SPECIAL ISSUE

THE IIOAB JOURNAL

VOLUME 7 : NO 5 : SEPTEMBER 2016 : ISSN 0976-3104

Institute of Integrative Omics and
Applied Biotechnology Journal

Dear Readers,

We extend a heartfelt welcome to each of you as we embark on an exciting exploration of Soft Computing's applications in the realms of Medical Data and Satellite Image Analysis through IIOAB scientific journal.

As the Editors of this issue, we are happy to witness the convergence of advanced computational techniques in unraveling the complexities of medical data analysis and satellite imagery interpretation.

Your expertise, dedication, and scholarly contributions play an instrumental role in revolutionizing how we understand and interpret medical data and satellite imagery. Your work in developing sophisticated algorithms, machine learning models, and intelligent systems enables us to derive meaningful insights from complex data sources, leading to advancements in healthcare diagnostics, environmental monitoring, disaster management, and more.

The interdisciplinary nature of these fields offers unprecedented opportunities for innovation and societal impact. Your relentless pursuit of excellence not only fuels scientific progress but also holds immense promise for improving healthcare outcomes, environmental sustainability, and disaster response strategies.

We encourage each of you to share your trailblazing insights, submit your transformative research, and actively engage in the vibrant discussions within our journal. Let us collectively foster an environment where knowledge flourishes, collaborations thrive, and innovations reshape the landscapes of medicine and satellite imaging.

Thank you for your unwavering dedication to advancing the frontiers of Soft Computing in medical data analysis and satellite image interpretation. We eagerly anticipate the transformative discoveries and insights that will emerge from your invaluable contributions.

Warm regards,
Editors-in-Chief

Swarnalatha P and B. K. Tripathy

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Ramanathan et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A NOVEL GENETIC NAND PAFT MODEL FOR ENHANCING THE STUDENT GRADE PERFORMANCE SYSTEM IN HIGHER EDUCATIONAL INSTITUTIONS

Ramanathan L.[1], Angelina Geetha[2], Khalid M.[3], Swarnalatha P.[4]
[1,4] *SCOPE, VIT University, Vellore, INDIA*
[2] *Department of Computer Science & Engineering, B.S. Abdur Rahman University, Chennai, INDIA*
[3] *NITMAS & TNU, Kolkata, INDIA*

ABSTRACT

The higher education system in India is curious about the success of students in education during their study. These educational institutions are adopting several methodologies to improve the quality of education and to improve the success rate of students each year. This is used as a primary objective in improving the academic excellence of a student in his/her higher educational level. The main aim of the study is to create a model that classifies the instances correctly to predict the performance of students using PAFT methodology. The PAFT methodology consists of several attribute of a student modeled into a three-tier model that is collected based on several level testing done on a particular student. The three tier model involves his complete academic details, his/her creative and other interpersonal skills and finally the level of interest towards the present educational approach. To classify the instances correctly, Genetic algorithm is modified in its mutation level with NAND gate. The proposed classifier eliminates the regression fit problem during the selection stage with the help of Tobit regression evaluation of each individual. The classifier is also compared with other techniques like genetic-OR, genetic-AND, Multi-Layer Preceptor and artificial neural network. This classifier optimizes correctly the attributes given as an input for its processing and better learning. The PAFT methodology combined with a genetic - NAND algorithm proves successful in terms of its classification rate. This could be inferred finally that this could be utilized in institutions in determining the performance and success ratio of students as a part of their knowledge management system. Also, this model could also be used for predicting the improvement level of students for the fore coming students based on the collected data set.

Received on: 30th-Nov-2015
Revised on: 22nd-Feb-2016
Accepted on: 26– March-2016
Published on: 10th–May-2016

KEY WORDS

Data mining; Genetic NAND; Knowledge Management; Students Perception; Academic Performance

***Corresponding author: Email:** lramanathan@vit.ac.in, **Tel:** +91 9894411988

INTRODUCTION

In Knowledge management, data mining system, various techniques are employed to extract and determine the meaningful value from a large dataset. In present scenario, data mining models provide a great concern and consideration in the society and information sectors. These data mining techniques receive a greater attention in data analysis and considered as a newly emerging tool for analysis [1]. The main aim of our study is modeling the perception, attention, behaviour and forms of thoughts (PAFT) of the students in the educational sector. This is being considered as a crucial and critical factor, since educational data mining has been coming up with various innovative and improved models [2]. These models are most helpful in modeling the PAFT of students to improve their academic performance in higher educational institutions in India.

Superby, et al., (2006) considered personal details of the student, Implication and perception of student behaviour to classify them into three groups, namely low-risk, probability of success, medium-risk and high-risk with high rate of dropping out. Also, significant, influential factor is identified in these three groups that impart major part in decision making for university entry and academic performance [3]. BK Baradwaj and S Pal, (2011) performed classification tasks for evaluating the division of students with a decision tree. The factors considered are test grade and marks, laboratory marks, attendance, marks [4, 5], seminar and assignment performance and general proficiency [4]. Ramesh, et al., (2013) considered grades obtained at elementary, secondary and higher secondary level, a community of students, group study and education of parents to evaluate their academic performance [6]. Other factors considered for educational data mining to increase the students' performance include: education of mother [7], family income of students [7], dietary habits [7] and education department [8]. Academic factors include CGPA [9, 5], maximum and minimum credits in current semester [10]. These factors are taken along with additional factors more related to increasing the academic performance. Santhi Thilagam and Ananthanarayana,

THE IIOAB JOURNAL ISSN: 0976-3104

2008 [11] proposed a multi objective function based on fuzzy implication and the dataset is collected through the computer activity done by the students. The algorithm failed to prove its effectiveness in comparing it with other algorithms. Amelia Zafra and Sebastián Ventura, 2012 [12] proposed multi-instance genetic algorithm that is used to predict the performance of student related to academics based on web based education. This research helps in identifying the pass rate and failure rate of each individual based on web based tests. The attribute considered for the student includes: total assignment number, total assignment time, total number of posts, total number of messages read, total time taken to complete the test, total questions, answered, seen and passed by the student. Ahmad Slim et al. (2014) [13] employs a Markov model in determining the performance rate of the student and applies linear regression [14] to avoid the regression fit problems. With the features used, the GPA is predicted from the intermediate test results and this helps in achieving the academic score effectively. Parneet Kaur et al. 2015 [15] compared various techniques like Multi-layer perception, SMO, J48, REP Tree, Naïve Bayes with the collected dataset. This dataset includes real world data from the high school environment that includes the personal portfolio of a student related to his name, attendance, internal grade, medium of instruction, school type, sex, private tuition, computer available at home, whether he/she is qualified or not. This is done to analyse the performance of student related to academics. The main drawback of this system is that the accuracy level of the proposed with various methods did not go more than 75%. The classical techniques and the evolutionary data mining technique called grammar based genetic algorithm (GGA) is used to build a model for predicting the performance of undergraduate students by using their past education and general record [16]. Firstly, a comparison of the accuracy of different classification algorithms used in data mining on multiple datasets and selection of an algorithm which has high accuracy for all classes to ensemble them with the help of a method by name, weak classifier to get a combination of multiple classifier is carried on[17]

Various authors suggested multiple steps [18] for building a data mining model, which involves a transformation process, data preparation with consolidation followed by data cleaning, transformation and reduction phase [19]. Berry and Linoff, (2000) defined an iterative data mining process with multiple steps that could well establish the success rate of students [20]. To analyse the data, data mining algorithm is used to understand the hidden patterns in the node and prediction of data. To do this Weka tool provides a better legitimacy in analysing the collected data [21] and support .java files is used in writing new algorithm. This tool is used as a standalone approach to read the .java files and implementing the algorithm for predicting the desired results. There are several supervised and unsupervised approaches to classify perfectly the instances; this research concentrates on improving the simplicity of an algorithm by analysing its internal parameters. This analysis is done in terms of errors occurring at each stage and rectifying it with suitable efficient technique. This helps in reducing the complexity of certain algorithm that could classify the instances correctly but with longer computation.

The research goal is to increase the prediction ability of the proposed model to improve the students in their academics. This paper also proposes a modified algorithm called Genetic NAND combined with PAFT model to form an efficient novel approach to improve the perception of learning ability in Higher Educational Institutions. The preliminary aim of this study is to collect the relevant data set related to the PAFT methodology by making a student to undergo several levels of testing. Thus, collecting suitably the relevant dataset related to study and this helps in collecting the independent variables with performance as a dependent variable. This possesses certain regression fitting problem in its selection of chromosome and that could be avoided with Tobit regression model. Also the use of Mahalanobis distance classifiers helps in finding the bit strings accurately based on the weighted distance between the selected chromosomes. Soon after this the functionality of NAND gate improves the mutation stage and avoids much complexity due to other logical gates. Here, the main outcome is the reliable success rate of the student with efficient prediction using the Genetic NAND PAFT model. The dataset is being handled with 30 different attributes that is collected from previous literatures and segregated based on PAFT model. This model is designed to work with large datasets, such that the prediction accuracy should no longer be less than previous exiting techniques. Also the simplicity of this approach helps in avoiding the hybridization of the genetic algorithm, thereby reducing the complexity of the genetic approach. This helps in avoiding the use of multi-instance learning, Neuro genetic algorithm, ensemble learning techniques etc.

The main contribution of the paper includes proposed novel genetic NAND algorithm for educational data mining, successive section discusses the results of the proposed method. Finally, it concludes that the proposed technique is efficient in improving the student's academic performance.

THE IIOAB JOURNAL ISSN: 0976-3104

MATERIALS AND METHODS

Initially, the students are evaluated using questionnaires based on 5-point Likert scale with secondary variables collected from various literatures. The questions are designed using PAFT model with three sections with multiple questions in each section. The students are allowed to self-evaluate themselves based on the questionnaires given before the final exam. Certain sets of questions are asked to instructors regarding their behavior and other factors specified in the next section. Soon after the completion of exams, the instructors are allowed to evaluate their students based on two criteria: whether they have improved or not. Finally, the questionnaires are collected, pre-processed for cleaning the dataset and the proposed novel classification algorithm is applied to it. This new data mining approach helps in increasing the prediction rate of the student's performance using PAFT methodology. This is beneficiary in the higher Educational Institution to predict the performance of the students in future years. To achieve this, the variables collected are grouped into three sets of factors and the proposed genetic NAND algorithm accurately classifies the selected data set to produce the qualified results. Of collected 1000 samples 200 are used to train the artificial neural network. The population focused here are the students from engineering college.

Proposed PAFT system

Student's psychology is one of the major factors being considered for PAFT system. PAFT system is modeled based on the three sets of factors that involves their Perception towards learning, Attention and behavior within the institution and finally their Form of Thoughts. The ultimate aim is to increase the prediction of student's behavior towards education and thereby increasing the perception of learning using PAFT workflow.

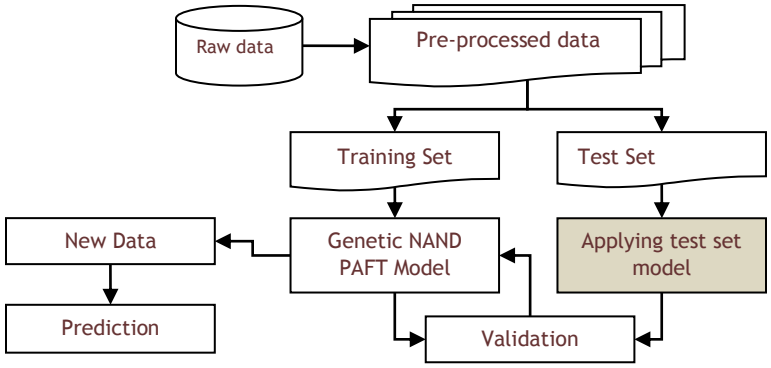


Fig: 1. Work flow of proposed PAFT model

This model has been divided based on final academic performance into three sets of factors that is being closely related to our study:

1. First factor involves stable group structural factors that relate to the personal history of the student's. This factor includes
    a. Identity of the student, academic details,
    b. Previous academic scores in primary, middle, higher secondary level, college level and
    c. Past history details of the student's family.
2. A second factor involves changing or process factors, which includes the involvement of students in relation with his studies or his behavior in studies. These factors are considered participation in other academic activities, technical relationship with his professors, marks or grade performance in periodic examination and in the final examination.
    a. Critical thinking [2]
    b. Communication and teamwork skills [2]
    c. Self-regulated learning [2]
    d. Innovative thinking [6]
    e. Systems-thinking [15]
    f. Non-routine problem-solving [15]
    g. Inquiry skills [15]
    h. Interpersonal skills [16]
    i. Skill demonstration [24]
    j. Technical capabilities include creative thinking [12],
    k. Being social [12],
    l. Patient and determined attitude [12],
    m. Leadership and problem management, crisis [12].
3. A third set of factors involves the perception of students: this set of factors considers
    a. Perception of students related to academic context
    b. His/her perception towards the course and professors

With PAFT methodology, a classification algorithm proposed is applied over these factors to find the rate of improvement in the student's performance in academics. The real-time data are collected from bachelor of computer science students from a self-

THE IIOAB JOURNAL ISSN: 0976-3104

financing institution in Tamil Nadu. Out of the collected data, 43% were female and the remaining are male. Each student is allowed to undergo certain training based on which the data are collected, the factor one is collected from the department staffs, factor two is collected by conducting certain interpersonal skill trainings based on which they are ranked and finally the factor three is collected through the questionnaire. The average age of participants, who undergone the training was nearly 21.

Novel genetic NAND PAFT model

To overcome the problems in genetic approach like over-fitting, high optimization time, etc. A logical approach has been designed to overcome the above constraints and to increase the accuracy rate and sensitivity level of the prediction model. To achieve this, the data has to go through various processing stages that involve data cleaning, data transformation and data classification. The complete approach of data classification using Genetic NAND Algorithm is mentioned in **Figure– 2**.

Initialization

Genetic NAND algorithm **[Figure– 3]** Involves three stages that include: Fitness stage, NAND stage, Shift stage. Initially, the parameters are initialized and the random population is generated for classifying the associated dataset in the search space. To solve this problem the relationship between the dependent variable and the independent variable needs to be improved. This is done using the following relation:

$$Y = \beta_0 + \beta_0 X_2 + \beta_3 X_3 + \ldots + \beta_N X_N + \zeta$$

Where, this explains the relationship between the independent variable and the dependent variable. X belongs to the set of independent variables and β is the parameter associated with Linear Square Method. The Tobit regression is based on the transformation between the independent variable with the relation X = X* when X*< C and X = X* when X* > C. From the precise value obtained from the independent variable all the non-linear points could be removed. This could be related as

$$Y^* = \beta' X + \zeta$$

Where W is the binary variable, X is the independent variable and ζ is the error factor.

Selection of chromosome

During the selection of chromosome, the independent variable makes a challenging task in selecting the chromosome [5n]. To avoid such regression fit problem in selecting proper chromosome, it is better to select a proper regression fit method.

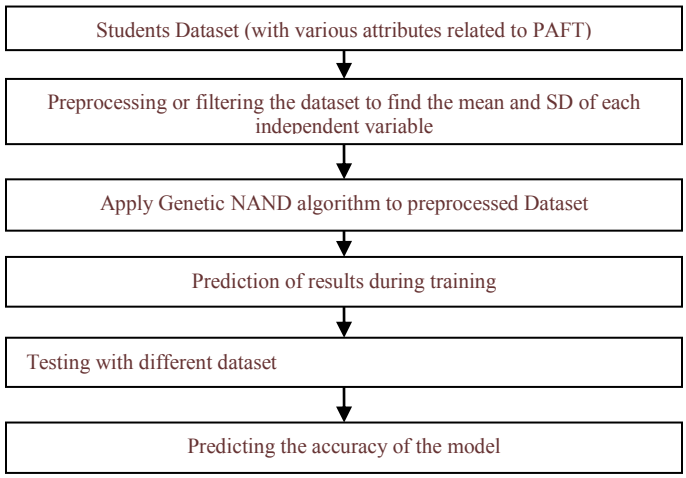


Fig: 2. Genetic NAND Approach

Fitness stage

During fitness stage, the objective function is selected or defined to calculate the best two individuals in the search space. This fitness function for a supervised approach is defined using Mahalanobis distance. In order to normalize the attributes in our system, Mahalanobis distance [17] function is taken into account which is defined as:

$$dij = [(x_i - x_j)^T M^T M (x_i - x_j)]^{1/2}$$

THE IIOAB JOURNAL ISSN: 0976-3104

Matrix ***M*** can be evolved by using the Mahalanobis distance function between the vectors x_i and x_j. The main aim of reducing the irrelevant dataset and for better functionality of GS NAND, matrix ***M*** is chosen to be diagonal or symmetrical. Thus the chromosome is represented as a bit strings formed using the binary representation of each matrix element that belongs to diagonal or upper triangular matrix.

$$M = \begin{pmatrix} m_{11} & m_{11} & \dots & m_{1d} \\ m_{21} & m_{21} & \dots & m_{1d} \\ \dots & \dots & \dots & \dots \\ m_{d1} & m_{d1} & \dots & m_{dd} \end{pmatrix}$$

The bit string chromosomes, $B(m_{ij})$ are represented using binary representation of m_{ij}.

This matrix ***M*** represents the distance function used for activation of neurons. This increases the accuracy and reduces the network error, thus making the distance function to be good. The fitness function of this matrix ***M*** is presented using mean squared error ***E***:

$$fitness = -\log_2 E$$

This function is chosen specifically for training the datasets, such that fitness increases when the rate of error value reduces. The main work of the fitness function is to increase the distance between the chromosomes when they are nearer to zero. This increases the evolutionary pressure at the latter stages of GSOR evolution.

The attributes used here are Age, Gender, Type of study (Full time/Part time), Field, Department, Year of study, Marks scored in Cycle Test , Marks scored in Final Exam, Interest in coming to college/school, A Good Knowledge of Subjects, Ability to clear exams, Interest in Homework, Interest in listening, Ability to learn, read and Write, Quality of work Problem, Solving Team work, Dependability on his friends, Ability to learn and adapt from mistakes, Student shows understanding and sensitivity to needs and differences of others (i.e. ethnicity, religion, language, etc.), demonstrates effective written communication, demonstrates effective Oral communication, The extent to which the student effectively listens, conveys, and receives ideas, information, and direction. Proficiency in English, Ability to Draw (Perspective, Freehand Pencil Drawing), Dimensional Imagination –Skill to Draw Dimensional, Technological Skills, Extracurricular activities.

NAND stage

This stage is used for generating new chromosome individuals, which are encoded and named as phenotype. The mating or crossover is carried out when the two best chromosome.

An individual from the fitness function is made NAND with each other to form an offspring represented as:

$$fx = \overline{fx_i . fx_j}$$

Where, fx_i and fx_j are the best chromosome individuals acquired using fitness function. This creates a two new offspring from the parent chromosomes after crossover using NAND is done. These two new offspring's are put into a new generation of the population and using this logical recombining, the process is likely to attain good individuals.

Shift Stage

The newly produced offspring's are subjected to bit flipping or shift to attain a better diversity within the population and to inhibit convergence within the search space. To achieve this linear NAND register is used to perform the mutation process, where the present state output depends on its previous stage. Thus, the generated mutation individual is added to the empty population P_E. Finally, if the P_E is equal to ***P***, process is completed and finally the iteration ends with a better classification due to good chromosome individuals.

