follows:

- ❖ Collecting CDR information
- ❖ Social network strength
- ❖ Direct trust
- ❖ Global reputation
- ❖ Automatic Threshold
- ❖ User involvement

## CDR (CALL DETAILS RECORD)

A call detail records (CDR) are unavailable because of the privacy. We get the CDR data from crawdad website and we preprocessed it. CDR contains the data about data that is Meta data. It contains data fields that describe a specific instance of a telecommunication transaction, but it does not include the content of the transaction. CDR is a file that contains information about recent system usage like identities of sources, the identities of destinations, the duration of each call. For billing purpose they maintain the information of the amount billed for each call, the total usage time in the billing period, the running total charged during the billing period and the total free time remaining in the billing period. The way of simplistic example, CDR describing a particular phone call might include the phone numbers of the calling and receiving parties, the start of the call, and duration of that call. The call detail records contain attributes such as:

- ❖ the phone number receiving the call and the phone number of the subscriber originating the call
- ❖ the starting time of the call and call duration of the call
- ❖ the identification of the telephone exchange
- ❖ call type and any fault condition encountered

Table 1. Structure of CDR

Table 1. Structure of CDR

ID	Calling Party	Called Party	Date and Time	Call Duration	Call Type	Fault Condition
123	9710410829	9003095132	24-01-2012 12.05pm	5min	Voice	Success
114	9965320235	9443799049	24-01-2012 1.30pm	7min	Voice	Success
189	9965245068	8695976414	24-01-2012 1.30pm	10min	Voice	Success
117	9597994023	9940527033	26-01-2012 03.00pm	15min	Voice	Success
145	9965216667	9094862745	27-01-2012 10.15am	8min	Voice	Success
179	9971213141	9003095132	27-01-2012 01.30pm	3min	Voice	Success

The TRAI (Telecom Regulatory Authority of India) and CRM (Customer Relationship Management) contains the information about age of the Subscriber Identity Module (SIM) of user. The legitimate callers will not change the number frequently. But the spam callers have the behavior of changing the number frequently.

## SOCIAL NETWORK FEATURES

A caller -callee social network graph is modeled in R Studio. R studio is advanced part of R tool. We can get the graph exactly in R studio. R Studio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for debugging, plotting, history and workspace management [15]. Whenever the new user enter into the VoIP network the social network graph will be analyzed between caller S and callee R if S calls R at least once.

The number of phone calls received at an exchange or call center in an hour;

X has a Poisson Distribution with parameter  $\lambda$  and

$$P(X = k) = f(k) = e^{-\lambda} \lambda^k / k! , k = 0, 1, 2, \dots$$

The random variable X indicates the number of successes in the whole interval.  $\lambda$  indicates the mean number of successes in the interval.



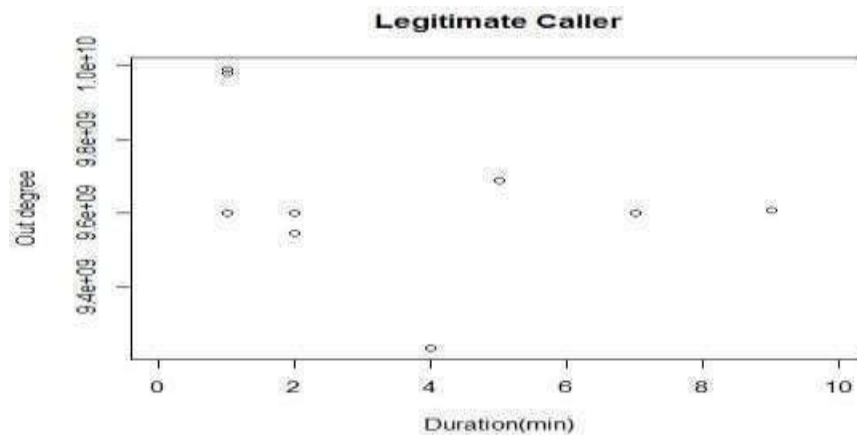
The following parameters are usually used for analyzing the relationship between the caller and the callee:

- ❖ Indegree
- ❖ Outdegree
- ❖ Call rate
- ❖ Call duration
- ❖ Call intervals between the calls

In-degree represents the number of other different users calling this user and out-degree represents the number of calls made from this user to other unique users. The spam callers have the behavior of uni-direction of communication so, it is unbalanced in/out degree. But, the legitimate callers have bi-directional relationship so that it is balanced in/out degree.

Call rate is the total number of calls made/ received by the caller and it can be categorized as in and out call rates. Call rate says the repetitive behavior of user based on the higher call rate, more frequently user calls the same people. Spam caller have higher call rate only for outgoing, they don't have the repetitive behavior. Call duration is the total duration of the calls made or received by the user. The call intervals is the intervals between the consecutive calls.

The legitimate callers have higher amount of call duration with their social network and have small amount of duration with outside of their social network as shown in [Figure- 3](#). Legitimate callers don't call the unknown numbers. But spam caller have the behavior of making calls to the unknown person. They make large number of calls with the small amount of call duration. The legitimate callers will call their social network repeatedly and they get reciprocal calls from their social network circle. So the number of unique outgoing calls i.e outdegree of the legitimate caller will be less.



**Fig: 3. Social network graph for legitimate user**

The SPIT callers have lower amount of call duration with outside of their social network as shown in [Figure- 4](#). Spam callers mostly call the unknown numbers. They make large number of calls with the small amount of call duration. The Spam callers will call unknown people repeatedly and they get very rare reciprocal calls. So the number of unique outgoing calls i.e outdegree of the spammer will be more.

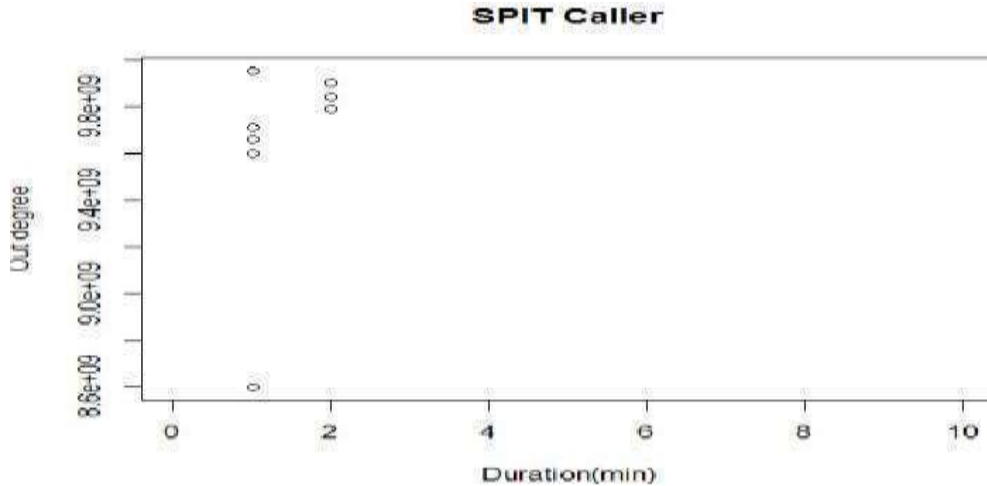


Fig: 4. Social Network Graph of a SPIT caller

### THRESHOLD VALUE

The analysis shown in **Figure- 5** is a research done by San Francisco State University which includes A-B call pairs of a legitimate caller. The scanner locates the call information on a spectrogram by pixel coordinates and separates the call sequences by a longer breathing interval. The difference between the time of each call is calculated and plotted. The mean and the standard deviation of the recorded calls are calculated by using the iterative procedure which tries to fit the distribution to the Gaussian plus a uniform distribution. The calculated average time is 135.4 seconds with a standard deviation of 8.8 second. The distribution ranges between 100 and 150 seconds which is clear from the graph in **Figure- 3**. The threshold value is set as 100 seconds for the call interval method of finding the spammer in the IP telephony. The mean call interval of the caller is checked by the SIP server before setting the call connection. The caller with lesser mean call interval is considered as spammer and such calls are blocked before setting up the call i.e the pre acceptance method.

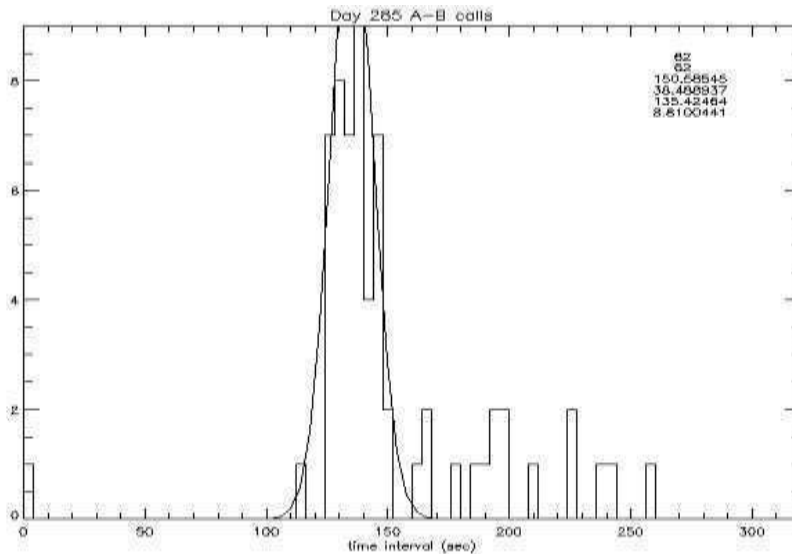


Fig: 5. Distribution of time intervals for Legitimate caller monitored for 285 days

### TYPES OF SMS MESSAGE

There are four classes of SMS message . These Classes identify the importance of an SMS message and also the location where it must be stored.

- **Class 0**  
This type of SMS message is displayed on the mobile screen without being saved in the message store or on the SIM card; unless explicitly saved by the mobile user.
- **Class 1**  
This message is to be stored in the device memory or the SIM card (depending on memory availability).
- **Class 2**  
This message class carries SIM card data. The SIM card data must be successfully transferred prior to sending acknowledgment to the service center. An error message is sent to the service center if this transmission is not possible.
- **Class 3**

This message is forwarded from the receiving entity to an external device. The delivery acknowledgment is sent to the service center regardless of whether or not the message was forwarded to the external device.

The OTP (one time password ) will be any one of the class based on the developer. A **one-time password (OTP)** is a password that is valid for only one login session or transaction, on a computer system or other digital device. A spam caller who makes bulk calls through IP telephony for advertisement or in call center will never use that number for authentication purpose. So IP Telephony numbers that receives OTP sms is taken as one of the parameter for segregating the legitimate caller.

## SPIT DETECTION ALGORITHM

*Algorithm . SPIT Detection*

1. Start the call detail log record
2. Assign call detail vector  $X$  for each caller
3. Define Matrix  $M$  with call interval element
4. Calculate  $\mu = \sum M$  of  $X$
5. Compare with the threshold value  $\beta$
6. If  $\mu < \beta$  then calculate the online reputation score
7. { If online reputation score is equal to NULL then report as spam caller
8. Else report as legitimate caller }
9. Else report as legitimate caller
10. Stop

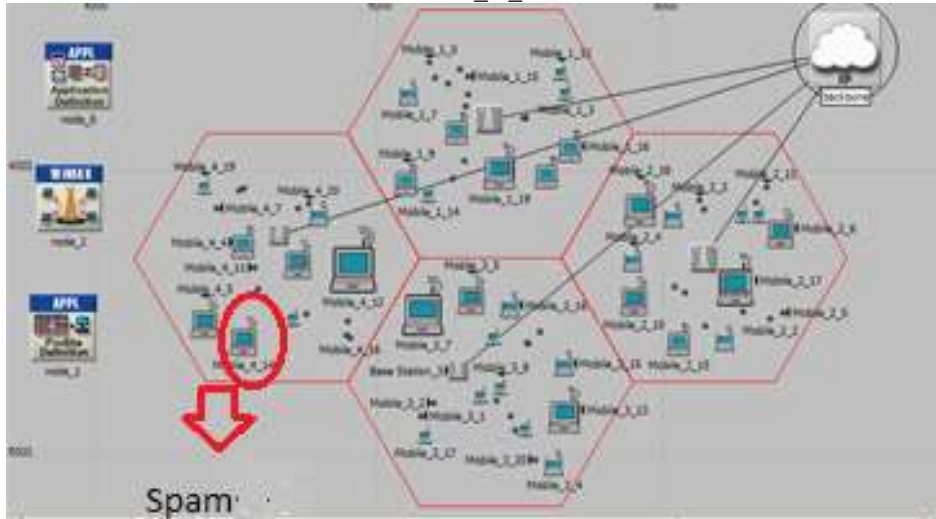
## SIMULATION RESULTS

The spit detection algorithm is stimulated in the WiMax environment . The Wimax wireless technology that uses the VOIP for transmitting the voice with the existing hardware and software. When spamming nature is introduced to one of the node in the network the performance of the network is found to degrade. The spammer is detected using the above algorithm and the saved bandwidth is calculated.

This scenario for the bandwidth calculation is a wireless network with four base stations and twenty subscriber stations (fixed node) is shown in fig 6. The topology consist of geographical overlay of four cells ,each with radius 3km and 20 subscriber stations. The nodes are placed in a random manner. Out of the 20, one of the node circled in red is simulated to generate unwanted spam calls. Main considerations made in our simulation while deploying wireless network are technology and topology.

**Technology :** WiMAX technology with subscriber node transmission power 0.5 w and base station transmission power 0.5w.

**Topology :** The simulated topology consist of 4 hexagonal cells with cell radius of 3km.Number of subscriber stations per cell is 20.Base stations are modeled with wimax\_3section\_bs\_atm2\_ethernet2\_slip4\_wlan\_router. The subscriber stations are modeled with wimax\_ss\_wkstn.



**Fig: 6 . WiMax Scenario taken for simulation**

This topology is designed for VOIP application that is both voice application. The configurations used in the above simulated scenario are profile config, application config and wimax config.

**Application Config :** This specifies the various application used in the project. The various application names are web browsing, FTP, databases, HTTP, remote login, voice and video conferencing. In the above simulation we have taken the voice and video conferencing.

**Profile Config:** It is used to create the traffic pattern for the application defined in the application config.

**Wimax config :** It is used to store profiles of physical and service classes which can be referenced by all wimax nodes in the network.

**Table 2.Simulation Parameter**

Parameter	Value
Network interface type	Phy/Wireless Phy/OFDMA
Propagation model type	Propagation/ OFDMA
Medium Access Control type	Mac/802_16/Base Station
Routing protocol	VOIP
Antenna model	Antenna/Omni Antenna
Link layer type	Logical Link layer
Frame size (msec)	5 (msec)
Duplex scheme	TDD
Packet Rate	4 packet/s

**Bandwidth Utilized by the spammer**

1 Hour	7200KB
3 hours	21600KB
9 hours	64800KB

1 day(app 12 hours )

86 MB

## CONCLUSION

The call interval method of spam detection is computed using Algorithm 1. Our system is used to analyze the call detail record to find whether the user is spam caller or legitimate caller. The CDR contains the private information of who, when and duration of every call. The CDR also has the information of all the sms delivered and sent of each user. The blocking of the spam caller before setting the call will save the bandwidth and avoid congestion of the network.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None.

## REFERENCES

- [1] Ricardo Morla, Muhammad Ajmal Azad. [2013] "Caller-REP: Detecting unwanted calls with caller social strength, ELSEVIAR, Pages 219–236.
- [2] Gamal A. Ebrahim. [2013] A VoIP SPAM Reduction Framework, *IEEE*.
- [3] Dirk Lentzen, Gary Grutzek, Heiko Knospe, Christoph Porschmann. [2011] Content-based Detection and Prevention of Spam over IP Telephony - System Design, Prototype and First Results, *IEEE*
- [4] Kentaroh Toyoda, Iwao Sasase. [2013] SPIT Callers Detection with Unsupervised Random Forests Classifier, *IEEE*
- [5] Yan Bail, Xiao Su, Bharat Bhargava.[2009] Detection and Filtering Spam over Internet Telephony - A User-behavior-aware Intermediate-network-based Approach, *IEEE*
- [6] He Guang-Yu, Wen Ying-You, and Zhao Hong. [2008] SPIT Detection and Prevention Method in VoIP Environment , *IEEE*
- [7] Dongwook Shin. [2006] Progressive Multi Gray-Levelling: A Voice Spam Protection Algorithm, *IEEE*
- [8] Mohammad Hossein Yaghmaee Moghaddam, Mina Amanian, Farideh Barghi, and Hossein Khosravi Roshkhari.[2014] A Survey of Different SPIT Mitigation Methods and a Presentation of a Comprehensive SPIT Detection Framework, *International Journal of Machine Learning and Computing*, 4( 2)
- [9] Christoph Sorge, Jan Seedorf. [2009] A Provider-Level Reputation System for Assessing the Quality of SPIT Mitigation Algorithms, *IEEE ICC*
- [10] Tetsuya Kusumoto, Eric Y. Chen, Mitsutaka Itoh. [2009 ] Using Call Patterns to Detect Unwanted Communication Callers, *IEEE*
- [11] Vijay A. Balasubramaniyan, Mustaque Ahamad, Haesun Park. [2007] Call Rank: Combating SPIT Using Call Duration, Social Networks and Global Reputation", CEAS, Fourth Conference on Email and Anti Spam, August 23, 2007
- [12] Fei Wang, Yijun Mo, Benxiong Huang. [2007 ] P2P-AVS: P2P Based Cooperative VoIP Spam Filtering, *IEEE*
- [13] <https://books.google.co.in/books?id=WN4JANNzkJoC&p>
- [14] [http://icsa.cs.up.ac.za/issa/2005/Proceedings/Research/091\\_Article.pdf](http://icsa.cs.up.ac.za/issa/2005/Proceedings/Research/091_Article.pdf)
- [15] <http://www.rstudio.com/products/rstudio>

# LOW POWER ARITHMETIC CIRCUITS USING FINFET DEVICE IN 32NM TECHNOLOGY

V. M. Senthil Kumar<sup>1\*</sup>, S. Saravanan<sup>2</sup><sup>1</sup>Department of Electronics and Communication Engineering, Vivekanandha College of Engineering for Women, Elayampalayam, Tiruchengode, Tamil Nadu, INDIA<sup>2</sup>Department of Electrical and Electronics Engineering, Muthyammal Engineering College, Rasipuram, Tamil Nadu, INDIA

## ABSTRACT

**Aim:** A low power computing device design is the objective of the work. The objective is met using FinFET and inexact computing methods. A FinFET based arithmetic circuits for the processing element units in DSP application is to be designed and proposed. **Materials and methods:** The circuits are modelled using predictive technology models in 32nm FinFET technology. The proposed arithmetic circuits are a half adder, a full adder and a multiplier. A 4-2 compressor based multiplier is proposed in this paper which reduces the number of operands and partial products. Using compressors for the multiplier reduces the number of interconnects and components. **Results:** The proposed FinFET based circuits reduces the leakage current which ultimately results in the reduction of power consumption. As CMOS circuits performance reduces when fabricated below 45nm technology, FinFET devices is the alternative solution. This work embasis on the design of circuits using FinFET. **Conclusion:** The proposed circuits are compared with the existing counterparts and found that the performance improvement is 96% on average with the existing CMOS circuits. The modelling and simulations were carried out using Synopsis HSPICE.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

FinFET, Adder, Multiplier,  
Compressor, Processing  
Element\*Corresponding author: Email: [vmspraneeth@gmail.com](mailto:vmspraneeth@gmail.com); Tel.: +91 9600562621

## INTRODUCTION

The CMOS based circuits for signal and image processing algorithm was dominating the integrated circuit world upto 90nm technology. The CMOS performance degrades due to leakage current and lower output swing below 90nm. The reduction of supply voltage is not effective. This reduces the driving capability when implemented for arithmetic and logic blocks used in Digital Signal Processing (DSP). The other issues involved are the Short Channel Effects, (SCE), Sub-Threshold Leakage (STL) device variations and gate dielectric leakage [1]. So the gate terminal control is to be taken care which is possible only by having multigate device. The multigate devices also work in different modes [2] and mock the operation of CMOS in short gated mode but with good operability conditions. When arithmetic circuits were designed by these devices the performance is better. Several works are carried out in the implementation of arithmetic circuit using FinFET. The result shows that the leakage and delay analysis are improved when compared to CMOS in an array multiplier is implemented. A low leakage MUX/XOR logics are implemented using symmetric and asymmetric FinFET in literature [3]. A 65% improvement in leakage and 45% in delay is reported.

## INEXACT COMPUTING/COMPRESSION

In digital processing where inexact computing becoming an attractive paradigm and VLSI implementation becomes much easier. Then complex arithmetic operations like Fourier analysis or discrete wavelet transform are to be performed the inexact computing becomes more efficient. The inexact computing circuit becomes viable for real time circuits through compressors. The truth table of the compressors is given in Table 1. It can be observed

that the carry output is equal to  $c_{in}$  in 24 out of 32 states while the carry is simplified for other 8 outputs. The value. The gate level compressor design for the truth table in table 1 is given in Figure 1(a). The respective multiplier is given in Figure 1(b).

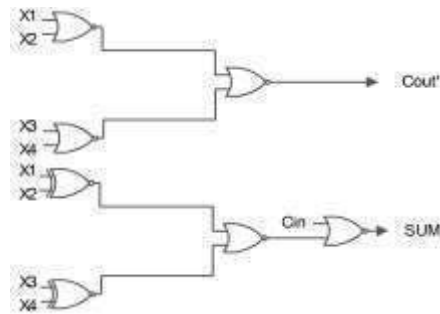


Figure. 1(a): Gate level Implementation of Compressor type 1

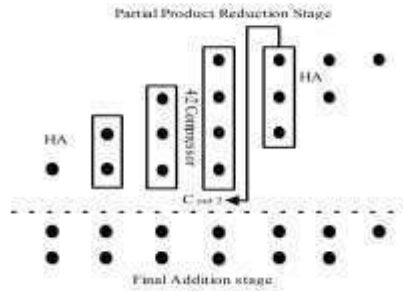


Figure. 1(b): Existing Compressor based Multiplier type 1

An another type of approximate compressor is shown in Figure 2(a) and its multiplier in Figure 2(b). The second compressor reduces the approximate error when compared to the first one. The truth table is given in Table 2.

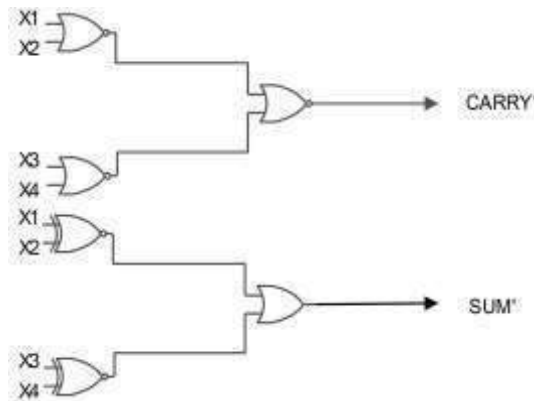


Figure. 2(a): Gate level Implementation of Compressor type 2

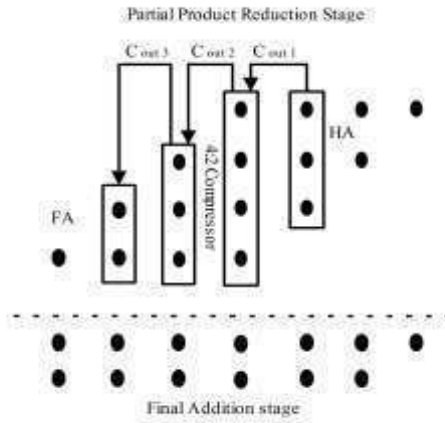


Figure. 2(b): Existing compressor based Multiplier type 2

TABLE I. TRUTH TABLE OF 4-2 COMPRESSOR TYPE 1

Cin	X4	X3	X2	X1	Cout	Carry	Sum
0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	1
0	0	0	1	0	0	0	1
0	0	0	1	1	1	0	0
0	0	1	0	0	0	0	1
0	0	1	0	1	1	0	0
0	0	1	1	0	1	0	0
0	0	1	1	1	1	0	1
0	1	0	0	0	0	0	1
0	1	0	0	1	0	1	0
0	1	0	1	0	0	1	0
0	1	0	1	1	1	0	1
0	1	1	0	0	0	1	0
0	1	1	0	1	1	0	1
0	1	1	1	0	1	0	1
0	1	1	1	1	1	1	0
1	0	0	0	0	0	0	1
1	0	0	0	1	0	1	0
1	0	0	1	0	0	1	0
1	0	0	1	1	1	0	1
1	0	1	0	0	0	1	0
1	0	1	0	1	1	0	1
1	0	1	1	0	1	0	1
1	0	1	1	1	1	1	0
1	1	0	0	0	0	1	0
1	1	0	0	1	0	1	1
1	1	0	1	0	0	1	1
1	1	0	1	1	1	1	0
1	1	1	0	0	0	1	1
1	1	1	0	1	1	1	0
1	1	1	1	0	1	1	0
1	1	1	1	1	1	1	1



Compressor circuits were initially used for generating partial products in BiCMOS technology [4] and BiCMOS circuit provided better power consumption but failed in the reduction of leakage current. Pass transistors were used for the design. A pass transistor based CMOS multiplier using compressor was reported in the literature [5] and [6]. The compressors have different configuration and best suited for the systems where the noise components are more.

TABLE II. TRUTH TABLE OF 4-2 COMPRESSOR TYPE 2

Cin	X4	X3	X2	X1	Cout'	Carry'	Sum'
0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	1
0	0	0	1	0	0	0	1
0	0	0	1	1	0	0	1
0	0	1	0	0	0	0	1
0	0	1	0	1	1	0	0
0	0	1	1	0	1	0	0
0	0	1	1	1	1	0	1
0	1	0	0	0	0	0	1
0	1	0	0	1	1	0	0
0	1	0	1	0	1	0	0
0	1	0	1	1	1	0	1
0	1	1	0	0	0	0	1
0	1	1	0	1	1	0	1
0	1	1	1	0	1	0	1
0	1	1	1	1	1	0	1
1	0	0	0	0	0	1	0
1	0	0	0	1	0	1	0
1	0	0	1	0	0	1	0
1	0	0	1	1	0	1	0
1	0	1	0	0	0	1	0
1	0	1	0	1	1	1	0
1	0	1	1	0	1	1	0
1	0	1	1	1	1	1	0
1	1	0	0	0	0	1	0
1	1	0	0	1	1	1	0
1	1	0	1	0	1	1	0
1	1	0	1	1	1	1	0
1	1	1	0	0	0	1	0
1	1	1	0	1	1	1	0
1	1	1	1	0	1	1	0
1	1	1	1	1	1	1	0

Even though the approximate computing reduces accuracy in hard real time systems where speed is pre-dominant factor accuracy can be compromised a little bit. The power and delay reduction can be achieved with significant accuracy [7]. Chip-Hong Chang et al [8] proposed a carry generator circuit to implement a 4-2 and 5-2 compressor cell to operate in tree structured parallel multiplier which operates at very low supply voltage i.e below 1V. Momeni et al [9] have proposed an approximate compressor for multiplication. The proposed design is implemented in an image processing application using Dadda multiplier. The paper reports that the two new approximate 4-2 compressors significantly reduces the power dissipation, delay and transistor count compared to an exact design. In multiplication intensive computing the compressors seem to be leading replacement for existing exact multiplication methods when multi operands are involved [10]. Realization of the compressors is normally based on XOR/XNOR gates. The decomposition of the XOR/XNOR leads to an optimized design presented in the literature [10]. Improvement in terms of delay about 17%, power by 13% and power-delay-product by 30% is reported. Apart from the modified versions in the existing compressors, high performance 5:2 compressor is designed and arithmetic circuits [11] in the literature [12] have been proposed. The design limits the

carry propagation to a single stage. Apart from the above research hybrid methods of combining magnetic tunnel junction (MTJ) and complementary metal–oxide–semiconductor (CMOS) are done in [13] for a compressor design. Joseph Whitehouse and Eugene John [14] proposed a FinFET based array multiplier with different technologies like 20nm, 16nm, 14nm, 10nm and 7nm. The reported performance is for the implementation of an array multiplier using standard cell 28 transistor full adders.

## REDUCTION OF LEAKAGE CURRENT AND LEAKAGE POWER USING FINFET TECHNOLOGY

In conventional CMOS devices below 45nm there are problems in driving capability due to leakage currents and other second channel effects like VT roll off and drain induced barrier lowering. At the onset of FinFET technology many works started moving towards FinFET technology from CMOS. Especially in the short channel territories like 65nm and 45nm FinFET promises to deliver superior levels of scalability needed to design integrated circuits for systems.

### Device Structure of FinFET

The device structure of FinFET [15] is shown in Figure 3 (a) and (b). The name signifies a fin like structure in the FET device formed on thin Silicon On Insulator (SOI) finger termed fin. The fin is protected by a silicon fin nitride deposit on a thin pad oxide during gate poly-SiGe etching. The fin acts as a channel and terminates both sides of source and drain. The single gate stacking arrangements on top of vertical gates allows three times surface area for the electrons to travel. Gate work-function tailoring is essential to adjust the threshold voltage. Therefore, for the gate material poly-SiGe has been chosen. The crucial geometric device dimensions are shown in Figure 3. A single poly silicon layer is deposited over a fin. Here fin itself acts as a channel and it terminates on both sides of the source and drain. FinFET consists of two gates: front gate and back gate, source and drain. Front gate is used to control the channel conduction and back gate is used to control the threshold voltage. The labels in the figure stands for  $L_g$ : printed gate length,  $L_{eff}$ : effective gate length which is determined by the distance of the junctions,  $T_{fin}$ : Height of the fin,  $W_{fin}$ : Width of the fin which is the distance between the gate oxides of the two gates.

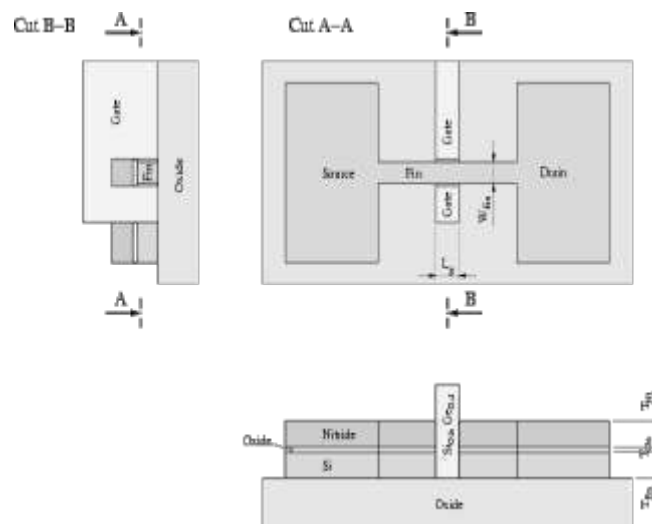


Figure. 3(a): Views of the FinFET Layout

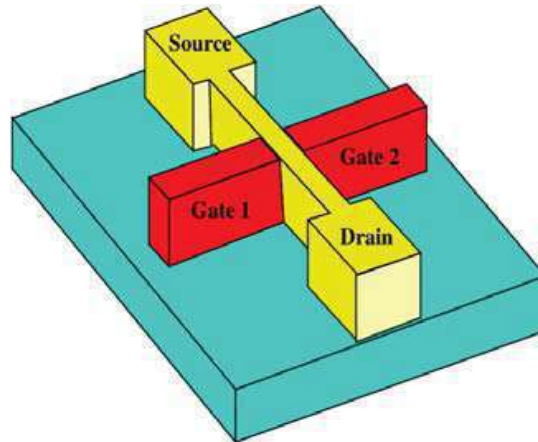


Figure. 3(b): FinFET Device Structure

The tradeoff of designing FinFET's as discrete components due to the issues involved in the manufacturing of similar GATE is rectified by integrated circuit design. From integrated circuits the FinFET technology is moved to the world of System on Chip (SoC) architecture where the complete system is embedded in integrated circuits. As SoCs require very much higher levels of integration and good driving capability FinFET in 35nm technology will be suitable.