THE IIOAB JOURNAL
ISSN: 0976-3104

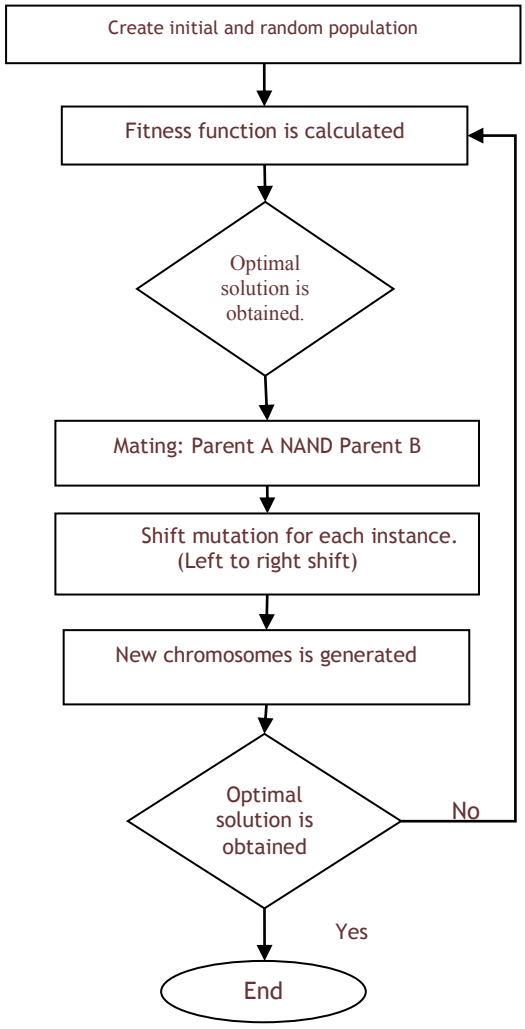


Fig:4. Genetic NAND Algorithm

RESULTS AND DISCUSSIONS

In this proposed approach, the students are allowed to undergo training based on their creative level and tested over several rounds. The students are also evaluated in terms of multiple choice questions that is being considered as an effective assessment technique for the third factor. This was conducted before the final exam date, where final year students were provided with this self-evaluation assessment. Certain questionnaires are asked to teachers regarding their improvement based on the defined set of factors. Finally, after the academic final exams the instructor gave the choice of improved or not improved decision to evaluate the student performance. Depending upon the decision of the instructor, out of collecting 1000 data samples 740 students were in improved state and rest in not improved state. The real time dataset is used as an input for the proposed technique to evaluate the performance of the students and their improvement towards academics. The dataset is filtered using an unsupervised discrete filter to remove all the duplicate sets, irrelevant and redundant dataset. All the 30 attributes were selected and given as an input to the model, since in dynamic environments like institutions the data is limited. This three set factor variable creates a full range view of substantial information regarding the prediction of student's performance in academics.

Initially, net beans are used for coding the GS-NAND algorithm and this file are converted to a weka classifier format (.arff). Then the dataset in .csv file is given as an input to weka classifier and redundant and duplicate data are removed using unsupervised filter. The output of which is given an input to testing GS- NAND classifier,

THE IIOAB JOURNAL ISSN: 0976-3104

which is followed by a set of patterns with predicted results of the given dataset **[Figure– 1]**. Then this model is applied to validate for finding out its accuracy for various other datasets and the results were presented accordingly. Finally the results are classified into two instances class 1 and class 0: class 1 refers to qualified level and class 0 refers to not qualified level.

PAFT variables are used efficiently during training the GS-NAND algorithm, since the pre-processing removes all the irrelevant individuals from the dataset. This increases the efficiency of the training, since the accuracy increased when compared with normal genetic and other RDFF algorithms. To attain this, the following parameters of GS-NAND were taken into consideration: Population Size: 100, target fitness = 0.9, maximum number of generations = 20. This GS-NAND classifier provides the following results for the specified Elite count shown in **[Table–1, Figure–1]**.

Table: 1. Training and testing results of GSNAND

Class	Elite program	Size	Training Fitness	Validation Fitness	Error	Program weights
No	0	7	0.906	0.934	0.08	1.516
No	1	9	0.854	0.828	0.21	0.363
No	2	5	0.800	0.817	0.14	1.152
No	3	9	0.860	0.850	0.28	0.080
No	4	18	0.916	0.906	0.25	0.619
Yes	0	4	0.921	0.922	0.07	1.585
Yes	1	10	0.869	0.806	0.15	0.619
Yes	2	15	0.800	0.815	0.21	0.363
Yes	3	14	0.885	0.882	0.13	0.708
Yes	4	16	0.859	0.877	0.19	0.663

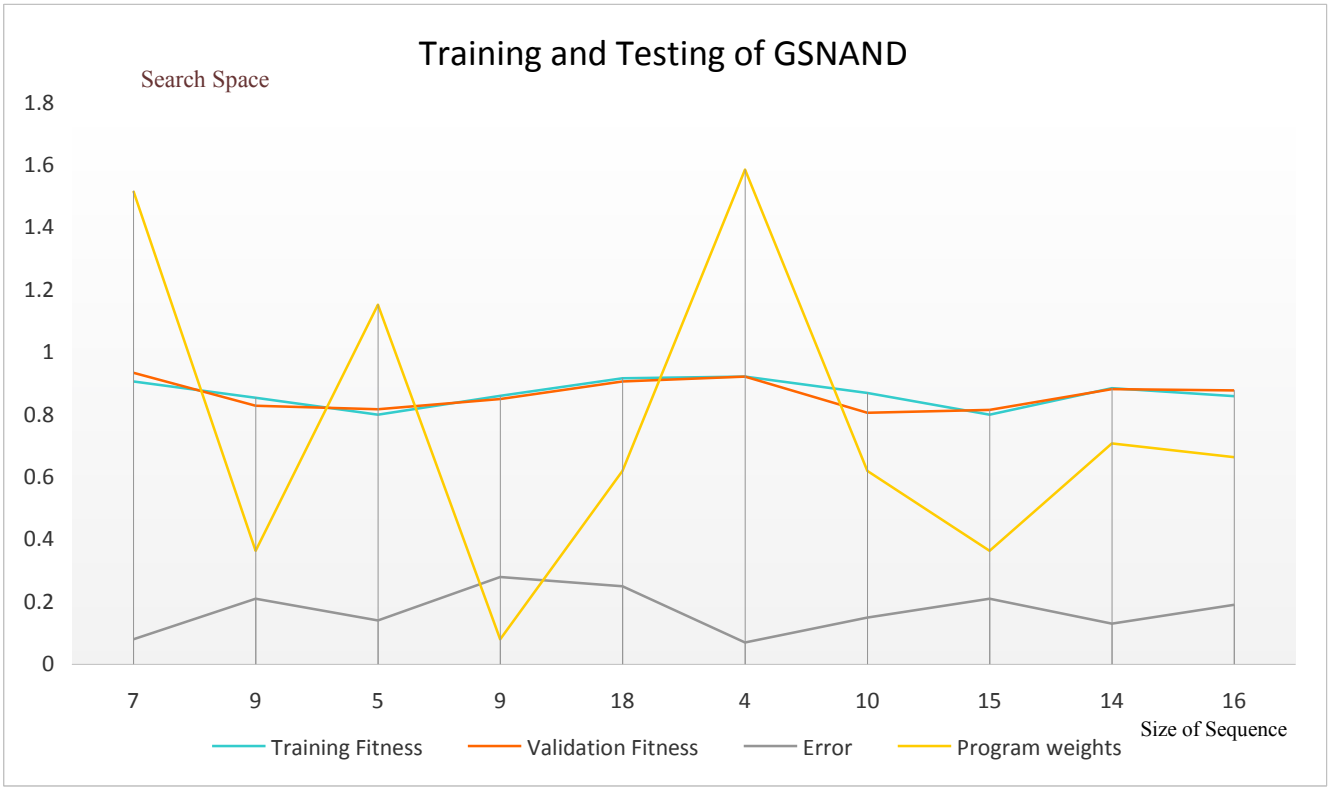


Fig: 1. Training and Testing Results of GSNAND

IIOAB JOURNAL ISSN: 0976-3104

From the above **Table**, it is found that the training and testing fitness were found to be good i.e. it is greater than 9.0 for all the five elite programs of two different classes. Also, the error is reduced significantly using this proposed algorithm in the search space. This shows that the algorithm is performing well, since the rate of fitness in both testing and training phases seems promising. Thus the efficiency of the proposed model is defined in terms of comparing it with existing techniques shown in **[Table–2, Figure–2]**.

Table: 2. Evaluation of Test Performance of Classifiers for 1000 samples

Evaluation	Class	GS - OR	GS-AND	GA	MLP	Neural Network	Proposed GS-NAND
Correctly Classified Instances	-	832	844	820	790	890	980
Incorrectly Classified Instances	-	168	156	180	210	110	20
Mean absolute error	-	0.0324	0.0216	0.18	0.0434	0.1214	0.0216
Root mean squared error	-	0.2341	0.1006	0.4243	0.1152	0.1201	0.1006
True Positive Rate	1	0.854	0.859	0.759	0.042	0.94	1
	0	0.847	0.842	0.957	0.902	0.921	0.957
Precision	1	0.894	0.899	0.923	0.5	0.935	0.964
	0	0.892	0.857	0.772	0.521	0.945	1
Recall	1	0.654	0.662	0.759	0.042	0.784	1
	0	0.	0.957	0.931	0.962	0.942	0.957
F-measure	1	0.887	0.898	0.845	0.077	0.901	0.982
	0	0.892	0.899	0.845	0.676	0.903	0.978

From the above table, the ratio of the correctly classified instance to incorrectly classified instance result seems to be higher than previous techniques like Genetic Algorithm and Multilayer Perception. For the verification of proposed GAXOR approach, the classified instance is examined using primary data approach by asking an opinion about a particular student to his/her teacher. Through the collected primary data it is observed that about 840 samples correctly correlate with the classified value, whereas information regarding remaining students is not known to their teachers. From that, it is concluded that proposed GAXOR approach effectively classifies students' attitude effectively rather that existing approaches. Also, the system is compared with genetic OR and AND gates, it is found that due to the complexity of the gates the system performs inefficiently depending on the results given above. From the above instances, we can tell that the proposed GS-NAND algorithm outperforms better than l network and other tested algorithms. This is due to regression problems occurring in the selection of chromosomes or the individuals in the selection process and this is avoided with the proposed algorithm. The error rate has been reduced absolutely, thereby the accuracy has increased significantly to the level of 1 for improved (1) cases and 0.957 for not improved (0) cases. Likewise the False Positive rate is lesser than previous techniques which is 0.043 for improved (1) cases and 0 for not improved (0) cases. Similarly, the precision, recall and F-measure values are promising towards the proposed GSNAND PAFT model.

CONCLUSION

Thus, from the above experiments using Weka tool, it is found that the proposed GS-NAND PAFT model results were high and accurate when compared with previous classification techniques. Depending on the predicted result, it is concluded that the prediction rate obtained in validation or testing are highly significant in finding the academic improvement of students. Also, there is no existence of disparities since the value of false positive values is less. Thus, this model proved efficient in terms of its accuracy, precision and F-measure values. Also, it could be concluded that the PAFT factors will successfully influence the success rate of students in their academics in the fore coming years. In future, this model could be enhanced using hybridization with other machine learning algorithms. Three factors set could be increase to multiple factor set for acquiring more efficient model to predict efficiently the academic performance of the students.

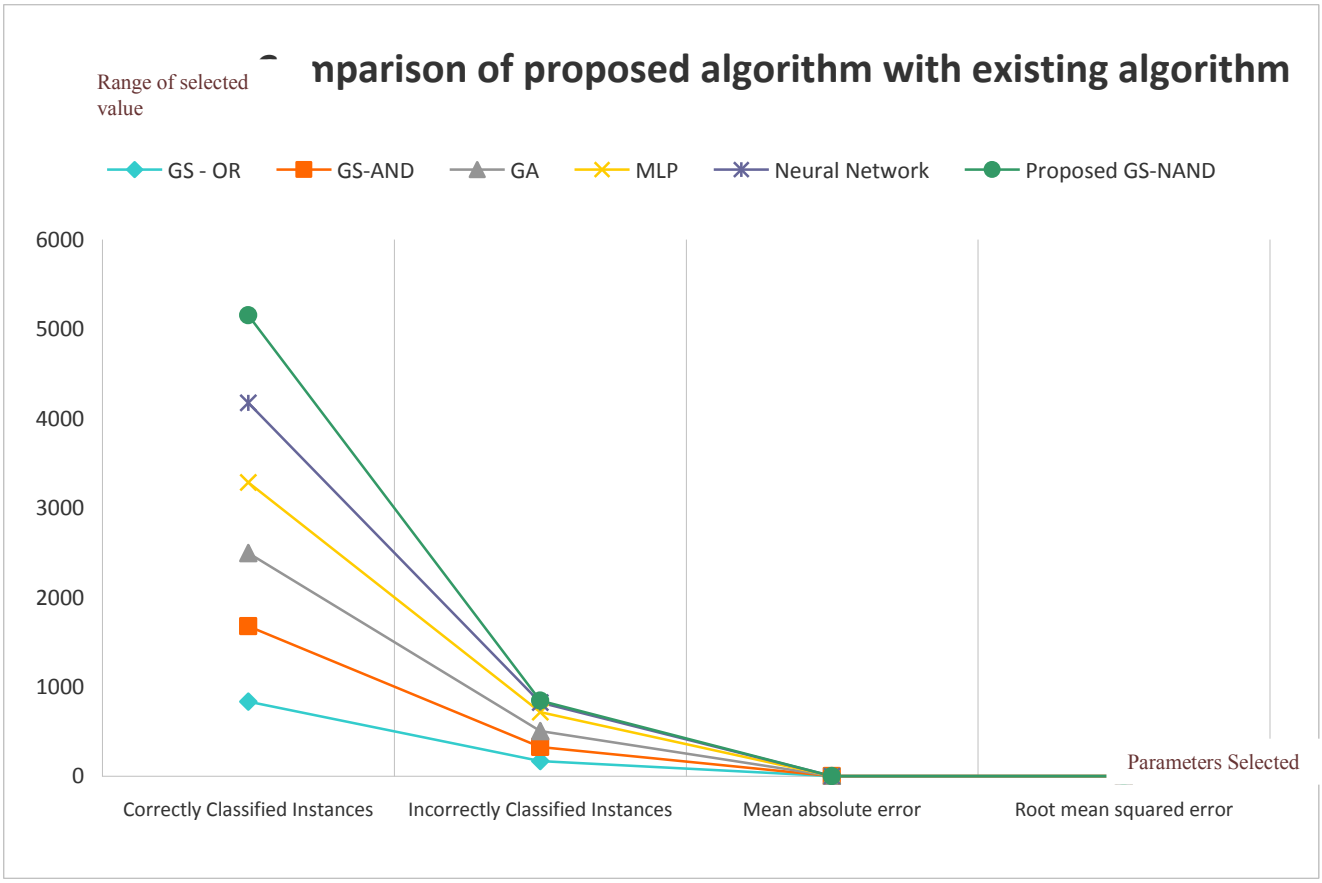


Fig: 2. Comparison of Proposed Algorithm with Existing Algorithm

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] Srecko Natek, Moti Zwilling. [2014] Student data mining solution: knowledge management system related to higher education institutions *Expert Systems with Applications* 41: 6400–6407.

[2] Corbett A, Anderson J. [1995] Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction* 4:253–278.

[3] Superby JF, Vandamme JP, Meskens N. [2006] Determination of factors influencing the achievement of the first-year university students using data mining methods, *Workshop on Educational Data Mining*.

[4] Bharadwaj BK, Pal S. [2011] Data Mining: A prediction for performance improvement using classification, *International Journal of Computer Science and Information Security*, 9(4): 136–140.

[5] Thilagam PS, Ananthanarayana VS. [2008] Extraction and optimization of fuzzy association rules using multi-objective genetic algorithm *Pattern Analysis and Applications*, 11(2):159–168.

[6] Ramesh V, Parkavi P, and Ramar K. [2013] Predicting Student Performance: A Statistical and Data Mining Approach. International *Journal of Computer Applications* 63(8): 35–39.

[7] Ayinde AQ, Adetunji AB Bello, M, Odeniyi OA. [2013] Performance Evaluation of Naive Bayes and Decision Stump Algorithms in Mining Students' Educational Data. *IJCSI*

International Journal of Computer Science Issues. 10(4): 147–151.

[8] Srimani PK, Kamath AS. [2012] Data Mining Techniques for the Performance Analysis of a Learning Model – A Case Study. *International Journal of Computer Applications* 53(5).

[9] Irfan Ajmal Khan, Jin Tak Choi. [2014] An Application of Educational Data Mining (EDM) Technique for Scholarship Prediction. *International Journal of Software Engineering and Its Applications* 8(12):31–42.

[10] Mohammed M, Abu Tair, Alaa M, El-Halees. [2012] Mining Educational Data to Improve Students' Performance: A Case Study. *International Journal of Information and Communication Technology Research* 2(2).

[11] Mohammed M, Abu Tair, Alaa M, El-Halees. [2012] Mining Educational Data to Improve Students' Performance: A Case Study. *International Journal of Information and Communication Technology Research* 2(2).

[12] Zafra A, Ventura S. [2012] Multi-instance genetic programming for predicting student performance in web based educational environments *Applied Soft Computing*, 12(8):2693–2706.

[13] Slim A, Heileman G L, Kozlick J, & Abdallah CT. [2014] Employing Markov Networks on Curriculum Graphs to Predict Student Performance. In Machine Learning and Applications (ICMLA), 2014 13th International Conference on (pp. 415–418). IEEE.

[14] Kau P, Singh M, Josan, GS. [2015] Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector, *Procedia Computer Science*, 57:500–508.

[15] Kongsakun K. [2013] An improved recommendation model using linear regression and clustering for a private university in Thailand. In Machine Learning and Cybernetics (ICMLC), 2013 International Conference on 4: 1625–1630). *IEEE*.

[16] L Ramanathan, A Geetha, and M Khalid.[2015] Mining Students' Record to Predict Their Performance in Undergraduate Degree, *Int J Appl Eng Res* ISSN, 10(1): 973–4562, L Ramanathan, S Dhanda, and D Suresh Kumar.[2013] Predicting students' performance using modified ID3 algorithm, *Int. J Eng Technol*.

[17] L Ramanathan, P Swarnalatha, K Vijayakumar, S. Kaushik, Prabu. [2014] Students Performance Prediction based on Multiple Classifiers, *Int. J Appl Eng Res* ISSN, 9(21): 973–4562.

[18] Partnership for 21st Century Skills. [2007] P21 Framework definitions. Retrieved from<http://www.p21org/storage/documents/P21 Framework Definitions.pdf/>.

[19] Turban E, Sharda R, Delen D. [2011] Decision support and business intelligent systems (9th ed.). Pearson.

[20] Berry M, Linoff G. [2000] Mastering data mining: The art and science of customer relationship management. Wiley.

[21] Sternberg, Robert J. [2006] The nature of creativity. *Creativity Research Journal* 18(1): 87–98.

APPENDIX

The samples of collecting 1000 dataset used above are shown here.