### PROPOSED COMPRESSOR BASED MULTIPLIER USING FINFET

The implementation of arithmetic circuit using FinFET is effective when analyzing the works available in the literature [16-18]. The conventional method have disadvantage due to the carry propagation and low speed of operation. On the other hand the CMOS based compressor circuit lags in driving capability due to leakage current. These problems are addressed in this paper by proposing a novel FinFET based compressor for multiplier. The design is well suited for any real time DSP application where the driving current should be high. Proposed design can even work under the supply voltage of 1V with good driving capacity. The circuit is faster and delay is reduced. The proposed architecture of Compressor based multiplier using FinFET is given in Figure 4.

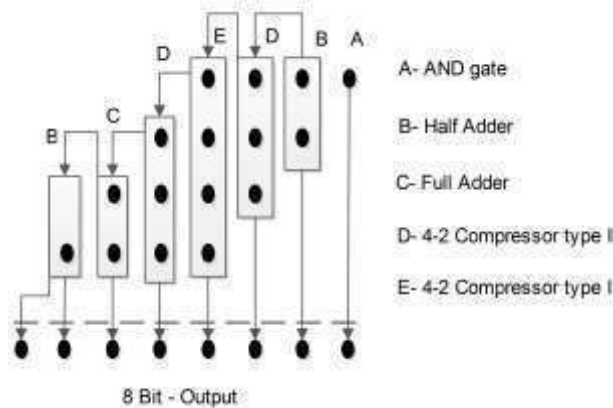


Figure. 4: Architecture of Compressor based Multiplier using FinFET

For the implementation of the multiplier architecture in Figure 4, 2 half adder, 1 full adder and 3 compressors are required.

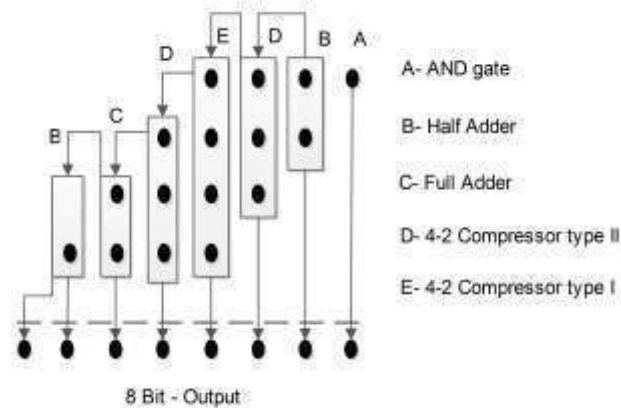


Figure. 5(a): Architecture of Compressor based Multiplier using FinFET

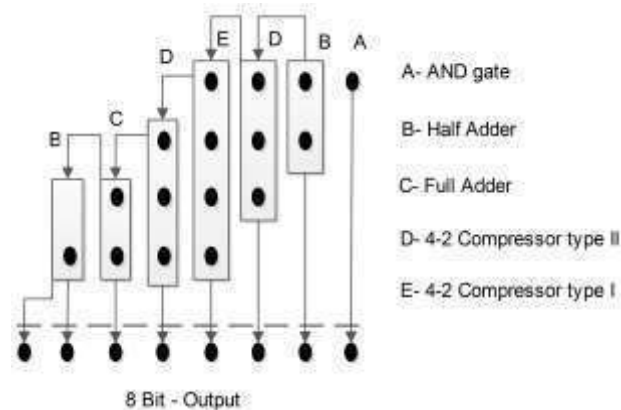


Figure. 5(b): Architecture of Compressor based Multiplier using FinFET

The proposed adder structure is designed using transmission gate which selectively block or pass the data from the input to the output. The proposed 6T half adder and 10T full adder is given in Figure 5.(a) and Figure 5.(b) respectively. The less number of transistors reduces the area and consume lesser power. When considering driving capability the proposed design in FinFET for the lesser transistors have better performance compared to the CMOS counterpart. The CMOS pass transistor logic have lesser output voltage swing due to the threshold voltage defeat problem is rectified in this design. The FinFET based adder not only reduces the power consumption, but may also lead to better switching in the case of cascaded operation such as ripple carry adder. A low  $V_{DD}$  operation is also possible, which can work at 0.7V.

## IMPLEMENTATION

A FinFET based AND gate implementation and shown in Figure 6(a) and 5(b). For the implementation predictive technology model for 32nm is used.

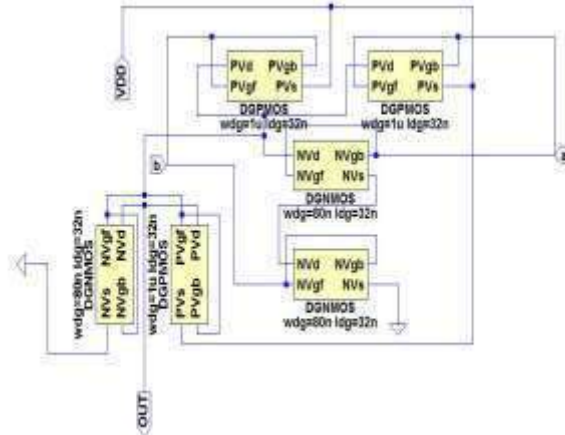


Figure. 6(a): FinFET AND Gate

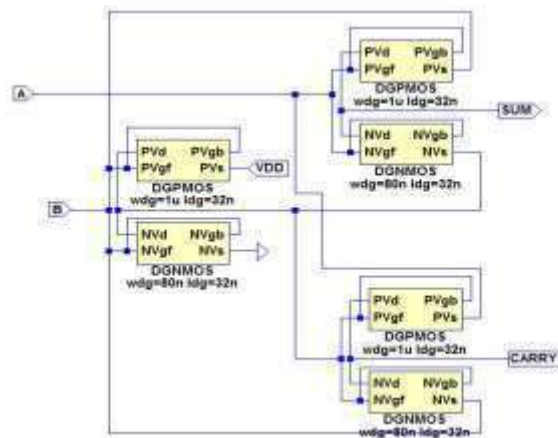


Figure. 6(b): FinFET Proposed Half Adder

The implementation of the proposed full adder is given in Figure 7. The supply voltage applied to the circuit is 1V and the proposed full adder implementation was carried out using FinFET Predictive Technology Model (PTM) in 32nm.

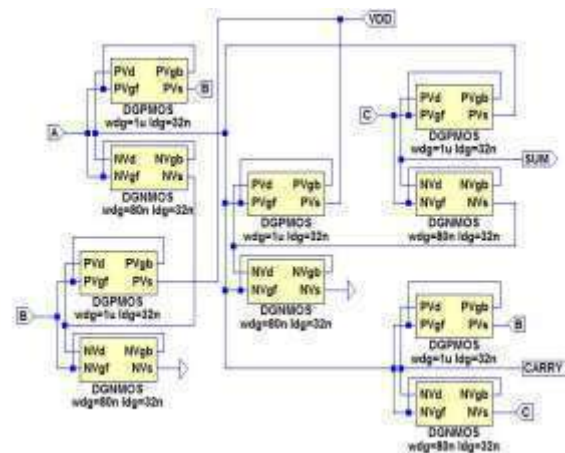


Figure. 7: FinFET Proposed Full Adder

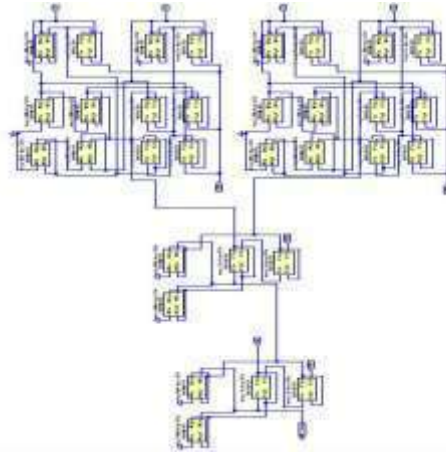


Figure. 8(a): FinFET Compressor type 1 (Sum block)

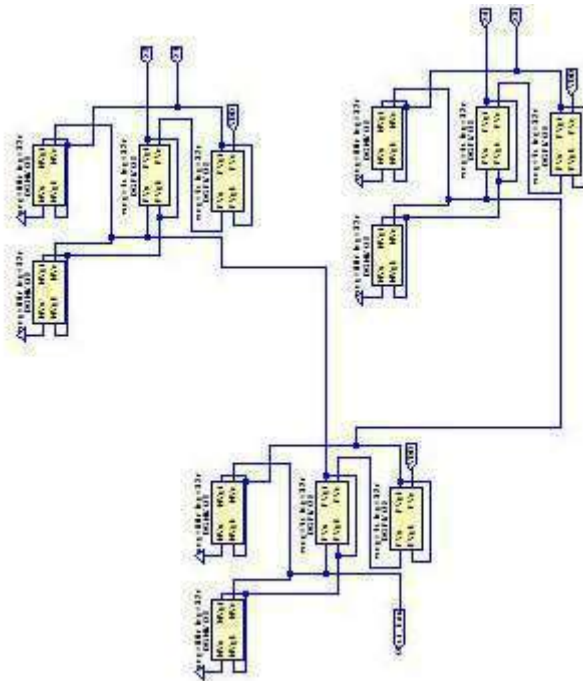


Figure. 8(b): FinFET Compressor type 1 (Carry block)

Figure 8(a) and 8(b) shows the FinFET based Compressor type 1 sum and carry block. Similarly compressor type 2 circuits are given in Figure 9. 4-bit implementation is carried out to validate the performance with respect to both compressor types.

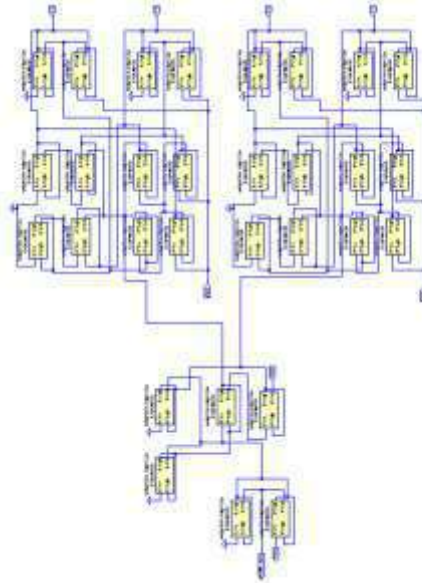


Figure. 9(a): FinFET Compressor type 2 (Sum block)

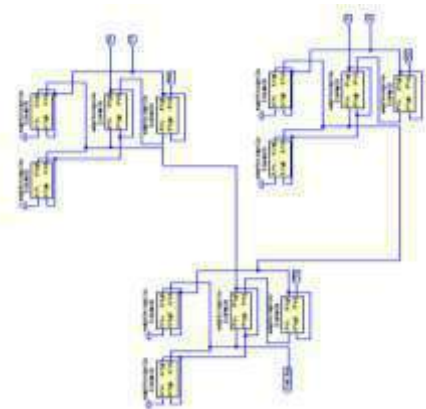


Figure. 9(b): FinFET Compressor type 2 (Carry block)

The conventional array multiplier implemented is shown in Figure 10. The compressor type 1 and type 2 based systolic array multiplier are given in Figure 11 and Figure 12 respectively.

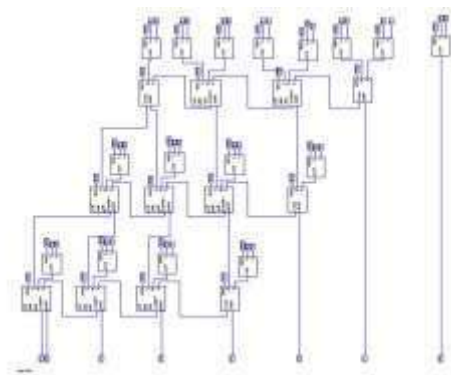


Figure. 10: Conventional Array Multiplier for Systolic Architecture

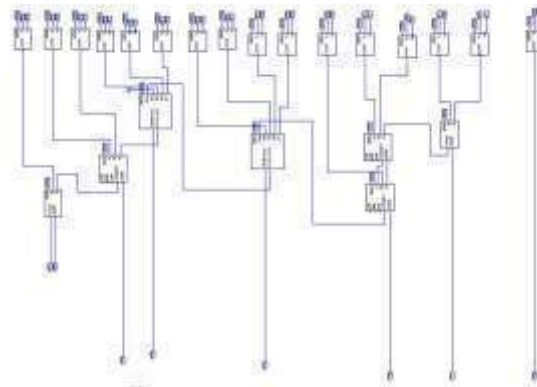


Figure. 11: Compressor Type 1 based Array Multiplier for Systolic Architecture

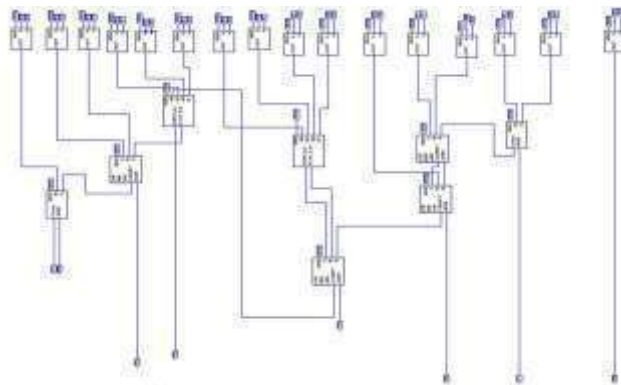


Figure. 12: Compressor Type 2 based Array Multiplier for Systolic Architecture

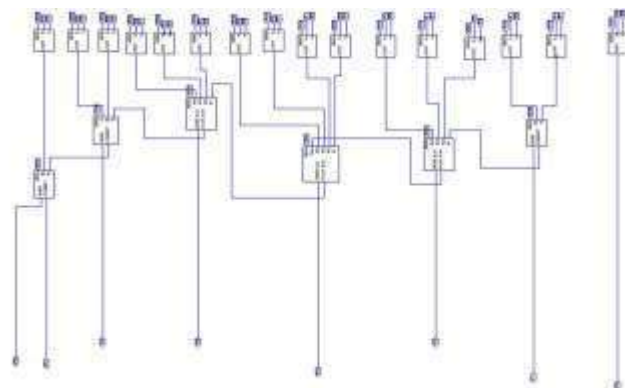


Figure. 13: Proposed Compressor Type 1 and Type 2 based Array Multiplier for Systolic Architecture

The proposed Compressor based Array multiplier for Systolic architecture is shown in figure 13. The proposed multiplier is a hybrid type, In which compressor type 1 and type 2 are used for implementation with the proposed 6T half adder and 10T full adder.



## RESULT AND DISCUSSION

The power consumption of different components implemented using CMOS and FinFET is shown in the Table 3. The average power is measured and tabulated. From the results it is observed that the power dissipation in FinFET is 95.46% compared to CMOS. Especially in comparison of CMOS with FinFET the half adder and full adder the reduction is nearly of 90% and 96 % respectively. The proposed method further improvises about 60% reduction in power for CMOS half adder and about 78% reduction for FinFET. Similarly for full adder its 85% in CMOS and 77% in FinFET. The results of compressor type 1 and compressor type 2 shows the improvement in power reduction by 97% and 98% while using FinFET instead of CMOS. The current analysis is given in the Table 3. for all the components of the proposed design.

**TABLE III. POWER ANALYSIS**

COMPONENT	AVERAGE POWER in $\mu$ W		% improvement in Power for proposed Circuits
	CMOS	FinFET	
AND GATE	49.4	1.24	97.49
HALF ADDER	22.9	2.41	89.48
PROPOSED HALF ADDER	9.1	0.528	94.20
FULL ADDER	81.2	2.55	96.86
PROPOSED FULL ADDER	12	0.562	95.32
COMPRESSOR1	49.4	1.24	97.49
COMPRESSOR2	49	1.27	97.41

**TABLE IV. CURRENT ANALYSIS**

COMPONENT	AVERAGE CURRENT in $\mu$ A		% improvement in Current for proposed Circuits
	CMOS	FinFET	
AND GATE	26.5	1.24	95.32
HALF ADDER	14.5	1	93.10
PROPOSED HALF ADDER	1.68	0.264	84.29
FULL ADDER	17.7	2.55	85.59
PROPOSED FULL ADDER	4.58	0.298	93.49
COMPRESSOR1	26.5	1.24	95.32
COMPRESSOR2	28.2	1.27	95.50

The energy consumed by different devices is shown in Table 4. The FinFET dominates and outperforms the CMOS in its consumption of energy. The measure of average power consumed is measured for a 4 bit multiplier in 32nm technology. Even though the existing circuits in literature show decreasing average power consumption as the process length is decreased, the proposed FinFET based design consumes least power among the multipliers with more speed. The power consumption is reduced when compared to the existing design. The circuit is further analyzed by executing the circuit using different systolic architecture. Power and energy are observed, performance of CMOS and FinFET based multiplier is shown in the Table 5. The table 5 presents the performance of different multipliers as shown in figure 9, 10, 11 and 12.

## CONCLUSION

FinFET based arithmetic circuits for the processing element units with less area and low power is proposed and the results found promising when compared to the existing methods. The proposed arithmetic circuit's half adder, full adder and a multiplier makes the important building blocks of various signal processing algorithms. To reduce the number of operands and partial products a 4-2 compressor based multiplier is proposed in this paper. When comes to complicated interconnects and more components the proposed design reduces the complexity and enhances the way the outputs are projected outside. The emphasis on low power is achieved at its maximum level in this work about 96%. The proposed circuits are modelled and simulated using Synopsis HSPICE.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

None

## REFERENCES

- [1] Prateek Mishra, Anish Muttreja and Niraj K. Jha, "FinFET Circuit Design", Nano electronic Circuit Design, Book Series, Springer-Verlag New York, 2011, pp. 23-33.
- [2] Anish Muttreja, Niket Agarwal and Niraj K. Jha, "CMOS Logic Design with Independent-gate FinFETs", 25th International Conference on Computer Design, 7-10 Oct. 2007, Lake Tahoe, CA, pp. 560 - 567.
- [3] Farid Moshgelani, Dhamin Al-Khalili and Come Rozon, "Low Leakage MUX/XOR Functions Using Symmetric and Asymmetric FinFETs", International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, 2013, Vol. 7, no. 4, pp. 354 - 359.
- [4] C.F.Law, S.S.Rofail and K.S.Yeo, "Low-power circuit implementation for partial-product addition using pass-transistor logic", IEEE Proceedings - Circuits, Devices and Systems, Aug 1999, Vol. 146, no. 3, pp. 124 - 129.
- [5] Norio Ohkubo, Makoto Suzuki, Toshinobu Shinbo, Toshiaki Yamanaka, Akihiro Shimizu, Katsuro Sasalu and Yoshinobu Nakagome, "A 4.4 ns CMOS 54 x 54-bit Multiplier Using Pass-Transistor Multiplexer", IEEE Journal of Solid-State Circuits, Mar. 1995, Vol. 30, no. 3, pp 251 - 257.
- [6] D. Radhakrishnan and A. P. Preethy, "Low power CMOS pass logic 4-2 compressor for high-speed multiplication", 43rd IEEE Midwest Symposium on Circuits and Systems, 08-11 Aug 2000, Lansing, MI, Vol. 3, pp. 1296 - 1298.
- [7] Naman Maheshwari, Zhixi Yang and Jie Han and Fabrizio Lombardi "A Design Approach for Compressor Based Approximate Multipliers", 28th International Conference on VLSI Design and 2015 14th International Conference on Embedded Systems, 3-7 Jan. 2015, Bangalore, pp. 209 - 214.
- [8] Chip-Hong Chang, Jiangmin Gu and Mingyan Zhang, "Ultra Low-Voltage Low-Power CMOS 4-2 and 5-2 Compressors for Fast Arithmetic Circuits", IEEE Transactions On Circuits and Systems—I: Regular Papers, Oct. 2004, Vol. 51, no. 10, pp. 1985 - 1997.
- [9] A.Momeni, J. Han, P.Montuschi and F. Lombardi, "Design and Analysis of Approximate Compressors for Multiplication", IEEE Transactions on Computers, 25 Feb 2014, Vol. 64, no. 4, pp. 984 - 994.
- [10] Abdoreza Pishvaie, Ghassem Jaberipur and Ali Jahanian, "Redesigned CMOS (4; 2) compressor for fast binary multipliers", Canadian Journal of Electrical and Computer Engineering, Summer 2013, Vol. 36, no. 3, pp. 111 - 115.
- [11] Shen-Fu Hsiao, Ming-Roun Jiang and Jia-Sien Yeh, "Design of high-speed low-power 3-2 counter and 4-2 compressor for fast multipliers", Electronics Letters, 19th Feb. 1998, Vol. 34, no. 4, pp. 341 - 343.
- [12] R. Menon and D. Radhakrishnan, "High performance 5:2 compressor architectures", IEEE Proceedings - Circuits, Devices and Systems, Oct. 2006, Vol. 153, no. 5, pp. 447 - 452.
- [13] Vahid Jamshidi, Mahdi Fazeli and Ahmad Patooghy, "A Low Power Hybrid MTJ/CMOS (4-2) Compressor For Fast Arithmetic Circuits", 18th CSI International Symposium on Computer Architecture and Digital Systems (CADSD), 7-8 Oct. 2015, Tehran, pp. 1 - 6.
- [14] Joseph Whitehouse and Eugene John, "Leakage and delay analysis in FinFET array multiplier circuits", IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS), 3-6 Aug. 2014, College Station, TX, pp. 3 - 6.
- [15] Y. K. Choi, T. J. King and C. Hu, "Nanoscale CMOS Spacer FinFET for the Terabit Era", IEEE Trans. Electron Device, 2002, Vol. 23, no. 1, pp. 25-27.
- [16] Aqilah binti Abdul Tahrim, Huei Chaeng Chin, Cheng Siong Lim and Michael Loong Peng Tan, "Design and Performance Analysis of 1-Bit FinFET Full Adder Cells for Sub threshold Region at 16 nm Process Technology", Journal of Nanomaterials, March 2015, vol. 2015, pp. 1-13.
- [17] Shyam Akashe and Vishwas Mishra, "Calculation of Power Delay Product and Energy Delay Product in 4-Bit FinFET Based Priority Encoder", Advances in Optical Science and Engineering, 03 June 2015, vol. 166, pp 283 - 289.
- [18] Neha Yadav, Saurabh Khandelwal and Shyam Akashe, "Design and analysis of FINFET pass transistor based XOR and XNOR circuits at 45 nm technology", International Conference on Control Computing, Communication & Materials (ICCCCM), 3-4 Aug. 2013, Allahabad, pp. 1 - 5.

# ABC FOR OPTIMAL ENERGY-AWARE ALLOCATION OF DATA CENTRE RESOURCES

C Ramesh\*, P Thangaraj

Dept. of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, T.N. INDIA

## ABSTRACT

Cloud computing provides workable solutions for long-term medical image archiving systems. A huge problem in present research is the effective dividing of resources present in hand. In this study, allocation protocols for efficient management of energy resources in cloud computing environments are detailed. The proposed energy-aware allocation technique divides Data Centres (DC) resources among client applications in such a way that improves energy efficiency of DCs as well as provides desired Quality of Service (QoS) for every client. Heuristic protocols are used for optimizing the division of resources such that energy efficiency of DC is improved. In the current paper, energy aware resources allotment technique is implemented in clouds via Genetic Algorithm (GA), Artificial Bee Colony (ABC), Best Fit Decreasing (BFD) as well as the suggested ABC with crossover and mutation of onlooker bees with employed bees technique. ABC possesses the benefits of few variables as well as rapid convergence speeds. Outcomes are obtained for energy consumption, quantity of Virtual Machines (VM) migration as well as make span.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Cloud Computing, Scheduling, Data Centres, Energy Aware Resource Allocation, Genetic Algorithm (GA) Artificial bee Colony (ABC) algorithm and Best Fit Decreasing (BFD)

\*Corresponding author: Email: [ramesh.c0909@gmail.com](mailto:ramesh.c0909@gmail.com)

## INTRODUCTION

Cloud computing is an evolution of several technologies that are brought together for altering organizations' approach to building information technological architecture. It is not an entirely innovative notion, however several already present technologies are put together to form cloud computing, making it accessible to the public. Recently, software as well as hardware have undergone intensive development and the benefit of approach is that resources which were in isolation previously may now be used in such a manner so that their advantages may now be enjoyed through one virtualized unit.

Several definitions are present for explaining cloud computing, though there is no one definition which is recognized as official. The several definitions may be summed up thus: "Pools of virtual resources which are useful as well as accessible and may be utilized as resources on demand with or without nominal fee". Cloud computing is understood as the idea of provision of computing as a service like how electricity has been provided as a shared pool with on-demand service. With advances in cloud computing, load balancing between virtual machines as well as preservation of energy are important challenges.

The fundamental premise of cloud computing is that users data is not stored locally, rather it is stored in DCs of the Internet. Organizations that offer cloud computing services are the ones who take care of the management as well as maintenance of DC. The users are capable of accessing the stored data at any point through usage of Application Programming Interface (API) offered by cloud providers via any terminal equipment linked to the Internet. Although clouds have highly simplified the procedure of dividing capacities, it presents various new problems in the domain of managing quality of service. QoS represents the levels of performance, dependability, as well as availability provided by applications as well as by platforms or architecture which hosts them. Quality of Service is basic for cloud users who have the expectation that provider will deliver the stated quality, as well as for cloud providers, who are required to discover the correct trade-off between quality of service levels as well as operation cost. But, discovering optimum trade-off is a hard decision issue, frequently worsened by SLA particularly mentioning QoS target as well as economical penalty related to SLA violation [1].

Virtualization offers effective solutions to the aims of cloud computing paradigms through facilitation of generation of VM over basic physical servers resulting in enhanced resources usage as well as abstraction. Virtualization implies the creation of virtual variant of devices or resources like servers, storage devices, networks or even operating systems wherein the method partitions resources into one or more implementation environments. Device, application, as well as end user interact with virtual resources as if they were one real logical resource. Factors which cloud providers ought to take into consideration are elasticity, scalability, live migration of VM as well as performance isolation. Live migration of virtual machines, the procedure of dynamic transfer of VMs over various servers on the fly, has confirmed to indicate a novel chance for enabling agile as well as dynamic resources management in modern DC. This is particularly important as DC networks are fraught with scalability as well as efficacy problems that are concerns for users as well as scholars. Resources allotment protocols take resources needs of VMs into consideration and change allotted resources, thereby ensuring it an on-demand elastic cloud. VM placement as well as migration are a crucial component of resources allotment in cloud DC.

Resources allotment is a huge problem in cloud computing environments. Resources allotment has several levels of problems such as scheduling tasks, computational performances, re-allocations, response times as well as costs efficacy. Accomplishing tasks with least cost is the most significant challenge in cloud computing. Resources allotment refers to the procedure of provision of services as well as storage space to certain tasks requested by users. Resources allocation is the primary technology of cloud computing that uses computational resources in the network for facilitating the implementation of complex jobs which need huge-scale computations. Resources allotment requires the consideration of several factors like load balancing, make span as well as power usage. Choosing desirable resource nodes for executing tasks in cloud computing is to be taken into consideration, and they are to be correctly chosen as per characteristics of the job. Particularly, cloud resources are needed to be allotted not only for satisfying QoS requisites mentioned by users through SLA but also for decreasing power usage.

Virtualization technologies permit cloud providers to handle the issue of energy inefficacy through creation of several VM instances on physical servers, thereby enhancing usage of resources as well as improving Return on Investment (ROI). The decrease in power usage may be attained through switch of idle nodes off, thereby decreasing idle energy usage. Furthermore, through usage of live migrations, VMs may be consolidated in a dynamic manner on minimum number of physical nodes as per their current resource requisites. But, effective resources management in clouds is important because modern service applications frequently experience extremely varying workloads leading to dynamic resource utilization pattern. Hence, aggressive consolidation of VM leads to performance decrease when applications encounter rising demands leading to increase in resources utilization. Guaranteeing dependable quality of service defined through SLA is crucial in cloud computing environment, and so, cloud providers face energy-performance tradeoff.

Scheduling refers to the procedure of allotment of jobs to available resources based on jobs' qualities as well as needs. The primary aim of scheduling is improved usage of resources with no adverse impact on services offered by clouds. There are two kinds of scheduling, which are resources as well as task scheduling. Given below are few requirements of scheduling in cloud computing: **Fair resource allocation** – Scheduling is performed such that allotment of resources is carried out fairly. **QOS** – Resources as well as tasks are scheduled so that quality of service is attained. **Resource utilization** – This refers to the level to which system resources are used. Excellent scheduling protocol offers maximal resource usage. **Energy consumption** – This refers to the level to which system resources are utilized. Excellent scheduling protocol conserves power usage [2].

Scheduling procedure in clouds are split into three phases which are: **Resource discovering and filtering**- DC Broker finds the resources available in the network system and gathers status data regarding the resources. **Resource selection** - Target resources are chosen on the basis of particular requisites of tasks as well as resources. This is the decision phase. **Task allocation** –Tasks are allotted to the chosen resources.

The issue of optimization is the most critical issue currently and lot of research has been carried out for solving it. Earlier, work was carried out on GA, ABC as well as hybrids of several EAs. Some works are present that compare their performance evaluations as well as suggest the optimal method for particular issues. ABC is the most recent protocol suggested by Dervis Karaboga in 2005, which simulates the intelligent activity of honey bees. It is similar to Particle Swarm Optimization (PSO) as well as Differential Evolution (DE) protocols and

utilizes solely common control variables like colony size as well as maximal cycle number. ANC as an optimization protocol offers a population-based search process wherein food positions are altered by the bees with time with the aim of the bees being to find the food position with greatest nectar quantity.

In this work, an energy-aware resources allocation is carried out through the BFD method along with optimization methods like GA as well as ABC and proposed the ABC with crossover and mutation of onlooker bee with employed bee. Section 2 provides an account of studies related to the field, section 3 discusses the methods and materials utilized, section 4 presents the outcomes obtained while section 5 provides the conclusion.

## RELATED WORK

Raju et al., [3] suggested an Energy-Aware Multi objective Chiropteran Algorithm (EAMOCA) by combining echolocation as well as hibernation characteristics for scheduling resources and for conservation of energy. Promoting energy conservation in cloud environments was attained in a deep manner. Through usage of performance measures like total energy utilized by physical resources, Service Level Agreement (SLA) violations (CPU performance) as well as VM migrations, the study modified the method via real time execution by initializing private clouds utilizing VMware.