	Gender	Type of study (Full time/Part time)	Field	Department	Year of study	Marks scored in Cycle Test	Marks scored in Final Exam	Interest in coming to college/school
21	MALE	FULL TIME	ENGINEERING	CSE	4	9.1	8.9	highly interested
21	FEMALE	FULL TIME	ENGINEERING	CSE	4	8.2	7.6	interested
21	MALE	FULL TIME	ENGINEERING	CSE	4	6.3	6.3	not interested
20	MALE	FULL TIME	ENGINEERING	CSE	3	5.4	5.8	highly uninterested
20	FEMALE	FULL TIME	ENGINEERING	CSE	3	7.4	7.5	neutral
20	MALE	FULL TIME	ENGINEERING	CSE	3	6.4	6.3	neutral
20	FEMALE	FULL TIME	ENGINEERING	CSE	3	5.6	5.2	highly uninterested
20	MALE	FULL TIME	ENGINEERING	CSE	3	8.6	8.4	interested
20	FEMALE	FULL TIME	ENGINEERING	CSE	3	9.5	9.2	highly interested

A Good Knowledge of Subjects	Ability to clear exams	Interest in Homework	Interest in listening	Ability to learn, read and Write	Quality of work	Problem Solving	Team work
extreme knowledge	very high	highly interested	highly interested	very high	very high	very high	very high
knowledgeable	high	interested	interested	high	high	high	high
less knowledgeable	low	not interested	not interested	low	low	low	low
poor knowledge	very low	highly uninterested	highly uninterested	very low	very low	very low	very low

THE IIOAB JOURNAL ISSN: 0976-3104

neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
less knowledgeable	low	not interested	not interested	low	low	low	low
knowledgeable	high	interested	interested	high	high	high	high
extreme knowledge	very high	highly interested	highly interested	very high	very high	very high	very high

Dependability on his friends	Ability to learn and adapt from mistakes	Student shows understanding and sensitivity to needs and differences of others (i.e. ethnicity, religion, language, etc.)	demonstrates effective written communication	demonstrates effective Oral communication
highly dependable	very high	highly sensitive	highly demonstratable	demonstratable
dependable	high	sensitive	demonstratable	less demonstratable
independable	low	less sensitive	less demonstratable	poor demonstration
highly independable	very low	insensitive	poor demonstration	poor demonstration
neutral	neutral	neutral	neutral	poor demonstration
dependable	neutral	sensitive	less demonstratable	poor demonstration
neutral	low	insensitive	less demonstratable	poor demonstration
highly independable	high	sensitive	demonstratable	demonstratable
neutral	very high	neutral	highly demonstratable	highly demonstratable

ABOUT AUTHORS

Ramanathan L *has received his B.E. in Computer Science & Engineering from Bharathidasan University, Trichirappali, India and M.E in Computer Science from Sathyabama University, Chennai, India. He is currently an Assistant Professor (Selection Grade) in VIT University, Vellore, India. His area of interest is in Data Mining and Database systems, Software Egg, Cloud Computing and Virtualization. He is currently pursuing his PhD in the field of Educational Data mining. His main focus is on Prediction, Classification and Clustering for Educational systems.*

Prof. Angelina Geetha *is currently Professor at B.S.Abdur Rahman University, Chennai, India since 1995. She is Ph. D. from ANNA University Chennai, India. She is having 21 years of teaching and R&D experience. She received his Bachelor from Mother Teresa Womens University, Kodaikanal,in 1994, Master and PhD in Computer Science Engineering from Anna University, Chennai, in 2001, 2008 respectively. He published more than 35 papers in different journals, conferences, patents etc. She is member of various professional bodies like: Life Member System Society of India, Indian Society for Technical Education, senior member of IEEEetc. Her area of interest is Data Mining and Software Engineering and communication Systems.*

Prof M Khalid *has received his PhD degree from IIT Bombay, India. He is currently the Director at Neotia Institute of Technology Management & Science (NITMAS) & TNU, Kolkata, India. He is an academician having more than 35 years' experience including Professor at IIT Bombay, India, Pro-Vice Chancellor and Director at Galgotia University, Greater Noida, India, and VIT University, Vellore, India. He is author of more than 70 research papers published in International journals and conferences. He has received many more awards and memberships to his credit.*

Prof. Swarnalatha Purushotham *is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D. degree in Image Processing and Intelligent Systems. She has published more than 50 papers in International Journals/International Conference Proceedings/National Conferences. She is having 14+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.*

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Ramu and Sevugan

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

AN APPROACH FOR EFFICIENT PRE-PROCESSING OF MULTI-TEMPORAL HYPERSPECTRAL SATELLITE IMAGERY

Swarna Priya Ramu[1*] and Prabu Sevugan[2]
[1]*School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, INDIA*
[2]*School of Computing Sciences and Engineering, VIT University, Vellore, Tamil Nadu, INDIA*

ABSTRACT

The recent advances of technology helps in accessing any data remotely. The observations of earth surface can also be done remotely with the help of high resolution satellite images. These remotely sensed data are used in various applications like urban monitoring, fire detection, flood prediction, oil spills, disaster monitoring, rock type mapping, road networks, change detection, etc. for continuous monitoring and accurate results, the data which is acquired has to be of high resolution and minimum errors. The multi-temporal satellite data gives the data in periodic basis which helps for continuous monitoring, but due to the earth's rotation, climatic changes, sensor characteristics, etc. there are too many distortions and noises which has to be removed before further processing for getting better results. Many researchers have put forth different methodologies for removing the different kinds of noises from the satellite data. Each and every methodology helps in removing a specific kind of noise. This paper proposes a method named Cellular Automata based Gaussian Filter for pre-processing of Multi-temporal Hyperspectral Satellite Images which could be used for removing the noises, filtering it and giving an enhanced image which forms as input for image registration. The performance is analyzed using the Peak Signal to Noise Ratio. The results specify that the proposed methodology is better than the traditional Gaussian Filter.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26th–March-2016
Published on: 12th–May-2016

KEY WORDS

Cellular Automata, Pre-processing, Noise Removal, Gaussian Filter, Hyperspectral Images and Multi-temporal Images.

***Corresponding author:** **Email:** swarnapriya.rm@vit.ac.in; **Tel.:** + 91-9486120585; **Fax:** +91-0416-2243092

INTRODUCTION

The satellite images acquired can vary in resolutions. Based on the resolution in which the data is acquired, they are broadly classified as Spatial Resolution, Spectral Resolution, Radiometric Resolution and Temporal Resolution. The Temporal Resolution satellite data is highly useful in detecting the changes in earth surface, for periodic monitoring of climatic changes, updating the map and predicting the natural disasters.

For these applications to yield best results, the remotely sensed data has to be accurate enough for further processing. But in reality, the data which is sensed remotely suffers from various distortions due to climatic conditions, earth rotation, reflection and refraction. These real time factors which cause the distortions cannot be avoided but the distortions can be removed at the initial stage before further processing. For removing these distortions, there are various traditional methodologies like low pass filter, median filter, high pass filter, etc. The filtering is the basic step which is done in any image processing application for removal of the base noise. Among the various filters available, the Gaussian Filter is commonly used in satellite images for noise removal.

This paper focuses on a pre-processing methodology where cellular automata is used along with the traditional Gaussian Filter and hence the basic white noise, speckle noise and all other reflection and refraction based noises are removed from the multi-temporal hyperspectral satellite images. The paper has been organized into five sections. The Section II presents a description about the earlier research works relevant to the image pre-processing methods. Section III involves the detailed description about the proposed pre-processing methodology for multi-temporal hyperspectral satellite imagery. Section IV presents the performance analysis of the proposed method. The conclusion and future work are discussed in Section V.

THE IIOAB JOURNAL
ISSN: 0976-3104

RELATED WORK

There are different uses of Cellular Automata (CA), for example, in complex frameworks' demonstrating, dissecting and controlling are: The Games of Life [1], Cell Automata in natural framework and environmental framework [2, 3], organic frameworks [4], cell automata in the activity framework [5], cell automata in machine learning and control [6] what's more, CA in cryptography [7].

Advanced Digital Image Processing assumes an essential part in actuality applications, for example, satellite TV, PC tomography and attractive reverberation imaging. It is additionally utilized in zones of examination and innovation, for example, natural data frameworks and astronomy [8].

CA is utilized as a part of different image processing assignments such as Image Filtering in preferred path over some current channels in denoising process [9, 10, 11], Border Detection in Computerized Images that give limits of pictures [12], CA in edge derivation [13, 14, 15], connected set morphology, thinning and thickening of images [11], Image segmentation which is a necessary pieces of image processing applications like restorative picture examination and photograph editing [16, 12] and in image enhancement on account of its element behavior [17].

Popovici and Popovici [9] proposed cellular automata based channel in which the Von Neumann neighborhood norm is used. On correlation with the Gaussian channel, this strategy performs better image upgrade utilizing CA. Selvapeter and Hordijk proposed uniform cellular automata rules for improving the performance of the filtering process [18]. Rosin used a three state automata for removing the noise [19].

The existing approaches are suitable for a basic 2D image where the noise is only a single type, the noise reduction level varies drastically when the level of distortion increases and also the computational time is higher.

PROPOSED METHODOLOGY

Cellular automaton

The cellular automaton is a model which is related to the computational theory. This is otherwise called as cellular spaces. The automaton is made up of a grid of cells which are finite in number and is usually called as states. Out of the available states, one state is considered to be initial state and the nearby cells are considered to be neighborhood and using some mathematical function the new state or next state is determined. The cellular automaton can be one dimensional, two dimensional or n dimensional.

Generally a two dimensional automaton which is deterministic in nature can be well suited for image processing applications. The CA is represented as a triplet as follows.

$$A = (S, N, \delta) \qquad (1)$$

Here 'S' represents a non-empty set which is called as the state set i.e. it is the set of the initial states and next states.

'N' represents the neighborhood cells or states and

δ is the mathematical function or rule which is the condition for moving from initial state to the next state.

This transition is represented by

$$\delta : S^N \rightarrow S \qquad (2)$$

The neighborhoods are found by using basic available norms. There are two commonly used norms namely Von Neumann neighborhood and Moore Neighborhood. The Von Neumann neighborhood is given by,

$$R^2 \ni x \rightarrow h(x) = |x|_1 = |x_1| + |x_2| \in R+ \qquad (3)$$

Moore Neighborhood is using the norm,

THE IIOAB JOURNAL
ISSN: 0976-3104

$$R^2 \ni x \to h(x) = |x| = \max\{|x_1|,|x_2|\} \in R+ \quad (4)$$

The cellular automata A can be either symmetric or asymmetric. The symmetricity is known by following the local rule where the final value is a constant using the following equation,

$$\delta(s1,s2,\ldots,sn) = \delta(S\sigma(1),S\sigma(2),\ldots\ldots) \quad (5)$$

where $s_1, s_2, \ldots, s_n \in S$

$\sigma \in S_N$

where N is the permutation group.

Gaussian filter

Gaussian Filtering is widely used in spatial domain for suppressing the noise but the signal is distorted as well at the same time. Gaussian Filters are widely used in computer vision applications too. There are various forms of Gaussian Filter which are as given below.
The initial variety of filter which was designed for noise suppression was 1D Gaussian filter which is given by,

$$G(x) = \frac{1}{\sqrt{2\Pi\sigma}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (6)$$

where σ is the standard deviation.

The standard deviation is usually calculated for the normal distribution which is always assumed to have the mean as 0.

The later filter which was used is a two dimensional filter which is given by,

$$G(x) = \frac{1}{\sqrt{2\Pi\sigma}} \exp(-(x^2+y^2)/2\sigma^2) \quad (7)$$

Where σ^2 represents the variance
x and y are the co-ordinates.

The 2D Gaussian functions are to be used for working with images. This is just the product of the basic 1D Gaussian function. Here the mean is considered to be (0, 0) and the standard deviation is calculated based on the normal distribution similar to the 1D filter.

Cellular automata based Gaussian filtering

The digital image is taken as the input and it can be assumed as a two dimensional array with m*n pixels. Since the image is a satellite image, the pixel can vary in the color. Hence the satellite image input can be considered as another triplet similar to that of the cellular automaton as,

$$(x,y,I) \quad (8)$$

where x and y are the pixel co-ordinates of the digital image.
I forms the color intensity of the corresponding pixel.

Now, since the image is represented as a triplet, it can be considered as a state in the cellular space of the array m*n. The non-empty set S is modeled as {#, 0, 1, …., I-1}
and the color of the pixel is represented by the states

{0,1,…,I-1}

where I=2 for a monochromatic image
I=16 for an image with 16 colors.
I= 256 for an image with 256 colors.

The Von Neumann norm is used as the neighborhood function in the method. Now, let (x_1, y_1, I_1) be considered as the pixel value and hence the nearby pixels will be (x_2, y_2, I_2), (x_3, y_3, I_3), …, (x_n, y_n, I_n). The transition function δ is given by,

$P^N \to P$ (9)

where P is the set of pixels.

THE IIOAB JOURNAL
ISSN: 0976-3104

After the noise is reduced using the cellular automata, Gaussian filter is performed using the weighted probability density function using,

$$W_{i,j} = \frac{1}{2\pi\sigma^2} exp\left\{\frac{-(i^2+j^2)}{2\sigma^2}\right\} \qquad (10)$$

where σ^2 is the variance and is considered to be greater than or equal to 1.
i and j are the pixels of the image.

The algorithm for the pre-processing of the multi-temporal hyperspectral satellite image set is given in the algorithm below.

Algorithm Cellular Automata Based Gaussian Filter

```
Procedure CANoiseGaussianFilter
    For a set of N multi-temporal images of B bands each
        For each image I in the set
            For each band B0 to BB
                Take the pixels at I(x,y) co-ordinate with I intensity.
                Identify the neighborhood of I(x,y) with I intensity using Von-Neumann Neighborhood Norm.
                Calculate the mean and use the transition function to create a new state N(x,y)
            end for
        Update the image band as per the new N(x,y)
        end for
        Update the image in the set.
    end for
    For set of updated N multi-temporal images of B bands each
        For each image I in the set
            For each band B0 to BB
                Smoothen using the weighted Gaussian filter as per the equation 10
            end for
        end for
    end for
end procedure
```

RESULTS

The algorithm is implemented and tested with a multi-temporal hyperspectral satellite image set having nine images of a same place taken on different dates which are of 27 bands. The original nine satellite images are given in the **Figure– 1**.

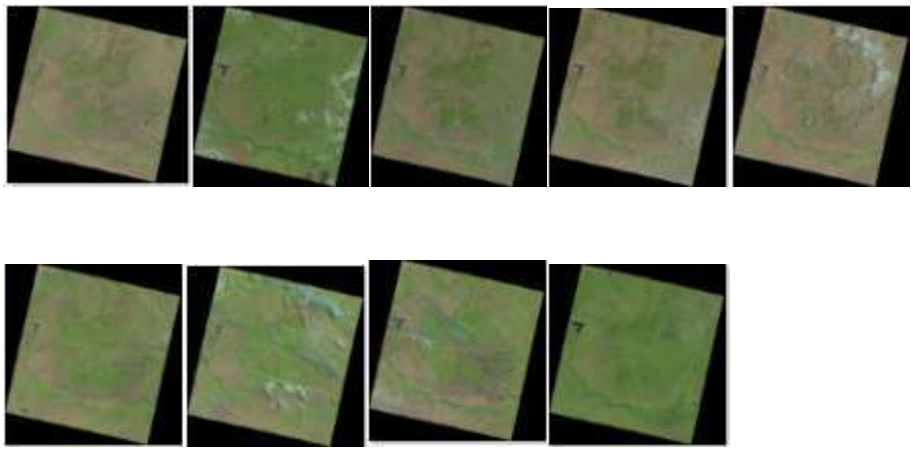


Fig: 1. Multi-temporal hyper-spectral satellite images

These input images are pre-processed with the algorithm and the resulting images with respect to each band of a single image is shown in the **Figure– 2**.

THE IIOAB JOURNAL
ISSN: 0976-3104

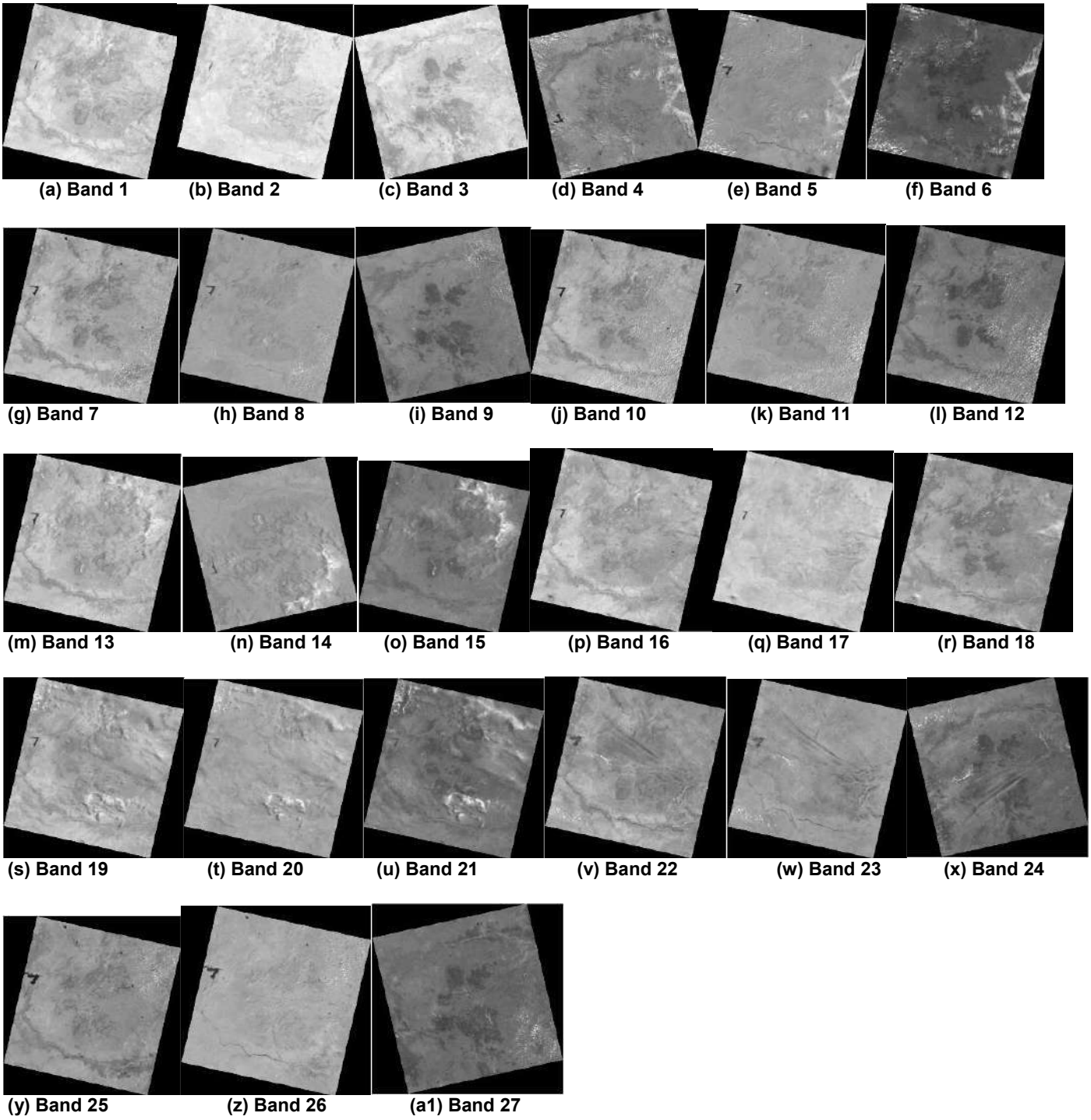


Fig: 2. Filtered images

Similarly all the other images in the set are also pre-processed. The Peak Signal to Noise Ratio (PSNR) values and Mean Square Error (MSE) obtained for each band of the hyperspectral satellite image is tabulated in the T**able–1** and 2 respectively and the results are compared with the existing Gaussian Filter Method.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table:1 PSNR values Vs band number of the image

Band No.	PSNR (%)	
	Proposed CA Based Gaussian Filter	Existing Gaussian Filter
1	38.8034	34.0132
2	39.1606	34.0132
3	38.9423	34.0132
4	30.7844	27.2317
5	30.8598	27.3217
6	30.7343	27.0132
7	33.0398	31.73
8	33.2244	31.82
9	33.1554	31.79
10	27.9696	21.8723
11	28.0001	21.9231
12	27.9378	21.8723
13	35.076	34.0132
14	35.2707	34.0132
15	35.2117	34.0132
16	37.6825	34.0132
17	37.9293	34.0132
18	37.7219	34.0132
19	36.0227	34.0132
20	36.0969	34.0132
21	35.7978	34.0132
22	33.7271	33.5211
23	33.8309	33.5217
24	33.661	33.4217
25	33.3715	32.0122
26	33.4710	32.1576
27	33.5596	32.9717

Table: 2. Band Number of the Image Vs Mean Square Error (MSE) for the Proposed Method

Band No.	Mean Square Error (MSE)	
	Proposed CA Based Gaussian Filter	Existing Gaussian Filter
1	28.6496	32.5432
2	7.8278	10.1234
3	8.2957	9.1234
4	54.2802	56.2309
5	53.3454	57.3456
6	54.9098	58.9876
7	32.2927	37.8767
8	30.9485	32.1234
9	31.4439	35.4312
10	103.7807	107.1234
11	103.0563	106.0123
12	104.5438	108.1234
13	20.206	22.2134
14	19.3204	22.7654
15	19.5844	19.9678

ISSN: 0976-3104

16	11.0874	13.4234
17	10.4748	12.234
18	10.9873	12.12
19	16.2483	17.154
20	15.9731	16.102
21	17.1118	17.789
22	27.566	28.987
23	26.9149	32.345
24	27.9886	34.1534
25	29.9178	36.6545
26	28.1126	38.7912
27	28.6496	41.3456

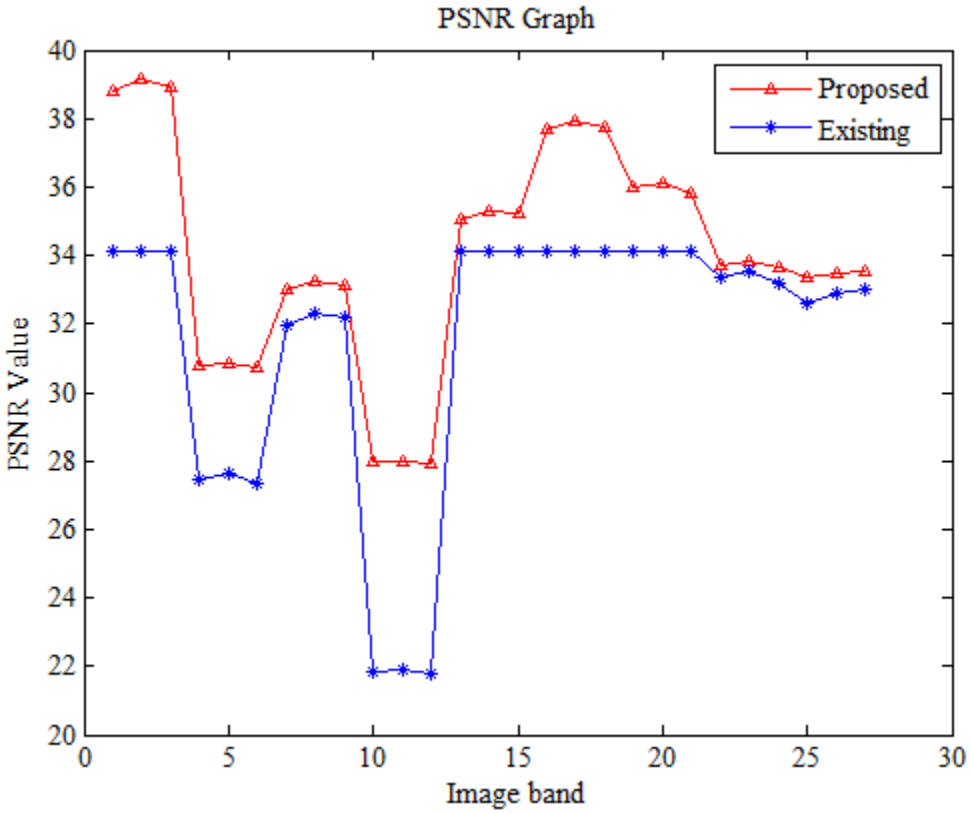


Fig: 3. Image band vs. PSNR value (%)

The performance analysis graph for the proposed cellular automata based Gaussian filter and the existing Gaussian filter is shown in the **Figure– 3**. The performance graph clearly shows that the proposed method gives a better PSNR value compared to the existing method.

The performance analysis graph comparing the MSE values for the proposed cellular automata based Gaussian filter and the existing Gaussian filter is shown in the **Figure–4**. The graph shows that the error rate is reduced than the existing method and the error rate is reduced further when the number of bands increases in the hyperspectral images.

ISSN: 0976-3104

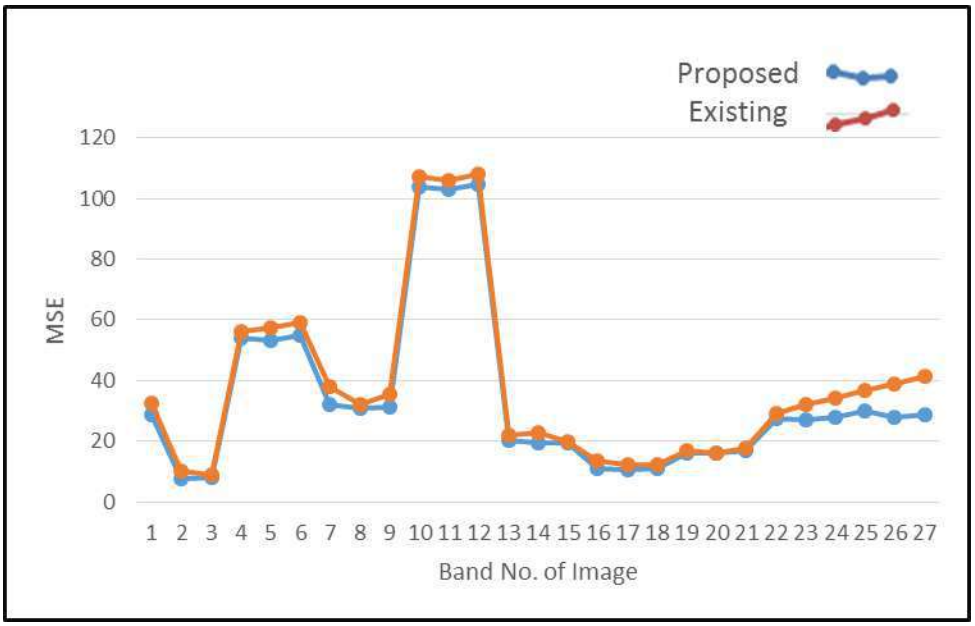


Fig: 4. Image band vs. MSE value

CONCLUSION AND FUTURE WORK

This paper proposes a cellular automation based Gaussian filter for pre-processing of multi-temporal hyperspectral satellite images. This algorithm initially removes the salt and pepper noise, noise caused due to the reflection and refraction, white noise and the speckle noise at the cellular automaton level. Further the Gaussian Filter function smoothens by the linear kernel using the normal distribution where the other random noises are suppressed still further. The performance analysis also clearly shows that the proposed methodology removes the noise at a better level which is proved by the increased PSNR value at each band. The future work is to remove the noises caused due to cloud cover and mask, generally any distortions due to haze which is not focused in this work.

ACKNOWLEDGEMENT

The first author thanks the School of Computing Sciences and Engineering, VIT University and Special thanks to Dean SCOPE, for his kind guidance and support. This work has been (Partially) supported by the research program in SCOPE, VIT University, India.

CONFLICT OF INTEREST

No conflict of interest

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] M Gardner.[1970] Mathematical games: The fantastic combinations of john conways new solitaire game life, Scientific American, 23(4):120–123.

[2] S Aassine and MCEl Jaı.[2002] Vegetation dynamics modelling: a method for coupling local and space dynamics, *Ecological modelling*, 154(3): 237–249.

[3] R Smith.[1991] The application of cellular automata to the erosion of landforms, Earth Surface Processes and Landforms, 16(3): 273–281

[4] H De Garis.[1996] Cam-brain the evolutionary engineering of a billion neuron artificial brain by 2001 which grows/evolves at electronic speeds inside a cellular automata machine (cam), Towards evolvable hardware, Springer, pp. 76-98.

[5] M Fukui and Y Ishibashi.[1996] Traffic flow in 1d cellular automaton model including cars moving with high speed, *Journal of the Physical Society of Japan*, 65(6): 1868–1870.

[6] FM Marchese.[2002] A directional diffusion algorithm on cellular automata for robot path-planning, Future Generation Computer Systems,. 18(7): 983–994.

[7] S Nandi, B Kar and P Pal Chaudhuri.[1994] Theory and applications of cellular automata in cryptography, Computers, *IEEE Transactions on*, 43(12): 1346–1357

[8] LS Davis.[1975] A survey of edge detection techniques," Computer graphics and image processing, 4(3): 248–270

[9] A Popovici and D Popovici.[2002] Cellular automata in image processing, in Fifteenth International Symposium on Mathematical Theory of Net works and Systems, vol. 1,

[10] PL Rosin.[2006] Training cellular automata for image processing, Image Processing, *IEEE*, Transactions on, 15(7): 2076–2087,

[11] A Shukla, S Chauhan, and S Agarwal.[2013] Training of cellular automata for image filtering, in Proc. Second International Conference on Advances in Computer Science and Application – CSA 2013,, pp. 86–95,

[12] YY Boykov and MP Jolly.[2001] Interactive graph cuts for optimal boundary & region Segmentation of objects in ND images, in Computer Vision, ICCV 2001. Proceedings. Eighth IEEE *International Conference on*, 1: 105–112,.

[13] CL Chang, Yj Zhang, and YY Gdong.[2004] Cellular automata for edge detection of images," in Machine Learning and Cybernetics, Proceedings of 2004 International Conference on, *IEEE,* 6: 3830–3834.

[14] T Kumar and G Sahoo.[2010]nA novel method of edge detection using cellular automata, *International Journal of Computer Applications*, 9(4): 0975–8887

[15] S Wongthanavasu and R Sadananda.[2000] Pixel-level edge detection using a cellular automata-based model, Advances in Intelligent Systems: Theory and Applications, 59: 343–351

[16] M Mitchell, JP Crutchfield, R Das et al.[1996] Evolving cellular automata with genetic algorithms: A review of recent work, in Proceedings of the First International Conference on Evolutionary Computation and Its Applications (EvCA96), .

[17] G Sahoo, T Kumar, B Raina, and C Bhatia.[2009] Text extraction and enhancement of binary images using cellular automata, *International Journal of Automation and Computing*, 6(3): 254–260,

[18] PJ Selvapeter and W Hordijk. [2009] Cellular automata for image noise filtering, in Nature & Biologically Inspired Computing, 2009. NaBIC World Congress on. *IEEE*, 2009, pp. 193–197,

[19] PL Rosin.[2010] Image processing using 3-state cellular automata, *Computer vision and image understanding*, 114(7):790–802.

ABOUT AUTHORS

Swarna Priya Ramu *received her Bachelor of Engineering in Computer Science and Engineering from Sona College of Technology affiliated under Periyar University securing a First Class with Distinction and a Master of Engineering in Software Engineering from Sona College of Technology affiliated to Anna University in 2002 and 2008 respectively. She is currently pursuing the Ph.D. Degree with the School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India. Her current research interests include Hyperspectral Image Processing, Classification, Machine Learning, Image Segmentation, and Image Registration.*

Dr. Prabu Sevugan *have completed Bachelor of Engineering in Computer Science and Engineering from Sona College of Technology (Autonomous) and Master of Technology in Remote Sensing from College of Engineering Guindy, Anna University Chennai and one more Master of Technology in Information Technology at School of Computer Science and Engineering, Bharathidasan University Trichy. Did his Doctoral studies on Integration of GIS and Artificial Neural Networks to Map the Landslide Susceptibility from College of Engineering Guindy, Anna University, Chennai. He was a Post-Doctoral Fellow at GISE Advanced research lab, Department of Computer Science and Engineering, Indian Institute of Technology Bombay. He has more than 45 publications in national and international journals and conferences. He organized 3 International Conferences which includes one IEEE Conference as chair and also participated in many workshops and seminars. He is a member of many professional bodies and senior member of IACSIT, UACEE and IEEE. He is having more than ten years of experience in teaching and research. Currently I am working as a Division Chair for Parallel and Distributed Computing, School of Computing Science and Engineering, VIT University Vellore.*

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Singh

ARTICLE OPEN ACCESS

PROCESSING LINKED FORMAL FUZZY CONTEXTS USING NON-COMMUTATIVE COMPOSITION

Prem Kumar Singh*
Centre for Mobile Cloud Computing Research, Faculty of Computer Science, University of Malaya, Kuala lumpur-50603, MALAYSIA

ABSTRACT

This paper focused on analysis of formal fuzzy contexts having common attribute set (Y) as follows: (X, Y, R_1) and (Z, Y, R_2). For this purpose given contexts are linked through composition and further analyzed using projection operator. Since the composition is not commutative i.e Due to this fact the knowledge discovered from the both the compositions i.e. $(X,Z,R_3 = R_1 * R_2)$ or $(Z,X,R_4 = R_2 * R_1)$ *is compared. One application of the proposed method is also shown for predicting the score of a cricket match.*
.

Received on: 25th-Nov-2015
Revised on: 1th-Feb-2016
Accepted on: 21th– Mar-2016
Published on: 16th –May-2016

KEY WORDS

Composed context; Formal Concept Analysis; Formal fuzzy context; Formal fuzzy concept; Knowledge representation.

***Corresponding author: Email:** premsingh.csjm@gmail.com, premsingh.csjm@yahoo.com

INTRODUCTION

Knowledge discovery from a given formal context is one of the major concerns for the research communities [1]. For this purpose Wille [2] introduced a mathematical model called as Formal Concept Analysis (FCA). This mathematical model was based on applied lattice theory and its properties. FCA receives the input in form of a binary matrix called as formal context. From the given formal context FCA discovers formal concepts having objects set with their common attributes closed with Galois connection. All the discovered formal concepts can be visualized as a hierarchical ordering in the concept lattice. Several algorithms are proposed for generating the concept lattice with its application in different contexts [3-4]. Further for handling the uncertainty and vagueness in linguistics words like (`young', `tall' etc.) precisely, FCA is linked with fuzzy [5], interval-valued [6-7], and bipolar fuzzy set [8-9]. Recent year's attention has been paid towards linking the formal fuzzy context [10] using composition [11-13], hedges [14] and, shared bond [15]. Motivated from these recent studies, in this paper, we focus on the knowledge processing tasks via linking the fuzzy contexts. Let us suppose a formal fuzzy contexts-*:* (X, Y, R_1) have the similar attribute set when compare to another formal fuzzy context-(Z,Y, R_2). In this case these two contexts can be linked via composition i.e. $(X,Z,R_3 = R_1 * R_2)$ *or* $(Z,X,R_4 = R_2 * R_1)$ [16]. The composition of contexts contains set of objects (X) from the first context and set of attributes (Z) from another context. In this case a question arises that how to process the linked fuzzy contexts using composition.

Some attention has been paid towards the knowledge discovery task from the linked fuzzy contexts. In these type of dynamic cases it may not be necessary to work with all formal fuzzy concepts. We need some of the formal concepts which fulfill our requirements based on given objects and attribute set [17-18]. This motivates us to study the operation on fuzzy relations which project it on selected dimensions (like objects and attribute set) for the adequate analysis. In this case the projection of fuzzy relation and its applications with formal context [19-20], formal fuzzy context [21] looks useful. The properties of projection provides us an alternative way to project the computed fuzzy relation on the chosen subset of selected dimensions of the given context [22]. In this paper we try to utilize this vital property of projection operator to handle the composed fuzzy contexts.

THE IIOAB JOURNAL
ISSN: 0976-3104

Recently, the analysis of composed context is discussed with an illustrative example [21]. However, the composition is not commutative i.e. $\underset{\sim}{R}_1 * \underset{\sim}{R}_2 \neq \underset{\sim}{R}_2 * \underset{\sim}{R}_1$. Due to that it provides two different fuzzy formal contexts. To reveal some interested concepts through $(X, Z, \underset{\sim}{R}_3 = \underset{\sim}{R}_1 * \underset{\sim}{R}_2)$ *or* $(Z, X, R_4 = \underset{\sim}{R}_2 * \underset{\sim}{R}_1)$ is discussed in this paper. Further the comparative study between the knowledge discovered via non--commutative composition is shown with an illustrative example. The proposed method is unique when compare to other available approaches in following aspects:

(1) The proposed method analyze the formal fuzzy contexts using both non--commutative composition i.e. $\underset{\sim}{R}_1 * \underset{\sim}{R}_2 \neq \underset{\sim}{R}_2 * \underset{\sim}{R}_1$ and,

(2) Second it compares the knowledge discovered from both the non-commutative composed contexts with an example.

Other parts of the paper is composed as follows: The related notions of FCA with fuzzy setting is given in section 2. The proposed algorithm and its illustration is shown in section 3 and section 4, respectively. Section 5 contains application of the proposed algorithm to project the cricket score followed by conclusions, acknowledgment and references.

FORMAL CONCEPT ANALYSIS IN THE FUZZY SETTING

The In this section we recall some basic definition of FCA in the fuzzy setting as given below:

Definition 1. (Formal fuzzy context). A fuzzy formal context is a triplet F= (X, Y,) where X is set of objects, Y is set of attributes and is a fuzzy binary relation on X Y. For any $x \in X$ $y \in Y$, $\underset{\sim}{R}(x, y) \in L$, to which object x has attribute y (L is a support set of some complete residuated lattice L).

Definition 2. (Fuzzy set representation of object). Each object x X in a fuzzy formal context K can be represented by a fuzzy set (A) as A(x) = $\{(y_1(\mu_1)), (y_2(\mu_2)), \ldots, (y_m(\mu_m))\}$, where $\{y_1, y_2, \ldots, y_m\}$ is the set of attributes in K and is the membership of x with attribute Y in K. Then A (x) is called the fuzzy representation of object X. Similarly we can define it for attribute B(y).

Definition 3. (Residuated Lattice). A residuated lattice L = $(L, \wedge, \vee, \otimes, \rightarrow, 0, 1)$ is the finite structure of truth values of object and its properties. L is complete residuated lattice iff:

(1). $(L, \wedge, \vee, 0, 1)$ is a complete lattice.

(2). $(L, \otimes, 1)$ is commutative monoid.

(3) $\otimes$ and $\rightarrow$ are binary operations called multiplication and residuum, respectively i.e. $a \otimes b \leq c$, iff $a \leq b \rightarrow c$ for any *a*, *b*, *c* ∈ *L*.

The operators and are defined distinctly by Lukasiewicz, Gödel, and Goguen independently. as given below:

Definition 4. (Formal fuzzy concept). For any L-set A LX of objects and L- set of B LY attributes, we can define a L- set of attributes and L- set of objects as follows:

- $A^{\uparrow}(y) = \wedge_{x \in X}(A(x) \rightarrow \underset{\sim}{R}(x, y))$
- $B^{\downarrow}(y) = \wedge_{y \in Y}(B(y) \rightarrow \underset{\sim}{R}(x, y))$

$A^{\uparrow}(y)$ the truth degree of attribute y is covered by all objects from A and is the truth degree of object x has all attributes from B. Then $A^{\uparrow}(y)$ a pair of (A, B) LX LY is a fuzzy formal concept if and and defines sub-super concept hierarchy as (A1, B1) ≤ (A2, B2) then A1 ⊆ A2 (or B2 ⊇ A1) in L(F).

Definition 5. (Fuzzy concept lattice). All the concepts from the formal fuzzy context F are denoted as L(F) and they denies partial order relation: if (A1, B1) ≤ (A2, B2) then A1 ⊆ A2 or B2 ⊇ A1). Thus, (L(F),≤) is a fuzzy concept lattice. It defined by several researchers separately. Together with this ordering in the complete lattice there exist an infimum and a supremum for any two formal concepts given by [2-3]:

- $\wedge_{j \in J}(A_j, B_j) = (\bigcap_{j \in J} A_j, (\bigcup_{j \in J} B_j)^{\pi})$
- $\vee_{j \in J}(A_j, B_j) = ((\bigcup_{j \in J} A_j)^{\uparrow\downarrow}, \bigcap_{j \in J} B_j)$

Classical logic is special case of FCA with fuzzy setting in which the L-set contains only two elements i.e 1 and 0. The hierarchical order visualization of formal concepts in the concept lattice is an attentive output for the applications of FCA [4]. To get this output from the linked formal fuzzy contexts via their attribute set is major concern of researchers. The issue includes the connection of these two context and their adequate analysis. For this purpose in this paper a method is proposed to link these contexts based on composition. Further the composed context is analyzed with projection on objects and attribute set. The followings are given in the next section.