Zhang et al., [4] studied the scheduling process for cooperative implementation in mobile cloud computing. Mobile applications were denoted by sequences of fine-grained jobs developing a linear topology, and all of them were implemented either on mobile devices or offloaded on cloud sites for implementation. Design aim was the minimization of power utilized by mobile devices, while fulfilling time deadline. The authors formulate the minimal-energy job scheduling issue as a restricted shortest route issue on directed acyclic graphs and adapted canonical “Lagrange Relaxation based Aggregated Cost (LARAC)” protocol for resolving the issue approximately.

Zhang et al., [5] offered a theoretical model of energy-optimal mobile cloud computing under stochastic wireless channels. The aim was the conservation of energy for mobile devices, through optimal execution of mobile applications in mobile devices or offloading onto clouds. Moreover, for energy-optimal implementation scheme of applications with minimal output data, the study derived a threshold policy that states that data consumption rates, defined as ratio between data size (L) as well as delay constraint (T), was contrasted to a threshold that relies on both power usage as well as wireless channel models. In the end, numerical outcomes propose that a considerable quantity of power could be conserved for the mobile device through optimal offloading of mobile applications to clouds in certain situations.

Kliazovich et al., [6] suggested a scheduling solution, called e-STAB, that considers account traffic requisites of cloud applications offering energy effective task allotment as well as traffic load balancing in DC network. Efficient distribution of network traffic enhances QoS of running cloud applications through reduction of transmission-related delay as well as congestion-related packet loss. Evaluations on Green Cloud simulators, highlight advantages as well as efficacy of the suggested scheduling method.

Jiang et al., [7] suggested a resilient routing protocol for reaching greater network energy efficacy that has its basis in optimization problems. The authors attained extremely effective routing in energy-effective networks for cloud computing, the links of lesser usage were switched to sleep state for saving network energy. Simultaneously, the lesser link traffic was gathered to links with higher usage for enhancing link usage as well as to sleep maximal number of links. The study suggested an optimized link sleeping technique for maximizing quantity of sleeping links. Through targeting of network resilience, a weight-adaptive scheme was introduced for reducing link congestions as well as enhancing resilience of networks. Simulations indicated that the protocol was efficient as well as viable for achieving energy-effective networks for cloud computing.

Hosseinmotlagh et al., [8] suggested an optimum usage level of hosts to implement a particular set of instructions for minimizing power usage of hosts. The investigators suggested CM scheduling protocol on the basis of unsurpassed usage level for coming up with optimum power usage while fulfilling desired quality of service. Otherwise put, the suggested protocol performs regulation of allotted computational resources of VM on hosts that lead to achieving optimum energy levels in the hosts. Simulations prove that the suggested technique not only decreases overall energy usage of clouds by 60%, but also impacts turn-around time of real-time jobs by 94%. It

also improves acceptance rates of arrival jobs by 96%. Furthermore, it has a significant part in the acceptance of long jobs that have shorter deadlines.

Lin et al., [9] suggested a new protocol that begins from minimum-delay scheduling solutions and consequently carries out energy reductions through migration of jobs from local core as well as cloud and through application of dynamic voltage as well as frequency scaling method. Linear-time re-scheduling protocol was suggested for tasks migrations. Simulations demonstrated considerable energy reductions with application completion time restriction fulfilled.

Cheng et al., [10] suggested a power-saving job scheduling protocol on the basis of vacation queuing model for cloud computing system. Firstly, the authors utilize vacation queuing model with exhaustive services for modelling tasks scheduling of heterogeneous cloud computing systems. Then, on the basis of busy period as well as busy cycle under steady states, the investigators analysed expectation of task sojourn times as well as power utilization of computing nodes in the heterogeneous cloud computing systems. Consequently, the authors suggested a task scheduling protocol on the basis of similar jobs for reducing power usage. Simulations prove that the suggested protocol decreased power usage of cloud computing systems efficiently while fulfilling task performance.

Singh & Chana [11] suggested fuzzy-logic based power-aware autonomic resources scheduling model for clouds for energy effective scheduling of cloud computing resources in DC. The investigators tested the suggested model in CloudSim based simulation environments as well as real cloud environment. Outcomes of experiments prove that the suggested model outperforms with regard to resources usage as well as power utilization along with other quality of service variables.

Hung et al., [12] suggested a GA that is power-aware in scheduling of resources allotment (GAPA) which resolved the Static Virtual Machine Allocation Problem (SVMAP). Because of restricted resources like memory for implementing simulations, the authors generated workloads which contain samples of one-day timetable of lab hours in the university. The study evaluates GAPA as well as a base scheduling protocol (BFD) that ranks lists of VM in start time and utilizing BFD protocol, for resolving SVMAP. This leads to GAPA obtaining lesser overall energy usage than the base protocol in simulations.

Beloglazov & Buyya [13] suggested an effective resources management policy for virtualized cloud DC. The goal was the continuous consolidation of VMs leveraging live migrations as well as switching off idle nodes for minimal energy usage, while offering needed quality of service. The investigators presented outcomes of evaluations proving that dynamic re-allotment of VM brings about considerable power conservation, thereby proving this approach worth of further study.

## MATERIALS AND METHOD

Here, energy aware allocations of DC resources, BFD, GA as well as ABC with crossover and mutation of onlooker bees with employed bees are detailed. BFD finds optimal solutions while First Fit Decreasing (FFD) provides sub-optimal solutions. Benefits of GA [14] include the fact that it is able to resolve every optimization problem which might be described with chromosomes encoding; it is able to resolve issues with various solutions and is able to solve multidimensional, non-differential, non-continuous and non-parametrical problems. Genetic Algorithm needs minimal expertise to handle it and can be carried over easily to other models. The benefit of ABC rests in its simplicity, flexibility and resilience, utilization of less control parameters in contrast to other search techniques, simple hybridization with other optimization frameworks; able to handle objective cost stochastically and simple implementation with basic mathematical or logical operations.

## ENERGY-AWARE ALLOCATION OF DATA CENTRE RESOURCES

DC is home to several varied applications with varied resource needs as well as performance goals. Generally, cloud applications may be decomposed into 1 or more jobs implemented in 1 or more VMs. During runtime, schedulers handle the assignment of jobs to machines. These days, production DCs like Google's Cloud Back-end frequently implements huge quantities of jobs every day. Very huge-scale workloads hosted by DC not only uses considerable storage as well as computational power, but also large quantities of energy. Practically, operational cost on energy is not solely through the running of physical machines, but also from the cooling of all DCs. It is noted that power consumption makes up around 12% of the monthly operational cost for generic DCs. For huge organizations such as Google, 3% decrease in energy costs might convert to more than a million dollars in savings. Governmental organizations attempt the implementation of standards as well as regulations for promoting energy-effective (that is, Green) computing.

Current advances in virtualization have led to its distribution across data centres. Through supporting the movement of virtual machine between physical nodes, it permits dynamic migration of virtual machines as per performance requisites. If virtual machines do not utilize every resource which is provided, they may be logically resized as well as consolidated to minimal quantity of physical nodes, whereas idle nodes may be turned to sleep mode for eliminating idle power utilization as well as decrease overall energy utilization by DCs [15].

At present, resources allocation in Cloud DCs aim at the provision of high performance while also fulfilling SLA, with no focus on allotment of virtual machines to decrease power utilization. For exploring both performance as well as energy efficacy, three critical problems are to be handled: Firstly, excessive power cycling of servers can decrease dependability. Secondly, switching resources off in dynamic environments is problematic from the viewpoint of quality of service. Because of the diversity of workload as well as aggressive consolidation, few virtual machines might not get the needed resources at peak loads and thereby fail in meeting the required quality of service. Thirdly, guaranteeing SLA introduces problems into accurate application performance management in virtualized environments. Every issue requires efficient consolidation policies which can decrease power usage with no compromise to user-specified quality of service requisites.

### BEST FIT DECREASING (BFD)

Conventional BFD protocols rank incoming jobs in descending order on the basis of CPU requisites. After the ranking, job at the top of the list is chosen and placed on already utilized server with minimal CPU capacity. In the event of non-availability of resources on utilized servers, job is placed on new server which has minimal CPU capacity [16].

BFD protocols are famous for online bin packet, and are regarded as excellent for VMs placement in cloud environments. BFD protocols are regarded better compared to next fit, and first fit protocols with regard to worst case as well as average uniform case. Furthermore, BFD protocols may also be improved for managing multi-criteria optimizations as done. At present, there is no in-depth work present which tests BFD on the basis of various scenarios. Hence, a comparison is carried out for evaluating BFD methods on the basis of workload as well as migration methods.

### GENETIC ALGORITHM (GA)

Genetic Algorithms are search heuristics which mimic the procedure of natural evolution. The heuristics are routinely utilized for generating helpful solutions to optimization as well as search issues. Genetic Algorithms are part of the bigger class of Evolutionary Algorithms (EA) that create solutions to optimization issues through usage of methods which owe their inspiration to natural evolution, like inheritance, mutations, selections, as well as crossover. But, the adequate representation of possible solutions is critical for ensuring that the mutations of pairs of individuals (that is, chromosomes) results in novel, valid as well as meaningful individuals for the problems. Output schedules of jobs are array lists of populations (known as chromosomes or genotypes of the genome) that encode potential solutions to optimization issues, evolve towards better solutions. The major terms utilized in Genetic Algorithm include:

**Initial Population:** This refers to the set of every individual which is utilized in Genetic Algorithm for finding out the optimum solutions. All solutions in the population are known as individuals. All individuals are denoted as chromosomes for making them adequate for genetic operations. From initial population, individuals are chosen and few operations are employed for forming the subsequent generations. The mating chromosomes are chosen on the basis of certain conditions [17].

**Fitness Function:** Fitness functions are utilized for measuring quality of individuals in the population as per certain optimization goal. Fitness functions may be different for various cases. In certain cases, fitness functions may have their basis in deadlines, whereas in other cases, it may have their basis in budgetary restrictions.

**Selection:** It utilizes the proportion selection operator for determining the probability of several individuals genetic to the subsequent generation in the population. Proportional selection operators imply the probability that is chosen as well as genetic to subsequent generation groups is proportional to size of individual fitness.

**Crossover:** It utilizes single-point crossover operators. Here, solely one intersection was initialized in individual code, at that point, some of the pair of the individual chromosome is interchanged.

**Mutation:** Mutations imply that the value of some gene locus in the chromosome coding series was substituted by another gene value for generating a novel individual. Mutations are that which negate the values at mutated points with respect to binary coded individuals.

Genetic Algorithm functions thus:

1. Start
2. Set population with arbitrary potential solutions
3. Test all candidates
4. Iterate till (terminating criterion is fulfilled)
  - a. Choose parents
  - b. Recombine pairs of parents
  - c. Mutate the resultant offsprings
  - d. Test new candidates
  - e. Choose individuals for the subsequent generation;
5. End

#### PROPOSED ABC WITH CROSSOVER AND MUTATION OF ONLOOKER BEE WITH EMPLOYED BEE

In the suggested technique, two extra steps are included to generic ABC – which are crossover as well as mutation operators. The initial step of ABC is the generation of population. The populations created are utilized by employed bees. After this, crossover is employed. If crossover or probability fulfils, then 2 parents are chosen arbitrarily for performing crossover operation on them. Novel offspring is created after this. With best created offspring, substitution of the worst offspring is done if it is better than the worst parent with regard to fitness values. Here, crossover operator is employed on 2 randomized chosen parents from current population [18].

2 offspring created from crossover, one replaces worst parent, and the other parent is the same. Now, mutation is carried out on scout bee phase of ABC. Through usage of mutation operator, there is a possibility of altering local best position and the protocol might not rely on local solution. On the other hand, particles might make use of the other's benefit through sharing of information method. In the suggested technique, mutation is performed on the probabilistic way in every food searching operation for every iteration in the lifecycle of ABC.

Choosing of food source is carried out in an arbitrary manner. Food sources are chosen arbitrarily from food size and then mutation is carried out. In mutations, created offspring substitutes for older offspring. Mutations utilized here are uniform mutations. When carrying out mutations, they arbitrarily choose a food source and substitute one of the dimension values through arbitrary numbers created between upper as well as lower bounds of the food source. Suggested protocol is detailed below

ABC with Crossover and Mutation Operator Algorithm:

##### Step 1: Initialization

For  $i=0$  to maximum number of food sources do

For  $d=0$  to dimension size do

Arbitrarily set food source positions  $X_{ij}$

End for  $d$

Calculate fitness values of all food sources

End for  $i$

Iterate

##### Step 2: Employed Bee Stage

For  $i=0$  to maximum number of employed bees do

For  $d=0$  to dimension size do

Yield novel potential solutions

End for  $d$

Compute fitness values of all individuals

If fitness values of novel solutions are better than older ones, substitute the older ones with the new ones

End for  $i$

For  $i=0$  to maximum number of food sources do

Compute probability for all food sources

End for  $i$

##### Step 3: Crossover stage

If crossover condition fulfils

For  $i=0$  to maximum number of food sources

Choose 2 arbitrary individuals from current population for crossover

Employ crossover

Novel offspring created from parents because of crossover. Substitute worst parents with best novel offspring if it is better.

End of  $i$

##### Step 4: Onlooker Bee Stage

For  $i=0$  to maximum number of onlooker bees do

Based on probability  $P_i$ , select food source



For d=0 to dimension do  
Yield novel potential solutions for  
Food source positions Xij  
End for d  
Calculate fitness values of individual food sources  
If fitness values of novel potential solution is better than already present solution, substitute the older one  
End for i

**Step 5: Scout bees Stage**

If any food source is depleted, substitute it with novel arbitrary position created by scout bees

**Step 6: Mutation stage**

If mutation conditions are fulfilled then  
Choose arbitrary particle from current population for mutation  
Employ mutation for generating novel individuals novel offspring created because of mutations  
Novel set of sequence is yielded for offspring  
Calculate cost for the offspring  
Calculate fitness value for updated individual  
Memorize best food source as of yet  
Till (terminating conditions are fulfilled)

**RESULTS**

In this section, the energy consumption KWh-ABC and ABC with crossover and mutation of onlooker bee with employed bee are evaluated. An energy consumption, number of VM migration and makespan as shown in [Table- 1 to 3] and [Figure- 1 to 3].

Table: 1. Energy Consumption

Average Utilization Threshold %	Energy Consumption KWh - ABC	ABC with Crossover and Mutation of Onlooker bee with Employed bee
45	2.18	2.11
50	2.07	1.99
55	1.95	1.89
60	1.81	1.77
65	1.69	1.66
70	1.59	1.54
75	1.55	1.5
80	1.43	1.4
85	1.33	1.29
90	1.26	1.22
95	1.22	1.19
100	1.14	1.12

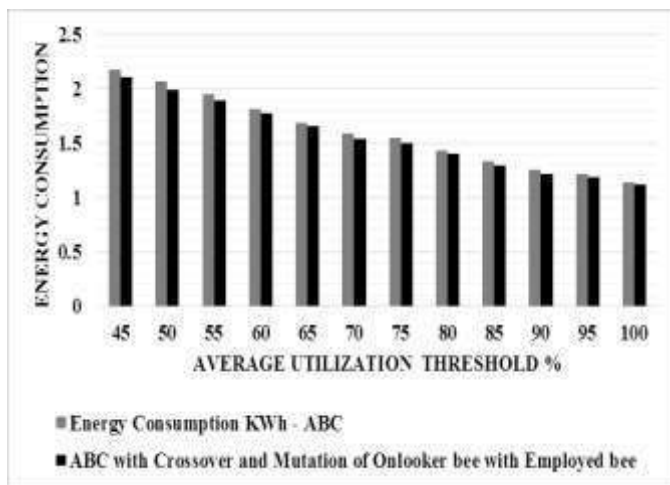


Fig: 1. Energy Consumption

From the [Figure- 1], it can be observed that the ABC with crossover and mutation of onlooker bee with employed bee has lower energy consumption by 3.26% for 45 average utilization threshold, by 1.79% for 65 average utilization threshold, by 3.05% for 85 average utilization threshold and by 1.76% for 100 average utilization threshold when compared with energy consumption KWh - ABC.

Table: 2. Number of VM migrations

Average Utilization Threshold %	Energy Consumption KWh - ABC	ABC with Crossover and Mutation of Onlooker bee with Employed bee
45	3231	3134
50	3126	3056
55	3006	2908
60	2901	2799
65	2911	2821
70	2751	2690
75	2670	2601
80	2647	2551
85	2599	2527
90	2514	2428
95	2424	2354
100	2347	2264

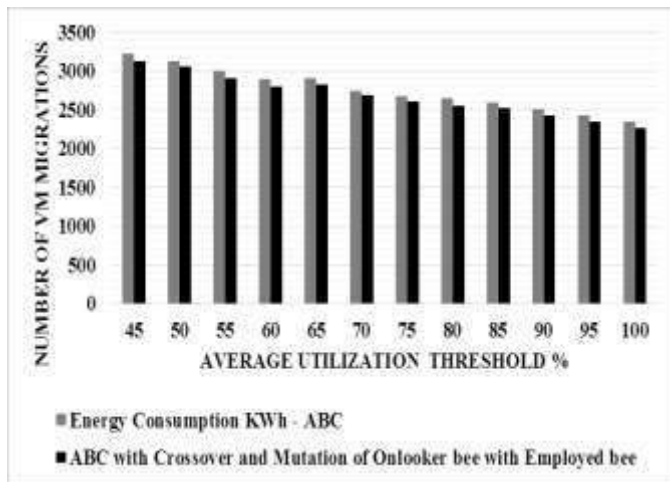


Fig: 2. Number of VM migrations

From the [Figure- 2], it can be observed that the ABC with crossover and mutation of onlooker bee with employed bee has lower number of VM migrations by 3.04% for 45 average utilization threshold, by 3.14% for 65 average utilization threshold, by 2.8% for 85 average utilization threshold and by 3.6% for 100 average utilization threshold when compared with energy consumption KWh - ABC.

Table: 3. Makespan

Number of jobs	Energy Consumption KWh - ABC	ABC with Crossover and Mutation of Onlooker bee with Employed bee
50	463	449
100	832	815
150	1312	1272
200	1687	1620
250	2202	2125
300	2689	2600

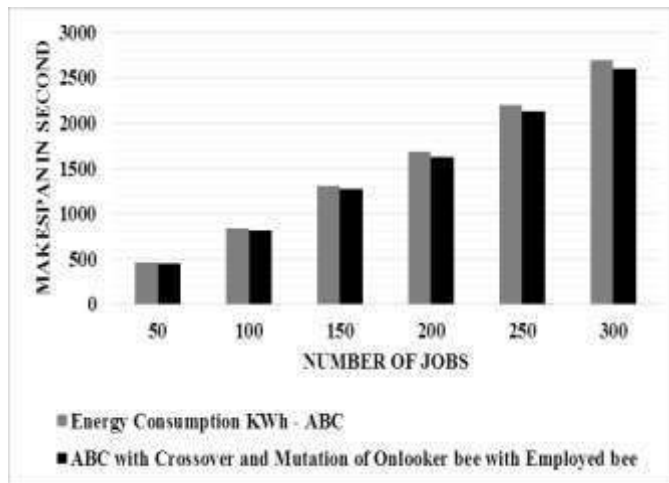


Fig: 3. Makespan

From the [Figure- 3], it can be observed that the ABC with crossover and mutation of onlooker bee with employed bee has lower makespan by 2.01% for 100 number of jobs, by 4.05% for 200 number of jobs and by 3.36% for 300 number of jobs when compared with energy consumption KWh - ABC.

## CONCLUSION

Virtual DS facilitates the sharing of resources between computational services as well as hosted appliances. Cloud provider ought to execute energy effective resources management methods for maximizing their returns as well as for improving their efficacy. In the current study, energy aware scheduling was included utilizing ABC, GA as well as BFD. Read coded crossover as well as mutation operators are employed after employed bee phase as well as scout bee phase of ABC. With usage of crossover, new offspring is created from initial population and it replaces the worst parent with best offspring and with mutation operator, a food source is arbitrarily chosen and one of its dimension values are replaced by an arbitrary number generated within lower as well as upper bounds of the food source. Outcomes probe that ABC with crossover as well as mutation of onlooker bees with employed bees leads to lesser makespan by 2.01% for 100 tasks, by 4.05% for 200 tasks as well as by 3.36% for 300 tasks in comparison to power consumption KWh - ABC.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Patel R, Patel H, Patel S. [2015] EFFICIENT RESOURCE ALLOCATION IN CLOUD COMPUTING. *International Journal For Technological Research In Engineering* 2(7)
- [2] Kaur R, Kingar S. [2014] Analysis of Job Scheduling Algorithms in Cloud Computing. *International Journal of Computer Trends and Technology (IJCTT)*–volume, 9.
- [3] Raju R, Amudhavel J, Kannan N, Monisha M. [2014, March]. A bio inspired Energy-Aware Multi objective Chiropteran Algorithm (EAMOCA) for hybrid cloud computing environment. In *Green Computing Communication and Electrical*

- Engineering (ICGCCEE), 2014 International Conference on (pp. 1-5). *IEEE*.
- [4] Zhang W, Wen Y, & Wu DO. [2013, April] Energy-efficient scheduling policy for collaborative execution in mobile cloud computing. In INFOCOM, 2013 Proceedings IEEE (pp. 190–7194). *IEEE*.
- [5] Zhang W, Wen Y, Guan K, Kilper D, Luo H, Wu DO. [2013] Energy-optimal mobile cloud computing under stochastic wireless channel. *IEEE Transactions on Wireless Communications* 12(9): 4569–74581.
- [6] Kliazovich D, Arzo ST, Granelli F, Bouvry P, Khan SU. [2013] e-STAB: Energy-efficient scheduling for cloud computing applications with traffic load balancing. In Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCoM), IEEE International Conference on and IEEE Cyber, Physical and Social Computing (pp. 7–713). *IEEE*.
- [7] Jiang D, Xu Z, Liu J, Zhao W. [2015] An optimization-based robust routing algorithm to energy-efficient networks for cloud computing. *Telecommunication Systems* 1–710.
- [8] Hosseinimotlagh S, Khunjush F, Samadzadeh R. [2015] SEATS: smart energy-aware task scheduling in real-time cloud computing. *The Journal of Supercomputing* 71(1): 45–766.
- [9] Lin X, Wang Y, Xie Q, Pedram M. [2015] Task scheduling with dynamic voltage and frequency scaling for energy minimization in the mobile cloud computing environment. *IEEE Transactions on Services Computing* 8(2):175–7186.
- [10] Cheng C, Li J, Wang Y. [2015] An energy-saving task scheduling strategy based on vacation queuing theory in cloud computing. *Tsinghua Science and Technology* 20(1):28-39.
- [11] Singh S, Chana I. [2016] EARTH: Energy-aware autonomic resource scheduling in cloud computing. *Journal of Intelligent & Fuzzy Systems* 30(3):1581–71600.
- [12] Hung N Q, Nien PD, Nam NH, Tuong NH, Thoai N. [2013] A Genetic algorithm for power-aware virtual machine allocation in private cloud. In Proc of the 2013 Int Conf on Information and Communication Technology. Piscataway: *IEEE* (pp. 183-193).
- [13] Beloglazov A, Buyya R. [2010, May] Energy efficient allocation of virtual machines in cloud data centers. In Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on (pp. 577–7578). *IEEE*.
- [14] Tabassum M, Mathew K. [2014] A genetic algorithm analysis towards optimization solutions. *International Journal of Digital Information and Wireless Communications (IJDIWC)* 4(1):124–7142.
- [15] Beloglazov A, Abawajy J, Buyya R. [2012] Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future generation computer systems* 28(5):755–7768.
- [16] Mustafa S, Bilal K, Madani SA, Tziritas N, Khan SU, Yang LT. [2015, December] Performance Evaluation of Energy-aware Best Fit Decreasing Algorithms for Cloud Environments. In 2015 IEEE International Conference on Data Science and Data Intensive Systems (pp. 464–7469). *IEEE*.
- [17] Ravichandran S, Naganathan DE. [2013] Dynamic Scheduling of Data Using Genetic Algorithm in Cloud Computing. *International Journal of Computing Algorithm* 2(1):127–7133.
- [18] Shrivastava A, Gupta M, Swami S. [2015] SPV and Mutation based Artificial Bee Colony Algorithm for Travelling Salesman Problem. *International Journal of Computer Applications* 116(14).

# PARTICLE SWARM OPTIMIZED FEATURE SELECTION FOR ALZHEIMER CLASSIFICATION

S Sountharajan\*, P Thangaraj, E Suganya

Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu. INDIA

## ABSTRACT

Alzheimer's disease (AD) refers to a neuro-degenerative chaos that is a general kind of dementia which leads to memory loss, and lack of cognitive functioning and so on. Magnetic Resonance Imaging (MRI) is popularly utilized for human body imagings. MRI is a fundamentally non-invasive method giving high degree clarity on the soft tissue inside the brain better than conventional Computed Tomography (CT), ultrasound, Positron Emission Tomography (PET), etc. SVM is a kind of ANN (artificial neural network) which has got training from supervised learning methods and had showed the benefits of decreasing the training-testing error and hence producing greater recognition precision. This paper investigates empirically PSO's (Particle Swarm Optimization's) effectiveness towards selection of features.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Alzheimer's disease (AD),  
Magnetic Resonance Imaging  
(MRI), Particle Swarm  
Optimization (PSO)

\*Corresponding author: Email: [thangarajp.cse@rediffmail.com](mailto:thangarajp.cse@rediffmail.com)

## INTRODUCTION

Alzheimer's disease (AD) is a typically occurring kind of dementia. The clinical symptoms are featured as cognitive decline along with degradation of everyday life as well as other symptoms related to neuropsychiatric or behavioural change. It is very difficult to diagnose AD in its early stage because there is no existing known biomarker. AD has a stealthy onset that can be of either genetic or environment oriented reasons. Distinguishing various kinds of dementia is complex. The transitional phase named as Mild Cognitive Impairment (MCI) occurs between usual and demented behavior. The disorder is diagnosed with symptoms of high decline in cognition than predicted for the particular age but at the same time no disruption is found in the routine activities.

The key symptom of MCI is featured by the decline in memory as well as the impairment of cognition. Researchers report the association of AD and MCI that MCI provides the threat of 10 to 64 percent in leading to AD [1]. AD is defined as continuous neurodegenerative illness and it is different from MCI by its symptom of progressive decline of routine activities. The occurrence of AD is found to increase drastically at 65 years of age and it is also studied that around 26 million of people in the globe are affected by AD, the number is also expected to increase rapidly by as much as four times by 2050. In order to treat or prevent AD, many researches have been undertaken to diagnose it in early stage itself. Hence, by detecting the change in the tissue of brain that reflects the pathological process of MCI can assist in preventing or postponing further progression or its conversion in to AD. This early detection and efficient intervention could possibly decrease further impairment.

The major clinical factor to identify and predict the advancement of AD is the universally accepted cognition test. The research with the objective to improve early detection is done with a tool kit which includes brain images, biomarker of Cerebrospinal Fluid (CSF) as well as biomarker of plasma.

The main reason to select features is to decide the quantity of feature subset which is necessary and accurate for the classifier to differentiate the link or action into usual or invasive. It is also found that the feature subset p is low when compared to subset m. FS investigates the combination space to obtain the best combination of features.

Normally FS is classified into three categories: First is the conventional exponential algorithm, where large quantities of subsets are assessed and the quantity grows in an exponential manner with dimensional growth. Here, exhaustive search is the classical algorithm used. The second category being sequential technique which adds or removes the feature sequentially, but the limitation is that the algorithm has the challenge of being trapped into local minima. The final category is the meta-heuristic algorithm that includes random search to ensure escape from local minima.

Particle Swarm Optimization (PSO) is a populace oriented stochastic technique to solve optimization problem. In PSO, the particle wanders in the search area to find optimal solution. A candidate solution is obtained from the present location of the particle. Every particle goes in search of optimal position by altering the velocity based on rules simulating the behavior of birds. PSO is a part of the category of swarm intelligent methods which is chosen for solving problem of optimization [2].

## RELATED WORK

Jongkreangkrai et al [3] enhances the AD classification performance by merging hippocampus and amygdala volume and thickness of entorhinal cortex. Its aim is to examine the helpful feature got through MRI to classify the AD affected patients with the help of Support Vector Machine (SVM). The samples of MR brain images weighed as T1 of 100 normal and AD patients were treated with the software Free Surfer to evaluate hippocampus and amygdala volume and thickness of entorhinal cortex of both the hemispheres of brain. Comparative volume of hippocampus and amygdala are computed to correct variations in each head size. SVM is deployed with different fusions of feature: (H: hippocampus relative volume, A: amygdala relative volume, E: entorhinal cortex thickness, HA: hippocampus and amygdala comparative volume and ALL: all features). Receiver operating characteristic (ROC) analysis is applied to assess the technique. AUC range of five fusions are 0.8575 (H), 0.8374 (A), 0.8422 (E), 0.8631 (HA) and 0.8906 (ALL) respectively. Even though "ALL" gave the maximum AUC, no numerically important difference had been noticed except in the case of "A". Result proved the feasibility of recommended features towards AD patient via computer aided classification.

Moradi et al [4] presents a new Magnetic Resonance Imaging (MRI) oriented technique to predict the conversion of MCI to AD within 1 to 3 years prior to medical diagnoses. Primarily, a new MRI bio-marker for MCI to AD progressions is designed with the help of semi supervised learning and then combined with age factor and cognitive metric of the subject with a supervised learning algorithm that results in the cumulative biomarker. Added value of this new feature to predict the MCI to AD change is formulated on the received data from Alzheimer's Disease Neuroimaging Initiative (ADNI) databases. With this ADNI data, the MRI biomarker attained a 10 fold cross validated region within the ROC curve (AUC) of 0.7661 in differentiating progressive MCI patient (pMCI) from stable MCI patient (sMCI). Cumulative biomarker according to MRI data altogether along with base cognitive metrics and patient's age reached a 10 fold cross validated AUC value of 0.9020 in differentiating MCI from sMCI. The output presents the demonstration of the probability of proposed technique to diagnose AD early and the significant part of MRI in predicting the conversion of MCI to AD. Moreover, the result showed that merging of MRI data with cognition test result enhances the precision of predicting the conversion of MCI to AD.

Zhang et al [5] suggested a technique that primarily deployed digital wavelet transform for extracting attributes and later applied Principal Component Analysis (PCA) for reducing the space of the feature. Then, kernel support vector machine (KSVM) is built along RBF kernel, utilizing Particle Swarm Optimization (PSO) for optimizing the attributes  $C$  and  $\sigma$ . Over fitting is avoided by the use of five-fold cross validation. The experiment used a data set of 90 brain images downloaded from the web site of Harvard Medical School. The fivefold cross-authentication classification output proved the classification precision as 97.78% which is greater than 86.22% by BP-NN and 91.33% by RBF-NN. The parameters are selected in comparison to PSO and other random selection technique. The result proved that PSO is more efficient to construct best KSVM.