THE IIOAB JOURNAL
ISSN: 0976-3104

PROPOSED METHOD

Let us suppose we have given a formal fuzzy context-F = *:* $(X, Y, \underset{\sim}{R})$. In any given fuzzy context following possible conditions may exists as shown in **[Table-1]**. Complete (or incomplete) represents that the information is available (or not-available) for the given sets [26--27]. To link the data set we should have the complete information about the given contexts [10-11, 28]. Then the given fuzzy contexts i.e. $(X, Y, \underset{\sim}{R}_1)$ and $(Z, Y, \underset{\sim}{R}_2)$ can be linked as follows:

Table 1: Some of the well-known conditions in a given formal fuzzy context

Conditions	Object	Attribute	Fuzzy Relation
a	Complete	Complete	Incomplete
b	Incomplete	Complete	Complete
c	Complete	Incomplete	Complete
d	Complete	Complete	Complete
e	Incomplete	Incomplete	Incomplete

Step 1. The given fuzzy contexts can be linked via composition $\underset{\sim}{R}_1 \times \underset{\sim}{R}_2 = \underset{\sim}{R}_3$ or $\underset{\sim}{R}_2 \times \underset{\sim}{R}_1 = \underset{\sim}{R}_4$. The composed fuzzy contexts $(X, Z, R_3 = \underset{\sim}{R}_1 * \underset{\sim}{R}_2)$ *or* $(Z, X, R_4 = \underset{\sim}{R}_2 * \underset{\sim}{R}_1)$ includes objects from first context and attributes from second. In this scenario it can be process as follows:

Step 2. The fuzzy relation for the given composed context $(X, Z, R_3 = \underset{\sim}{R}_1 * \underset{\sim}{R}_2)$ can be projected on objects set-(X) as follows: $\prod_{R3}(X) = \{(x), \max_z \mu_{R3}(x, z_k) \in X \times Z\}$. This provides the maximum degree of membership value from the tuples (x, z_k) in the given fuzzy formal context. Subsequently we can find the projection on attribute set $\prod_{R3}(Z) = \{(z), \max_X \mu_{R3}(x_i, z) \in X \times Z\}$. These projection on object and attribute set provides two fuzzy contexts based on object and attribute set of context $(X, Z, R_3 = \underset{\sim}{R}_1 * \underset{\sim}{R}_2)$.

Step 3. Similarly second composed fuzzy context $(Z, X, R_4 = \underset{\sim}{R}_2 * \underset{\sim}{R}_1)$ can be projected on objects and attribute set for knowledge discovery task as per step 2. This will also provide two distinct fuzzy context can be received based on projection. Further we can generate the formal concepts to reveal the interested pattern in the data set.

Step 4. The composition of fuzzy contexts is not commutative. Due to this fact we can compare the knowledge discovered from both the composed contexts for adequate analysis.

The above steps are shown as pseudo code in the **[Table-2]**. The proposed algorithm receives input as fuzzy contexts and link them using composition of matrix code as shown in steps 2 to 4. Then it computes the projection on objects using the steps 5 and 6. To generate the fuzzy concept lattice based on object concept as given in step 7. Similarly the proposed algorithm provides a projection on attribute set using the steps 9 and 10. To generate the fuzzy concepts for the attribute context at step 11. In the end the proposed algorithm provides two fuzzy concept lattice structure based on objects and attribute set from the composed context $(X, Z, R_3 = \underset{\sim}{R}_1 * \underset{\sim}{R}_2)$. Similarly the proposed method provide two fuzzy concept lattice for another (non -commutative) composition- $(Z, X, R_4 = \underset{\sim}{R}_2 * \underset{\sim}{R}_1)$. In this way the proposed method provides an alternative way to analyze the linked formal contexts based on their projection rather than one independent context. Moreover, it optimizes the possibility to find some interested formal concepts in less computational complexity.

Complexity: Let us suppose, number of objects in first fuzzy context $(|X|) = n$, and number of objects in second fuzzy context $(|Z|)=k$ where number of attributes in both of the context $(|Y|) = m$. Now suppose max $\{n, m, k\}=n$. Then in the computing the composition of fuzzy contexts will takes $O(n^3)$ time. Further, the proposed method finds the projection on objects and attributes set which takes complexity $O(n^2)$. Hence the overall computation takes $O(n^3 + n^2)$ time.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 2. Proposed Algorithm for processing the linked formal fuzzy contexts

Input : Given two formal fuzzy contexts(X, Y, R_1) and (Z, Y, R_2)
Output: Formal fuzzy concepts based on objects and attribute set

(1) Enter the matrix (X, Y, R_1) and (Z, Y, R_2)
(2) Compose the matrix $R_1 \cdot R_2$ as follows:

```
for (i=0;i<n;i++)
  for(j=0;j<m;j++)
      z[i][j]=0;
        for (k=0;k< max(n,m);k++)
          z[i][j]+= x[i][k]*y[k][j].
```

(3) Return the composed matrix $(X, Z, R_3 = \underset{\sim}{R}_1 * \underset{\sim}{R}_2)$
(4) End the for loop
(5) Project the composed context based on object set
for (i=0;i<n;i++)
$\prod_{R3}(X) = \{(x), \max_z \mu_{R3}(x, z_k) \in X \times Z\}$
(6) Return the context obtain using projection on object set.
(7) Generate the formal fuzzy concept lattice for this context.
(8) End for loop.
(9) Similarly project the composed context based on object set
for (k=0;j<k;k++)
$\prod_{R3}(Z) = \{(z), \max_X \mu_{R3}(x_i, z) \in X \times Z\}$
(10) Return the context obtain using projection on object set.
(11) Generate the formal fuzzy concept lattice for this context.
(12) End for loop.
(13) Similarly other composed context $(Z, X, R_4 = \underset{\sim}{R}_2 * \underset{\sim}{R}_1)$ can be processed.

KNOWLEDGE DISCOVERY USING COMPOSITION OF FUZZY CONTEXTS

In this section we illustrate the proposed algorithm through an example. Let us suppose a company want to analyze the suitability of some candidate to offer the employments-Domestichelper, Waiter, Accountant, Carsalesman based on the following knowledge-Computer Science, Accounting, Mechanical, Cooking as represented in **[Table-3]**. [11-12]. In the same time company receives some CV from the students--C_1, C_2, C_3, C_4, C_5 based on their knowledge-Computer Science, Accounting, Mechanical, Cooking as shown in **[Table-4]**. Now a problem arises with company that how to discover the knowledge from these given contexts to offer the employments.

Table: 3. Different knowledge required for the employment in a company i.e. $(X, Y, \underset{\sim}{R}_1)$

Employment	Computer Science	Accounting	Mechanical	Cooking
Domestic Helper	0.1	0.3	0.1	1.0
Waiter	0.0	0.4	0.0	0.7
Accountant	0.9	1.0	0.0	0.0
Car Salesman	0.5	0.7	0.9	0.0

Table: 4. CV received from the candidates based on the given knowledge i.e $(Z, Y, \underset{\sim}{R}_1)$

Employment	Computer Science	Accounting	Mechanical	Cooking
C_1	0.5	0.8	0.3	0.6
C_2	0.2	0.5	0.1	1.0
C_3	0.0	0.2	0.0	0.3
C_4	0.9	0.4	0.1	0.5
C_5	0.7	0.5	0.2	0.1

THE IIOAB JOURNAL ISSN: 0976-3104

Table :5.Composition of Table 3 and Table 4 i.e. $(X, Z, R_3 = R_1 * R_2)$

R_1*R_2	C_1	C_2	C_3	C_4	C_5
Domestic Helper	0.6	1.0	0.3	0.5	0.1
Waiter	0.9	1.0	0.6	0.8	0.4
Accountant	0.6	0.3	0.1	0.4	0.5
Car Salesman	0.4	0.2	0.5	0.2	0.3

To solve this problem we can link these two context using composition (X, Z, R_3=R_1*R_2) as shown in **[Table-5]**. From the composed contexts shown in **[Table-5]** we can generate the fuzzy formal concepts to analyze the company and candidate requirements which are as follows:

(1). Company requires a suitable candidate based on the advertisement shown in **[Table-3]**.
(2). Candidate requires a suitable job based on their knowledge shown in **[Table-4]**.

To solve the above mentioned problems, we require some of the formal concepts based on objects--(employments (X)) and attributes--(Candidates (Z)) set. This can be achieved through the proposed method. **[Table-5]** can be projected based on set of objects--(employment-X) as follows:
(Domestic helper, C_2)=1.0,
(Waiter, C_2)=1.0,
(Accountant, C_1)=0.6,
(Carsalesman, C_3)=0.5.
All computed fuzzy relations are shown in **[Table-6]**.

Table: 6. Projection of Table 5 on object set i.e. employment (X)

Employment	C_1	C_2	C_3	C_4	C_5
Domestic Helper	0.0	1.0	0.0	0.0	0.0
Waiter	0.0	1.0	0.0	0.0	0.0
Accountant	0.6	0.0	0.0	0.0	0.0
Car Salesman	0.0	0.0	0.5	0.0	0.0

The generated formal fuzzy concepts from **[Table-6]** are:

1. $\{\{1.0/\text{Domestic helper} + 1.0/\text{Waiter} + 1.0/\text{Accountant} + 1.0/\text{Carsalesman}\}, \varnothing\}$,
2. $\{\{1.0/\text{Accountant}\}, 0.6/C_1\}$,
3. $\{\{1.0/\text{Domestichelper} + 1.0/\text{Waiter}\}, 1.0/C_2\}$,
4. $\{\{1.0/\text{Carsalesman}\}, 0.5/C_3\}$,
5. $\{\varnothing, 1.0/C_1 + 1.0/C_2 + 1.0/C_3 + 1.0/C_4 + 1.0/C_5\}$.

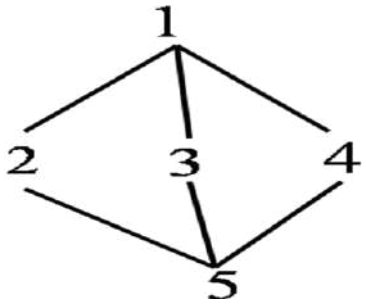


Fig: 1. Fuzzy concept lattice generated for Table-6

All the above generated concepts are shown in **Figure-1**. From that following information can be extracted:

THE IIOAB JOURNAL ISSN: 0976-3104

A. Concept 2 includes that the candidate-C_1 is suitable for Accountant.
B. Concept 3 includes that candidate- C_2 is suitable for Domestichelper or Waiter.
C. Similarly, concept 4 includes that candidate-C_3 is suitable for Carsalesman.

The above information is helpful for the company.

Similarly, projection on attributes (candidate-Z) of the context shown in **[Table-5]** can be computed as follows:
(Waiter, C_1)= 0.9,
(Domestic helper, C_2)=1.0,
(Waiter, C_2)=1.0,
(Waiter, C_3)=0.6,
(Waiter, C_4)=0.8,
(Accountant, C_5)=0.5.
All the computed fuzzy relations using the projection on attributes is shown in **[Table-7]**.

Table:7. Projection of Table 5 on attribute set i.e. candidate (Z)

Employment	C_1	C_2	C_3	C_4	C_5
Domestic Helper	0.0	1.0	0.0	0.0	0.0
Waiter	0.9	1.0	0.6	0.8	0.0
Accountant	0.0	0.0	0.0	0.0	0.5
Car Salesman	0.0	0.0	0.0	0.0	0.0

The generated formal fuzzy concepts from **[Table-7]** are:

1. $\{\{1.0/\text{Domestic helper} + 1.0/\text{Waiter} + 1.0/\text{Accountant} + 1.0/\text{Carsalesman}\}, \varnothing\}$,
2. $\{\{1.0/\text{Domestichelper} + 1.0/\text{Waiter}\}, \{1.0/C_2\}\}$,
3. $\{\{1.0/\text{Waiter}\}, \{1.0/C_1 + 1.0/C_2 + 1.0/C_3 + 1.0/C_4\}\}$,
4. $\{\{1.0/\text{Accountant}\}, \{0.5/C_3\}\}$,
5. $\{\varnothing, \{1.0/C_1 + 1.0/C_2 + 1.0/C_3 + 1.0/C_4 + 1.0/C_5\}\}$.

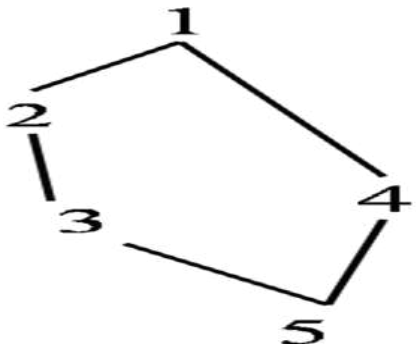


Fig: 2. Fuzzy concept lattice generated for Table 7

The formal concept generated from **[Table-7]** is shown in **Figure-2**From which following information can be extracted:

(A) Concept 2 includes that candidate-C_2 is suitable for Domestic helper or Waiter.
(B) Concept 3 includes that almost each candidates are suitable for Waiter. Among them candidate-C_2 is highly eligible.
(C) Similarly, concept 4 includes that candidate-C_5 is suitable for Accountant.

The above information is useful for the candidate. The obtained conclusions from **Figures-1 and 2** shown in **[Table-8]**.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 8. Conclusions obtained from Figure 1and Figure-2

Candidate	C_1
C_1	Accountant or Waiter
C2	Domestichelper or Waiter
C3	Car Salesman
C_4	Not Specific for any Job
C_5	Accountant

Similarly we can analyze another non-commutative composition of the given contexts i.e. $R_2 * R_1$ as shown in **[Table-9]**.

Table:9. Composition of Table 3 and Table 4 i.e. $(Z, X, R_4 = \underset{\sim}{R}_2 * \underset{\sim}{R}_1)$

R_2*R_1	Domestic helper	Waiter	Accountant	Car salesman
C_1	0.5	0.5	0.4	0.4
C_2	0.8	0.7	0.0	0.0
C_3	0.3	0.2	0.2	0.7
C_4	0.2	0.1	0.5	0.5
C_5	0.4	0.3	0.8	0.8

The projection of composed fuzzy context shown in **[Table-9]** on objects (i.e. candidate) is as follows:
(C_1, Domestichelper)= 0.5,
(C_1, Waiter)= 0.5,
(C_2, Domestichelper)= 0.8,
(C_3, Carsalesman)= 0.7,
(C_4, Accountant)= 0.5,
(C_4, Carsalesman)= 0.5,
(C_5, Accountant)= 0.8,
(C_5, Carsalesman)= 0.8.
These relations are shown in **[Table-10]**.

Table 10. Projection of Table 9 on objects set i.e. Candidate (Z)

R_2*R_1	Domestic helper	Waiter	Accountant	Car salesman
C_1	0.5	0.5	0.0	0.0
C_2	0.8	0.0	0.0	0.0
C_3	0.0	0.0	0.0	0.7
C_4	0.0	0.0	0.5	0.5
C_5	0.0	0.0	0.8	0.8

The generated formal fuzzy concepts from **[Table-10]** are:

1. $\{\varnothing, \{1.0/C_1 + 1.0/C_2 + 1.0/C_3 + 1.0/C_4 + 1.0/C_5\}\}$.
2. $\{0.5/C_1 + 0.5/C_2,\ 1.0/\text{Domestichelper}\}$,
3. $\{0.5/C_1 + 0.5/C_2,\ 1.0/\text{Domestichelper} + 1.0/\text{Waiter}\}$,
4. $\{0.5/C_4 + 0.8/C_5,\ 1.0/\text{Accountant} + 1.0/\text{Carsalesman}\}$,
5. $\{0.7/C_3 + 0.5/C_4 + 0.8/C_5,\ 1.0/\text{Accountant}\}$,
6. $\{\{1.0/\text{Domestic helper} + 1.0/\text{Waiter} + 1.0/\text{Accountant} + 1.0/\text{Carsalesman}\}, \varnothing\}$.

THE IIOAB JOURNAL
ISSN: 0976-3104

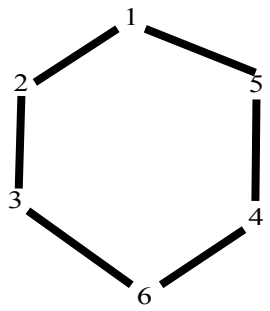


Fig: 3. Fuzzy concept lattice generated for Table 10

The concepts generated from the **[Table-10]** are shown in the concept lattice of Fig. 3. From that following information can be extracted:
(A) Concept 2 includes that the for the employment of Domestichelper candidates C_1 and C_2 are suitable .
(B) Concept 3 includes that for the employment of Domestichelper or Waiter candidate C_1 is suitable.
(C) Concept 4 includes that for the employment of Accountant or Carsalesman candidates C_4 and C_5 are suitable .
(D) Concept 5 includes that for the employment of Carsalesman candidates-C_3, C_4 and C_5 are suitable.
Above conclusions are helpful for the candidates.
Similarly, the projection of composed context shown in **[Table-9]** can be computed on attributes (i.e. employment) as follows:
(C_1, Domestichelper)= 0.8,
(C_2, Waiter)= 0.7,
(C_3, Accountant)= 0.8,
(C_4, Carsalesman)= 0.8.
These computed relations are represented in the **[Table-11]**.

Table: 11. Projection of Table 9 on attribute set i.e. Employment (X)

R_2*R_1	Domestic helper	Waiter	Accountant	Car salesman
C_1	0.0	0.0	0.0	0.0
C_2	0.8	0.7	0.0	0.0
C_3	0.0	0.0	0.0	0.0
C_4	0.0	0.0	0.0	0.0
C_5	0.0	0.0	0.8	0.8

The generated formal fuzzy concepts from **[Table-11]** are:

1. $\{\varnothing, \{1.0/C_1 + 1.0/C_2 + 1.0/C_3 + 1.0/C_4 + 1.0/C_5\}\}$.
2. $\{0.8/C_2,$ 1.0/Domestichelper $\}$,
3. $\{0.7/C_2,$ 1.0/Domestichelper+1.0/Waiter $\}$,
4. $\{0.8/C_5,$ 1.0/Accountant+1.0/Carsalesman$\}$,
5. $\{\{1.0/$ Domestic helper $+1.0$/Waiter+1.0/Accountant+1.0/Carsalesman$\}, \varnothing\}$.

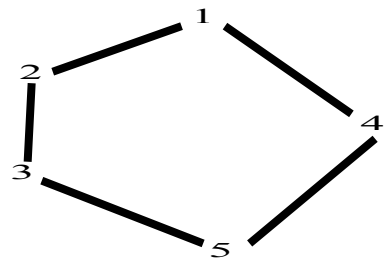


Fig:4. Fuzzy concept lattice generated for Table 11

THE IIOAB JOURNAL
ISSN: 0976-3104

Concept lattice builds by concept generated from **[Table-11]** is shown in Fig. 4. From that following information's can be concluded:
(A) Concept 2 includes that for the employment of Domestichelper candidates-C_2 is suitable.
(A) Concept 3 includes that for employment of Domestichelper or Waiter candidates-C_2 is suitable.
(C) Concept 4 includes that for employment of Accountant or Carsalesman candidates-C_5 is suitable.
The above information is useful for the company for shortlisting the CV. The obtained conclusions from **Fig. 3 and Fig. 4** are shown in **[Table-12]**.

Table 12. Conclusions obtained from Fig. 3 and Fig. 4

Candidate	C_1
C_1	Domestichelper or Waiter
C2	Domestichelper or Waiter
C3	Car Salesman
C_4	Accountant or Carsalesman
C_5	Accountant or Carsalesman

Now we need to compare the knowledge discovered from the both the composition shown in **[Table-9]** and **[Table-12]**. For this purpose the comparative study via knowledge discovered from both the composition is shown in **[Table-13]**. .

Table: 13. Conclusions obtained from Fig. 3 and Fig. 4

Candidates	Conclusion from R_1*R_2	Conclusion from R_2*R_1	Comparative knowledge discovered
C_1	Accountant or Waiter	Domestichelper or Waiter	Waiter
C2	Domestichelper or Waiter	Domestichelper or Waiter	Domestichelper or Waiter
C3	Car Salesman	Car Salesman	Car Salesman
C_4	Job cannot be derived	Accountant or Carsalesman	Accountant or Carsalesman
C_5	Accountant	Accountant or Carsalesman	Accountant

From **Table 13** we can derive the following informations:
(A) Candidate C_1 is suitable for Accountant, Domestichelper or Waiter.
(B) Candidate C_2 is suitable for Domestichelper or Waiter.
(C) Candidate C_3 is suitable for Carsalesman.
(D) Candidate C_4 is suitable for Accountant or Carsalesman.
(E) Candidate C5 is suitable for Accountant or Carsalesman.

The above information shown in Table 13 accelerates the company work while choosing the candidates for the given employment. Similarly the proposed method can be applied in in searching advanced query in mobile cloud computing [29-30], fuzzy homomorphism [31], keyword extraction [32], and processing the link data bases (Algal Image Database of India (AIDI) available online at *http://indianalgae.co.in*) [33]. In the next section one real life application of the proposed method is demonstrated.

APPLICATION OF THE PROPOSED METHOD FOR PROJECTING THE CRICKET SCORE

The proposed method can be applied for predicting the score of a cricket or Football games as well. For example: Suppose we want to measure the relationship between condition of pitch and runs scored based on following parameters [13]:
(1). Conditions of Pitches: Good Wicket, Fair Wicket, Sporting Wicket, Green Wicket, Crumbling Wicket, Rough Wicket.
(2). Speed of Bowling: Fast, Medium, Spin, and
(3) Runs Scored: Low Score, Average Score, High Score.

THE IIOAB JOURNAL
ISSN: 0976-3104

The relation between first two parameters i.e. conditions of pitches and speed of bowling has been shown in **[Table-14]**. Similarly the **[Table-15]** shows a fuzzy relation based on condition of pitches and runs scored.

Table: 14. A Fuzzy context based on speed of bowling and condition of pitches (R_1)

	Fast	Medium	Spin
Good Wicket	0.6	0.8	0.7
Fair Wicket	0.5	0.6	0.8
Sporting Wicket	0.4	0.9	0.6
Green Wicket	0.1	0.2	0.7
Crumbling Wicket	0.9	0.1	0.1
Rough Wicket	0.5	0.6	0.2

Table:15. A Fuzzy context based on speed of bowling and run scored (R_2)

	Low Score	Average Score	High Score
Good Wicket	0.4	0.8	0.7
Fair Wicket	0.3	0.8	0.8
Sporting Wicket	0.2	0.7	0.8
Green Wicket	0.8	0.6	0.4
Crumbling Wicket	0.9	0.1	0.1
Rough Wicket	0.5	0.6	0.2

The contexts shown in **[Table-14]** and **[Table-15]** are related via similar attribute set i.e. conditions of pitches. In this case how to analyze these fuzzy contexts and generate the fuzzy formal concept is major concerns. This problem can be solved using the proposed method in this paper using the composition as shown in **[Table-16]**. This composed context (shown in **[Table-16]**) can be projected based on runs scored and speed of bowling as shown in **[Table-17]** and **[Table-18]** respectively.

Table 16. Composition of Fuzzy context shown in Table 14 and Table 15

	Low Score	Average Score	High Score
Fast	0.7	0.6	0.6
Medium	0.6	0.8	0.8
Spin	0.7	0.8	0.8

Table 17. Projection of context shown in Table 16 on speed of bowling

	Low Score	Average Score	High Score
Fast	0.7	0.0	0.0
Medium	0.0	0.8	0.8
Spin	0.0	0.8	0.8

Table 18. Projection of context shown in Table 16 on runs scored

	Low Score	Average Score	High Score
Fast	0.7	0.0	0.0
Medium	0.0	0.8	0.8
Spin	0.7	0.8	0.8

From Table 17 and Table 18 we can find following information:

1. If speed of bowling is fast or spin then runs scored will be low.\
2. If speed of bowling is medium or spin then runs scored is average or high.

THE IIOAB JOURNAL ISSN: 0976-3104

Similarly the proposed method can be applied in several games like football match for prediction of number of goals. In future work will be focused on extending the proposed method to interval-valued formal fuzzy context [7-9] with its application.

CONCLUSION

This paper discussed a method to discover the knowledge from the linked contexts i.e.-(X, Y, R1), (Z, Y, R2) using composition i.e. R1* R2 or R1* R2 with complexity O(n3+n2). It is well known that the composition is not commutative i.e. $. Hence the knowledge discovered from both the composed contexts are compared using their projection on objects and attributes set as shown in **[Table-13]**. The analysis derived from the proposed method is extensive and adequate with [21].

CONFLICT OF INTEREST

Author accepts that there is no conflict of interests.

ACKNOWLEDGEMENT

Author thanks the Dr. Abdullah Gani for his continuous encouragement and suggestions. Author also thanks the anonymous reviewers for their useful suggestions and comments.

.

FINANCIAL DISCLOSURE

Author acknowledges the financial support from Mobile Cloud Computing research project funded by Malaysian Ministry of Higher Education under the University of Malaya High Impact Research Grant with reference UM.C/625/1/HIR/MOE/FCSIT/03.

REFERENCES

[1] Poelmans J, Ignatov DI, Kuznetsov S, Dedene G. [2013] Formal concept analysis in knowledge processing: A survey on applications, *Experts Systems with Applications*, 40(16) : 6538-6560.

[2] Wille R. [1982] Restructuring lattice theory: an approach based on hierarchies of concepts, *Ordered Sets*, NATO Advanced study Institutes 83: 445--470.

[3] Ganter B, Wille R.[1999] *Formal Concept Analysis: Mathematical Foundations*, Springer-Verlag, Berlin.

[4] Ch. Aswani Kumar, Prem Kumar Singh.[2014] Knowledge Representation using Formal Concept Analysis: A Study on Concept Generation, *Global Trends in Knowledge Representation and Computational Intelligence, IGI Global Publishers*, pp. 306-336.

[5] Burusco A, Fuentes--Gonzalez R. [1994] The study of the L-fuzzy concept lattice, *Matheware and Soft Computing* 1(3): 209-218.

[6] Djouadi Y, Prade H. [2010] Interval--Valued Fuzzy Formal Concept Analysis, *Lecture Notes in Computer Science*, 5722 : 592-601

[7] Prem Kumar Singh, Ch. Aswani Kumar, Jinhai Li. [2015] Knowledge representation using interval--valued fuzzy concept lattice, Soft Computing, 20(4):1485-1502,DOI:10.1007/s00500-015-1600-1.

[8] Bloch I. [2011] Lattices of fuzzy sets and bipolar fuzzy sets, and mathematical morphology, *Information Sciences*, 181 (10):2002-2015.

[9] Prem Kumar Singh, Ch. Aswani Kumar. [2014] Bipolar fuzzy graph representation of concept lattice, *Information Sciences*, 288:437-448.

[10] Qi J J, Liu W, Wei L. [2012] Computing the set of concepts through the composition and decomposition of formal contexts, in: *Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian*, pp. 1326-1332.

[11] Alcalde C, Burusco A, Fuentes-Gonzalez R. [2012] Composition of L--fuzzy contexts, in: *Proceedings of 10th International Conference on Formal Concept Analysis* 2012, pp.1-14.

[12] Alcalde C, Burusco A, Fuentes-Gonzalez, R. [2012] Some results on the composition L- fuzzy contexts, *Communications in Computer and Information Science*, 298: 305-314.

[13] Hussain M. [2010] Fuzzy Relations, *Master Thesis*, Blekinge Institute of Technology School of Engineering, Department of Mathematics and Science 2010.

[14] Bartl E, Belohlavek R, Vychodil V. [2008] Composition of fuzzy relation with hedges, in: *Proceedings of IEEE World Congress on Computational Intelligence* 2008, pp. 1100-1105.

[15] Babin M A, Kuznetsov SO. [2011] On links between concept lattices and related complexity problems, in: *Proceedings of ICFCA 2012, LNAI Springer*, 5986: 138-144.

[16] Ragab MG, Emam EG. [1995] On the min--max composition of fuzzy matrices, *Fuzzy Sets and Systems*, 75:83-92.

[17] Prem Kumar Singh, Ch. Aswani Kumar. [2012] A method for reduction of fuzzy relation in a fuzzy formal context, *Communications in Computer and Information Science*, 283:343-350.

[18] Prem Kumar Singh, Abdullah Gani, [2015] Fuzzy concept lattice reduction using Shannon entropy and Huffman coding, *Journal of Applied Non-Classical Logics*, 25(2): 101-119, DOI :10.1080/11663081.2015.1039857

[19] Ch. Aswani Kumar. [2011] Knowledge discovery in data using formal concept analysis and random projections, *International Journal of Applied Mathematics and Computer Science*, 21(4): 745-756.

[20] Yuan M, Li W, Zhangang L. [2009] Projection granular space in formal concept, in: *Proceedings of World Congress on Software Engineering* 2009, IEEE, pp.94-98.
[21] Prem Kumar Singh, Ch. Aswani Kumar. Analysis of composed context through projection, *International Journal of Data Analysis Techniques and Strategies*, In 2016 Press.
[22] Klir J, Yuan B. [2008] *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall PTR, New Jersey.
[23] Belohlavek R.[2004] Concept lattices and order in fuzzy logic, *Ann. Pure Appl. Logic*, 128:277-298.
[24] Belohlavek R, Vychodil V.[2005] What is fuzzy concept lattice, in: *Proceedings of 3rd International Conference on Concept Lattices and Their Applications* 2005, pp.34-45.
[25] Pocs J. [2012] Note on generating fuzzy concept lattices via Galois connections, *Information Sciences* 185:128-136.
[26] Krupka M, Lastovicka J. [2012] Fuzzy Concept lattice with incomplete knowledge, *Communications in Computer and Information Science*, 299:171-180.
[27] Li J, Mei C, Lv Y.[2013] Incomplete decision contexts: approximate concept construction rule acquisition and knowledge reduction, *International Journal of Approximate Reasoning*, 54(1): 191-207.
[28] Medina J. [2012] Multi-adjoint property--oriented and object--oriented concept lattices, *Information Sciences*, 190:95-106.
[29] Sarnovsky, M., Butka, P., Pocsova, J., Cloud Computing as a Platform for Distributed Fuzzy FCA Approach in Data Analysis. In: *Proceedings of IEEE 16th International Conference on Intelligent Engineering Systems INES* 2012, Lisbon, Portugal, pp. 291-296.
[30] Todoran, I., Glinz, M. [2014] Quest for Requirements: Scrutinizing advanced search queries for Cloud services with fuzzy Galois lattices. In: *Proceedings of International Conference on IEEE 10th World Congress on Services* 2014, pp. 234-241.
[31] Prem Kumar Singh and Ch. Aswani Kumar, A note on constructing fuzzy homomorphism map for the given fuzzy formal context. *Advances in Intelligent Systems and Computing*, 258: 845-855.
[32] S Khan, QM Ilyas, W Anwar.[2009]. Contextual advertising using keyword extraction through collocation. In: Proceedings of seventh International Conference on Frontiers of Information Technology, Article 09, doi: 10.1145/1838002.1838081
[33] Lalit Kumar Pandey, Krishna Kr Ojha, Prem Kumar Singh, Chandra Shekhar Singh, Shubham Dwivedi, EA Bergey [2016] Diatoms image database of India (DIDI): a research tool, *Environmental Technology & Innovation*, Elsevier, 5:148-160. doi:10.1016/j.eti.2016.02.001

ABOUT AUTHOR

Dr. Prem Kumar Singh *is a Post-Doctoral Fellow at Faculty of Computer Science and Information Technology, University of Malaya-Kuala Lumpur. He holds a PhD degree in Computer Science from VIT University-Vellore, India. His current research interests are Graph Theory, Set theory, Formal Concept Analysis, and Machine Intelligence. He has published more than 20 refereed research papers so far in various national, international journals and conferences. He is also reviewer of several peer reviewed journals. His research output can be found at Google Scholar: https://scholar.google.co.in/citations?user=FFmAj_MAAAAJ*

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Hudedagaddi and Tripathy

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

APPLICATION OF SPATIAL IFCM IN LEUKAEMIA CELLS

Deepthi P Hudedagaddi* and Balakrushna Tripathy

SCOPE, VIT University, Vellore, Tamil Nadu-632014, INDIA

ABSTRACT

Managing nature of images induced by noise with the fusion of fuzzy algorithms has been a challenge. Traditional Fuzzy C Means (FCM) calculation is amazingly noise sensitive and fails in giving great results. Subsequently, as an answer for this issue Tripathy et. al presented a change of fuzzy c means which includes calculation that consolidated spatial data and intuition. The spatial capacity is the summation of the considerable number of estimations of the participation elements of the pixel's neighborhood which is under consideration. This methodology ended up being a right answer for the noise affectability issues when contrasted with FCM. This calculation incorporates an intuitionistic component in the enrollment capacity of the current spatial FCM (sFCM). Intuitionism manages the wavering segment that emerges because of less data and insufficient learning. This is superior to anything existing calculations. This spatial IFCM(sIFCM) has been connected on leukaemia images.

Received on: 09th-Sep-2015
Revised on: 27th-Sep-2015
Accepted on: 25th– March-2016
Published on: 16th–May-2016

KEY WORDS

Fuzzy C Means; Intuitionistic Fuzzy set; Clustering; Spatial Fuzzy C Means, Intuitionistic Fuzzy C Means

***Corresponding author: Email:** deepthiph@gmail.com **Tel:** +91-9986387435

INTRODUCTION

Analysis of images is a noteworthy part for diagnosis and treatment of diseases, research studies and more. At present, the calculation is assuming a vital part because of the expanding size and number of medicinal images. The choice of strategies fundamentally relies on imaging modalities, its particular application and different variables. For example, the brain tissue has distinctive necessities from other organs. Restorative image segmentation computerizes the particular radiological capacity. The objective of segmentation is to investigate the representation of an image into important and simpler divisions. It alludes to post a computerized image into various sections which are fundamentally developed with sets of pixels. Every pixel in the locale of interest comprises of some essential qualities and processed property, known as power, surface and shading. Yet, no single division strategy yields satisfactory results for each therapeutic image.

In image understanding and vision based machine intelligence, image segmentation plays an important role by dividing an image into multiple homogeneous segments that are more suitable to further analysis. Among different image segmentation methods, the most popular and intensively explored ones are the approaches using data clustering algorithms. The targets of maximizing the similarities of pixels in each segment and minimizing the similarities of pixels from different segments in image segmentation problems also constitute exactly the requirement of traditional data clustering if we regard the pixels as the data points to be clustered. However, the visual similarities between pixels in an image and their calculation need to be well designed. Partitioning pixels into spatial homogeneous regions should be considered in the clustering algorithms.

Spatial data mining (SDM), or learning revelation in spatial database, alludes to the extraction of understood information, spatial relations, or different examples not expressly put away in spatial databases. SDM comprises of removing information, spatial connections and some other properties which are not unequivocally put away in the database. SDM is utilized to discover verifiable regularities, relations between spatial information and/or non-spatial information. We can consequently see the immense significance of spatial connections in the examination process. Worldly perspectives for spatial information are likewise a main issue, however they are infrequently considered. Information mining strategies are not suited to spatial information since they do not bolster area information nor the certain connections between items. Thus, it is important to grow new strategies including

THE IIOAB JOURNAL
ISSN: 0976-3104

spatial connections and spatial information taking care of. Computing these spatial connections is tedious, and an immense volume of information is created by encoding geometric area. Spatial information mining incorporates different errands and, for every undertaking, various distinctive techniques are frequently accessible, whether computational, factual, visual, or some blend of them. Conventional FCM algorithm fails to provide appropriate results on images in the presence of noise. Spatial FCM (sFCM) and spatial IFCM(sIFCM) [1,2], a bi-step process, incorporates spatial information of the pixel in consideration. Though it fails in completely removing the distortion of noise, the algorithm proves to be less sensitive to noise as compared to traditional FCM. However,it does not eliminate the distortion caused by noise completely. In addition to the uncertain based clustering methods, hybrid clustering algorithms have been developed [3-7].

SPATIAL CLUSTERING

Cluster investigation is generally utilized for information examination, which arranges an arrangement of information things into gatherings (or groups) so that things in the same gathering are like each other and not quite the same as those in other groups. Many diverse grouping techniques have been produced in different exploration fields, for example, insights, design acknowledgment, information mining, machine learning, and spatial analysis [5].

Clustering is a noteworthy unsupervised learning procedure. Fuzzy C-Means clustering is an understood delicate division technique and it suitable for therapeutic image segmentation. In any case, this customary calculation is computed by iteratively minimizing the separation between the pixels and to the bunch focuses. Spatial relationship of neighboring pixel is a guide for segmentation of images. These neighboring pixels are exceedingly related the same element information. In spatial space the enrollment of the neighbor focused are indicated to get the group dissemination insights. In view of this insights to figure the weighting work and connected into the enrollment function [6].

Spatial clustering techniques can be apportioning or hierarchical, density-based, or framework based. Regionalization is a remarkable kind of spatial clustering that tries to total spatial things into spatially contiguous gatherings (i.e., areas) while redesigning an objective limit. Various geographic applications, for instance, climate zoning, scene examination, remote recognizing picture division, frequently require that groups are geographically circumscribing. Existing regionalization frameworks that rely on a clustering thought can be planned into three events: (1) multivariate (non-spatial) gathering took after by spatial planning to amend clusters into regions (2) gathering with a spatially weighted uniqueness measure, which considers spatial properties as a variable in confining bunches and (3) contiguity constrained grouping that approves spatial contiguity in the midst of the gathering strategy[8].

EXISTING METHODS

Fuzzy models are incorporated in analyzing spatial data.

Fuzzy Clustering

Fuzzy c-mean in view of fuzzy sets [9] is an algorithm proposed by James C. Bezdek [10,11]. In fuzzy clustering (additionally alluded to as soft clustering), information components can fit in with more than one cluster, and connected with every component is an arrangement of enrollment levels. These demonstrate the quality of the relationship between that information component and a specific cluster. Fuzzy clustering is a procedure of relegating these enrollment levels, and after that utilizing them to allocate information components to one or more clusters [12,13].

1. Assign initial cluster centers or means for c clusters.
2. Calculate the distance d_{ik} between data objects x_k and centroids v_i using Euclidean formula

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

3. Generate the fuzzy partition matrix or membership matrix U:

If $d_{ij} > 0$ then

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \quad (2)$$

Else

THE IIOAB JOURNAL
ISSN: 0976-3104

$\mu_{ik} = 1$

4. The cluster centroids are calculated using the formula

$$V_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m x_j}{\sum_{j=1}^{N} (\mu_{ij})^m} \quad (3)$$

5. Calculate new partition matrix by using step 2 and 3
6. If $\|U^{(r)} - U^{(r+1)}\| < \varepsilon$ then stop else repeat from step 4.

Usually, for all experimental purpose, m is considered to be 2 and ε to be 0.02.

Spatial Fuzzy C Means

Chuang. et al. [1] clarified that a traditional FCM algorithm does not completely use the spatial data in the image. They altered the present fuzzy c-means (FCM) calculation and created FCM that joins spatial data into the participation capacity for clustering. The spatial function is given by the summation of the membership values in the neighboring of every pixel under consideration. The benefits of the new strategy are: (1) it yields more homogeneous clusters than those of other strategies, (2) it diminishes the spurious blobs, (3) it uproots boisterous spots, and (4) it is less sensitive to noise than other systems. This procedure is an intense strategy for image segmentation and works for both single and several information on spatial data. On comparable lines, spatial IFCM was additionally developed by Tripathy et.al by presenting the intuitionsitic feature.