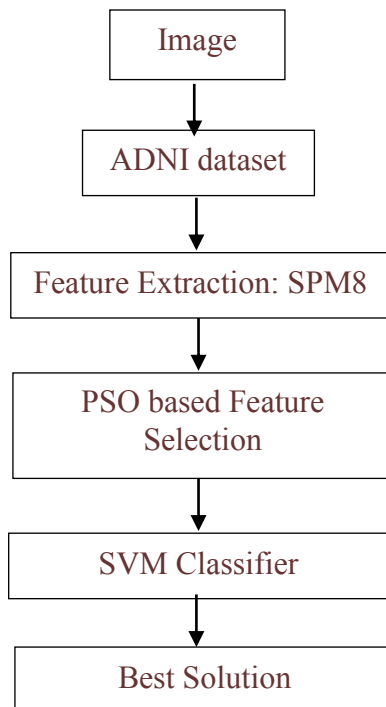
Alzheimer's disease (AD) is a well-known precursor to dementia. The treating of AD is more effective if it is detected at the early stage. Being the pre-cursory phase of AD, Mild cognitive impairment (MCI) is considered as better focus point for early diagnosis, but its diagnosis is much difficult because of cognitive deterioration's subtlety. Lee et al [6] develops a technique for automation of the detection of AD and MCI with SVM as well as Diffusion Tensor Imaging (DTI) data. The method applied two SVM models: first is to classify AD and MCI and next is to classify AD and Normal Control (NC). Both the model uses the Fractional Anisotropy (FA) and the

mode of anisotropy (MO) values of DTI were utilized as attributes. In the two models, MO values gave the result of optimal performance than FA values. In an individual assessment the AD-MCI classifier shows sensitivity of 69.2 %, specificity of 100 % and precision of 89.7 %, and the AD-NC classifier shows sensitivity of 84.6 %, specificity of 90.9 % and precision of 87.5 %. The output proves satisfactory and recommends classification of DTI on the basis of SVM is most strong while detecting MCI and AD in earlier stage.

Yang et al [7] proposes an MRI-based classification structure to differentiate AD and MCI patient from healthy subjects with the use of several attributes as well as various classifiers. As the first stage, the features are extracted (volume and shape) from the given MRI data with sequential image processing steps followed by the application of principal component analysis (PCA) to transform the collection of attributes of probably co-related parameters into a small set of value of linearly non-correlated parameters, minimizing the dimensions of features space. In the end, a new data mining structure merging with SVM, PSO to classify AD/MCI is designed. The hybrid technique is compared with conventional classifiers, namely SVM and SOM (Self Organizing Map) were trained for the classification of patients. It is noted that for the suggested structure the precision of classification had enhanced to 82.35% and 77.78% among the patient with AD as well as MCI. The results attained up to 94.12% and 88.89% in AD and MCI through fusing of features of volume and shape with the use of PCA. Current result suggests that new multivariate techniques of pattern matching attain a clinically relevant precision for a priori prediction of the development from MCI to AD.

## MATERIALS AND METHOD

[Figure- 1] depicts the flowchart of the proposed method. The technique is given as follows:



**Fig: 1. Flowchart for Proposed Methodology**

### ALZHEIMER'S DISEASE NEURO-IMAGING INITIATIVE (ADNI) DATABASE

This study is done with the data got from Alzheimer's Disease Neuro-imaging Initiative (ADNI) database (<http://adni.loni.usc.edu/>). The main objective of ADNI is to evaluate whether MRI, PET, and other existing biomarkers, clinical and as well as neuropsychological evaluation can be merged to calculate the progression of MCI and early AD. Determining responsive and particular biomarkers of early AD development is needed to help the researcher and physiologists to design novel treatment and measure the efficiency along with decrease in time and costs of trials.

## FEATURE EXTRACTION

Statistical parametric mapping means building and evaluating the extensive numerical procedures for testing the hypotheses on functional imaging of data. These notions were included in the SPM software which is designed to analyze the sequences of brain image data that could be the series of images from various cohorts or time-series from one subject. A huge benefit of SPM is its uni-variant system unlike other multivariate methods where huge quantity of observation is required. Even though wide research had been undertaken in the domain of features extraction, no investigation was performed in the domain of AD CAD since the image represents huge volume of data and imaging study has restricted subjects [8]. To classify AD, feature vector is of great dimension and methods to minimize the input space dimensionality are necessary to improve the classification precision in total.

## PARTICLE SWARM OPTIMIZATION (PSO)

PSO is a populace-based method that was built by Eberhat and Kennedy. PSO is a successful and well known global search method. This is the most appropriate protocol to handle the issues related to features selection because of the given factors: assesses the fitness function of the complete swarm in every step. The position of the given two best particles is used to update the  $V_i$  of ith particle: (1) the best position travelled by a particle ( $p_B$ ) and (2) the best position of the neighbors of ith particle that had been travelled till then ( $n_B$ ). If the complete swarm is considered as neighbourhood region, then it becomes the global best and is called as " $g_B$ ". From the given information, the enhanced equation may be deduced as: Here, rand() is an arbitrary number generator whose value is between [0, 1]. The rand() is performed while they occur. c is called the "inertia weight". While c is lower than 1, the particle helps exploitations over explorations; else if c is greater than 1, the particle helps explorations over exploitations. The parameter  $a_1$  and  $a_2$  are non-negative constants termed as "acceleration coefficient" [9]. The position of the particle swarm is modified based on formula (3) and (4). They will get near to one another from several directions. The PSO executes the process repeatedly till the stopping factor is reached. It is seen that v<sub>max</sub> (the maximum velocity) must be decided prior, with the objective to keep the optimizer in a considerable range.

## FEATURE SELECTION BASED PSO

The idea of PSO for best features selection [10] issue is taken into consideration. Take a huge feature space complete with features subset. Every features subset may be taken as a location in such space. Consider that there exists N features, then it is found that there will be  $2^N$  types of subsets. The subset with minimum distance and high quality classification is known as the best location. Then, a particle swarm is inserted into the features space and each particle takes one location. With the aim to reach the optimal position, the particles start flying in the space. In due course of time, they alter the positions, interact among themselves and go in search of the local as well as global optimum positions. In time, they ought to converge at a good probably optimum position. This capacity for exploration ensures the better performance of feature selection and hence the discovery of best possible subset.

The position of the particle is given as binary bits of string of length N, wherein N is the complete quantity of entities. Each bit representing a feature, the value '1' refers to the respective feature being chosen and '0' if not chosen. Every position is a feature subset. The velocity of every particle is given as positive integer, ranging within 1 and V<sub>max</sub>. It states the count of particle bits that need to be modified at a specific time to be similar to the global optimum position. The count of various bits among 2 particles is related to the variation in their positions respectively.

Next to the velocity updation, the position of the particles is modified by the novel velocity. Consider that the novel velocity is V, and the quantity of various bits between the present particle and gbest is g x. Two cases are found to occur while the position is being updated:

- 1)  $V \leq g \times$ . Here, the velocity of the particle is less or equal to the difference in position between the particle and gbest. Distinct from the gbest, V bits of the particles are modified in random. Instead of being similar to gbest, the particles move towards the global optimum with continuous exploration of the search area.
- 2)  $V > g \times$ . Here, the velocities of the particles overrun the location difference between the particle and gbest. Also, the particles are modified to be similar to gbest, it is further changed in random ('random' means 'exploration ability') modification (V- g x) of bits outside of the various bits between the particle and gbest. Hence even after reaching the best position, the particle moves further in other directions, for more searches.

The high velocity V<sub>max</sub> acts as a limitation to monitor the global exploring ability of a particle swarm. A greater V<sub>max</sub> helps global explorations, at the same time small V<sub>max</sub> stimulates the local exploitations. If V<sub>max</sub> is very less, then the particle faces mode challenge to escape from the local optima whereas with high V<sub>max</sub>, the particle may fly rapidly away from a good solution.

## SUPPORT VECTOR MACHINE (SVM)

SVM is a kind of ANN, which is trained by the use of supervised learning, shows the benefit on minimizing errors on training as well as testing, which results in getting better recognition precision. [11]. But, some features data is linearly nonseparable. Also, in few circumstances, the features will not be properly separable, specifically in the case of border between the classifications. In



order to permit flexibility to some extent in dividing the groups, SVM uses a costs parameter, represented as  $C$ , to monitor the tradeoff between permitting training defects and forced rigid margin. Costs function with  $C$  is given as follows, wherein  $\zeta_i$  denotes a slack variable,

$$\text{cost} = C \sum_{i=1}^N (\zeta_i) \tag{1}$$

Mapping the patterns in a high dimensional features space was developed by fusing features to generate a kernel matrix. The kernel matrix is normally built with a kernel function that has 2 patterns as argument and gives output value. This research employs a radial basis function (RBF) kernel. One- against-rest is used to assemble the classifier which distinguishes individual class. This technique contains of building 1 SVM in each class and training is given to differentiate sample of 1 class from the others. Generally, unknown pattern's classification is performed based on the highest output of all the SVM.

$$k(x_i, y_i) = e^{-\gamma \|x_i - y_i\|^{Fit_p}}, i = j = 1, 2, \dots, n \tag{2}$$

wherein  $x_i$  represents the input vector,  $y_i$  represents the  $j$ th prototype vector, and  $Fit_p$  correctly-classified / total number of testing data. Lastly, the best solution can be given by usage of Lagrange technique,

$$L_p \equiv \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \zeta_i - \sum_{i=1}^m \alpha_i \{y_i (w \cdot x_i + b) - 1 + \zeta_i\}, \tag{3}$$

$$L_D \equiv \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \alpha_i \alpha_j y_i y_j^k (x_i, y_j)$$

where  $\|w\|$  is the Euclidean norm of  $w$ ,  $\alpha_i$  which denotes Lagrange multipliers,  $L_p$  represents the Lagrange function, and  $L_D$  represents the dual solution of  $L_p$ .  $C$  as well as  $\gamma$  are utilized to monitor the tradeoff between training error as well as generalization capability in SVM with RBF kernel. Hence PSO is used to obtain the best fusion of  $C$  and  $\gamma$ .

The usage of an SVM to classify the image is an instance for linear discrimination. In the base model, it is a binary classifier, that implies that it splits the space into where the MR image is distributed into 2 classes through identification of a dividing hyper plane. The inspiration in implementing SVM is that it applies the principles of structural risk minimization that has the objective to obtain a hyper plane that increases the length between training classes.

## RESULTS

[Table- 1] shows the summary of results. [Figure- 2 to 4] shows the classification accuracy, sensitivity and specificity respectively.

Table: 1. Summary of Results

	C=1, gamma =0.01	C=10, gamma =0.01	C=100, gamma =0.01	C=1, gamma =0.001	C=10, gamma =0.001	C=100, gamma =0.001
Classification Accuracy	84.67	77.33	86.67	86	81.33	86.75
AD- Sensitivity	0.8667	0.7111	0.8889	0.9111	0.7556	0.8696
MCI - Sensitivity	0.8333	0.8333	0.8667	0.8667	0.8667	0.9167
CS-Sensitivity	0.8444	0.7556	0.8444	0.8	0.8	0.8
AD- Specificity	0.8889	0.875	0.9	0.9072	0.898	0.9192
MCI - Specificity	0.9167	0.8462	0.9286	0.9167	0.875	0.9268
CS- Specificity	0.9468	0.8913	0.9583	0.949	0.9149	0.9406

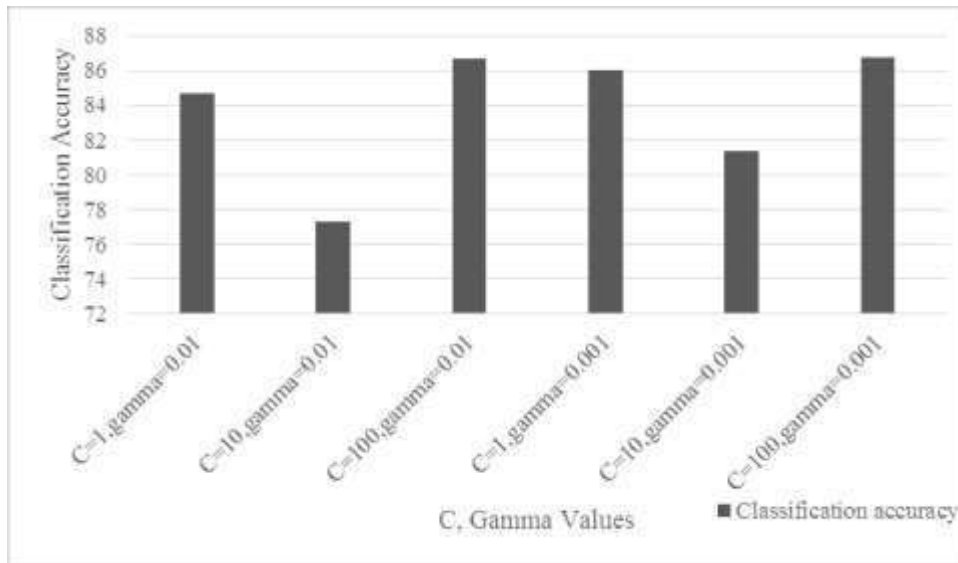


Fig: 2. Classification Accuracy

[Table- 1] and [Figure- 2] shows that the classification accuracy of C=100, Gamma=0.001 performs better by 2.43% than C=1, Gamma=0.01, by 11.48% than C=10, Gamma=0.01, by 0.09% than C=100, Gamma=0.01, by 0.87% than C=1, Gamma=0.001 and by 6.45% than C=10, Gamma=0.001.

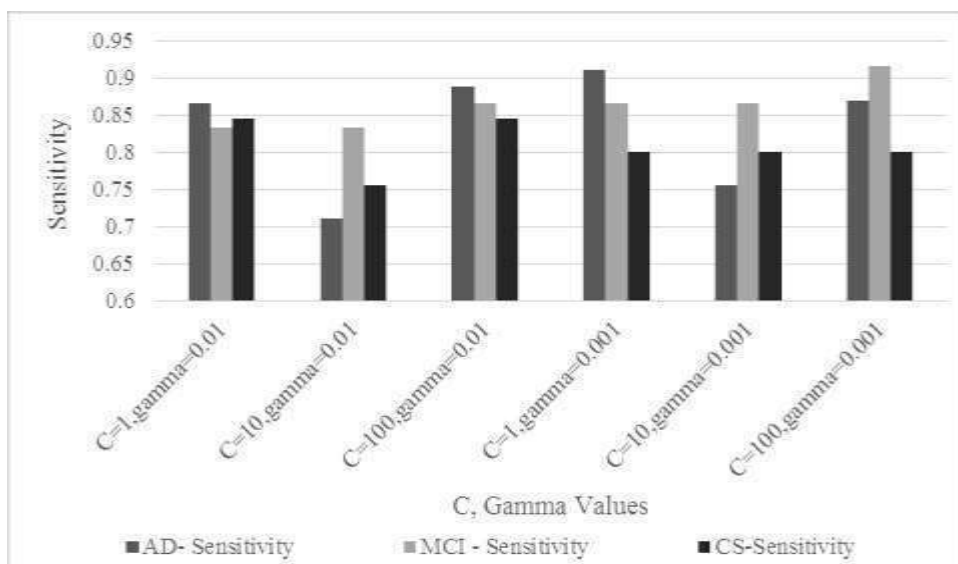


Fig: 3. Sensitivity

[Table- 1] and [Figure- 3] shows that the sensitivity of AD sensitivity performs better than MCI and CS for C=100, Gamma=0.01. Results show that the average values for C=100, Gamma=0.01 performs better by 2.16% than C=1, Gamma=0.01, by 12.24% than C=10, Gamma=0.01, by 0.86% than C=1, Gamma=0.001, by 7.08% than C=10, Gamma=0.001 and by 0.53% than C=100, Gamma=0.001.

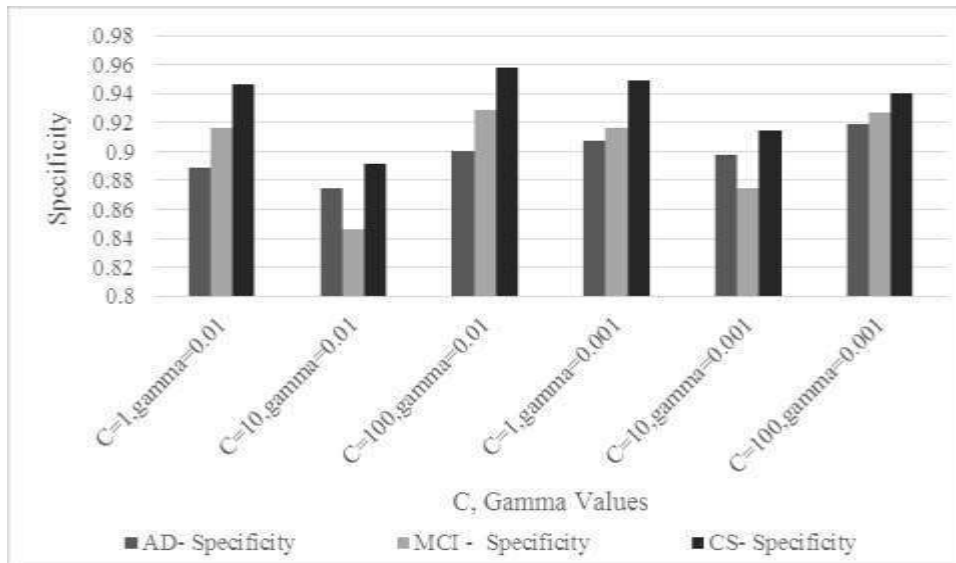


Fig: 4. Specificity

[Table- 1] and [Figure- 4] shows that the specificity of AD specificity performs better than MCI and CS for C=100, Gamma=0.01. Results show that the average values for C=100, Gamma=0.01 performs better by 1.25% than C=1, Gamma=0.01, by 6.5% than C=10, Gamma=0.01, by 0.5% than C=1, Gamma=0.001, by 3.62% than C=10, Gamma=0.001 and by 0.01% than C=100, Gamma=0.001.

## CONCLUSION

Alzheimer's Disease (AD), a general kind of dementia, is noticed very often in the people of old age. In order to detect AD, one of the best imaging methods is MRI that permits the quantitative estimation of features of brain to assess AD through a non-invasive method. PSO conducts search through a particle swarm that updates from each round of search. In order to find the best solution, every particle flies in the direction of previously known best position (pbest) and the best global position in the swarm (gbest). To assess the performance of every obtained solution, SVM is constructed. To experiment AD, MCI and CS specificity and Sensitivity are utilized Result shows that the classification accuracy of C=100, Gamma=0.001 performs better by 2.43% than C=1, Gamma=0.01, by 11.48% than C=10, Gamma=0.01, by 0.09% than C=100, Gamma=0.01, by 0.87% than C=1, Gamma=0.001 and by 6.45% than C=10, Gamma=0.001.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Alzheimer's disease facts and figures, [2012] [http://www.alz.org/alzheimers\\_disease\\_facts\\_figures.asp?type=homepage](http://www.alz.org/alzheimers_disease_facts_figures.asp?type=homepage).
- [2] Wang S, Phillips P, Yang J, Sun P, Zhang Y. [2016] Magnetic resonance brain classification by a novel binary particle swarm optimization with mutation and time-varying acceleration coefficients. *Biomedical Engineering/Biomedizinische Technik*.
- [3] Jongkreangkrai C, Vichianin, Y, Tocharoenchai C, Arimura H. [2016, March]. Alzheimer's Disease Neuroimaging Initiative. Computer-aided classification of Alzheimer's disease based on support vector machine with combination of cerebral image features in MRI. In *Journal of*

- Physics: Conference Series* 694(1): 012036). IOP Publishing.
- [4] Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Alzheimer's Disease Neuroimaging Initiative. [2015] Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage* 104:398-412.
  - [5] Zhang Y, Wang S, Ji G, Dong Z. [2013] An MR brain images classifier system via particle swarm optimization and kernel support vector machine. *The Scientific World Journal*.
  - [6] Lee W, Park B, Han K. [2015] SVM-Based Classification of Diffusion Tensor Imaging Data for Diagnosing Alzheimer's Disease and Mild Cognitive Impairment. In International Conference on Intelligent Computing(pp. 489-499). Springer International Publishing.
  - [7] Yang ST, Lee JD, Chang TC, Huang C H, Wang J. J, Hsu WC, Li KY. [2013]. Discrimination between Alzheimer's disease and mild cognitive impairment using SOM and PSO-SVM. *Computational and mathematical methods in medicine*.
  - [8] Ramírez J, Górriz JM, Salas-Gonzalez D, Romero A, López M, Álvarez I, Gómez-Río M. [2013] Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences* 237:59-72.
  - [9] Zhang Y, Wang S, Ji, G. [2015] A comprehensive survey on particle swarm optimization algorithm and its applications. *Mathematical Problems in Engineering* 38.
  - [10] Kennedy J. [1997, April]. The particle swarm: social adaptation of knowledge. In *Evolutionary Computation, 1997. IEEE International Conference on* (pp. 303-308). IEEE.
  - [11] Cortes C, Vapnik V. [1995] Support-vector networks. *Machine learning* 20(3): 273-297.

# AN EFFICIENT DATA PROCESSING ARCHITECTURE FOR SMART ENVIRONMENTS USING LARGE SCALE MACHINE LEARNING

A Mahesh\*, P Manimegalai

Karpagam University, Coimbatore, Tamil Nadu. INDIA

## ABSTRACT

The evolution of semiconductor devices used in embedded systems are requires an Internet connection for data processing and analytics in an automation system. The increasing number of everyday embedded devices that are interconnected over the Internet leads to the need of new software solutions for managing them in a proficient, scalable, and elegant way. In addition, these devices may produce a large volume of data, this can lead to classic Big Data problem that need to be stored and processed. Huge volume of information are being produced by the Social networking users and this data can be high in velocity and the traditional database fails to store the data which is generated, hence this raises two main issues, storing the information generated by the users/IoTs and processing this big data for analytics. In this project we address the above issues by implementing an efficient, fault tolerant and data immutability architecture for big data processing in addition to that applied large scale machine learning techniques such as pattern recognition for the Intelligent Reports

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

IoT, Big Data, Machine Learning Algorithms, Decision Making

\*Corresponding author: Email: [legacymahesh@gmail.com](mailto:legacymahesh@gmail.com), [manimegalai.vairavan@gmail.com](mailto:manimegalai.vairavan@gmail.com)

## INTRODUCTION

The Internet of Things (IoT) is the network of physical objects or "things" embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data. Processing the wide variety of data which the sensors may produce during the data acquisition on the smart environments in high velocity, these data provides the classic Big Data problem, Big Data is a technology where the processing or querying can't be done with traditional Relational Database Management System (RDBMS), the characteristics of the Big Data are: Volume, Velocity and Variety. In which Volume describes the size of the data sets which are produced in an embedded systems usually these size of the datasets varies for few millions of Bytes to PB (Peta Bytes). Velocity describes the speed in which the data arrives to the processing environment this varies from 10 messages per second to 1000 messages per second. Variety talks about the different format of data such as structured, semi-structured and un-structured data.

Acquiring the data from the sensors are collected through an API (Application Programming Interface), in the present decade, however the data from an automation system is in high velocity, however the smart device or an electronic sensor were connected through Wifi - WiMax, Zig-bee and many, such technology usually provided an associated IP address in IPV6. As we know that the address space of the IPV6 is 128 bits long, so we can add the device to IPV6 network by providing the address to that device being connected to the internet this is what we called Internet of things (IoT). To process these data to fast-decision making we need an architecture to process the same, Lambda Architecture [1], is a solution for the processing the big data for pattern extraction. In this proposed work, we implement the smart-home as a case study with appropriate machine learning algorithm is mapped from Spark MLlib for decision making or prediction or classification in the smart-home [2]. For example, to close the windows curtains automatically, based on the intensity of the light. To sense the intensity of the light is done using an LDR (Light Dependent Resistor). The value of the LDR is applied to the System that is built using Raspberry Pi. Raspberry Pi has an associated motor controller to open the curtain if the intensity inside a Smart Home is lesser than a threshold value.

The main contribution in this paper is, to develop an architecture/framework that suitably parallel processing of sensor data for quick decision making.

The remainder of this paper is organized as follows, Section II describes the Background of related technologies used in this paper. Section 3 describes the architecture for parallel processing of the sensor data. Section 4 describes the implementation details of our work. Section 5 describes the related work and finally Section 6 concludes.

## BACKGROUND

### RASPBERRY PI 2

The Raspberry Pi is a series of credit card-sized single-board computers developed in the United Kingdom by the Raspberry Pi Foundation with the intent to promote the teaching of basic computer science in schools and developing countries. The original Raspberry Pi and Raspberry Pi 2 are manufactured in several board configurations through licensed manufacturing agreements with Newark element, RS.

Components and Egoman. The hardware is the same across all manufacturers. The firmware is closed-source. Several generations of Raspberry Pi's have been released. The first generation (Pi 1) was released in February 2012 in basic model A and a higher specification model B. A+ and B+ models were released a year later. Raspberry Pi 2 model B was released in February 2015 and Raspberry Pi 3 model B in February 2016. These boards are priced between 20 and 35 US\$. A cut down "compute" model was released in April 2014 and a Pi Zero with smaller footprint and limited IO (GPIO) capabilities released in November 2015 for 5 US\$.

All models feature a Broadcom system on a chip (SOC), which includes an ARM compatible CPU and an on chip graphics processing unit GPU (a VideoCore IV). CPU speed ranges from 700 MHz to 1.2 GHz for the Pi 3 and on board memory range from 256 MB to 1 GB RAM. Secure Digital SD cards are used to store the operating system and program memory in either the SDHC or MicroSDHC sizes. Most boards have between one and four USB slots, HDMI and composite video output, and a 3.5 mm phono jack for audio. Lower level output is provided by a number of GPIO pins which support common protocols like I2C. Some models have an 8P8C Ethernet port and the Pi 3 has on board WiFi 802.11n and Bluetooth.

**[Figure- 1]** The Foundation provides Debian and Arch Linux ARM distributions for download, and promotes Python as the main programming language, with support for BBC BASIC (via the RISC OS image or the Brandy Basic clone for Linux), C, C++, PHP, Java, Perl, Ruby, Squeak Smalltalk and more also available.

**[Figure- 1]** shows the sample Raspberry Pi Board and layout<sup>1</sup>



**Fig: 1. Sample Raspberry Pi 2 Board**

### SMART HOME

Home automation is the residential extension of building automation and involves the control and automation of lighting, heating, ventilation, air conditioning (HVAC), appliances, and security. Modern systems generally consist of switches and sensors connected to a central hub sometimes called a "gateway" from which the system is controlled with a user interface that is interacted either with a wall mounted terminal, mobile phone software, tablet computer or a web interface.

While there are many competing vendors, there are very few world-wide accepted industry standards and the smart home space is heavily fragmented. Popular suites of products include X10, Ethernet, RS-485, ZigBee and Z-Wave, or other proprietary

<sup>1</sup> [https://en.wikipedia.org/wiki/Raspberry\\_Pi](https://en.wikipedia.org/wiki/Raspberry_Pi)

protocols all of which are incompatible with each other. Manufacturers often prevent independent implementations by withholding documentation and by suing people.

## SPARK MLlib

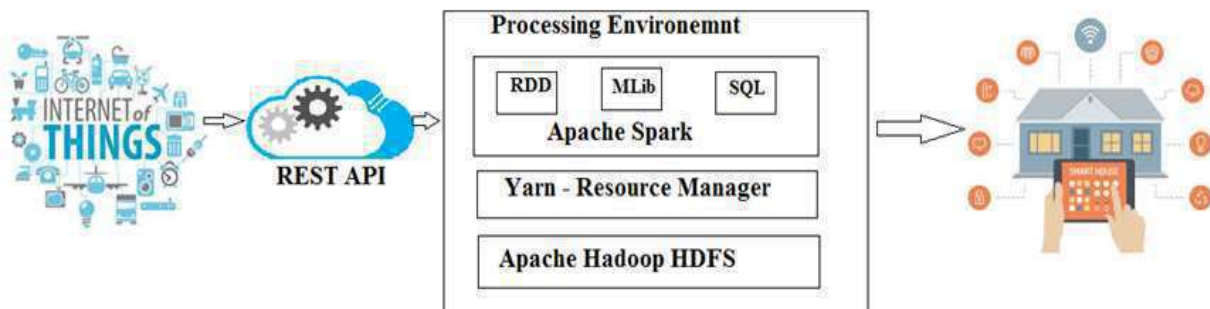
MLlib is Spark's machine learning (ML) library [3]. Its goal is to make practical machine learning scalable and easy. It consists of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as lower-level optimization primitives and higher-level pipeline APIs.

- Summary statistics, correlations, stratified sampling, hypothesis testing, random data generation
- Classification and regression: state vector machines, logistic regression, linear regression, decision trees, naive Bayes classification
- Collaborative filtering techniques including alternating least squares (ALS)
- Cluster analysis methods including k-means, and Latent Dirichlet Allocation (LDA)
- Dimensionality reduction techniques such as singular value decomposition (SVD), and principal component analysis (PCA)
- Feature extraction and transformation functions
- Optimization primitives such as stochastic gradient descent, limited-memory BFGS (L-BFGS). It divides into two packages:
  - spark.mllib contains the original API built on top of RDDs.
  - spark.ml provides higher-level API built on top of DataFrames for constructing ML pipelines.

Using spark.ml is recommended because with DataFrames the API is more versatile and flexible. But we will keep supporting spark.mllib along with the development of spark.ml. Users should be comfortable using spark.mllib features and expect more features coming. Developers should contribute new algorithms to spark.ml if they fit the ML pipeline concept well, e.g., feature extractors and transformers.

The main difference between Hadoop MR and Spark MLlib will come from underlying frameworks. In case of Mahout it is Hadoop MapReduce and in case of MLlib it is Spark. To be more specific - from the difference in per job overhead

If Your ML algorithm mapped to the single MR job - main difference will be only startup overhead, which is dozens of seconds for Hadoop MR, and let say 1 second for Spark. So in case of model training it is not that important. Things will be different if the algorithm is mapped to many jobs. In this case we will have the same difference on overhead per iteration and it can be game changer.



A Cloud Environment where the data gets persisted and processed for fast decision making.

Fig. 2. An Architecture for Smart Home

## SYSTEM ARCHITECTURE

[Figure-2] shows the complete System Architecture for processing the IoT data. This architecture receives the data that are emitted by IoT. A RESTful API is designed with MQTT protocol and send to the processing environment via the API. In the processing environment Apache Spark Streaming receives the data from MQTT Streaming Source, Apache Spark Streaming receives the data emitted by IoT, and MQTT Streaming and creates a mini batch of data is send to Apache Spark Core for Parallel Processing on top of Apache Hadoop Yarn. The data flow diagram of this process is shown in [Figure- 6].

Apache Spark Streaming creates many mini batch of dataset that can be manually configured in the Apache Spark Streaming Context. This allows to process the real-time data seamlessly. For each mini batch of data, Apache Spark Core applies the feature extraction on the mini batch of data, the extracted features are stored in the HDFS

in the *libsvm* format. Followed by, Apache Spark MLlib takes the pre-processed data from the HDFS. The required ML algorithm will be applied on the datasets which are complex and in real-time.

The data for the analytics is extracted from HDFS, then feature extraction is applied for the pattern recognition and finally the training model is built based on the extracted features using some Machine Learning tools. There are several technology alternatives available to build a training model on the real-time data for activity detection on the Smart Home application.

## IMPLEMENTATION

In this section, the most suitable analytics tool, for analyzing, extracting the patterns and reduce the usage of energy in a Smart Home. In this paper, a rule based activity detection has been explored for the IoT dataset [4] and Random Forest [5] classification technique has been applied using Apache Spark Streaming to evaluate the time taken to build the training model

DataSet

A batch of time-series datasets has been stored in the HDFS. This data is being used as a training data for building the classification model. In this paper, this has been processed by Apache Spark Core. To test a new data against the model to calculate the accuracy of the classification and also to predict the unknown classes for certain data, Apache Spark Streaming is used. The predicted values of the test data are retrieved from the HDFS. Two different smart houses were deployed and data over the months has been collected. The duration of the deployment for “House A” was 82 days with around 22.5 million collected data points. For “House B”, [4] had deployed the system for 62 days with about 20 million collected data points.

[Figure-3] shows the sample pre-processed data in *libsvm* format.

```

IoT-libsvm.txt
7 1:185 2:118 3:16 4:0 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
7 1:185 2:118 3:16 4:0 5:0 6:0 7:0 8:0 9:0 10:2 11:2 12:0 13:0 14:2 15:191
7 1:185 2:116 3:40 4:0 5:0 6:0 7:0 8:0 9:0 10:2 11:2 12:0 13:0 14:2 15:191
7 1:185 2:114 3:47 4:0 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:228 2:182 3:0 4:198 5:0 6:4 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:228 2:184 3:0 4:198 5:0 6:4 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:228 2:180 3:0 4:214 5:0 6:4 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:221 2:112 3:31 4:222 5:0 6:4 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:222 2:112 3:5 4:222 5:0 6:4 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:222 2:116 3:13 4:236 5:0 6:4 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:222 2:122 3:6 4:228 5:0 6:4 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:222 2:124 3:3 4:248 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:223 2:126 3:15 4:252 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:223 2:130 3:49 4:258 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:224 2:130 3:0 4:262 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:224 2:130 3:0 4:254 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:191
8 1:226 2:132 3:2 4:232 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:229 2:138 3:163 4:238 5:4 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:231 2:142 3:37 4:238 5:0 6:0 7:0 8:0 9:0 10:2 11:2 12:0 13:0 14:2 15:192
8 1:233 2:150 3:2 4:234 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:234 2:154 3:2 4:230 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:234 2:160 3:6 4:226 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:234 2:162 3:0 4:206 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:235 2:162 3:0 4:206 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:236 2:154 3:0 4:184 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:237 2:148 3:0 4:152 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:238 2:168 3:0 4:170 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:192
8 1:261 2:02 3:0 4:12 5:0 6:0 7:0 8:0 9:0 10:0 11:0 12:0 13:0 14:2 15:190
8 1:262 2:00 3:0 4:28 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:190
8 1:262 2:78 3:0 4:28 5:0 6:0 7:0 8:0 9:0 10:2 11:0 12:0 13:0 14:2 15:190
  
```

Fig. 3. IoT libsvm Pre-processed data

## FEATURE EXTRACTION

The time series data consists of a sequential set of activities accompanied with another sequential set of sensor readings. In order to train the machine learner, it is essential to build a training set which includes training instances that match the current activity with the features extracted from the sensor data. Therefore, the sequential sensor data needs to be divided into time windows so that each window can be provided as an input to the learning algorithm with the activity during this window as an output.

[Figure-4] The time series data is divided into equally sized time slots. For each time slot, the features extracted from sensor data are taken as an input and the current activity as an output. The time slots define the boundaries of the feature extraction algorithm. The algorithm runs for each time slot and extracts the features required for the machine learner. For each sensor, a feature is defined that represents the arithmetic mean of the readings related to this sensor during the time slot as shown in the following equation.



$$F_x = \frac{\sum_{i=1}^n r_i}{n}$$

Where  $F_x$  is the feature that represents sensor  $x$ ,  $n$  is the number of sensor readings during the time slot, and  $r_i$  is the sensor reading  $i$ .

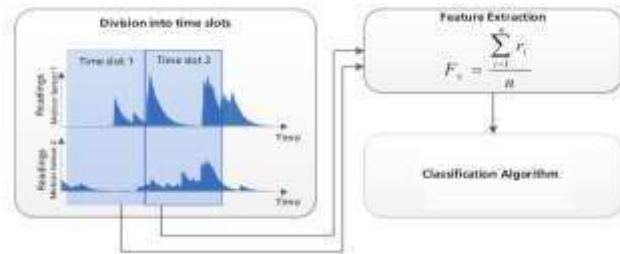


Fig: 4. Feature Extraction Dataset

## RESULTS

Spark MLlib supports random forests for binary and multiclass classification and for regression, using both continuous and categorical features. spark.mllib implements random forests using the existing decision tree implementation.

Random Forest Classification [5] on the real-time IoT dataset [4] using Apache Spark MLlib is executed as follows. Feature Extraction Algorithm is applied on the time-series data as mentioned in Section B, and data cleaning is performed to execute Random Forest in Spark MLlib. Spark accepts input training files in *libsvm* format.

[Figure-5] and [Figure-6] 90% of the real-time IoT data has been given for training and the remaining 10 % of the real-time IoT data has been given for testing the unknown instances and the accuracy of the model built is evaluated. The depth of the tree would be 9 level, and the number of trees are constructed is varied from 100 to 500.

The performance of the proposed method as follows.

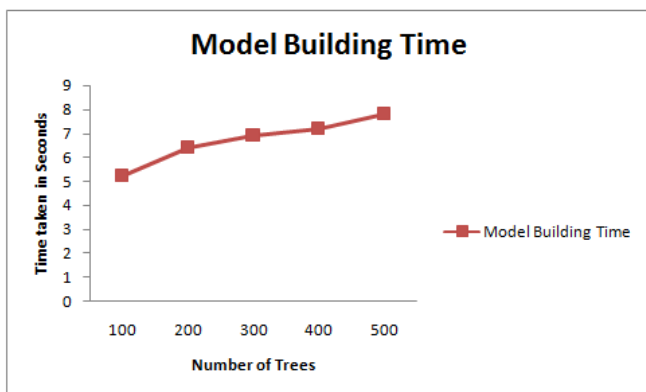


Fig: 5. Time taken for building the Model

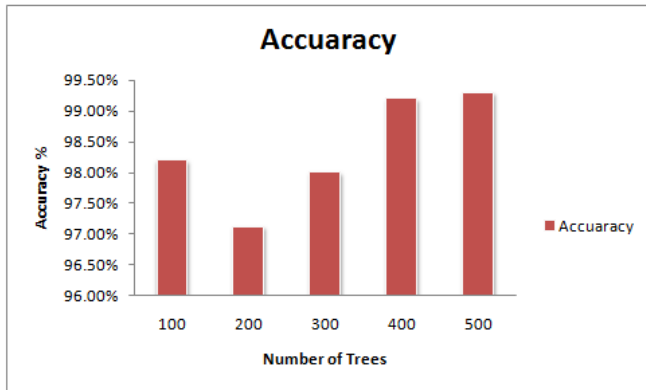
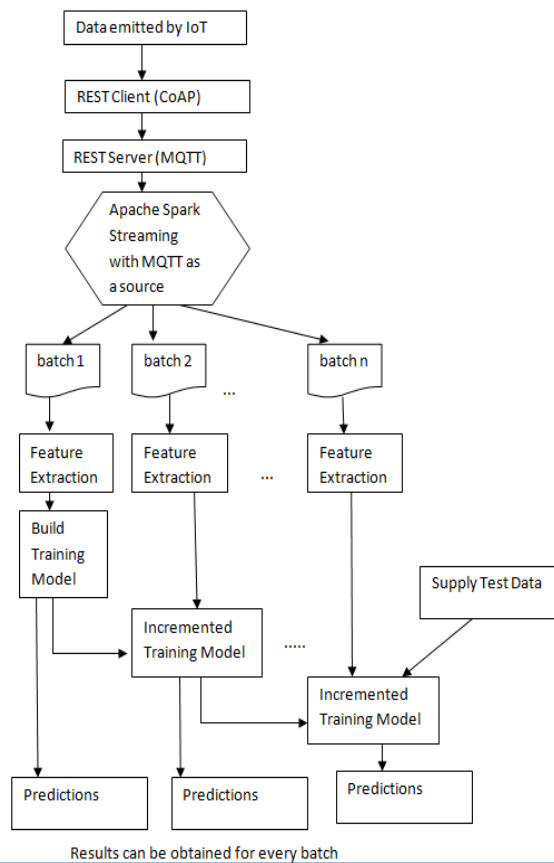


Fig: 6. Accuracy of the Proposed Approach

## RELATED WORKS

[Figure-7] AllJoyn, is a system that allows devices to communicate with other devices around them. A simple example would be a motion sensor letting a light bulb know that no one is in the room it is lighting, so it can shut itself off. AllJoyn is an open source software framework that makes it easy for devices and Mobile Applications to discover and communicate with each other. Developers can write applications for interoperability regardless of transport layer, manufacturer, and without the need for Internet access. The software is openly available for developers to download, and runs on popular platforms such as Linux and Linux-based Android, iOS, and Windows, including many other lightweight real-time operating systems.

Real, digital and virtual worlds are converging to create environments that make cities, energy, transport and many other services smart. In this context, the Internet of Things (IoT) is a very challenging paradigm that enables things, i.e., embedded devices, to be ideally available anytime and anywhere to anyone. Integrating heterogeneous systems is the main goal of IoT. In this work, Lambda Architecture is used for processing and recognizing patterns. Batch Layer of LA is implemented using Apache Hadoop, Similarly, Speed Layer of LA is implemented using Apache Storm.



**Fig: 7. Data Flow**

The idea of energy efficient Smart Homes or Home Automation System has attracted many researchers in academic as well as in industrial domains [6-8]. The main challenge in this domain is to capture and process the data which are emitted by each sensors in a smart home. Energy efficient smart homes are similar applications in which helps to reduce the amount of energy consumed by a particular home. An activity based energy advisor will work in non-intrusive way. Detecting the serious of activities on a particular home will helps to monitor where the energy is getting wasted through a alert message received by the smart phone send by the central server.

**[Figure- 8]** Alhamoud et al. developed a SMARTENERGY.KOM an activity based intelligent system for energy saving in smart home. The system architecture of SMARTENERGY.KOM is shown.

The major sensors to sense the information in a home such as temperature, light and humidity can be sensed by this system and sent to the central server through a Raspberri Pi. Central server is responsible for collecting these data from the Raspberri Pi which acts as a gateway of the sensors and to process these data to reduce the energy usage in smart home. This work uses Traditional MySQL as a database for storing the sensor data, However, Relational Database will becomes to process if the data grows and also the number of smart homes increases [9-10].

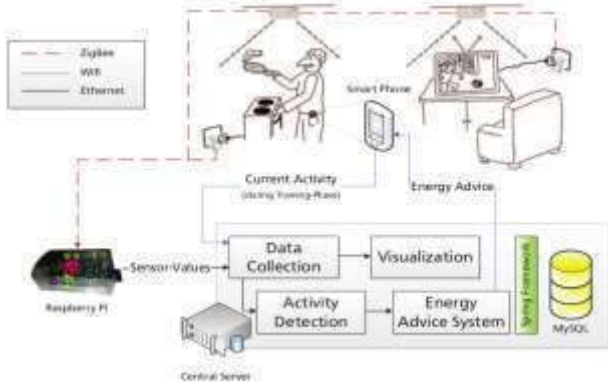


Fig: 8. Smart Home Application

[Figure- 9] Alhamoud et al. have used a standard Random Forest classification for activity detection in the smart home. This algorithm has been implemented in a WEKA. However, this algorithm can be distributed to multiple servers for quick processing and decision making. A. Alhamoud, et. al. had deployed 2 smart homes HomeA and HomeB where the researcher Alhamoud was the user of HomeA and for the HomeB a non-technical user activities were recorded. The sensors connected in both the homes are as follows.

Room	In House A	In House B
Kitchen	Coffee machine, Radio, Electric kettle, Lamp above the hob, Lamp above the kitchen sink, Oven, Fridge	Radio, Electric kettle, Electric iron, Cooking mixer, Electric bread cutter
Living room	Projector, Audio system, Lamp	Television, Satellite receiver, Lamp
Office room	PC, PC accessories (screen, loud-speaker, etc.)	-
Bedroom	-	Lamp, Radio alarm clock
Technical room	Warm water	-

Fig: 9. Appliances at Each House

## CONCLUSION

This paper has presented a novel architecture for efficiently processing large scale data using machine learning algorithms. MQTT streaming has been implemented in Apache Spark Streaming for collecting real-time IoT data. Various mini batches of datasets are processed in parallel with the cluster. A cluster consisting of 3 machines has been set up. The performance of the proposed approach has been analyzed for incrementally increasing the number of trees constructed in the Random Forest for activity classification. The proposed approach proves that it provides better performance as compared to the existing implementation.

As a future of this work, extending the data processing and analyzing using Apache Spark H2O (A Deep Learning tool). Deep Learning enables the training model and feature extraction to be selected from large scale data and constructs the model as its own.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Villari M, Celesti A, Fazio M, Puliafito A. [2014]. Alliovn lambda: An architecture for the management of smart environments in iot. In *Smart Computing Workshops (SMARTCOMP Workshops), 2014 International Conference on* (pp. 9-14). IEEE.
- [2] Dixit A, Naik A. [2014] Use of Prediction Algorithms in Smart Homes. *International Journal of Machine Learning and Computing*, 4(2): 157.
- [3] Kabir, M. H. Hoque, M. R. Seo, H. & Yang, S. H. [2015] Machine Learning Based Adaptive Context-Aware System for Smart Home Environment. *International Journal of Smart Home* 9(11):55-62.
- [4] Alhamoud A, Ruettiger F, Reinhardt A, Englert F, Burestahler D, Böhnstedt D, Steinmetz R. [2014]. Smartenergy. kom: An intelligent system for energy saving in smart home. In *Local Computer Networks Workshops (LCN Workshops) 2014 IEEE 39th Conference on* (pp. 685-692). IEEE.
- [5] [https://www.stat.berkeley.edu/~breiman/RandomForests/cc\\_home.htm](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm)
- [6] Choi S, Kim E, Oh S. [2013]. Human behavior prediction for smart homes using deep learning. In *2013 IEEE RO-MAN* (pp. 173-179). IEEE.
- [7] Barker S, Mishra A, Irwin D, Cecchet, E, Shenoy, P, & Albrecht, J. [2012]. Smart\*: An open data set and tools for enabling research in sustainable homes. *SustKDD, August*, 111-112.
- [8] <https://www.raspberrypi.org/documentation/>
- [9] <http://spark.apache.org/docs/latest/mllib-guide.html>
- [10] <https://www.smarthome.com/>

# A COMPARITIVE FRAMEWORK FOR FEATURE SELCTION IN PRIVACY PRESERVING DATA MINING TECHNIQUES USING PSO AND K-ANONUMIZATION

S Mohana\*, SA Sahaaya, Arul Mary

M.I.E.T Engineering College, Jayaram College of Engineering &Tech, Tamil Nadu, INDIA

## ABSTRACT

The trend of technological era leads to accumulate and utilization of enormous quantity of private details of individuals using internet, which eventually lead to disclose their personal identities. Privacy preserving of data must uphold from revealing sensitive data during the disclosure of the individual's data. Privacy preserving should be incorporated as mining of these datums and the domain deals with this known as Privacy Preserving Data Mining. In the proposed framework, an attribute suppression technique is employed using Particle swarm optimization algorithm and a generalization technique for anonymization is proposed. Also the same work is done using k anonymization and the results are compared for classification accuracy, Precision and recall. In the proposed system Genetic Algorithm and Particle Swarm Optimization takes the common population for evaluation and the results are compared. An optimal generalized feature set is acquired by the PSO and k anonymization technique and is which is used for classification task. The end results of classification are compared with average classification accuracy, average precision and average recall.

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Privacy Preserving Data Mining (PPDM), K-Anonymity Generalization, Particle swarm optimization (PSO), Classification accuracy)

\*Corresponding author: Email: [mohana.p3@gmail.com](mailto:mohana.p3@gmail.com), [samjessi@gmail.com](mailto:samjessi@gmail.com)

## INTRODUCTION

Alzheimer's Data mining process is digging out useful information from the huge capacity of data. Data mining tasks have been classified as association rule mining, clustering, classification and prediction. Gathering data for mining, may leads to the collection of the private data which identifies personal details of individuals must be confined without affecting the data mining process. The main objective of PPDM is to aggregate the information available in the data by not leaking the individual information of the participants. There are two main privacy models. Privacy-preserving data mining (PPDM) algorithms are built in such a way that the private data on which data mining is employed should not publicize the client private information. PPDM is broadly classified in to non-interactive model and interactive model. In the non-interactive model, the database is sanitized and revealed to public. The interactive model deals with accessing the desired data by asking multiple questions to the database and getting the answers from the database. Many methods for privacy preserving have been proposed so far. Some of the methods to preserve privacy are, preserving privacy while publishing the data, changing the results of data mining for preserving privacy and changing or restricting the results of a query to preserve the privacy either online or offline. Perturbation, Randomization, k-anonymity, t-d closeness and l-diversity are some of the methods used for privacy preserving while publishing the data [1].

Anonymization of data converts a dataset in to a form that maintains privacy using k-anonymity, so that individually identifiable information is covered. K Anonymization converts data to equivalence classes where of the classes has a set of K- fields indistinguishable from one another.

Generalization technique is used in k-anonymity, where some values are replaced with less specific but retaining the meaning of the value and some of the values are suppressed. For example, country can be generalized to state, age is generalized to age-range such as 30-35 yrs or young, middle, old etc which leads to less identification of individual's data. Data loss can be minimized through optimization of an aggregated value in all

features and records.

Particle Swarm Optimization (PSO) is population based heuristic search technique, which is usually employed for solving the NP-hard problems. In Particle Swarm Optimization, a particle which can be idle because of the stagnant way and also suboptimal solutions were got because of early congregation. During optimization problem, reducing or maximizing an objective function is the difficulty encountered. Global optimization is the best set of permissible environment achievable for an objective in given constraints. Many fusions of algorithms integrating GA are proposed to overcome the limitations of PSO.

In this paper, classification accuracy is compared to investigate the effect of anonymization for the Adult dataset with respect to k anonymization and PSO. PSO is used to select the features of the data set. Performance of classifier is analyzed with PSO selection and K Anonymization for different levels. Sections 2 presents the Literature review of previous work; Section 3 describes the how PSO is used to select to select features from the anonymous data. Section 4 reports on evaluation of various aspects of the proposed work and concludes the paper.

## RELATED WORK

There are two different approaches for privacy preserving in data mining suggested by Wu [2]. perturbation is used as a first approach in which perturbing the data using a random process is employed. Cryptographic methods for multi party computing is used as Second approach used. After preserving the privacy association rule mining was applied to the data. The data flow in uni-direction in the processor boards was taken and the author compared the impact of privacy preserving in rule mining. The results proved that cryptographic methods were better than data perturbation method.

PPDM needs accurate models for data aggregation without accessing the precise information in the data record of individuals. Perturbation-based PPDM approach was widely used for preserving data before publishing the data. This approach was limited because it used trust only on data miners. Therefore, Li et al [3] suggested Multilevel Trust (MLT-PPDM). The trustful data miner was one, which access less the perturbed copy of data. Malicious miner could access the data more times in various means and jointly used them to infer the original data which was not published by the owner of the data. This attack was reduced by correlating the copies of data among different miners and trust level was created based on the correlation level. This trust level was used when miner requested data. Since many users are unwilling to share the personal data, many users give incorrect information. This may affect the results of data mining because of not having sufficient amount of correct information [4]. The dimensionality of the information is also large and selection of good privacy preserving is needed for the success of data mining.

Kadampur et al [5] proposed a strategy to protect the privacy of data during decision tree construction of data mining process. With the numeric attributes of data a specific noise is added and then given to second party to construct the decision tree by CART algorithm. The best split point was found based on the info gain measures. The decision obtained from the original tree was compared with the decision constructed by the second party using the obfuscated data. The comparison proved that both the trees were similar. Therefore this method preserved the privacy.

Lindell et al [6] analyzed how multiparty secure communication is applied to privacy preserving in data mining. Authors implemented two different approaches for privacy preserving. In the first approach, original data was divided into many partitions and distributed to multiple parties. While performing mining, these partitions were united by not allowing each party to see the individual data stored in other parties. In this multiparty secure communication the goals were set by using the properties such as privacy, correctness, fairness, independence of inputs and guaranteed output delivery. In the second method, from the original data only statistical information is calculated and released for data mining. The mining results were compared for these two approaches. Results proved that multiparty secure communication was better than statistical information for data mining.

Sumana and Hareesh [7] proposed a k-anonymity model by generalization and suppression to protect the identities of individuals while releasing truthful information. This k-anonymity model protected against the identity disclosure, but it could not protect against the individual attribute disclosure. The  $\ell$ -diversity method solved this problem by using equivalence classes and each class had at least  $\ell$  well-defined values for each sensitive attribute. The author proposed a complete  $(\alpha, k)$  model by using the distance between two distributions

with a threshold  $\alpha$ . Results proved that complete model preserved the privacy is better than simple k-anonymity model.

Sharing the patient data is often needed for the purpose of research. But the identity of individuals must be protected. Most commonly used methods to hide the personal details are k-anonymity and l-diversity models. Tamas et al [8] analyzed these two models with the details of cancer patients which were published to health professionals. After implementing these models, discernibility was used to compare the performance. Result proved that l-diversity was better for single sensitive attribute and k-anonymity model was better for multiple sensitive attribute.

Most of the privacy preserving algorithms are based on various privacy and utility assumption. Bingchun et al [9] proposed creation of decision tree from the anonymated data directly. This method avoided the data preparation by the ID3 algorithm. Experimental results showed that proposed decision tree from the k-anonymated data performed efficiently for classification problems.

In k-anonymity model, generalization technique is used to swap a sensitive value with a less specific value and maintaining semantically consistent value, and suppression was used to hide a value at all. Generalization was commonly used, because suppression may lessen the quality of the data mining results if not used efficiently. But in generalization every quasi-identifier needs to consider the hierarchy of the domain. Therefore, Kisilevich et al [10] proposed multidimensional suppression for generating classification trees. Multidimensional suppression was used based on the attribute values without using the domain hierarchy trees. Experiments were conducted with 10 different data sets and the results proved that classification accuracy was improved up to 5.3% than manual classification and classification tree from generalized data.

Mandapati et al [11] proposed a Hybrid Evolutionary Algorithm using Particle Swarm Optimization (PSO) Genetic Algorithm (GA) and for PPDM. While preserving privacy, the entire existing EA algorithm produced solutions which were restricted to specific problems like the cost function evaluation. Authors proposed both the GA and PSO with the same population and, k-anonymity was used to generalize the actual dataset. The hybrid optimization found the optimal generalized feature set and the improved the success of mining.

Slava Kisilevich et al [12] proposed a method to achieve k-Anonymity of Classification Trees by Using Suppression (kACTUS) where kACTUS performs an efficient multi-dimensional suppression, in which, suppression is done only on certain records based on other attribute values, without manually-producing domain hierarchy trees. Results proved that kACTUS' predictive performance was good than the k-anonymity. Also, average the accuracies of TDS, TDR and kADET are lower than kACTUS in 3.5%, 3.3% and 1.9% correspondingly regardless of usage of manually defined domain trees.

Goryczka et al [13] proposed m-privacy in the anonymity model which preserves the privacy constraint against in any group of m number of colluding providers of data. Heuristic algorithms were used for data aware anonymization which provides the m-privacy efficiently. Experiments were conducted on the real data sets and the efficiency of the proposed algorithm was compared with base line algorithms which provided m-privacy.

## MATERIALS AND METHOD

**[Table- 1]** The Adult dataset from UCI machine learning Repository is used for assessment. There are 48,842 rows, containing both categorical and integer attributes derived from Census information from the year 1994. There are about 32,000 rows with 4 numerical columns, and the column contains age {17 – 90}, fmlwgt {10000 – 1500000}, hrsweek {1 – 100} and edunum {1 – 16}. k- anonymization is employed in age column and native country. Original attributes of the Adult dataset is shown in **Table- 1**.

**Table: 1. Attributes of the Adult Dataset**

Age	native-country	Class
39	United-States	<=50K
50	United-States	<=50K
38	United-States	<=50K
53	United-States	<=50K
28	Cuba	<=50K
37	United-States	<=50K



49	Jamaica	<=50K
52	United-States	>50K
31	United-States	>50K
42	United-States	>50K

### K-ANONYMIZATION

[Figure- 1] In k-anonymity, the data is changed to equivalence classes , in which each class consist of a set of k- records that diverges from K others. suppression & Generalization techniques are employed to lessen the minute sign of the pseudo-identifiers. The features are generalized to a series so as to lessen the microscopic view, for example, street is generalized as city and it prevents the disclosure of individual's information. Suppression is used to remove the value of the attribute in order to reduce the identification risk with the records available publically and the example is shown. Because of its easiness the k-anonymity is a popularly used technique and also many techniques are existing to practice anonymization [14].

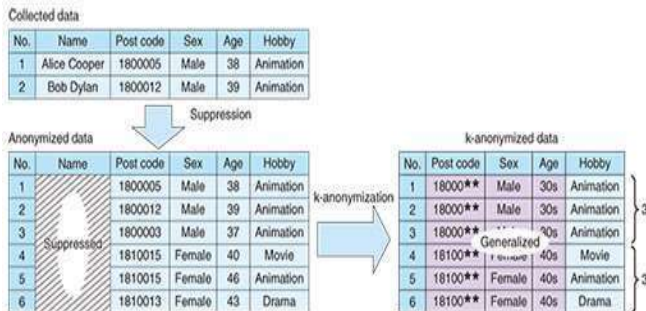


Fig. 1. 3-Anonymous Table

### THE PARTICLE SWARM OPTIMIZATION

[Figure- 2] It is a well-built optimization technique based on the behavior and attitude of swarms. It shares the idea of group communication to solving of problem. It makes use of a amount of particles that characterize a swarm going around in the search space affording the best result. every particle is consider as a point in a K-dimensional space which alters its "flying" having its own flying knowledge as well as the flying knowledge of supplementary particles. [11] best balue can be obtained by keeping track of its coordinates in the result space which are associated with the finest solution which is known as personal best , **pbest**. **hbest is the value** obtained so far by any particle in the neighborhood of that particle.

The notion of PSO falls in moving each particle in the direction of its pbest and the hbest position.

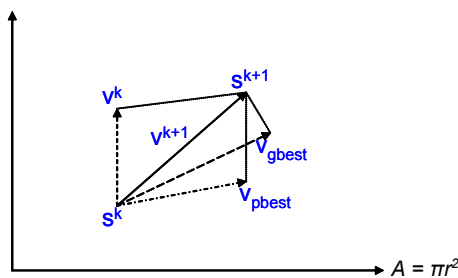


Fig. 2. modifying the searching point

- S<sup>k</sup>: current searching point
- S<sup>k+1</sup>: modified searching point
- V<sup>k</sup>: current velocity
- V<sup>k+1</sup>: modified velocity
- V<sub>pbest</sub>: velocity based on pbest
- V<sub>gbest</sub>: Velocity based on hbest

Every particle aims to adjust its place using information from the existing positions, the existing velocities, the distance involving the present position and pbest, the distance involving the present position and hbest. The modification of the particle's position can be mathematically modeled according the following equation :

$$V_{ik+1} = wV_{ik} + c_1 rand1(\dots) \times (p_{besti} - S_{ik}) + C_2 + rand2(\dots) \times (g_{best} - S_i^k) \dots (1)$$

Where

- V<sub>ik</sub>: velocity of agent i at iteration k,
- W: weighting function,

$C^i$ : weighting factor,  
 $rand$ : uniformly distributed random number between 0 and 1,  
 $S_i^k$ : current position of agent  $i$  at iteration  $k$ ,  
 $Pbest_i$ :  $Pbest_i$  of agent  $i$ ,  
 $Gbest$ :  $Gbest$  of the group

The following weighting function is usually utilized in (1)

$$W = W_{max} - \frac{(W_{max} - W_{min}) \times iter}{maxiter} \dots\dots\dots (2)$$

where  $W_{max}$  = initial weight,  
 $W_{min}$  = final weight,  
 $maxiter$  = maximum iteration number,  
 $iter$  = current iteration number.

$$S_i^{k+1} = S_i^k + V_i^{k+1} \dots\dots\dots (3)$$

In the present paper, the 'Adult' dataset available in the UCI machine learning repository is used. Adult Dataset provides the 1994 Census information. The dataset contains 48842 instances, with both categorical and integer attributes. There are about 32,000 rows and 4 numerical columns present in the Adult Data set. The columns and their ranges are: age[17 - 90], fnlwgt [10000 - 1500000], hrsweek[1 - 100] and edunum[1 - 16]. The age column and the native country were aggregated using the principles of K anonymization. Table I and II show the original data and the modified attribute data. Using 10 fold cross validation the original and the K anonymized dataset are classified.

The Naïve Bayes classifier is popular because it is more efficient. It also has the benefit of having fine classification accuracy and is employed in a number of areas. Using Bayes theorem the classifier model is formulated as:

$$p(C|F_1, \dots, F_n) = \frac{p(C) p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}$$

**Table: 2. The K-Anonymous Dataset**

Age	Job	Class
Old	Employed	Good
Young	Employed	Bad
middle age	Employed	Good
middle age	Employed	Good
middle age	Employed	Bad
Young	Employed	Good
middle age	Employed	Good
Young	Employed	Good
Old	Employed	Good
Young	Employed	Bad

**Table: 3. The Original Attributes Of Adult Dataset**

Age	native-country	Class
39	United-States	<=50K
50	United-States	<=50K
38	United-States	<=50K
53	United-States	<=50K
28	Cuba	<=50K
37	United-States	<=50K
49	Jamaica	<=50K
52	United-States	>50K
31	United-States	>50K
42	United-States	>50K

Anonymization is achieved using attribute suppression technique and generalization using Particle swarm optimization algorithm. Also the same work is done using k anonymization for different levels of K and the results are compared for classification accuracy, Precision and recall using Bayesian classifier. Both K Anonymization and PSO work with the same population in the

system proposed and the results are compared. An optimal generalized feature set is acquired by the PSO and k anonymization technique which is used for classification task. The end result of classification are compared with average classification accuracy, average precision and average recall. In this paper it is proposed to compare the classification accuracy of Naive Bayes anonymized dataset for PSO and k anonymization. As the anonymization complexity increases it is observed that the classification accuracy of K-Anonymization outrages the classification accuracy of PSO.

## RESULTS

[Figure- 3] The classification accuracy obtained from Naïve Bayes is shown. It is shown that the classification accuracy of k Anonymization outrages the classification accuracy of PSO.