It implies when two close pixels are considered, relation between them is generally high. Since the neighboring pixels offer comparative force, the likelihood of them gathering into a same group is comparatively high. The spatial FCM calculation exploits this criteria. A spatial function is characterized as

$$h_{ij} = \sum_{k \in NB(x_j)} u_{ik} \quad (4)$$

where $NB(x_j)$ refers to the neighborhood pixels of x_j. A 5x5 equally weighted mask centered on pixel x_j has been used for this purpose. The spatial function h_{ij} represents the degree of likeliness that x_j is in the i^{th} cluster. The value of the spatial function for a pixel is high for a particular cluster if most of the neighborhood pixels belong to the same cluster. It is included in the membership function as:

$$u'_{ij} = \frac{u_{ij}^{p} \cdot h_{ij}^{q}}{\sum_{c=1}^{m} u_{ij}^{p} \cdot h_{ij}^{q}}]) \quad (5)$$

here p and q indicate the relative weightage of the initial membership and the spatial function respectively. In case of a noisy image, the spatial function reduces the number of misclassified pixels by taking the neighboring pixels into account.

The sFCM clustering algorithm has two-steps. For every iteration, the conventional FCM algorithm is followed in the first step. Here the distance formula being used is radial-based kernel distance. Later, the centroid and the membership functions are updated. In the second step, the spatial function h_{ij} is calculated and then the new membership function (5) is computed.

Intuitionsitic FCM

The Intuitionistic fuzzy c-means proposed by T. Chaira [14,15] brings in to account a new parameter that helps in increasing the accuracy of clustering. This parameter is known as the hesitation value.

1. Assign initial cluster centres or means for c clusters.

2. Calculate the distance d_{ik} between data objects x_k and centroids v_i using Euclidean formula(1).

3. Generate the fuzzy partition matrix or membership matrix U:

If $d_{ij}>0$ then compute μ_{ik}

Else

$\mu_{ik}=1$

1. Compute the hesitation matrix π

THE IIOAB JOURNAL
ISSN: 0976-3104

2. Compute the modified membership matrix U' using
   $\mu'_{ik} = \mu_{ik} + \pi_{ik}$

3. The cluster centroids are calculated using the formula
   $V_i = \frac{\sum_{j=1}^{N} (\mu'_{ij})^m x_j}{\sum_{j=1}^{N} (\mu'_{ij})^m}$

4. Calculate new partition matrix by using step 2 to 5

If $\|U^{(r)} - U^{(r+1)}\| < \varepsilon$ then stop else repeat from step 4. ε is considered to be 0.02.

.

Spatial IFCM

The spatial Intuitionistic Fuzzy C means (sIFCM) algorithm developed by Tripathy et.al[2] is an extended version of IFCM where the membership function incorporates the spatial function.

The algorithm of spatial IFCM(sIFCM) is as follows.

1. Provide the initial values for the centroids v_i where i =1,…,c

2. Compute the membership function as follows:
   $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}}}$

   for all i = 1,…,c and j = 1,…,N

3. Compute the hesitation value as:
   $\pi_{ij}(x) = 1 - u_{ij}(x) - (1 - u_{ij}(x))/(1 + \lambda u_{ij}(x))$
   for all i = 1,…c , λ>0 and j = 1,…,N

4. Compute the membership function as:
   $\mu'_{ik} = \mu_{ik} + \pi_{ik}$
   for all i = 1,…c and j = 1,…,N

5. Calculate the spatial function as
   $h_{ij} = \sum_{k \in NB(x_j)} u_{ik}$
   for all i = 1,…c and j = 1,…,N

6. Compute the new membership function which incorporates the spatial function as:
   $u'_{ij} = \frac{u_{ij}^p h_{ij}^q}{\sum_{k=1}^{c} u_{kj}^p h_{kj}^q}$

7. Set $u_{ij} = u_{ij}$'' for all j = 1,…,N and i = 1,…c

8. Calculate the new centroids as follows:

$v_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}$

   for i=1,..,c

9. If $|u_{ij}(new) - u_{ij}(old)| < \in$ then stop, otherwise go to step 2.

THE IIOAB JOURNAL
ISSN: 0976-3104

Usually m=2 and ∈ =0.02.

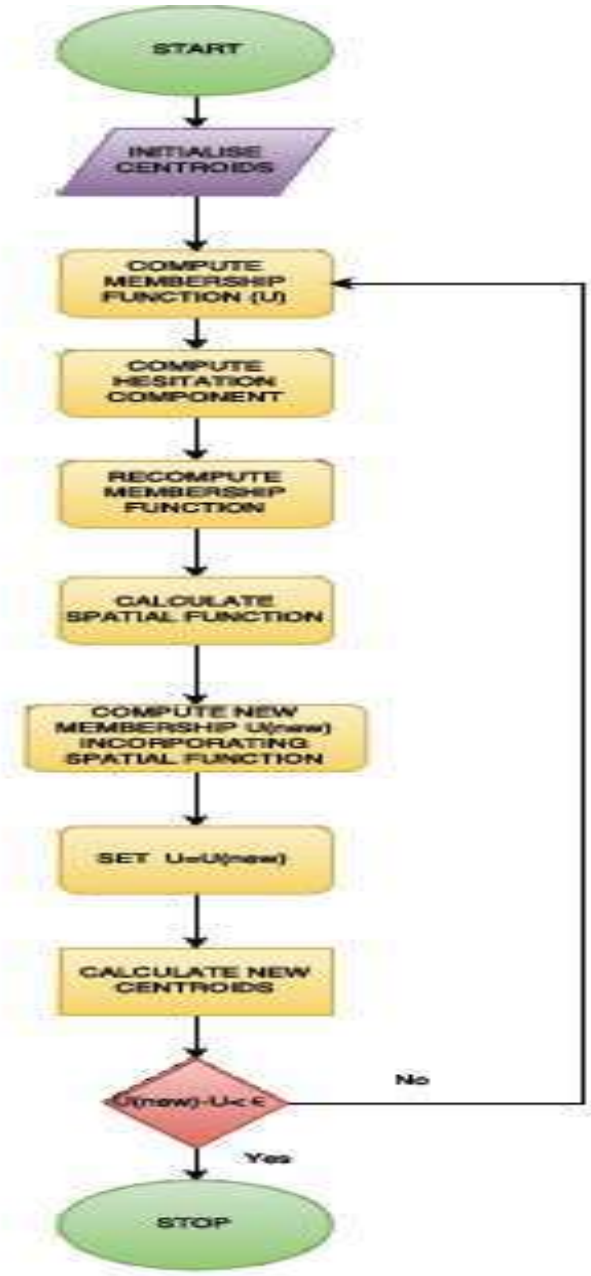


Fig: 1. Spatial IFCM

sIFCM ON LEUKAEMIA IMAGE

It is a well established fact that early detection of cancer cells and estimating their rate of growth plays a crucial role in detection and treatment of cancer/leukaemia. In this paper, we have considered leukaemia cells for the study. Two types of cluster validity functions, fuzzy partition and feature structure, are often used to evaluate the performance of clustering in different clustering methods. The representative functions for the fuzzy partition are partition coefficient V_{pc}[2, 9] and partition entropy V_{pe}[10].

THE IIOAB JOURNAL
ISSN: 0976-3104

The idea of these validity functions is that the partition with less fuzziness means better performance. Hence, the best clustering is achieved with maximum value of V_{pc} and minimum value of V_{pe}. Disadvantages of V_{pc} and V_{pe} are that they measure only the fuzzy partition and lack a direct connection to the featuring property. Other validity functions based on the feature structure are available.

$$V_{pc} = \frac{\sum_j^N \sum_i^c u_{ij}^2}{N} \qquad (6)$$

and

$$V_{pe} = -\frac{\sum_j^N \sum_i^c [u_{ij} \log u_{ij}]}{N} \qquad (7)$$

Clustering is said to be efficient if sample in a cluster are extremely. A good clustering result generates samples that are compacted within one cluster and samples that are separated between different clusters. Minimizing V_{xb} is expected to lead to a good clustering.

$$V_{xb} = \frac{-\sum_j^N \sum_i^c u_{ij} \forall x_j - v_i \forall^2}{N * \left(min_{i \neq k} \left\{ \left| v_k - v_i \right|^2 \right\} \right)} \quad (8)$$

DB and D indices indicate the proximity of clusters within and in between. Hence, larger D value and lower DB value indicates a good clustering[16].

As a follow on with the development of spatial fuzzy c means algorithm, Deepthi et. al[17] applied the developed sFCM on leukaemia images to demonstrate it's working. This clustering of leukaemia cells helps in providing information regarding the growth of cancerous cells. This further helps in diagnosis.

Table: 1. Performance indices of sIFCM on leukaemia image

Index	FCM	IFCM	$sIFCM_{1,1}$	$sIFCM_{1,2}$	$sIFCM_{2,1}$
V_{pc}	0.2118	0.2074	0.4762	0.2241	0.2281
V_{pe}	0.0178	0.02585	3.56E-005	1.68E-005	5.90E-006
V_{xb}	0.0168	0.02096	0.05929	0.0069	0.0063
DB	0.467	0.4906	0.4126	0.4208	0.4188
D	2.2254	2.0529	3.5866	3.0203	2.9655

The results in **Table 1** show that the sIFCM succeeds in providing better results than conventional FCM. It is distinctly visible through higher D value and lower DB values, different combinations of p and q values in sIFCM overpower the conventional FCM and IFCM. The partition entropies are also proved to be lower. However, the partition coefficient values are higher in sIFCM. As per the cluster validity measures, V_{xb} should be minimum to prove it the algorithm is providing better clustering. In this measure also, the application of sIFCM is better and hence proves the clusters provided are efficient.

THE IIOAB JOURNAL
ISSN: 0976-3104

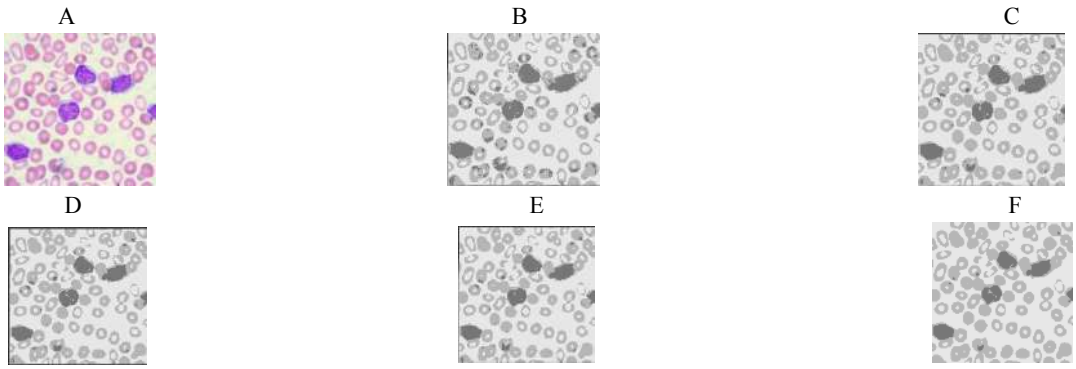


Fig: 2. (a) Original image, Segmented images of leukaemia using (b)FCM; (c)IFCM (d)sIFCM$_{2,1}$; (e)sIFCM$_{1,1}$; (f)sIFCM$_{1,2}$

The segmented images with the application of sIFCM provide better clarity and understanding of presence of leukemia cells than that of conventional FCM and IFCM. As an extension, the sIFCM as shown above, is applied to leukaemia image. The leukaemia cells are clustered based on sIFCM and the images are shown. On trial and error method, different values were provided for p and q values in spatial membership function. The results of three combinations are given. However, the appropriate result depends on the image and the application.

CONCLUSION AND FUTURE WORKS

The drawbacks in one clustering algorithm paves a way for developing novel algorithms or modifications for the same algorithm. However, it can be seen that if developing algorithm for one application is one arena, finding the applications of same algorithm in various other domains is another arena. This paper has made an attempt to apply the spatial IFCM algorithm which was developed incorporating the drawbacks of IFCM to a cancer cell images. In this manner, algorithms developed should also be reused for several other applications in different domains. The results obtained are also better than conventional algorithms. This application would help in diagnosis or estimating the growth and effect of these cancerous cells. Also, spatial data clustering has provided a wide platform for researchers. The developed algorithms are to be exploited completely and various applications in different domains are to be found.

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] KS Chuang, HLTzeng, S Cheren, Jay Wu,Tzong-Jer Chen.[2006] Fuzzy c-means clustering with spatial information for image segmentation *Computerized Medical Imaging and Graphics*, vol. 30,

[2] Tripathy BK, AvikBasu,and SahilGovel.[2014] Image segmentation using spatial intuitionistic fuzzy C means clustering. Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. *IEEE*

[3] B.K.Tripathy and A.Ghosh.[2013] Data Clustering Algorithmsusing Rough sets, Handbook of research on Computational Intelligence for engineering, Science and Business, *IGI Global Publications*, pp.297-327.

[4] BK.Tripathy, A Ghosh, GK Panda.[2012] Kernel Based K-Means Clustering Using Rough Set, Proceedings of 2012 International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, (2012), Coimbatore, INDIA, pp.1 -5.

THE IIOAB JOURNAL
ISSN: 0976-3104

[5] BK Tripathy, A Tripathy, K Govindarajulu, R Bhargav.[2014] On Kernel Based Rough Intuitionistic Fuzzy C-means Algorithm and a Comparative Analysis. In Advanced Computing, Networking and Informatics, *Springer International Publishing*. 1: 349-359.

[6] BK Tripathy and P Swarnalatha.[2014] A Comparative Study of RIFCM with Other Related Algorithms from Their Suitability in Analysis of Satellite Images using Other Supporting Techniques, Kybernetes, .43(1): 53-81.

[7] B.K.Tripathy and R. Bhargav: Kernel Based Rough-Fuzzy C-Means, PReMI, ISI Calcutta, December, LNCS 8251, (2013), pp.148-157.

[8] DianshengGuo, Jeremy Mennis.[2009] Spatial data mining and geographic knowledge discovery—An introduction", Computers, Environment and Urban Systemspp.403-408

[9] LA Zadeh,[196] *Fuzzy Sets, Information and Control*, .8(11): 338 – 353.

[10] Bezdek JC. Cluster validity with fuzzy sets. *J Cybern* 1974;3:58–73.

[11] Bezdek JC.[1975] Mathematical models for systematic and taxonomy. In:proceedings of eigth international conference on numerical taxonomy,San Francisco;, p. 143–66.

[12] P Swarnalatha, BK Tripathy, PL Nithin and D Ghosh.[2014] Cluster Analysis Using Hybrid Soft Computing Techniques, CNC- 2014 International Conference of Network and Power Engineering ,Proceedings of Fifth CNC-2014,pp. 516-524.

[13] R Bhargav, BK Tripathy, A Tripathy, R Dhull, E Verma, and P Swarnalatha. [2013]Rough Intuitionistic Fuzzy C-Means Algorithm and a Comparative Analysis, *ACM conference, Compute* 2013, ACM 978-1-4503-2545-5/13/08.

[14] Chaira, T. (2011). A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images. Applied Soft Computing, 11(2), 1711-1717.

[15] KT Atanassov. [1986] Intuitionistic Fuzzy Sets, Fuzzy sets and Systems, 20(1):87-96.

[16] JC Dunn. [1973]A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, pp. 32-57.

[17] Deepthi P Hudedagaddi,B.K. Tripathy, Application of Spatial FCM in Cancer Detection Cells, Proceedings of NCMCS-2015,Upendranath College of Engineering,India.

ABOUT AUTHORS

Deepthi P Hudedagaddi is pursuing her Masters at Vellore Institue of Technology. She is working fuzzy clustering techniques on spatial data.

Dr. Tripathy, a triple gold medalist, is a senior professor in VIT University. He has supervised 19 PhD s, 13 M. Phil s and 02 M.S degrees. He is a senior member of IEEE, ACM, ACEEE and CSI. and is associated with over 60 international journals,published 320 articles, two research volumes and two books.He is working on Rough sets, Fuzzy sets, Social networks, Data mining, Soft Computing, E-Learning, Granular computing, Multi criteria decision making, Neighbourhood systems, SIoT and Soft Sets.

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Gopinath and Prabu

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

CLASSIFICATION OF THYROID ABNORMALITIES ON THERMAL IMAGE: A STUDY AND APPROACH

M. P. Gopinath* and S. Prabu

School of Computer Science and Engineering, VIT University Vellore, Tamil Nadu, INDIA

ABSTRACT

Thermal distribution in human body is a natural indicator of abnormalities. Thermal imaging is a non – invasive screening method for monitoring the distribution of body temperature. The objective of this paper is to compare different classification techniques which suites for classification of Thyroid Data set from UCI Machine Learning Repository. In the view of study we propose classification of Thyroid Abnormalities using thermal image. The proposed technique is based on the following computational methods; the median filter for preprocessing, Otsu's technique for segmentation, Gabor filter and Gray-Level Co-Occurrence Matrix is used to extract the feature from the selected Region of Interest and classifier to classify inputs into normal or abnormal using Decision tree. The experiment is carried out with 51 Thermal image which consist of 30 abnormal (hyper and hypo) and 21 normal from a real human neck. The classification accuracy was significantly good and finally possible future directions are suggested.

Received on: 09th-Sep-2015
Revised on: 27th-Sep-2015
Accepted on: 28th– April-2016
Published on: 16th –May-2016

KEY WORDS

Thermal imaging; Feature extraction; Feature selection; Segmentation; classification

***Corresponding author:** Email: mpgopinath@vit.ac.in **Tel:** +91-9840697916

INTRODUCTION

Medical Digital Infrared Thermal Imaging (DITI) is a new non – destructive adjunctive diagnostic technique that allows engineering projects to examine to visualize and measure the skin temperature by scanning the skin surface[1], scanning device converts the heat emitted by the skin and convert to signals that are captured as image. This color image maps the body temperature known as thermo gram. Skin tissue temperature is extracted by Penne's bio heat equation which is general heat diffusion equations [2]:

$$\rho C \frac{\partial T}{\partial t} = \nabla \left(k \nabla T \right) + W_b C_b \left(T_a - T \right) + q_m \qquad [1]$$

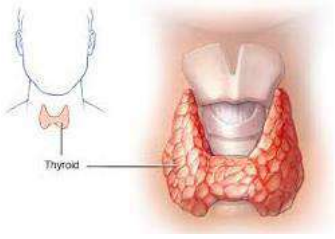


Fig 1a. Normal Thyroid Gland location

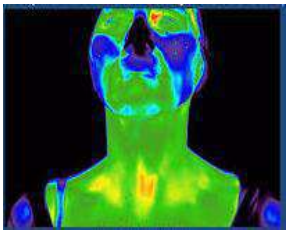


Fig 1b. Thermal image of Neck shows even distribution heat

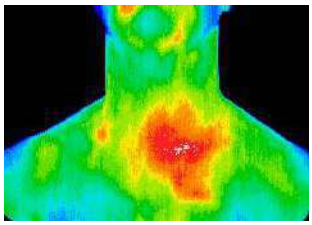


Fig 1c. Thermal Image of Neck shows uneven distribution of heat

ISSN: 0976-3104

Thermal image system forms an image using infrared radiation and operates in wavelength as long as 14,000nm (14µm). The principal of temperature rises is the amount of radiation emitted by an object also increases. Every object has its own absorption and emission of thermal energy when any eccentricity happen it shows the stress of that object. Inflammation in human tissue shows the difference in distribution of heat around the body as any rapid changes happens automatically blood flow will increase in that particular place shows the abnormality of that part which can be identified by Penne's bio heat equation.