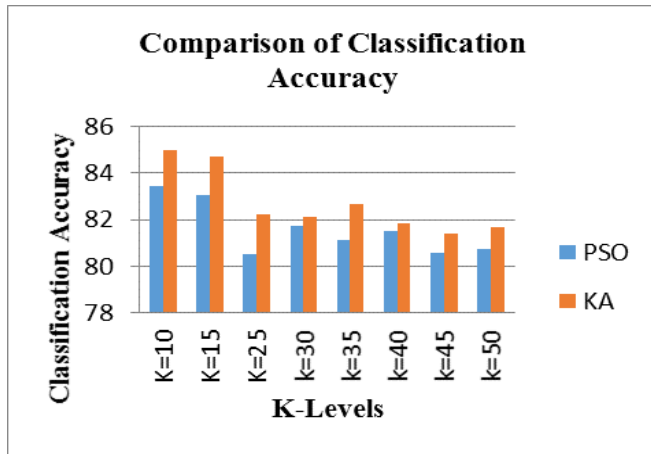


Fig: 3. Classification accuracy of Naïve Bayes for PSO and K Anonymization

[Figure- 4] The Precision value of the classification accuracy is shown in [Figure- 4]. It is shown that the precision of k anonymization outrages the Precision value of

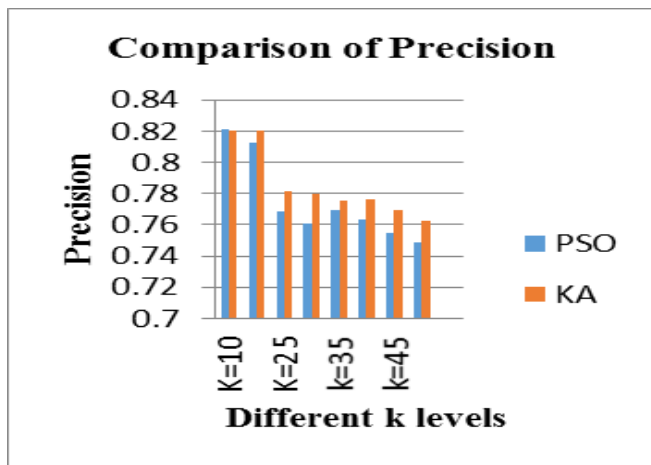


Fig: 4. Comparison of Precision for PSO and K Anonymization

[Figure- 5] The Recall value of the classification accuracy is shown in [Figure- 5]. It is shown that the Recall value of k Anonymization outrages the Recall value of PSO

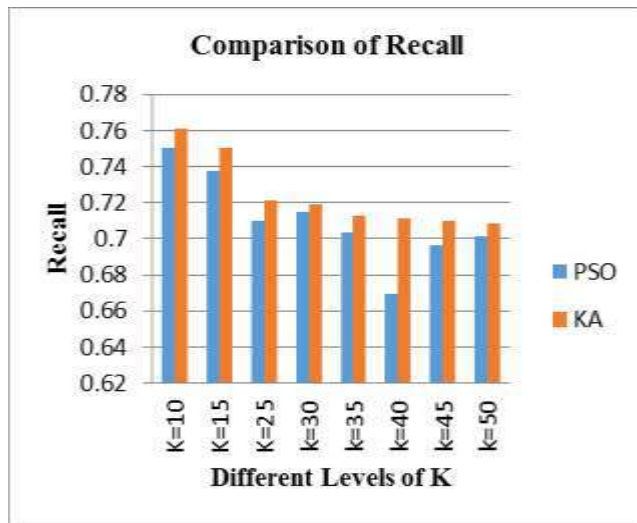


Fig: 5. Comparison of Recall for PSO and K Anonymization

## CONCLUSION

In this work, it is proposed to implement feature selection by using PSO and K anonymization technique. To validate the results classification accuracy of Particle Swarm Optimization (PSO) and k anonymization technique is compared. K-anonymization outrages PSO feature selection which is evaluated in terms of Classification accuracy, Precision and Recall. K-anonymity is accomplished by generalization and suppression of the original dataset. For different levels of k-anonymity experiments were performed and the results achieved are evaluated.

## CONFLICT OF INTEREST

The authors declare no conflict of interests.

## ACKNOWLEDGEMENT

None

## FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Machanavajjhala A, Kifer D, Gehrke J, Venkatasubramanian M. [2007]. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1): 3.
- [2] Wu CW. [2005] Privacy preserving data mining with unidirectional interaction. In *2005 IEEE International Symposium on Circuits and Systems* (pp. 5521-5524). IEEE.
- [3] Li Y, Chen M, Li Q, Zhang W. [2012] Enabling multilevel trust in privacy preserving data mining. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1598-1612.
- [4] Wang J, Luo Y, Zhao Y, Le J. [2009] A survey on privacy preserving data mining. In *2009 First International Workshop on Database Technology and Applications* (pp. 111-114). IEEE.
- [5] Kadampur MA. [2010] A noise Addition scheme in Decision tree for privacy preserving data mining. *arXiv preprint arXiv:1001.3504*.
- [6] Lindell, Y, & Pinkas, B. [2009]. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality* 1(1):5.
- [7] Sumana M, Hareesha KS. [2010]. Anonymity: An Assessment and Perspective in Privacy Preserving Data Mining. *International Journal of Computer Applications*, 6(10).
- [8] Gal TS, Chen Z, Gangopadhyay A. [2008] A privacy protection model for patient data with multiple sensitive attributes. *IGI Global*, 28-44.
- [9] Bingchun L, Guohua, L. [2011] The Classification of k-anonymity Data. In *Computational Intelligence and Security (CIS), 2011 Seventh*

- International Conference on* (pp. 1374-1378).  
IEEE.
- [10] Kisilevich, S, Rokach, L, Elovici, Y, & Shapira, B. [2010]. Efficient multidimensional suppression for k-anonymity. *IEEE Transactions on Knowledge and Data Engineering*, 22(3): 334-347.
  - [11] Mandapati S, Bhogapathi RB, Chekka RB. [2013] A Hybrid Algorithm for Privacy Preserving in Data Mining. *International Journal of Intelligent Systems and Applications* 5(8):47.
  - [12] Kisilevich, S, Elovici, Y, Shapira, B, & Rokach, L. [2009]. kACTUS 2: privacy preserving in classification tasks using k-anonymity. In *Protecting Persons While Protecting the People* (pp. 63-81). Springer Berlin Heidelberg.
  - [13] Goryczka S, Xiong L, Fung BC. [2014] -Privacy for Collaborative Data Publishing. *Ieee Transactions On Knowledge And Data Engineering* 26(10):2520-2533.
  - [14] El Emam, K, & Dankar, F. K. [2008]. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association* 15(5): 627-637.

## A FUSION TECHNIQUE TO CLASSIFY GLAUCOMA FROM FUNDUS IMAGES

M Nageswara Rao<sup>1\*</sup>, M Venu Gopala Rao<sup>2</sup><sup>1</sup>Research scholar, Dept. E.C.E, K L University, INDIA<sup>2</sup>Dept. of E.C.E, K L University, INDIA

## ABSTRACT

*Glaucoma is a common cause of blindness and it is increasingly becoming more severe when taking into consideration the aging population. As the dead retinal nerve fibres are not healable, earlier detection as well as prevention of glaucoma is crucial. Resilient and automatic mass-screening will assist in the extension of symptoms-free life for the patients. A new automatic appearance-based glaucoma classification system which does not rely on segmentation-based measures is suggested here. It employs image fusion, which refers to the procedure of fusing images obtained from various sources for acquiring improved situational awareness. The goal of fusion of source images is the combination of highly relevant data from sources into one composite image. Genetic Algorithm (GA) is utilized for features selection. Features extraction methods utilized are Discrete Wavelet Transform (DWT) utilized for multiresolution fusion as well as Local Binary Patterns (LBP) for texture features. Outcomes prove that the suggested model attains excellent glaucoma classification.*

Published on: 2<sup>nd</sup> -December-2016

## KEY WORDS

*Glaucoma, Image Fusion, Wavelet Transform, Local Binary Patterns (LBP), Genetic Algorithm (GA), Random Forest (RF), Bagging and Boosting*

\*Corresponding author: Email: [nagmedikonda@gmail.com](mailto:nagmedikonda@gmail.com)

## INTRODUCTION

Glaucoma is a common cause of blindness amongst retinal diseases with 13% of the cases getting affected. The changes happen in retinal structure that gradually results in loss of peripheral vision and in the end leads to blindness if it is not treated in time. No cure is present for Glaucoma, however, its earlier detection as well as treatment helps in preventing loss of vision. Because the procedure of manual diagnosis is expensive as well as error-prone, effort has been made toward automated detection of Glaucoma in its earlier stage [1].

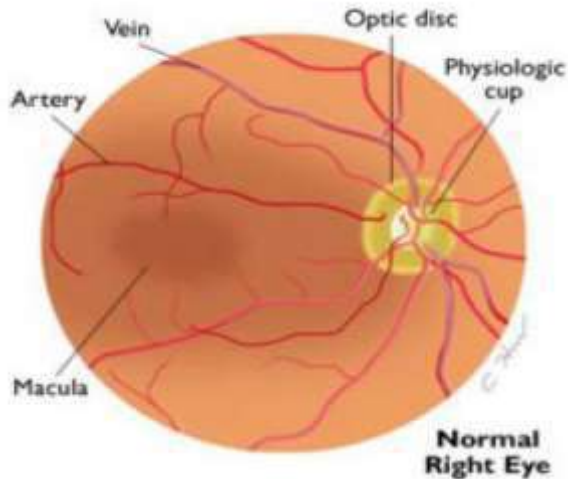
Glaucoma refers to a set of eye diseases that are related to concurrent functional failures in visual field. Changes in structure are symptomized by a slow diminishing neuro-retinal rim denoting degeneration of axons as well as astrocytes of the Optic Nerve. Because any lost capacity of the optic nerve is not recoverable, earlier detection as well as care is crucial for patients to retain their vision.

Two major kinds of Glaucoma include: 1) Primary Open Angle Glaucoma (POAG) as well as (ii) Angle Closure Glaucoma (ACG). The former progresses in a slow manner and at times without any significant loss in vision for several years. Treatment involves medication if an early diagnosis is made. The latter requires surgery as a small portion of the outer edge of the iris is to be removed.

In latest studies, a lot effort has been put into automatic diagnosis of glaucoma on the basis of computer vision. The structure of glaucoma analysis systems is dependent on the types of image cue and image modality utilized. Amongst structural image cues learnt for diagnosing glaucoma, cues based on optic discs as well as cups are significant. Optic discs are situated near ganglion nerve fibres congregating in the retina. Optic cup is where the depression occurs in the optic disc from where the fibre comes out of the retina through the Optic Nerve Head (ONH). The borders of the cup and disc structures are required to be found since it helps the assessment of glaucoma cues like disc and cup asymmetry and high Cup-to-Disc Ratio (CDR), described as the ratio between vertical cup diameter to vertical disc diameter. Value of CDR is assessed by planimetry from color fundus images after outlining the optic discs and cups physically. As the process of manual annotation of cup and disc for every image involves

more labour and time, this work proposes automation through computer vision methods to segment the cups and discs from fundus images.

The term Fundus refers to the part of the organ opposite to the opening. The image [Figure 1] below shows the interior surface of the eye which includes retinal blood vessel, macula, fovea, optic disc as well as cup. Different factors from the image are required to be analysed to figure out the glaucoma disorders. ONH, CDR, rim disc ratio as well as Retinal Nerve Fibre Layer Height (RNFLH) are few of the significant factors in the assessment of ONH as well as Retinal Nerve Fibre Layer (RNFL). As glaucoma has impact on the optic disc as well as cup by converting the cup to disc ratio as well as rim to disk ratio, appropriate segmenting of the characteristics is necessary for detecting glaucoma [2].



**Fig: 1. Left Retinal Color Fundus Image**

Colour Fundus Imaging (CFI) is a modality through which glaucoma may be evaluated. It has also grown into a recommended modality for wide-scale screening of retinal diseases and moreover it has been already used for huge-scale screening of diabetic retinopathy. With this technology, fundus image can be acquired in a non-invasive way that is best suited to huge-scale screenings. In these programmes, automatic system determines any symptoms of doubt which indicate the presence of a disease and this improves effectiveness as only those images that contain those symptoms need ophthalmologist's attention. Many efforts had been made for automatic detection of glaucoma from a 3-D image. But, due to the high cost, this facility is not available at basic health care centres and hence could not be used properly for wide screening.

Combining multiple images' information of one scene is known as Image fusion where the images are obtained from various sensors, captured at various timings, and have various spatial or spectral features. The aim of this fusion is the retaining of the beneficial feature of every image. Because of the existing multi-sensor data in various fields, researchers are more focussed towards image fusion for its application in several fields. For instance, in multi-focus imaging, one or more objects might be focused in a certain image, however other objects in the scene might be focused on in another image. For remotely sensed images, few have excellent spectral data while the others have excellent geometric resolution.

Within the field of bio-medical imaging, two extensively utilised modalities are viz., the Magnetic Resonance Imaging (MRI) as well as the Computed Tomography (CT) scan do not reveal similar information on the structure of brain. CT scan is best suited to take the image of hard tissues or bone structures whereas MR imaging is best for the soft brain tissues which helps to detect diseases that affect the skull base. As the required information cannot be received through a single image, these images are complementary.

By fusing the images the best features of individual image are obtained. The integration of these features is a huge advantage. The primary stage in image fusion is registering of the image that brings the constituent images to general coordinate system, because image fusion is useful only if common objects in image have similar geometrical configuration with regard to size, region as well as orientation. This can also be named as the pre-

processing stage. Next, the images are put together to form a single image with a careful selection of various features taken from various images.

Fusion technique includes a simple technique of pixel averaging to even more complex techniques like principal component analysis and wavelet transform fusion. Various methods for fusing the image can be differentiated based on whether image fusion occurs in the spatial domain or transformed to another domain.

In this work, various feature extraction, feature selection and classifier techniques are proposed. And image fusion in Glaucoma is discussed. Section 2 reviews associated literature review. Section 3 describes methods applied and Section 4 details experimental result. Section 5 includes conclusion of the paper.

## RELATED WORK

Xu et al., [3] presents an unsupervised technique to segment optic cups in fundus images to detect glaucoma with no use of extra training images. The method adopts the super pixel framework and domain prior wherein the super pixel task of classification is designed as a Low-Rank Representation (LRR) problem along with an effective closed-form solution. Moreover, the author develops an appropriate scheme for automatic selection of the key criteria in LRR and obtains the end results for every image. Assessed on the common ORIGA dataset, the result shows that the method achieved optimal performance in comparison to existing methods.

Niwas et al., [4] suggests the discriminate features are chosen by various feature selection protocols to detect Primary Angle-Closure Glaucoma (PACG) on the basis of Anterior Segment Optical Coherence Tomography (AS-OCT) images. A new condition was recommended for further selection of more dependable attributes. The method depends on the selection of top-ranking attributes in every algorithm and the ranking combination to select the best feature.

Guo et al., [5] presents an analysis of fundus images based in a computer aided system to automate the classification and grading of cataract that is extremely helpful for reducing the workload of experienced ophthalmologists as well as assist cataract patients in developing regions to detect the disease in time receive treatment from health care providers. The wavelet transform as well as sketch based method are examined for extracting the best suited feature to classify and grade the condition of cataract from the fundus image.

Raja & Gangatharan [6] presents an automatic technique to detect glaucoma from the fundus image with the help of hyper-analytic wavelet transform. Enhancing the directional selectivity and preservation of the information on phase is done with the image hyper-analytic wavelet transform. Prior to the transforming, the required pre-processing steps like gray scale conversion as well as equalization of histogram are executed. Next the magnitude as well as phase spectra of the image are assessed. This confirmed that the accuracy of classification was increased by 5 %.

Niwas et al., [7] focuses on the application of repetitive features to diagnose complicated diseases like Angle-Closure Glaucoma (ACG) with the use of anterior segment optical coherence tomography image. Supervised [Minimum Redundancy Maximum Relevance (MRMR)] as well as unsupervised algorithms for [Laplacian score (L-score)] features selection are investigated with various ACG methods. The total accuracy has proven the utility of repetitive attributes by L-score technique in enhanced ACG diagnosis as opposed to minimally redundant attributes by MRMR technique.

Gupta & Awasthi [8] presents image fusion technique which provides optimal result with Discrete Wave Packet Decomposition (DWPT) as well as optimizes results with GA and then compared them with Intensity Hue Saturation (IHS) utilized to fuse the images. Finally the performance of proposed technique is assessed with its mean, standard deviation, entropy, variance, mutual information, Peak Signal to Noise Ratio (PSNR) as well as structure similitude.

Bock et al., [9] suggested an innovative automatic system for detecting glaucoma which functions with affordable as well as typically utilized digitized colourful fundus image. After the pre-processing step particular to glaucoma, various general types of features are compressed with appearance oriented dimension reduction techniques. Consequently, a probabilistic two-stage classification strategy merges the feature kinds for extracting the new Glaucoma Risk Index (GRI) which exhibits a remarkable performance on detecting glaucoma. With sample group of 575 fundus images, 80 % classification accuracy is attained in the five fold cross validation setup. The GRI achieves 88 % of competitive Area under ROC (AUC) in comparison to existing topography oriented glaucoma probability score of scanning laser tomography with AUC of 87 %.

## METHODOLOGY

In this section, feature extraction based wavelet transform and LBP are described. Feature selection based GA is discussed. The RF, bagging and boosting classifiers are described.

### FEATURE EXTRACTION

Feature extraction is a crucial step in fusing images. In the process of extracting features the raw image is transformed into frequency domain as well as sampling with wavelet transform function. Texture values of features extracted in horizontally, vertically and diagonal transform vector forms are provided by wavelet transform function. Merging these texture values, a feature matrix is created. Value of integer wavelet transform function is used for the feature extraction. Integer wavelet transform function



refers to a family of wavelet transform functions. The values of transforms usually produce the values of filter in whole number [10].

### WAVELET TRANSFORM

Wavelet coefficient measured by a wavelet transform represents modification in the time series at a specific resolution. With consideration of the time series at different resolutions one can filter out and perform processing of actual features of the image. The term wavelet thresholding is described as disintegration of the data or the image into wavelet coefficient when compared to the detail coefficient with the given threshold value, and diminishing the coefficient near to zero to remove the impact of features in the given data. From the altered coefficients, the image is rebuilt. This procedure is also called as inverse discrete wavelets transform. At the time of thresholding, a wavelet coefficient is compared against given threshold and is marked as zero if the magnitude is lesser than the given threshold; or else it is maintained or it is changed as per the threshold rule. Threshold differentiates the coefficients because of feature containing significant information on signals. Choosing the appropriate threshold is most important because it has significant role in the elimination of features as fusing of images result in sharpness reduced images otherwise known as smoothed images.

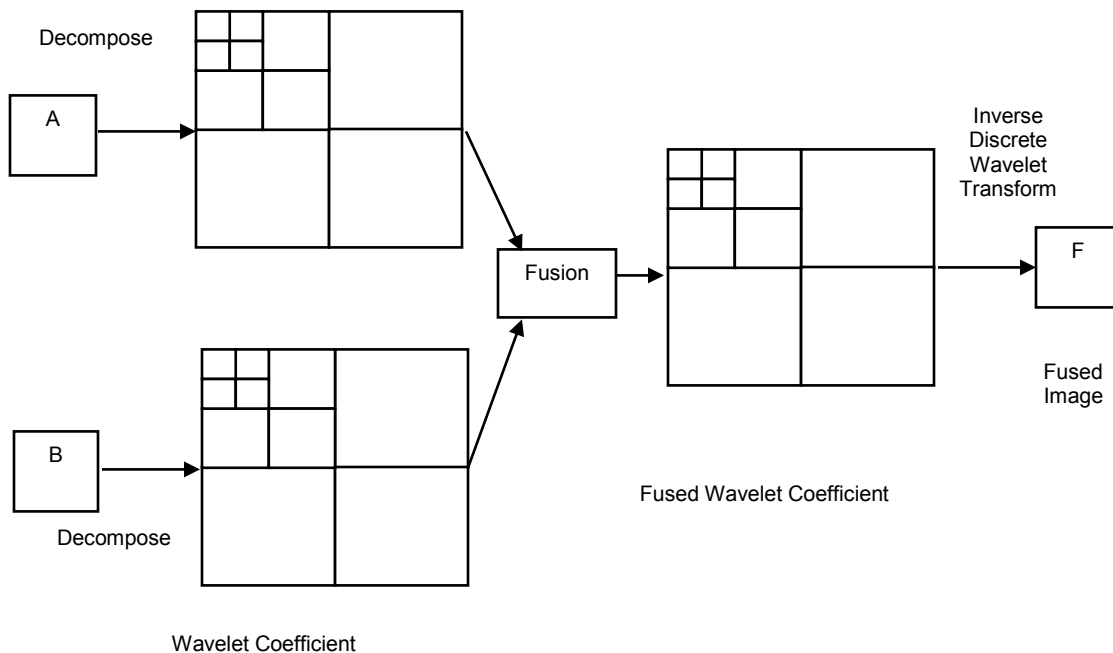
A signal analysis technique identical to image pyramid is the discrete wavelet transform. The major distinction is that the image pyramid leads to a complete collection of transform coefficients, whereas the wavelet transform leads to a non-redundant image representation. The discrete 2D wavelet transform is calculated by the recursive use of low pass as well as high pass filter in every direction of the input image after which sub-sampling is done. One of the key limitations of the wavelet transform while used for fusing of image is its common shift dependency which means that an easy shift of the input signal leads to absolutely different transform coefficient. This leads to inconsistent image fusion when brought up in image sequence fusion. To defeat this limitation of shift dependency the input image must be split into the shift invariant representations. [Figure 2] shows the wavelet based image fusion.

The segmentation of image is done by horizontal and vertical straight line and this indicates the first-order of DWT, with the given image being divided into four sections as LL1, LH1, HL1 and HH1. Common image fusion process with DWT:

**Step 1:** Implementation of Discrete Wavelet Transform on all source images for creating wavelet lower decomposition.

**Step 2:** Fusion of every decomposition stage with various fusion rules such as simple average, simple maximum, simple minimum, etc.

**Step 3:** Execute Inverse Discrete Wavelet Transform on fused decomposed level, for rebuilding of fused end image F.



**Fig: 2. Wavelet Based Image Fusion**

Two major group of transforms, continuous and discrete. Of specific attention is the DWT that is a spatial-frequency decomposition which will provide a flexible multi-resolution image analysis. In single dimension (1-D) the principle notion of the DWT is for representing the signal as a superposition of wavelet. Assume that a discrete signal is denoted by  $f(t)$ , then the wavelet decomposition is given as in equation (1) [11]:

$$f(t) = \sum_{m,n} C_{m,n} \psi_{m,n}(t), \quad (1)$$

Where  $\psi_{m,n}(t) = 2^{-m/2} \psi[2^{-m}t - n]$ ,  $m$  and  $n$  represent integers. There exists a special choice of  $\psi$  so that comprises an ortho-normal basis, therefore the wavelet transform coefficient can be got by the computation:

$$C_{m,n} = \langle f, \psi_{m,n} \rangle = \int \psi_{m,n}(t) f(t) dt \quad (2)$$

To build a multi-resolution analysis, a scaling function  $\phi$  is required along with the expanded and translated version of it  $\phi_{m,n}(t) = 2^{-m/2} \phi[2^{-m}t - n]$ . As per the features of the scale space spanned by  $\phi$  and  $\psi$  the signal  $f(t)$  can be decomposed in its coarse component and detail of different sizes by projection onto the respective space. Hence for finding such decomposition in an explicit manner, additional coefficient  $a_{m,n}$ , is needed at every scale. At every scale  $a_{m,n}$ , and  $a_{m-1,n}$ , describes the estimates of the function  $f$  at resolution  $2^m$  and at the coarser resolution  $2^{m-1}$  correspondingly, where the coefficient  $C_{m,n}$ , describes the loss of information while moving from an estimation to other. To acquire the coefficient  $C_{m,n}$ , and  $a_{m,n}$ , at every scale and position, a scaling function is required that is like equation (2). The approximation coefficient and wavelet coefficient can be got:

$$a_{m,n} = \sum_k h_{2n-k} a_{m-1,k}, \quad (3)$$

$$C_{m,n} = \sum_k g_{2n-k} a_{m-1,k}, \quad (4)$$

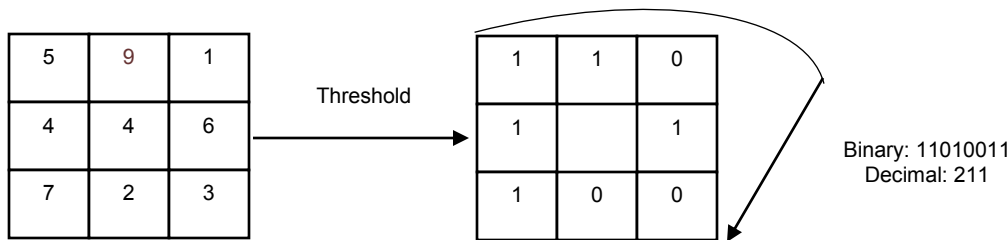
$h_n$  is a low pass FIR filter whereas  $g_n$  is associated high pass FIR filter. In order to rebuild the actual signal the analysis filter can be chosen from a bi-orthogonal set that has an associated collection of synthesis filters. The synthesis filters  $h\sim$  and  $g\sim$  could be utilized for appropriate reconstruction of the signal with the rebuilding equation

$$a_{m-1,1}(f) = \sum_n \left[ \tilde{h}_{2n-1} a_{m,n}(f) + \tilde{h}_{2n-1} C_{m,n}(f) \right] \quad (5)$$

Equations (3) and (4) are applied by filtering as well as down sampling. Conversely, equation (5) is applied by initial up sampling as well as consequent filtering

### LOCAL BINARY PATTERNS (LBP)

The actual LBP operators label the pixel of one image by decimal number known as LBP code that encodes the local structure around every pixel. It continues like this as shown in [Figure- 4]. It compares every pixel with its eight neighbours in a 3x3 neighbourhood by deducting the center pixel value; The finalised negative value is coded with 0 and the rest with 1; Through the clock wise concatenation of all the binary codes a binary number is obtained and its corresponding decimal value is applied for labelling. The derived binary numbers are referred to as LBP codes [12].



**Fig. 2. An example of the basic LBP operator**

The drawback of the base LBP operator is that its small 3x3 neighbourhood may not obtain dominant attributes in the large-scale structures. To handle with textures at various scales, the operator generalizes later to utilize neighbourhood of various sizes. A local neighbourhood is referred to a collection of sampling points spaced even on a circle that is at the centre of the pixel to be labelled, as well as the sampling point which does not fall among the pixels are interpolated with bilinear interpolation and hence allows for any radius and any quantity of neighbourhood sampling points.

Mathematically, given a pixel at  $(x_c, y_c)$ , the result LBP is shown in decimal form as equation (6):

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(i_p - i_c) 2^p \quad (6)$$

$i_c$  and  $i_p$  are corresponding gray-level value of the central pixel whereas P surrounding pixel in the circle neighbourhood with a radius R, and function  $s(x)$  is defined in equation (7):

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (7)$$

With the above description, the principle LBP operator is invariant to monotonic gray-scale transformation protecting pixel intensity order in the local neighbourhood. The LBP histogram label estimated over an area is utilized as a descriptor of texture.

The operators  $LBP_{(P,R)}$  produce  $2^P$  various output values, relating to  $2^P$  distinct binary pattern constructed by P pixels in the neighbourhood.

## FEATURE SELECTION

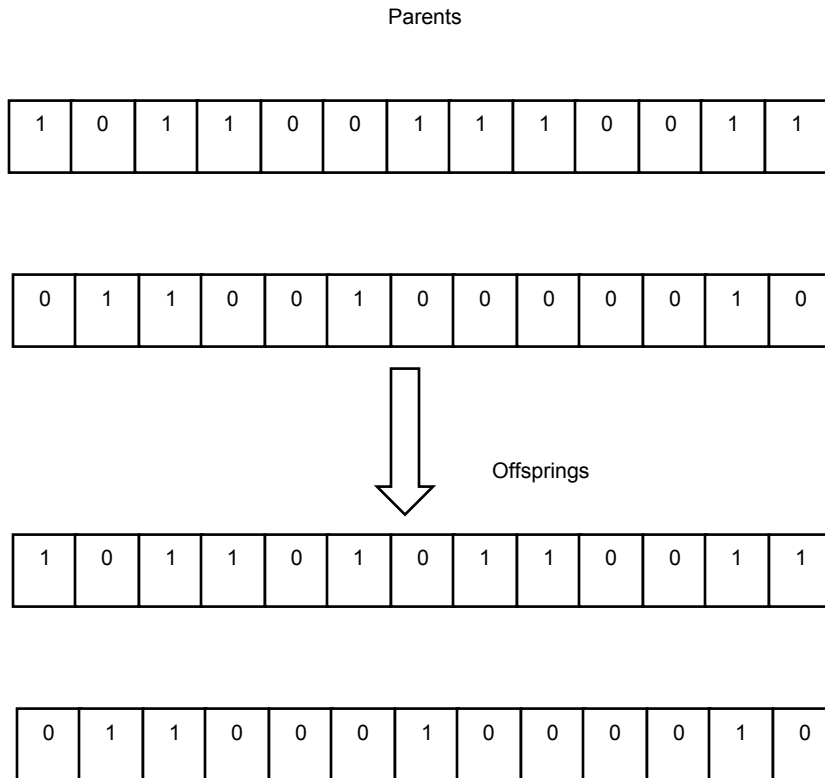
The process of feature selection may be taken as the process of finding and eliminating non-relevant, repetitive and random class co-related attributes. The feature selection problem is NP-hard. Hence the best solution is not assured to be found unless thorough search in the feature space is executed. Evolutionary algorithm such as GA has been extensively used for feature selection. GA is a random search method with the ability to explore effectively in large search space that is normally needed in case of selecting the attribute. Moreover, GA performs a global search whereas many other search algorithms perform local search. GA is a search algorithm which owes its inspiration to the idea of natural selection. The principle notion is the evolving of a populace of individuals, where every individual solution is a candidate solution for a given issue. Feature selection methods contain potential advantages such as: 1) Reduces the quantity of training data required for achieving learning, 2) Creates training model with enhanced accuracy on prediction, 3) Learned knowledge is more compact, simpler and easier to interpret, 4) Learning requires reduced execution time, and 5) Decreases storage requirement [13].

## GENETIC ALGORITHM (GA)

GA Optimization is performed by natural genetic substance exchange between the parents. Offspring is produced from parent gene. Offspring's fitness is assessed. Breeding is permitted only between the fittest individuals. GA is utilized in various areas like optimization of function, authenticating and controlling systems, processing of image, controller parameter optimization, multi-objective optimization, etc.

GA manipulates a populace of possible solutions for the issue to be addressed. Normally every solution is encoded as a binary string which is equal to the genetic substance of an individual in nature. All solutions are related to fitness values that reflect its goodness in comparison with another solution in the populace. Individual solution's fitness value and the survival chance and reproduction in the subsequent generation are directly proportional. Recombining of genetic substance in genetic algorithm is replicated by a crossover technique that will exchange parts among the strings. One more operation named mutation leads to sporadic and random bit alteration in string. Mutations have direct similarity in environment and play the roles of developing lost genetic substance. GA has application in many areas which includes image processing [14].

GA algorithm has a coding strategy and selection, crossover and mutation operator. An object is encoded with a classical binary code in the coding scheme. Roulette wheel method is applied for selection operator. In crossover step, single-point crossover produces new unit by swapping parts from the parent string at various crossover points. In binary coding, mutation operates by changing 0 to 1, or from 1 to 0. The crossover rate and mutation rate are pre-defined. GA has different phases progressing to a specified number of generations. This research focuses on, the chromosome that is coded in binary string bit with its size respective to the features and feature *i* is selected and is denoted by '1' or not by '0'. [Figure- 5] is a GA encoded sequential representation.



**Fig: 3. Example of GA Encoding for feature selection**

GA is capable of swift and efficient finding of an adequate solution for a difficult problem through probing a search space that is wide and complicated. [Figure- 6] shows the GA's flow diagram. Four attributes that helps to describe a GA problem are representation of the potential solution, the fitness function, the genetic operator to assist in identifying the best or near best solution and particular knowledge of problems like variables.

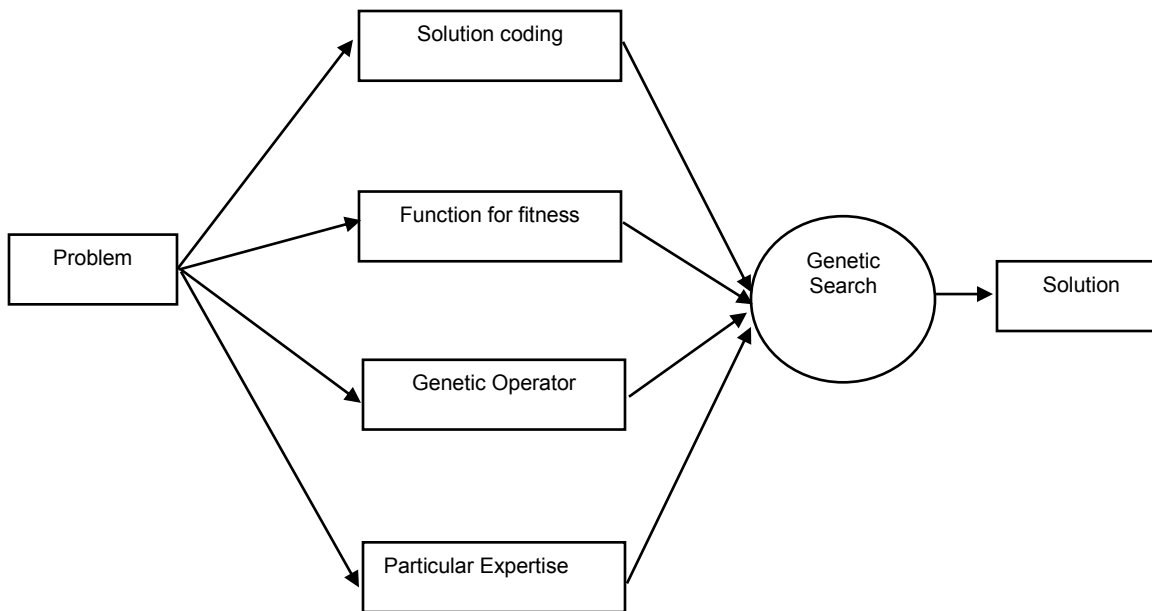


Fig: 4. Flow diagram of GA

GA provides solutions to the given issues through the genetic search part. The four major phases of genetic search are mentioned as:

**Initialization:** The randomly generated individual candidate solutions form the preliminary population. The size of population is based on the issue but normally comprises several hundred probable solutions which encompass the entire area of possible solutions in the search space.

**Evaluation:** The fitness function assesses every separate candidate in the populace and assists the procedure of selecting for the subsequent generation that depends on the fitness value.

**Selection:** A part of the available population is chosen for breeding a novel generation at the time of every subsequent generation. Individual candidate solution's fitness determines the candidate solution selection. Roulette wheel, rank-based, and tournament selection are few of the common selection methods.

**Genetic Operator:** Crossovers and mutations are two genetic operators which yield the novel populace for the subsequent generation after the selecting of the better individuals. A couple of "parent" solutions are chosen for breeding from the earlier selection pool. A novel solution is developed by generating a "child" solution with crossover and/or mutation. Novel candidate solution is chosen and process proceeds till a novel populace of solutions of proper size is produced.

## CLASSIFIERS

Classification refers to a data mining (machine learning) technique applied for prediction of group membership for data examples. Classification is the issue of detection of the class of data using existing known class and this is also known as supervised classification. Hence the requirement is that novel individual item is placed into a group depending on quantitative information on one or more metrics or features, also depending on the learning set which contains the already determined groups. The classification where no expert is available for predicting is known as unsupervised classification. A framework is defined in which the collection of attributes are defined where a number of classifiers and algorithms are also available. This work focuses on the RF, bagging and boosting classifiers.

## RANDOM FOREST (RF)

The RF algorithm based image fusion system is where the input image is classified into blocks of size  $M \times N$  pixels. The input data set is created by the extraction of spatial and frequency domain attributes. At the time of training, RF algorithm employs the feature information to generate the random tree of poor learners for ensemble decision. At the next stage, the prediction of poor learner in the form of decision tree is combined using the majority voting scheme [15].

The advantages of RF method are the generation of different kinds of random tree with decreased variance which means the poor learners. This kind of poor learner is required basically for ensemble. RF-based ensemble method can compensate for the drawback of a tree with the advantage of other. Hence, RF algorithm could efficiently use the diversity of arbitrary trees as opposed to individual methods. RF algorithm arbitrarily creates several tree classifiers and collates the predictions. The tree, thus

generated is trained on bootstrap sample of the training data set and utilizes random subsampling of attributes. This scheme provides resilience against excess training and over fitting of inputted data. RF method performs better with high dimensional features space, specifically in a small data set that has a highly complicated structure.

## BAGGING

Bagging technique is applied for enhancing the result of machine learning classification algorithms. This approach was designed by Leo Breiman and the name originates from "bootstrap aggregating". If classification is to be done into two potential class, the classification protocol generates a classifier  $H : D \rightarrow \{-1, 1\}$  based on the training dataset of example descriptions (in the case played by a document collection)  $D$ . The bagging approach produces a series of classifiers  $H_m, m=1, \dots, M$  with regard to alterations of the training dataset. The classifiers are merged into a compound classifier. Predicting the compound classifier is presented as a weighted combination of every single classifier prediction in equation (13):

$$H(d_i) = \text{sign} \left( \sum_{m=1}^M \alpha_m H_m(d_i) \right) \quad (13)$$

The above stated formula could be understood as a procedure for voting. An instance  $d_i$  is sorted to the class which receives major quantity of votes from classifiers. This is the theory of classifier voting. Parameters  $\alpha_m, m=1, \dots, M$  are decided such that more precise classifier has more effect on the end prediction than low accuracy classifier. The precision of base classifier  $H_m$  will be slightly high when compared to a random classification's precision. Hence these classifiers are known as feeble classifiers [16]. If the learning procedure is able to be influenced by the classifier  $H_m$  directly, the minimization of classification error by  $H_m$  is possible by maintaining parameters  $\alpha_m$  as constant.

## BOOSTING

The Boosting algorithm applied for classification and regression are AdaBoost and AdaBoost R2 correspondingly. The two algorithms subsequently generate a sequence of neural networks in which the training instances which are incorrectly predicted by the preceding neural networks play more vital role in the training of a latter network. The component prediction is merged through weighted averaging for regression task and weighted voting for classification task in which the weight is decided by the algorithm itself [17].

The AdaBoost algorithm was devised in 1995 by Freund and Schapire, implemented to solve most of the working complexities of previous boosting algorithms which is the objective of this work. Pseudo code for AdaBoost is mentioned in a generic form as given by Schapire and Singer. This algorithm is given a training dataset as input  $(x_1, y_1), \dots, (x_m, y_m)$  where each  $x_i$  belongs to certain domain or instance space  $X$ , and each label  $y_i$  is in a label set  $Y$ . For major part of this work, it is supposed that  $Y = \{-1, +1\}$ . AdaBoost, works with a poor or base learning algorithm iteratively in a sequence three of rounds  $t = 1, \dots, T$ . The major idea of the algorithm is to sustain a distribution or set of weights over the training set. Weights of this distribution on training instance  $i$ , on round  $t$  is represented as  $D_t(i)$ . In the beginning, every weight is set as equal but on every round the weight of wrongly classified instance is improved and hence the base learner focuses on the tough instance in the training set.

## RESULTS AND DISCUSSION

This section deals with the evaluation of Correlation based Feature Selection (CFS) - RF, CFS -Bagging, CFS - Adaboost, GA - RF, GA -Bagging and GA -Adaboost techniques. 360 normal images as well as 150 Glaucoma images are utilized. The classification accuracy, specificity, sensitivity and F-measure for both abnormal and normal are shown in the [Table- 1] and [Figure- 4 to 7]. The sample images 7 to 9 are shown in [Figure- 10 to 13].



Fig: 5. Sample Image 1



Fig: 6. Sample Image 2

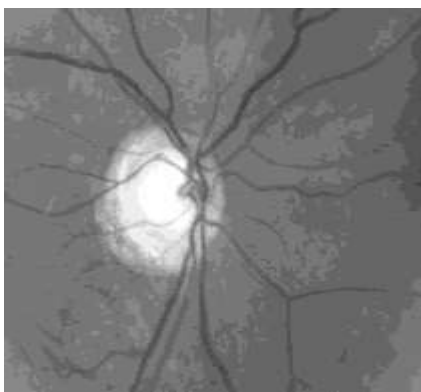


Fig: 7. Sample Image 3

Table: 1. Summary of Results

	CFS - Random forest	CFS - Bagging	CFS - Adaboost	GA - Random forest	GA - Bagging	GA -Adaboost
Classification Accuracy	85.1	87.25	89.02	89.61	90.59	91.18
Specificity for Abnormal	0.8722	0.8944	0.9111	0.9139	0.9194	0.9278
Specificity for Normal	0.8	0.82	0.84	0.8533	0.8733	0.8733
Sensitivity for Abnormal	0.8	0.82	0.84	0.8533	0.8733	0.8733

Sensitivity for Normal	0.8722	0.8944	0.9111	0.9139	0.9194	0.9278
F measure for Abnormal	0.7595	0.791	0.8182	0.8284	0.8452	0.8534
F measure for Normal	0.892	0.9083	0.9213	0.9255	0.9324	0.9369

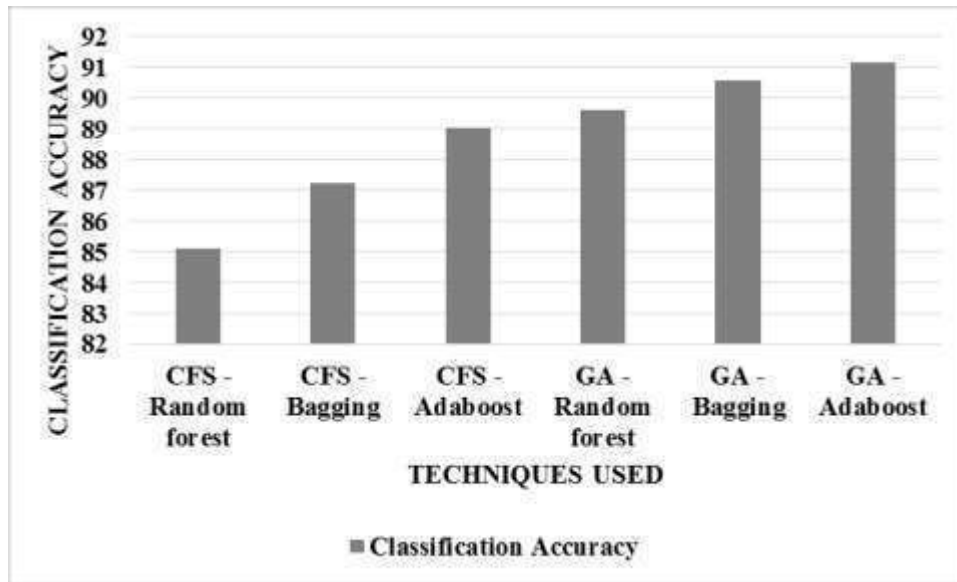


Fig:8 . Classification Accuracy

From the [Figure- 10], it can be observed that the GA -Adaboost has higher classification accuracy by 6.89% for CFS - RF, by 4.4% for CFS -Bagging, by 2.39% for CFS -Adaboost, 1.73% for GA - RF and by 0.64% for GA - Bagging.

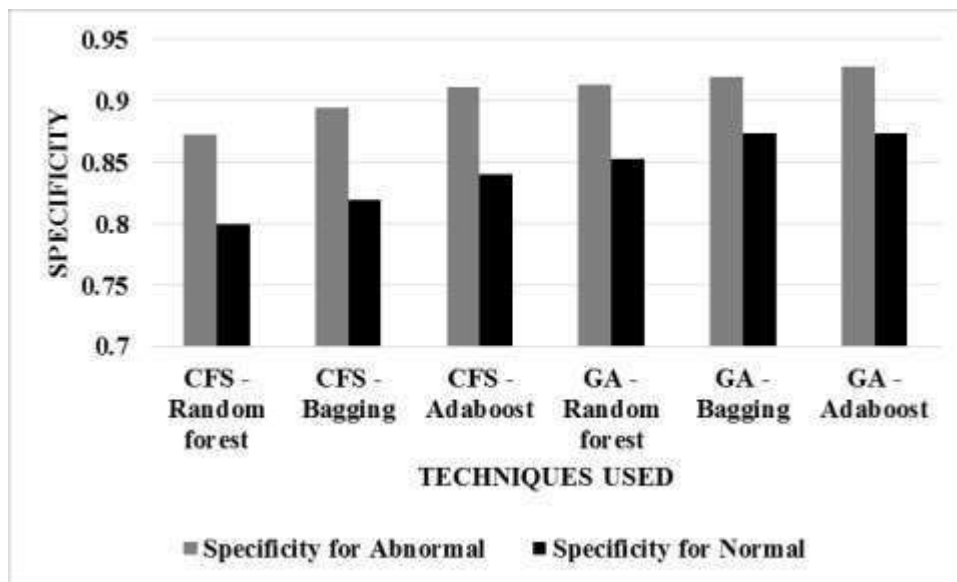


Fig:9. Specificity

From the [Figure- 11], it can be observed that the GA -Adaboost has higher specificity for abnormal by 6.17% for CFS - RF, by 3.66% for CFS -Bagging, by 1.81% for CFS -Adaboost, 1.5% for GA - RF and by 0.9% for GA - Bagging. The GA -Adaboost has higher specificity for normal by 8.76% for CFS - RF, by 6.29% for CFS - Bagging, by 3.88% for CFS -Adaboost, 2.31% for GA - RF and by same value for GA -Bagging.



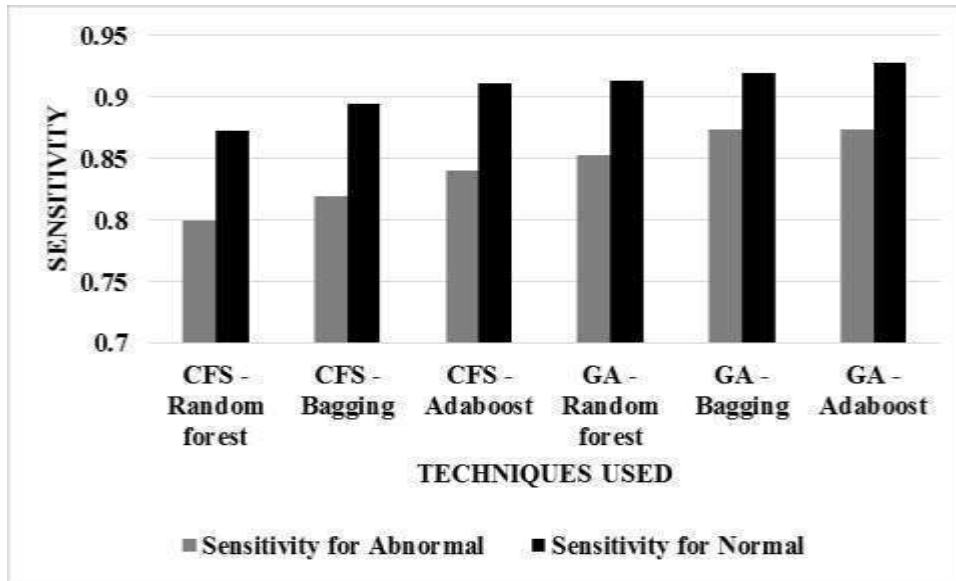


Fig: 10 . Sensitivity

From the [Figure- 12], it can be observed that the GA -Adaboost has higher sensitivity for abnormal by 8.76% for CFS - RF, by 6.29% for CFS -Bagging, by 3.88% for CFS -Adaboost, 2.31% for GA - RF and by same value for GA -Bagging. The GA -Adaboost has higher sensitivity for normal by 6.17% for CFS - RF, by 3.66% for CFS - Bagging, by 1.81% for CFS -Adaboost, 1.5% for GA - RF and by 0.9% for GA -Bagging.

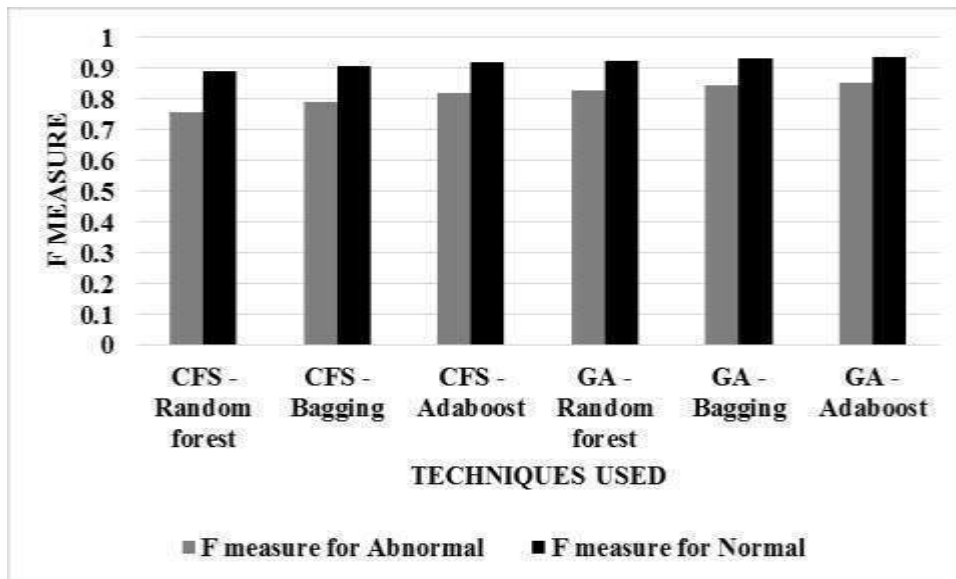


Fig:11. F Measure

From the [Figure- 13], it is observed that GA -Adaboost has greater F measure for abnormal by 11.64% for CFS - RF, by 7.58% for CFS -Bagging, by 4.21% for CFS -Adaboost, 2.97% for GA - RF and by 0.96% for GA - Bagging. The GA -Adaboost has higher F measure for normal by 4.91% for CFS - RF, by 3.09% for CFS - Bagging, by 1.67% for CFS -Adaboost, 1.22% for GA - RF and by 0.48% for GA -Bagging.

### CONCLUSION

The work provides an innovative automated classification system using digital fundus images. In opposition to traditionally implemented segmentation-based metrics, this is completely data-driven and uses image-based feature which is innovative in the area of glaucoma recognitions. This work evaluated a number of different combinations of image-based features and classifier schemes on a data set of 360 normal images as well as 150 Glaucoma images. Results proved that GA -Adaboost shows higher classification accuracy by 6.89% for CFS - RF, by 4.4% for CFS -Bagging, by 2.39% for CFS -Adaboost, 1.73% for GA - RF and by 0.64% for GA - Bagging.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Kumar PSJ, Banerjee MS. [2014] A Survey on Image Processing Techniques for Glaucoma Detection. *International Journal of Advanced Research in Computer Engineering & Technology* (IJARCET), 3.
- [2] Deepashri KM, Santhosh K V. [2015] Glaucoma Detection by Image Fusion from Fundus Color Retinal Images: A Review. *International Journal of Control Theory and Applications* 8(3): 1147-1152.
- [3] Xu Y, Duan L, Lin S, Chen X, Wong DW K, Wong TY, Liu J. [2014] Optic cup segmentation for glaucoma detection using low-rank superpixel representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 788-795). Springer International Publishing.
- [4] Niwa SI, Lin W, Bai X, Kwoh CK, Sng CC, Aquino MC, Chew PTK. [2015] Reliable feature selection for automated angle closure glaucoma mechanism detection. *Journal of medical systems*, 39(3): 1-10.
- [5] Guo L, Yang JJ, Peng L, Li J, Liang Q. [2015] A computer-aided healthcare system for cataract classification and grading based on fundus image analysis. *Computers in Industry* 69: 72-80.
- [6] Raja C, & Gangatharan, N. [2015] Incorporating Phase Information for Efficient Glaucoma Diagnoses Through Hyper-analytic Wavelet Transform. In *Proceedings of Fourth International Conference on Soft Computing for Problem Solving* (pp. 329-343). Springer India.
- [7] Niwas, S. I, Lin, W, Kwoh, C. K, Kuo, C. C. J, Sng, C. C, Aquino, M. C, & Chew, P. T. [2016] Cross-examination for angle-closure glaucoma feature detection. *IEEE journal of biomedical and health informatics* 20(1): 343-354.
- [8] Gupta R, Awasthi D. [2014, September] Wave-packet image fusion technique based on genetic algorithm. In *Confluence The Next Generation Information Technology Summit* (Confluence), 2014 5th International Conference- (pp. 280-285). IEEE.
- [9] Bock R, Meier J, Nyúl LG, Hornegger J, Michelson G. [2010] Glaucoma risk index: automated glaucoma detection from color fundus images. *Medical image analysis*, 14(3), 471-481.
- [10] Malviya, P, & Saxena, A. [2014] An Improved Image Fusion Technique based on Texture Feature Optimization using Wavelet Transform and Particle of Swarm Optimization (POS). *International Journal of Computer Applications* 101(6).
- [11] Yang Y. [2011] Multiresolution image fusion based on wavelet transform by using a novel technique for selection coefficients. *Journal of Multimedia* 6(1): 91-98.
- [12] Wady S H, & Ahmed HO. [2016] Ethnicity Identification based on Fusion Strategy of Local and Global Features *Extraction. Int. J. of Multidisciplinary and Current research*, 4.
- [13] Karegowda AG, Jayaram MA, Manjunath AS. [2010] Feature subset selection problem using wrapper approach in supervised learning. *International journal of Computer applications*, 1(7):13-17.
- [14] Erkanli S, Oguşlu E, Li J. [2012] Fusion of visual and thermal images using genetic algorithms. INTECH Open Access Publisher.
- [15] Kausar N, Majid, A. [2016] Random forest-based scheme using feature and decision levels information for multi-focus image fusion. *Pattern Analysis and Applications* 19(1):221-236.
- [16] Machová K, Barcak F, Bednár P. [2006] A bagging method using decision trees in the role of base classifiers. *Acta Polytechnica Hungarica*, 3(2):121-132.
- [17] Zhou ZH, Wu J, Tang W. [2002] Ensembling neural networks: many could be better than all. *Artificial intelligence* 137(1): 239-263.

# OPTIMIZED FEATURE SELECTION FOR BREAST CANCER DETECTION

S Devisuganya<sup>1\*</sup>, RC Suganthe<sup>2</sup><sup>1</sup>Velalar College of Engineering and Technology, Erode, Tamilnadu, INDIA<sup>2</sup>Kongu Engineering College, Tamilnadu, INDIA

## ABSTRACT

Breast cancer is a common life-threatening cancer affecting woman. Mammography is an effective screening tool; radiologists use for breast cancer detection. The major phase in diagnosing breast cancers is features extraction and selection. Detecting tumours naturally requires extraction of features as well as their classification. This work presents a mammogram framework for detection of cancer. Pseudo Zernike Moments and Gaussian Markov Random Field (GMRF) are used for feature extraction. To reduce the dimensionality of the feature set, a Hybrid Shuffled frog-PSO algorithm is proposed. The work focuses on improving classification performance through feature selection. Experimental results demonstrate the effectiveness of the proposed method in improving the classification of the mammograms

Published on: 2<sup>nd</sup> -December-2016

### KEY WORDS

Pseudo Zernike Moments,  
Gaussian Markov Random Field,  
Shuffle frog algorithm, Particle  
Swarm Optimization

\*Corresponding author: Email: [devisuganya.cse@gmail.com](mailto:devisuganya.cse@gmail.com)

## INTRODUCTION

Alzheimer's disease Breast cancer is most common cancer among women. Mammographic images are X-ray images of breast region. The commonly used diagnostic technique including biopsy, mammography, thermography and ultrasound image. Among these techniques mammography is best approach for early detection. In early stage visual clues are subtle and varied in appearance, it makes diagnosis difficult. The abnormalities are hiding by breast tissue structure. Breast cancer detection and classification of mammogram images is the standard clinical practice for the diagnosis of breast cancer.

Mammography is the efficient tool available for the detection of breast cancer before physical symptoms appear. The earlier the cancer detection is challenging and difficult task. The biopsy is a standard approach for cancer detection manually under a microscope. But biopsy is difficult and time consuming task [1]. Breast cancer is considered a major health problem in western countries. A recent study from the National Cancer Institute (NCI) estimates that, in the United States, about 1 in 10 women will develop breast cancer during their lifetime. Moreover, in such country, breast cancer remains the leading cause of death for women in their 40s.

Although manual screening of mammographies remains the key screening tool for the detection of breast abnormalities, it is widely accepted that automated Computer Aided Diagnosis (CAD) systems are starting to play an important role in modern medical practices [2]. Early detection of breast cancer increases the survival rate and increases the treatment options. Screening mammography, x-ray imaging of the breast, is currently the most effective tool for early detection of breast cancer. Screening mammographic examinations are performed on asymptomatic woman to detect early, clinically unsuspected breast cancer [3]. Early detection via mammography increases breast cancer treatment options and the survival rate. However, mammography is not perfect.

Detection of suspicious abnormalities is a repetitive and fatiguing task. Screening mammography is widely used for early detection of breast cancer. Biopsy is invasive procedure and makes patient discomfort. Digital mammography is proven as efficient tool to detect breast cancer before clinical symptoms appear. Digital mammography is

currently considered as standard procedure for breast cancer diagnosis. Various artificial intelligence techniques such as artificial neural network and fuzzy logic are used for classification problems in the area of medical diagnosis [4].

Feature selection is an important in breast cancer detection and classification. After features extraction, not all features are used to differentiate between normal and abnormal patterns. The advantage of limiting input features to make accuracy and reduce computation complexity. Many features are extracted from digital mammograms, they include region-based features, shape-based features, texture based features selection, and position based features. Texture features classify normal and abnormal in digital mammogram patterns. Feature classifies masses as benign or malignant using selected features. There are various methods used for mass classifications and some popular techniques are artificial neural networks and linear discriminating analysis [5].

Feature extraction is the first step in breast cancer detection. Texture feature is important for image classification. Various techniques have been used for computing texture features [1]. Grey-Level Co-Occurrence Matrices (GLCMs) is a powerful tool for image feature extraction. Gray level pixel distribution described by statistics like probability of two pixels having particular gray level at particular spatial relationships. This spatial information is provided as two dimensional gray level matrices. Image feature extraction is important step in mammogram classification. These features are extracted using image processing techniques. Several features are extracted from digital mammograms including texture feature, position feature and shape feature etc. [4].

Most of the limitations of conventional mammography can be overcome by using digital image processing. Thus, in order to improve the correct diagnosis rate of cancer, image enhancement techniques are often used to enhance the mammogram and assist radiologists in detecting it. Some of the efficient enhancement algorithm of digital mammograms based on wavelet analysis and modified mathematical morphology. Adopt wavelet-based level dependent thresholding algorithm and modified mathematical morphology algorithm to increase the contrast in mammograms to ease extraction of suspicious regions known as Regions of Interest (ROIs) are used [6].

Some scholars applied data mining techniques to predict diagnosis for digital mammography. Data mining techniques offer precise, accurate, and fast algorithms for such classification using dimensionality reduction, feature extraction, and classification routines. Neural networks have improved accuracy rate for the classification of benign and malignant patterns in digitized mammography. Feature selection is also commonly used in machine learning. It has already seen application in statistics, pattern recognition, and data mining. The aim of feature selection is to filter out redundant or irrelevant features from the original data.

Feature selection, a pre-processing step in the data mining process, is the step to select and extract more valuable information in massive related materials. It can improve the model's performance as well as reduce the effort of training the model [7]. Feature selection is a main point that should be taken under consideration when implementing a CADx system for recognizing breast tissue. Selecting the most significant features that have the capability to describe and maximize the differences between different tissues in an ample way. Feature selection is an important factor that directly affects the classification result.

Most systems extract features to detect abnormalities and classify them as benign or malignant. The classification of malignant and benign is still a challenging problem for researchers. There are various feature extraction methods that serve to condense input data and to reduce redundancies by highlighting important characteristics of the image. The features of digital images can be extracted directly from the spatial data or from a different space after using a transform such as Fourier transform, wavelet transform or curvelet transform [8].

This study proposes feature selection based on shuffled frog and PSO. The remaining sections organized as: Section 2 reviews the related work in literature. Section 3 explains the methods which are used in the proposed work. Section 4 discusses the experiment results and section 5 concludes the proposed work.

## RELATED WORK

Rehman et al., [9] proposed diverse features based breast cancer detection (DF-BrCanD) system to detect breast cancer that may be considered as a second opinion. The authors have used phylogenetic trees, statistical features and local binary patterns to generate a set of diverse and discriminative features for subsequent classification. Finally, Support Vector Machine with RBF kernel is used for the classification of mammographic images as

cancerous and non-cancerous. The performance of the proposed DF-BrCanD system is analyzed using standard database for screening mammography through experimental comparison based on various performance measures. The authors showed that the proposed DF-BrCanD system is quite effective in detecting breast carcinoma.

Patel and Sinha [10] introduced a novel approach for accomplishing mammographic feature analysis through detection of tumor, in terms of their size and shape with experimental work for early breast tumor detection. The objective is to detect the abnormal tumor/tissue inside breast tissues using three stages: Pre-processing, Segmentation and post processing stage. By using pre-processing noise are remove and then segmentation is applied to detect the mass, after that post processing is applied to find out the benign and malignant tissue with the affected area in the cancers breast image. Size of tumor is also detected in these steps. The occurrences of cancer nodules are identified clearly.

Ganesan et al., [11] presented a one-class classification pipeline for the classification of breast cancer images into benign and malignant classes. Because of the sparse distribution of abnormal mammograms, the two-class classification problem is reduced to a one-class outlier identification problem. Trace transform, which is a generalization of the Radon transform, has been used to extract the features. Several new functional specific to mammographic image analysis have been developed and implemented to yield clinically significant features. Classifiers such as the linear discriminant classifier, quadratic discriminant classifier, nearest mean classifier, support vector machine, and the Gaussian Mixture Model (GMM) were used.

Deshpande et al., [12] made an attempt to build classification system for mammograms using association rule mining based on texture features. The proposed system used most relevant GLCM based texture features of mammograms. New method was proposed to form associations among different texture features by judging the importance of different features. Resultant associations can be used for classification of mammograms. Experiments were carried out using MIAS Image Database. The performance of the proposed method was compared with standard Apriori algorithm. The authors also investigated the use of association rules in the field of medical image analysis for the problem of mammogram classification.

Sanae et al., [13] presented an efficient classification of mammograms using feature extraction. In this approach the authors proposed to use comprehensive statistical Block-Based features, derived from all sub-bands of Discrete Wavelet decomposition. The classification of these features was performed using the Support Vector Machine (SVM). The evaluation of the proposed method was applied on Digital Database For Screening Mammography (DDSM). The system classifies normal from abnormal cases with high accuracy rate (96%). Comparative experiments have been conducted to evaluate the proposed method.

Kim [14] proposed a new classification technique that is based on support vector machines with the additional properties of margin-maximization and redundancy-minimization in order to further increase the accuracy. The author have conducted experiments on publicly available data set of mammograms and the empirical results indicated that the proposed technique performed superior to other previously proposed support vector machines-based techniques.

Thangavel and Velayutham [15] proposed a novel unsupervised feature selection method using rough set based entropy measures. A typical mammogram image processing system generally consists of image acquisition, pre-processing, segmentation, feature extraction and selection, and classification. The proposed unsupervised feature selection method was compared with different supervised feature selection methods and evaluated with fuzzy c-means clustering in order to prove the efficiency in the domain of mammogram image classification.

Aroquiaraj and Thangavel [16] proposed a novel unsupervised feature selection in mammogram image, using tolerance rough set based relative reduct. And also, compared with Tolerance Quick Reduct and particle swarm optimization (PSO) - Relative Reduct unsupervised feature selection methods. A typical mammogram image processing system generally consists of mammogram image acquisition, pre-processing of image segmentation, feature extraction, feature selection and classification. The proposed method is used to reduce features from the extracted features and the method is compared with existing unsupervised features selection methods. The proposed method is evaluated through clustering and classification algorithms in K-means and WEKA.

Wong et al., [17] proposed an effective technique to classify regions of interests (ROIs) of digitized mammograms into mass and normal tissue regions by first finding the significant texture features of ROI using binary PSO

(BPSO). The data set used consisted of sixty-nine ROIs from the MIAS Mini-Mammographic database. Eighteen texture features were derived from the GLCM of each ROI. Significant features are found by a feature selection technique based on BPSO. Experimental results showed that the significant texture features found by the BPSO based feature selection technique can have better classification accuracy when compared to the full set of features. The BPSO feature selection technique also has similar or better performance in classification accuracy when compared to other widely used existing techniques.

## MATERIALS AND METHOD

This section discuss about Pseudo Zernike Moments and Gaussian Markov Random Field (GMRF) which are used for feature extraction. Hybrid Shuffled frog-PSO algorithm, IG for Feature selection and C4.5, Random Forest, Adaboost for Classifier.

### PSEUDO ZERNIKE MOMENTS

The Zernike moments computation of an input image has 3 steps – computation of

- radial polynomials,
- Zernike basis functions and
- Zernike moments by projecting image onto Zernike basis functions.

The kernel of pseudo-Zernike moments is orthogonal pseudo-Zernike polynomials set defined over polar coordinate space in a unit circle. The 2-dimensional pseudo-Zernike moments of order  $p$  with repetition  $q$  of an image intensity function is defined as:

$$Z_{pq} = \frac{p+1}{\pi} \int_{-\pi}^{\pi} \int_0^1 V_{pq}^*(r, \theta) f(r, \theta) r dr d\theta;$$

$$|r| \leq 1$$

where pseudo-Zernike polynomials  $V_{pq}$  of order  $p$  are defined as:

$$V_{pq}(r, \theta) = R_{pq}(r) e^{jq\theta}; \quad j = \sqrt{-1}$$

and the real-valued radial polynomials,  $R_{pq}(r)$ , is given as:

$$R_{pq}(r) = \sum_{k=0}^{p-|q|} (-1)^k \frac{(2p+1-k)!}{k!(p+|q|+1-k)!(p-|q|-k)!} r^{p-k}$$

Where  $0 \leq |q| \leq p$ .

As pseudo-Zernike moments are defined regarding polar coordinates  $(r, \theta)$  with  $|r| \leq 1$ , computation of pseudo-Zernike polynomials requires a linear transformation of image coordinates  $(i, j)$ ,  $i, j = 0, 1, 2, \dots, N-1$  to a suitable domain  $(x, y) \in \mathbb{R}^2$  inside a unit circle. Two commonly used cases of transformations. Based on these, following discrete approximation of continuous pseudo-Zernike moments' integral [18].

$$Z_{pq} = \lambda(p, N) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} R_{pq}(r_{ij}) e^{-jq\theta_{ij}} f(i, j),$$

$$0 \leq r_{ij} \leq 1$$

where most general image coordinate transformation to interior of unit circle is given by;

$$r_{ij} = \sqrt{(c_1 i + c_2)^2 + (c_1 j + c_2)^2}, \quad \theta_{ij} = \tan^{-1} \left( \frac{c_1 j + c_2}{c_1 i + c_2} \right)$$

### GAUSSIAN MARKOV RANDOM FIELD (GMRF)

Let  $x = (x_1, x_2, \dots, x_n)^T$  be a Gaussian random field with mean  $\mu$  and covariance matrix  $\Sigma$ , that is,  $x \sim N(\mu, \Sigma)$ . The precision matrix of  $x$  is denoted by  $Q$  and  $Q = \Sigma^{-1}$ . Gaussian random field  $x$  is said to be a Gaussian Markov Random Field (GMRF) regarding labeled undirected graph  $G = (V, \mathcal{E})$ , if nodes are  $V = \{1, \dots, n\}$  and edges;

$$\mathcal{E} = \{\{i, j\} \in V \times V : Q_{ij} \neq 0 \text{ and } i \neq j\}.$$

If  $\{i, j\} \in \mathcal{E}$ , then  $i$  and  $j$  are said to be neighbors and is written as  $i \square j$ . Further, notation  $x_{ij}$  to refer to sub-vector of  $x$  corresponding to nodes  $i, i+1, \dots, j$ . By definition, any GRF is a GMRF, generally regarding a fully connected graph  $G$ .

In practice, use of GMRFs is confined to situations where neighborhood size is small so that precision matrix is sparse. The precision matrix's non-zero pattern is related to conditional independence structure of GMRF by  $x_i \perp x_j \mid x_{-ij} \Leftrightarrow Q_{ij} = 0, i \neq j$ .

Here,  $x_{-ij}$  denotes all elements of  $x$  except elements  $i$  and  $j$ . As a consequence of correspondence between non-zero pattern of  $Q$  and conditional independence structure of GMRF, GMRF is specified regarding its conditional moments.

A mammographic image  $Y$  is modeled by a finite lattice GMRF. Each pixel in image lattice  $L$  is represented by a random variable  $y_{ij}$  where  $Y = \{y_{ij} : 0 \leq i \leq M-1, 0 \leq j \leq M-1\}$  and  $L = \{(i, j) : 0 \leq i \leq M-1, 0 \leq j \leq M-1\}$ . In a GMRF assumption of image  $Y$  with respect to a certain neighborhood system  $\eta$ ,  $Y$  is reshaped to a single vector  $y = [y_1, y_2, \dots, y_M^2]$  in lexicographic order [18].

### INFORMATION GAIN (IG)

Information Gain is supervised univariate feature selection algorithm of the filter model which is a measure of dependence between the feature and the class label. It is one of the most powerful feature selection techniques and it is easy to compute and simple to interpret. Information Gain (IG) of a feature  $X$  and the class labels  $Y$  is calculated as

$$IG(X, Y) = H(X) - H(X|Y)$$

Entropy ( $H$ ) is a measure of the uncertainty associated with a random variable.  $H(X)$  and  $H(X|Y)$  is the entropy of  $X$  and the entropy of  $X$  after observing  $Y$ , respectively.

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i)).$$

The maximum value of information gain is 1. A feature with a high information gain is relevant. Information gain is evaluated independently for each feature and the features with the top- $k$  values are selected as the relevant features. This feature selection algorithm does not eliminate redundant features [19].

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i | y_j) \log_2(P(x_i | y_j))$$

### PROPOSED SHUFFLED FROG ALGORITHM-PARTICLE SWARM OPTIMIZATION FOR FEATURE SELECTION

Particle swarm optimization algorithm is an optimization algorithm based on group and fitness. The system initializes particles (representing potential solutions) as a set of random solutions, which has two features of position and velocity. The fitness values of particles are decided by particle positions. Particles move in the solution space; the moving direction and distance are determined by the speed vector and new speed, position are updated from personal best position  $p_{best}$ , global best position  $g_{best}$  and the current particle velocity; particles search and pursue the optimal particle based on fitness values in the solution space, and gradually converge to the optimal solution. Assuming in a  $d$ -dimensional search space, there is a group composed of

$n$  particles, where of generation  $t$  particle  $i$  ( $i = 1, 2, \dots, n$ ), position coordinates  $x_i^t = (x_{i1}, x_{i2}, \dots, x_{id})$ , velocity  $v_i^t = (v_{i1}, v_{i2}, \dots, v_{id})$  personal best position  $p_i^t = (p_{i1}, p_{i2}, \dots, p_{id})$  and global best position  $p_g^t = (p_{g1}, p_{g2}, \dots, p_{gd})$ . For particle  $i$  dimension  $d$  generation  $t$ , its iterative formula can be expressed as:

$$v_{id}^{t+1} = \omega \mathcal{G}_{id}^t + c_1 r_1 (p_{id}^t - x_{id}^t) + c_2 r_2 (p_{gd}^t - x_{id}^t)$$

$$x_{id}^{t+1} = x_{id}^t + \mathcal{G}_{id}^{t+1}$$

where

$\mathcal{G}_{id}^t$  - Current velocity,

$\mathcal{G}_{id}^{t+1}$  - New speed of particle  $r$  after iteration  $t$ ,

$\omega$  - Inertia weight,

$C_1, C_2$  - Acceleration (learning) factors,

$r_1, r_2$  - Uniformly distributed random numbers between 0 and 1,

$x_{id}^t$  - Current position of particle i,

$x_{id}^{t+1}$  - new position of particle i after iteration t.

Shuffled frog leaping algorithm is a biological evolution algorithm based on swarm intelligence. The algorithm simulates a group of frogs in the wetland passing thought and foraging by classification of ethnic groups. In the execution of the algorithm, F frogs are generated at first to form a group, for N-dimensional optimization problem, frog i of the group is represented as  $(x_i^1, x_i^2, \dots, x_i^N)$  then individual frogs in the group are sorted in descending order according to fitness values, to find the global best solution  $P_x$ . The group is divided into m ethnic groups, each ethnic group including n frogs, satisfying the relation  $F=m \times n$ . The rule of ethnic group division is: the first frog into the first sub-group, the second frog into the second sub-group, frog m into subgroup m, frog m+1 into the first sub-group again, frog m+2 into the second sub-group, and so on, until all the frogs are divided, then find the best frog in each subgroup, denoted by  $P_b$ ; get a worst frog correspondingly, denoted by  $P_\omega$ . Its iterative formula can be expressed as:

$$D = rand() * (P_b - P_\omega)$$

$$P_{new-\omega} = P_\omega + D_i, -D_{max} \leq D_i \leq D_{max}$$

where  $rand()$  represents a random number between 0 and 1,

$P_b$  represents the position of the best frog,

$P_\omega$  represents the position of the worst frog,

D represents the distance moved by the worst frog,

$P_{new-\omega}$  is the improved position of the frog,

$D_{max}$  represents the step length of frog leaping.

In the execution of the algorithm, if the updated  $P_{new-\omega}$  is in the feasible solution space, calculate the corresponding fitness value of  $P_{new-\omega}$ , if the corresponding fitness value of  $P_{new-\omega}$  is worse than the corresponding fitness value of  $P_\omega$ , then use  $P_\omega$  to replace  $P_b$  and re-update  $P_{new-\omega}$ ; if there is still no improvement, then randomly generate a new frog to replace  $P_\omega$ ; repeat the update process until satisfying stop conditions.

Exploration and exploitation has been a contradiction in the search process of swarm intelligence algorithms. Exploration stresses searching for a new search region in the global range, and exploitation is focused on fine search in local search area. Although particle swarm optimization algorithm is simple and its optimization performance is good, in the entire iterative process, exploration capability is strong and exploitation capability is weak in early period, at this time if particles fall on the neighbourhood of the best particle, they may flee the neighbourhood of the best particle, due to too strong exploration capability; exploration capability is weak and exploitation capability is strong in later period, at this time if particles encounter local optima, the speed of all particles may be rapidly reduced to zero instead of flying, leading to convergence of particle swarm to local optima; the iterative mechanism and ethnic group division lead to strong exploitation and weak exploration in early period, and strong exploration and weak exploitation in later period.

Based on the analysis, in the update process of the algorithm, in order to ensure the diversity of particles, particle swarm and frog group sharing part of the particles, we propose particle sharing based particle swarm frog leaping hybrid optimization algorithm. The idea is as follows: divide the total number of particles N into two sub-groups of numbers N1 and N2, where the first sub-group uses shuffled frog leaping algorithm to optimize, the second sub-group uses the standard particle swarm optimization algorithm to optimize, and N, N1 and N2 satisfy  $N \leq N1 + N2$ , so the number of shared particles is  $N1 + N2 - N$  [20].

## CLASSIFIER

Classification models are monitored methods that are initially trained on a dataset of samples known as training sets. The performance of the algorithms is then evaluated on distinct training sets. The features that are extracted are inputs for the classifiers. The performance of three classifiers is examined on datasets.

### C4.5

C4.5 is an extension of Iterative Dichotomizer (ID3) algorithm that was designed by Quinlan to deal with issues that cannot be handled by the ID3 algorithm. These include avoidance of over fitting the data; reduced error pruning, rule post-pruning, handling continuous attributes and handling data with missing attribute values. It attempts to build a decision tree with a measure of the



information gain ratio of each feature and branching on the attribute which returns the maximum information gain ratio. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate [21].

Typically, C4.5 assigns the frequency of the correct counts at the leaf as the probabilistic estimate. For notational purposes, TP is the number of true positives at the leaf, FP is the number of false positives, and C is the number of classes in the data set. Thus, the frequency based probabilistic estimate can be written as [22]:

$$P_{leaf} = TP / (TP + FP)$$

## RANDOM FOREST

Random Forest (RF) is an approach which has been proposed by Breiman for classification tasks. It mainly comes from the combination of tree-structured classifiers with the randomness and robustness provided by bagging and random feature selection. The classification is performed by sending a sample down is each tree and assigning it the label of the terminal node it ends up in. At the end the average vote of all trees is reported as the result of the classification. Random forest is very efficient with large datasets and high dimensional data [21].

The principle of RF is the aggregation of a large ensemble of decision trees. During training, each individual tree in the ensemble is fitted by sampling the training data with replacement (bootstrap) and growing the tree to full depth on the training sample. The optimal data split at each tree node is determined by randomly choosing  $m$  of the available  $P$  input variables and selecting the one which splits the node best. In this work, node splitting was guided by the Gini cost function

$$G(N) = 1 - \sum_{k=1}^2 p^2(\omega_i)$$

which measures node impurity using  $p(\omega_i)$  as the fraction of features in class  $i$  at node  $N$ . The best split was the one which decreases node impurity the most. Further, by calculation of the mean decrease in Gini (MDG) for each variable over all trees, RF allow to obtain a variable importance ranking. The final RF classification score is determined by collecting the votes of each of the  $n$  trees in the forest for either class and outputting a vote ratio. As the method is based on decision trees, the splits in the nodes are always parallel to the coordinate axes of the features [23].

## ADAPTIVE BOOSTING (ADABOOST)

Adaptive Boosting (AdaBoost) is the popular ensemble method to enhance prediction accuracy of the base learner. Multiple classifiers are generated with this AdaBoost learning algorithm to utilize them to build as a best classifier. This requires less user knowledge for computing for improving accuracy over data sets. Also it is used for maintaining a set of weights over the training set. The training set  $(x_1, y_1), \dots, (x_n, y_n)$  where each  $x_i$  belongs to instance space  $X$  and each  $y_i$  is in the label set  $Y = \{-1, +1\}$ . The steps for AdaBoost are as follows [24]:

1. Assign  $N$  example

$$(x_1, y_1), \dots, (x_n, y_n); x_i \in \{-1, +1\}$$

2. Initialize the weights of  $D_1(i) = 1/N, i = 1, \dots, N$

3. For  $k = 1, \dots, K$

4. Train weak learner using distribution  $D_k$

5. Get weak hypothesis  $h_k : X \rightarrow R$  with its error:  $\epsilon_k = \sum_{i=h_k(x_i) \neq y_i} D_k(i)$

6. Choose  $\alpha_k = R$

7. Update

$$D_{k+1}(i) = \frac{D_k(i) \exp(-\alpha_k y_k h_k(x_k))}{Z_k}$$

where  $Z_k$  is the normalization factor.

8. Output the final hypothesis:

$$H(x) = \text{sign} \left( \sum_{k=1}^K \alpha_k h_k(x) \right).$$

## RESULTS

To evaluate the proposed technique, 150 normal mammogram image and 25 images with calcification obtained from MIAS dataset were used. Features are extracted using Pseudo Zernike Moments and Gaussian Markov Random Field technique. Features are selected using IG and the proposed hybrid SF-PSO. Classification is achieved using C4.5, random tree and AdaBoost techniques. Results are presented in this section. [Table- 1] and [Figure- 1], shows the classification accuracy.

Table: 1. Classification Accuracy

Techniques	IG	Hybrid SF-PSO
C4.5	84	94.97
Random tree	84.57	95.53
Boosting	85.14	97.21

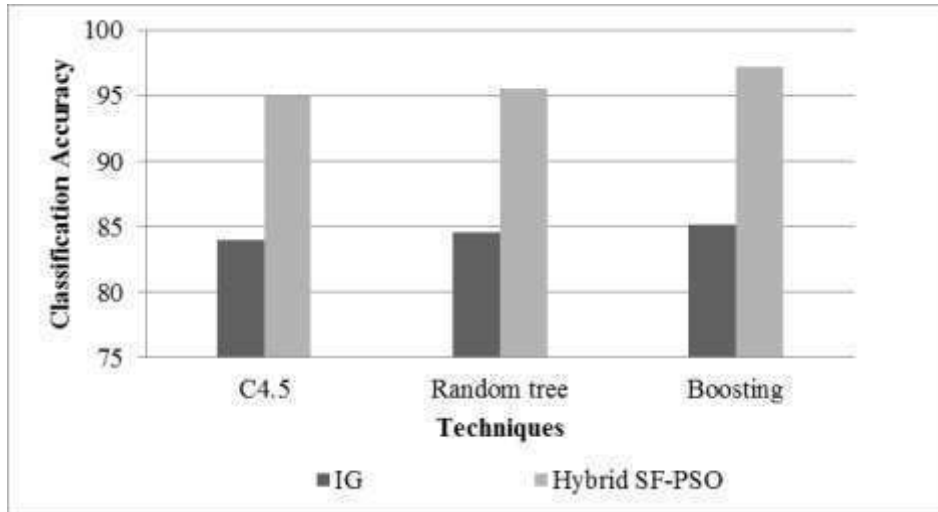


Fig: 1. Classification Accuracy

From [Table- 1] and [Figure- 1], it can be observed that the classification accuracy has improved for hybrid SF-PSO than IG by an average of 12.56%. For C4.5, hybrid SF-PSO has improved classification accuracy by 12.26% than IG. Similarly for Random tree, hybrid SF-PSO has improved classification accuracy by 12.17% than IG and for Boosting, hybrid SF-PSO has improved classification accuracy by 13.24% than IG. [Table- 2] and [Figure- 2] shows the sensitivity.

Table: 2. Sensitivity

Techniques	IG	Hybrid SF-PSO
C4.5	0.76	0.88
Random tree	0.8	0.92
Boosting	0.8	0.92

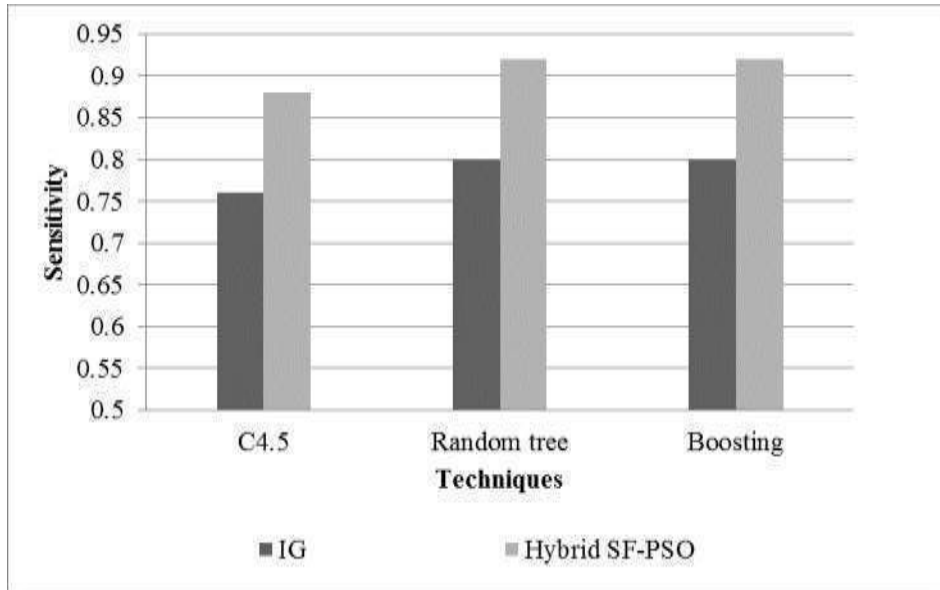


Fig. 2. Sensitivity

From [Table- 2] and [Figure- 2], it can be observed that the sensitivity has improved for hybrid SF-PSO than IG by an average of 14.17%. For C4.5, hybrid SF-PSO has improved sensitivity by 14.63% than IG. Similarly for Random tree, hybrid SF-PSO has improved sensitivity by 13.95% than IG and for Boosting, hybrid SF-PSO has improved sensitivity by 13.95% than IG. [Table- 3] and [Figure- 3] shows the specificity.

Table: 3. Specificity

Techniques	IG	Hybrid SF-PSO
C4.5	0.8533	0.961
Random tree	0.8533	0.961
Boosting	0.86	0.9805

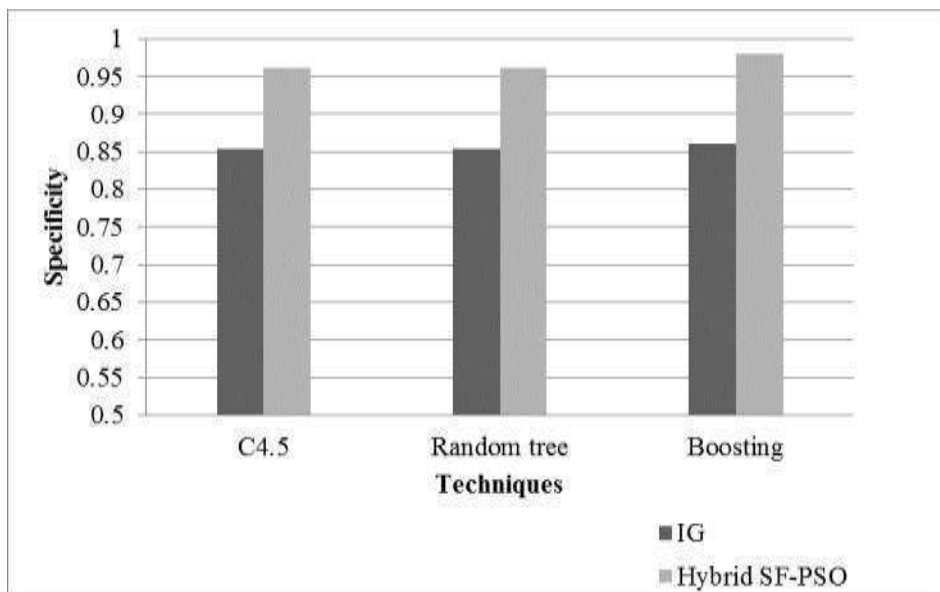


Fig. 3. Specificity

From [Table- 3] and [Figure- 3], it can be observed that the specificity has improved for hybrid SF-PSO than IG by an average of 12.29%. For C4.5 and Random tree, hybrid SF-PSO has improved specificity by 11.87% than IG. Similarly for Boosting, hybrid SF-PSO has improved specificity by 13.09% than IG. [Figure- 4] shows the Percentage of features selected – Hybrid SF-PSO.

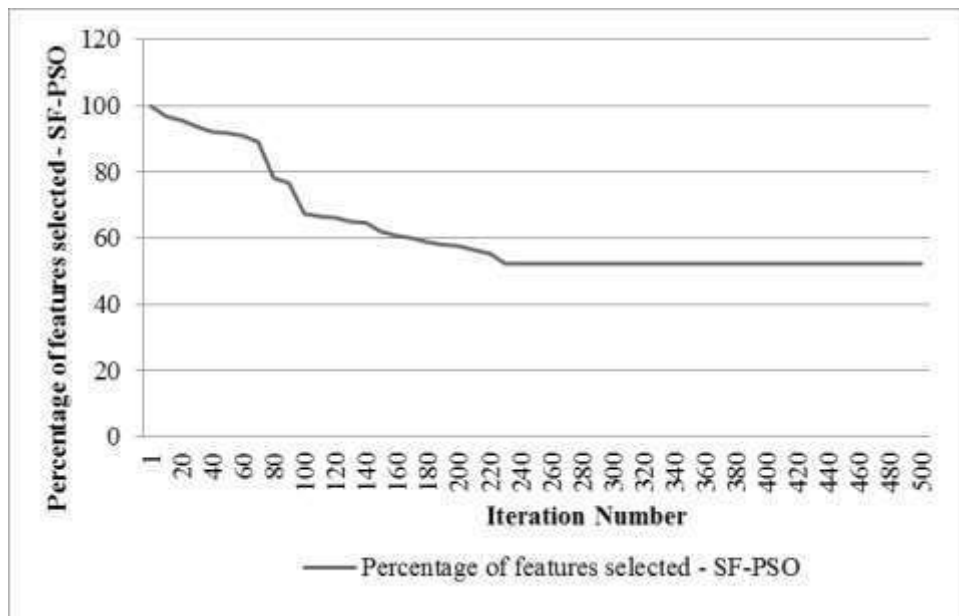


Fig: 4. Percentage of features selected – Hybrid SF-PSO

In iteration 230, 52% of features are selected which forms the optimal feature subset.

### CONCLUSION

Breast cancer is one of the most common cancers among women around the world. Mammography is one of the best breast cancer detection methods. But, in some cases, radiologists face problems in detecting tumours. A feature extraction method for finding the most significant coefficients was proposed and implemented to classify a set of mammogram images. This study presented a new approach to segment breast cancer mass in mammograms. The study focuses on improving classification performance through feature selection. It is seen that of the various classification techniques C4.5 outperforms other algorithms with highest accuracy. Using AdaBoost followed by domain adjusted post-processing such as false positive filtering, our approach achieved promising preliminary results. The classification accuracy has improved for hybrid SF-PSO than IG by an average of 12.56%. For C4.5, hybrid SF-PSO has improved classification accuracy by 12.26% than IG. Similarly for Random tree, hybrid SF-PSO has improved classification accuracy by 12.17% than IG and for Boosting, hybrid SF-PSO has improved classification accuracy by 13.24% than IG.

### CONFLICT OF INTEREST

The authors declare no conflict of interests.

### ACKNOWLEDGEMENT

None

### FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

## REFERENCES

- [1] Nithya R, Santhi B. [2011] Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *International Journal of Computer Applications*, 28(6): 21-25.
- [2] Bosch A, Munoz X, Oliver A, Marti J. [2006] Modeling and classifying breast tissue density in mammograms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) 2*: 1552-1558). IEEE.
- [3] Sampat M.P, Markey MK, Bovik AC. [2005] Computer-aided detection and diagnosis in mammography. *Handbook of image and video processing 2*(1):1195-1217.
- [4] Nithya R, Santhi B. [2011] Comparative study on feature extraction method for breast cancer classification. *Journal of Theoretical and Applied Information Technology*, 33(2):1992-1986.
- [5] Ramani R, Vanitha NS. [2014] Computer Aided Detection of Tumours in Mammograms. *International Journal of Image, Graphics and Signal Processing* 6(4): 54.
- [6] Arpana MA, Kiran P. [2014] Feature Extraction Values for Digital Mammograms. *International Journal of Soft Computing and Engineering (IJSCE)* 4(2): 183-187.
- [7] Luo ST, Cheng BW. [2012] Diagnosing breast masses in digital mammography using feature selection and ensemble methods. *Journal of medical systems*, 36(2):569-577.
- [8] Eltokhy MM, Faye I. [2014] An Optimized Feature Selection Method For Breast Cancer Diagnosis in Digital Mammogram using Multiresolution Representation. *Appl. Math* 8(6): 2921-2928.
- [9] Rehman AU, Chouhan N, Khan A. [2015] Diverse and Discriminative Features based Breast Cancer Detection using Digital Mammography. In *2015 13th International Conference on Frontiers of Information Technology (FIT)* (pp. 234-239). IEEE.
- [10] Patel BC, Sinha GR. [2014] Mammography feature analysis and mass detection in breast cancer images. In *Electronic Systems, Signal Processing and Computing Technologies (ICESC), 2014 International Conference on* (pp. 474-478). IEEE.
- [11] Ganesan K, Acharya UR, Chua CK, Lim CM, Abraham KT. [2014] One-class classification of mammograms using trace transform functionals. *IEEE Transactions on Instrumentation and Measurement* 63(2): 304-311.
- [12] Deshpande DS, Rajurkar AM, Manthalkar RM. [2013] Medical image analysis an attempt for mammogram classification using texture based association rule mining. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on* (pp. 1-5). IEEE.
- [13] Sanae B, Mounir AK, Youssef F. [2014] Statistical block-based DWT features for digital mammograms classification. In *Intelligent Systems: Theories and Applications (SITA-14), 2014 9th International Conference on*(pp. 1-7). IEEE.
- [14] Kim S. [2014] Margin-maximised redundancy-minimised SVM-RFE for diagnostic classification of mammograms. *International journal of data mining and bioinformatics*, 10(4), 374-390.
- [15] Thangavel K, Velayutham C. [2012] Rough set based unsupervised feature selection in digital mammogram image using entropy measure. In *Biomedical Engineering (ICoBE), 2012 International Conference on* (pp. 10-16). IEEE.
- [16] Aroquiaraj IL, Thangavel K. [2013] Mammogram image feature selection using unsupervised tolerance rough set relative reduct algorithm. In *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on* (pp. 479-484). IEEE.
- [17] Wong MT, He X, Nguyen H, Yeh WC. [2012] Particle swarm optimization based feature selection in mammogram mass classification. In *Computerized Healthcare (ICCH), 2012 International Conference on* (pp. 152-157). IEEE.
- [18] Devisuganya, S, & Suganthe, R. C. [2016] Breast Cancer Detection: A Framework to Classify Mammograms.
- [19] Porkodi R. [2014] comparison of filter based feature selection algorithms: An overview. *international journal of innovative research in technology& science*, 2(2): 108-113.
- [20] Lenin K, dranath Reddy BR, Kalavathi MS. [2014] Particle Sharing Based Particle Swarm Frog Leaping Hybrid Optimization Algorithm for Solving Optimal Reactive Power Dispatch Problem.
- [21] Oleiwi ASA. [2014] Classification of Mammography Image Using Machine Learning Classifiers and Texture Features.
- [22] Chawla NV. [2003, August] C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the ICML* (Vol. 3).
- [23] Lesniak JM, Hupse R, Blanc R, Karssemeijer, N & SzékelyG. [2012] Comparative evaluation of support vector machine classification for computer aided detection of breast masses in mammography. *Physics in medicine and biology*, 57(16): 5295.
- [24] Ramani R, Vanitha NS. [2015] Computer Aided Detection Of Tumors in Mammograms using Optimized Support Vector Machines. *ARPN Journal of Engineering and Applied Sciences*, 10( 4).