Thermal imaging or Thermography is the mapping of temperature distribution on the surface of the object or component. This technique is based on Infrared radiations. **Figure -1b** shows the normal distribution of heat ranges under control, **Figure -1c** shows the abnormal emission of heat in specified area which can be classified as thyroid abnormalities depends on various other factor Any object > 0 k emits electromagnetic radiations. At ambient temperatures and above are predominately within the infrared band of electromagnetic spectrum. Using appropriate detectors these radiations can be converted to suitable electrical signals and displayed on the monitor. Infrared cameras are manufactured predominately in two wavelength ranges 3-6 µm and 7-14 µm for medical application Sensitivity of FLIR camera vary with various versions of camera or models. Thermal sensitivity is smallest temperature difference that can be detected by the camera. Eg. For an infrared camera with uncooled micro bolometer detector the temperature sensitivity may be 0.01 ^{0}C @ 30 ^{0}C whereas for an infrared camera with semiconductor type of detector the temperature sensitivity may be 25mK @ 30 ^{0}C. Thermal image this captured from heat emitted from human body. Thyroid gland is responsible for all metabolic activity which increase blood flow and hormonal change which can be sensed by thermal camera. FLIR Thermal cameras with sensitivity of 0.01c with temperature range from -200c to +1200c are used for capturing thermal image [3].

Thyroid disorder is due to variation in TSH chemicals (hormones) that help the body to control metabolism. Thyroid hormone is normally produced in response to another hormone released by the pituitary gland. There are four main types of thyroid disease hyperthyroidism (too much thyroid hormone), hypothyroidism (too little thyroid hormone), benign (noncancerous). A thyroid disorders is an abnormal growth of cells within the thyroid gland or inside the throat, which can be cancerous (malign). It is defined as any intracranial tumor created by abnormal and uncontrolled cell division. Benign type of thyroid disorders can lead to Hyperthyroidism, Hypothyroidism, goiter, and thyroid nodules (benign/malignant). Detection of thyroid disorders in the earliest stage is the key for its successful treatment. The mean temperature with standard deviation of neck shows the thyroid abnormalities like hyperthyroid when temperature distribution of 36.63 ± 0.56 °C, hypothyroid when temperature distribution of 34.93 ± 0.32 °C, for normal 35.76 ± 0.49 °C [4,5]

The paper is organized as follows section 2presents various steps involved in thermal image diagnosis system with implementation of algorithm step by step specified in **Figure 2**, Section 3 presents different types of classification algorithm using UCI Thyroid data set repository, Section 4 Analysis the result of classification algorithm under specificity, sensitivity and accuracy, Section 5 presents the implementation and result discussion of proposed methods and shows the classified thyroid abnormalities in table6, Section 6 presents the conclusion and future work.

MEDICAL INFRARED THERMAL IMAGE DIAGNOSIS SYSTEM

Figure 2 shows the steps followed in proposed system. The development of thermal image diagnosis system helps in screening, prognosis and surgical procedures. Median filter is used for preprocessing Thermal image that taken by FLIR camera, enhancement using histogram equalization. Region of Interest (ROI) extracted manually Otsu Thresholding technique. Features were extracted and classified using Decision tree.

IIOAB JOURNAL
ISSN: 0976-3104

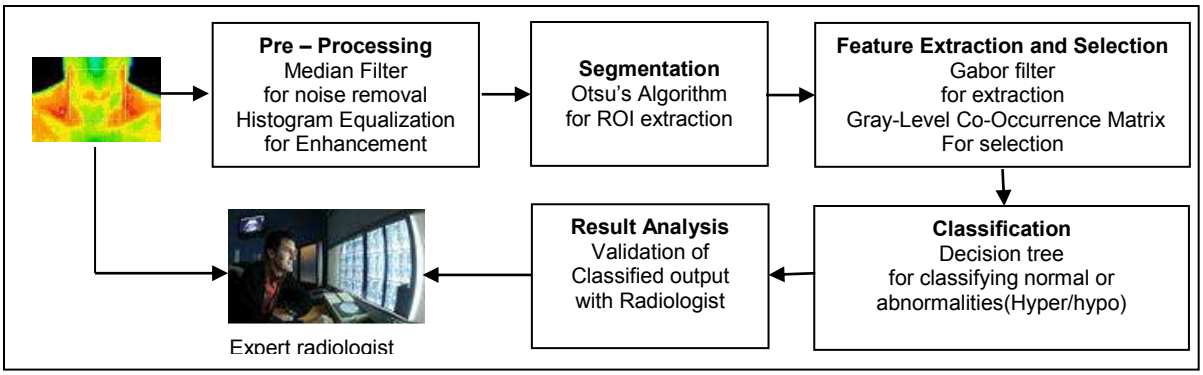


Fig: 2. Typical methodology of Thermal Image diagnosis system

Preprocessing

Filters is a method for modifying the image and remove noise from images which are classified as Linear and Non – Linear Filters. Linear filter will not preserve edges and losses fine minute details. Non – Linear filter preserve the information without any significant loss [6]. Median filters are one of the most popular nonlinear filters for removing salt and pepper noise through the image pixel by pixel, replacing the each value by the median value of center neighborhood shown in **Figure- 3**.

$$V(m,n) = median\left(y(m-k,n-l),(k,l)W\right) \qquad [2]$$

Where W chosen window

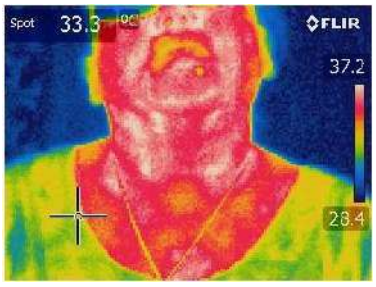


Fig 3a: Thermal image

Fig 3b: Preprocesed image

Image Enhancement

Image enhancement is used for sharpening, of image features such as edges or contrast to make an analysis. Histogram modeling is continuous process function[7]. Therefore the image intersection contains continuous intensity levels and the transformation function *f* which maps an input image *A(x,y)* onto output image *B(x,y)* is continuous within interval (0,1). The output probability densities of image enhancement is equal fraction of maximum number of intensity levels in the input image D_m. the transfer function necessary to achieve this result is simple. Result is shown in **Fig5**.

$$d\ (D_A)\ =\ D_M * P_A(D_A) \qquad [3]$$

Therefore

$$f(D_A) = DM\int_0^{DA} pA(U)du = DD_m * F_A(D_A) \qquad [4]$$

IIOAB JOURNAL
ISSN: 0976-3104

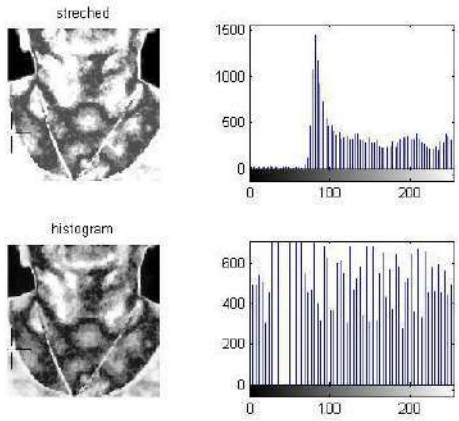


Fig :4a. Histogram Equalization (Streached)

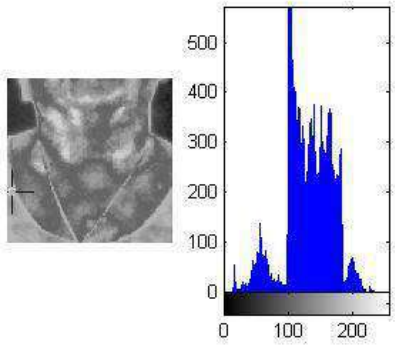


Fig: 4b. Histogram Equalization

Image Segmentation

Region of Interest is extracted manually from the region from binary image through Otsu Thresholding method[8] which is used to extract an object from its background by assigning an intensity value T through pixel by pixel is classified as an object point when ROI is segmented and shown in **Figure-7**.

Interclass variance is defined as

$$\sigma^2_{within}(T) = \omega B(T)\sigma^2_B(T)\sigma^2_0(T) \quad [5]$$

Where

$$\omega B(T) = \sum_{i=0}^{L-1} P(i) \quad [6]$$

(0,L-1] the range of intensity levels

$$\omega 0(T) = \sum_{i=0}^{L-1} P(i) \quad \omega 0(T) = \sum_{i=0}^{L-1} P(i) \quad [7]$$

$\sigma^2_B(T) = the\ varience\ of\ \ pixel\ in\ background\ below\ threshold$

$\sigma^2_0(T) = the\ varience\ of\ \ pixel\ in\ foreground\ above\ \ threshold$

$$\sigma^2_{Between}(T) = \sigma^2 - \sigma^2_{within}(T) = n_B(T)\left[\mu_B(T) - \mu\right]^2 + \left[\mu_0(T) - \mu\right]^2 \quad [8]$$

μ^2 $-is\ the\ combined\ varience$

$\mu - is\ the\ combined\ mean\ of\ the\ pixel\ Xi$

$$\sigma^2_{Between}(T) = n_B(T)n_0(T)\left[\mu_B(T) - \mu_0(T) - \mu\right]^2 \quad [9]$$

The standard deviation for any region with N pixels of intensity i = 1, 2, 3, ….N is given by

$$s = \sqrt{\sum_{i=1}^{N}(x_i - x)^2 / N}$$

Where x is the mean of x_i by substituting this we get

$$\sigma^2_{Between}(T) = n_B(T)n_0(T)\left[S_B(T) - S_0(T)\right]^2 \quad [10]$$

ISSN: 0976-3104

S_B *is the standard deivation for background pixels*

S_0 *is the standard deivation for object pixels*

Using the threshold value T, the given input image *f(x, y)* is transformed to binary image *g(x, y)* as:

$$g(x,y) = \begin{cases} i\,if\ f(x,y) \geq T \\ 0\,otherwise \end{cases}$$

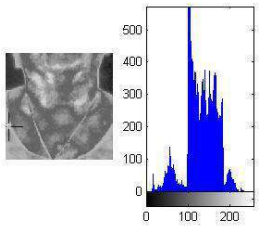


Fig: 5a. Input image for Segmantation **Fig: 5b. Segmented Image** **Fig: 5c.Thresholded Image**

Feature Extraction and Selection

This filter extract distinguishable texture features from the Gabor filtered image or global information image [9, 10]. It usually comprises of a complex exponential centered at a given frequency modified by a Gaussian envelop. This filter has real and imaginary parts because of the complex exponential. The 2D Gabor function provides the local spectral energy density, which particularly concentrates around the given position and frequency.

$$G(x,y,k_x,k_y) = \exp\left\{\frac{-(x-X)^2+(y-Y)^2}{2\sigma^2}\right\} e^{j(k_x^X + k_y^Y)} \qquad [11]$$

Where,
x, y - spatial coordinates
K_x, k_y - frequency coordinates
X, Y - spatial localizations of the Gaussian window.
It is self-similar and can be generated from wavelet by dilation and rotation. The pixel intensity of the global feature image is used to calculate feature vectors. The Gabor filters capture the spatial dependence with less ability and has low accuracy rate.

Gray-Level Co-Occurrence Matrix (GLCM) is a statistical method that includes the spatial relationship of pixels also known as gray-level spatial dependence matrix. The spatial relationship can be defined as the pixel of interest and the pixel to its direct right. The second order statistical texture features are extracted using GLCM. A GLCM is a matrix where the number of rows and columns are equal to the number of gray levels in the image. Based on the GLCM, parameters namely energy, contrast, entropy and correlation are computed. The texture feature extraction using GLCM involves following steps

- The R, G, B planes of images are separated.
- The GLCM matrices are calculated for each plane.

ISSN: 0976-3104

Parameters of GLCM

Features	Formula
Energy	$\sum_i \sum_j P^2(i,j)$
Entropy	$\sum_i \sum_j P(i,j) \log P(i,j)$
Contrast	$\sum_i \sum_j (i-j)^2$
Homogeneity	$\sum_i \sum_j \frac{P(i,j)}{1+\lvert i-j \rvert}$

The statistical features, energy, entropy, correlation and contrast are computed for each GLCM matrix. The feature vector is calculated by using the means and variances of all parameters. GLCM is calculated from different combination of pixel brightness values occur in a pixel pair. It has high discrimination accuracy and requires less computational time.

CLASSIFICATION ALGORITHM

Decision Tree

A decision tree is a tree structure like graph, where each internal node indicates a test on an attribute [11]. Each branch represents an outcome of the test condition and each leaf node holds a class label. The decision tree classifier consist of two phases

- Growth phase or Build phase
- Pruning phase

The tree is constructed by repeatedly splitting the training set based on local optimal criteria. The second phase, handles the problem of over fitting the data in the decision tree. It generalizes the tree by removing the noise and outliers. Pruning phase increases the accuracy of the classification. It handles non-parametric data and does not require any design and training process. It provides hierarchical association between input variables to prediction class membership and offers a set of decision rules. These rules are easy to interpret and computational efficiency is high. The computational complexity occurs, when there exist undecided values and correlated outcomes.

ISSN: 0976-3104

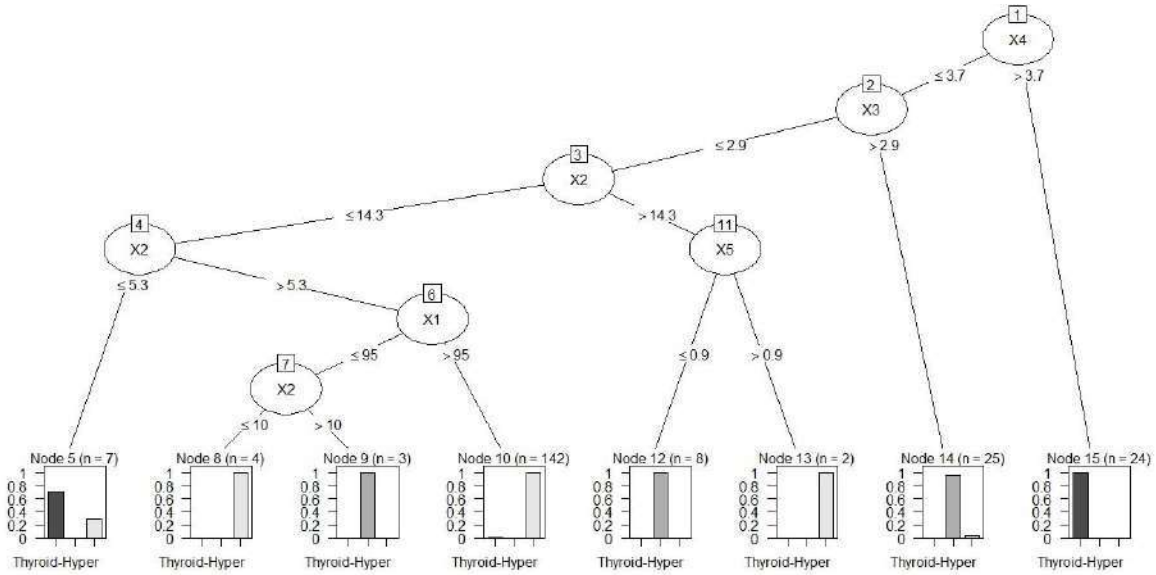


Fig: 6. Classification of data set using Decission tree

Naive Bayes

A Naive Bayes classifier is a family of probabilistic classifier works by applying Bayes' theorem [12]. It is simple method for generating classifiers that assign labels of finite set to vector of features that are selected will be treated as Independent features. For example thyroid gland is normal it produce even distribution of heat and have shape of butterfly. It considers each of these attribute as independent to the probability of classifying thyroid as normal, irrespective of possible correlation between features.

Conditional probability model to a given problem classified by vector $X = (x_1, ..., x_n)$ representing n features

$p(C_k \mid x_1, ..., x_n)$ k possible outcomes of class

$$p(C_k \mid X) = \frac{p(C_k)p(X \mid C_k)}{P(X)}$$

Using chain rule

$$p(C_k, x_1, \ldots, x_n) = p(C_k)\ p(x_1, \ldots, x_n \mid C_k) \qquad [12]$$

Under the independent assumption

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

Where $Z = p(x)$ the scaling factor depends on $x_1, \ldots, x_n$

ISSN: 0976-3104

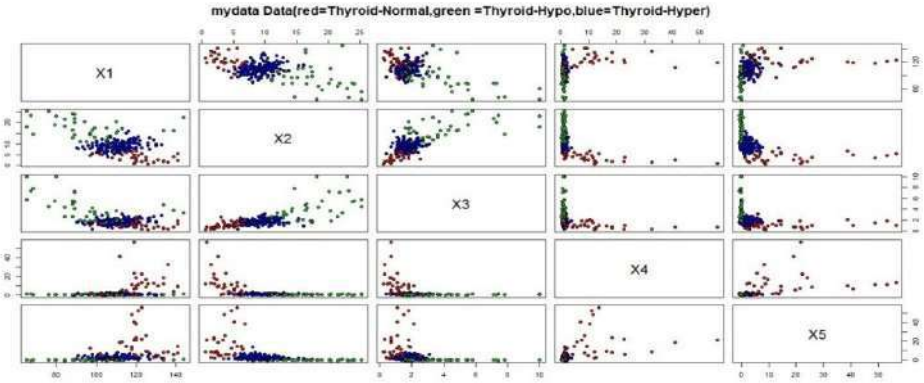


Fig: 7. Classification of data set using Naïve Bayes Classifier

Support Vector Machine (SVM)

SVM is a binary classifier and classifies two instance classes by identifying the maximum separation hyper plane. SVM in simpler form is termed as linear classifier and a non-linear SVM can be created by increasing the dimensionality of the feature space. A SVM constructs a hyper plane or set of hyper planes in a high or infinite dimensional space [13]. It uses a non-parametric with binary classifier approach and handles input data effectively and efficiently. The performance and accuracy of SVM depends on the hyper plane selection and kernel parameter. The result transparency in SVM is low and the structure of algorithm is difficult to understand. Bolster Vector Machine (SVM) models are a nearby cousin to established multilayer perceptron neural systems. Utilizing a portion capacity, SVM"s are an option preparing strategy for polynomial, spiral premise capacity and multi-layer perceptron classifiers in which the weights of the system are found by taking care of a quadratic solving so as to program issue with straight requirements, instead of a non-raised, unconstrained minimization issue as in standard neural system preparing. In the speech of SVM writing, an indicator variable is called a trait, and a changed ascribe that is utilized to characterize the hyper plane is known as a component. The errand of picking the most suitable representation is known as highlight determination. An arrangement of elements that portrays one case is known as a vector. So the objective of SVM displaying is to locate the ideal hyper plane that isolates groups of vector in a manner that cases with one class of the objective variable are on one side of the plane and cases with the other classification are on the other size of the plane. The vectors close to the hyper plane are the bolster vector.

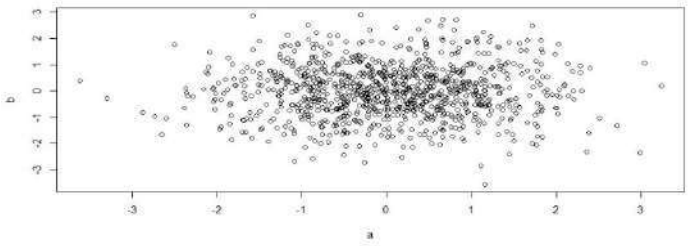


Fig :8. Classification of data set using SVM Classifier

Artificial Neural Network

An Artificial Neural Network (ANN), more often than not called neural network (NN), is a numerical model or computational model that is enlivened by the structure and/or useful parts of natural neural systems. A neural network comprises of an interconnected gathering of manufactured neurons, and it forms data utilizing a connectionist way to deal with calculation [14]. By and large an ANN is a versatile framework that progressions its structure taking into account outer or inward data that moves through the system amid the learning stage. Current neural systems are non-straight measurable information demonstrating apparatuses. Neural system models in manmade brainpower are typically alluded to as manufactured neural systems (ANNs); these are basically straightforward scientific models characterizing a capacity f: X → Y or a circulation over X or both X and Y, yet now and then models are likewise personally connected with a specific learning calculation or learning tenet. A

ISSN: 0976-3104

typical utilization of the expression ANN show truly implies the meaning of a class of such capacities. Simulated Neural Networks (ANN) has been produced as speculations of numerical models of organic sensory system.

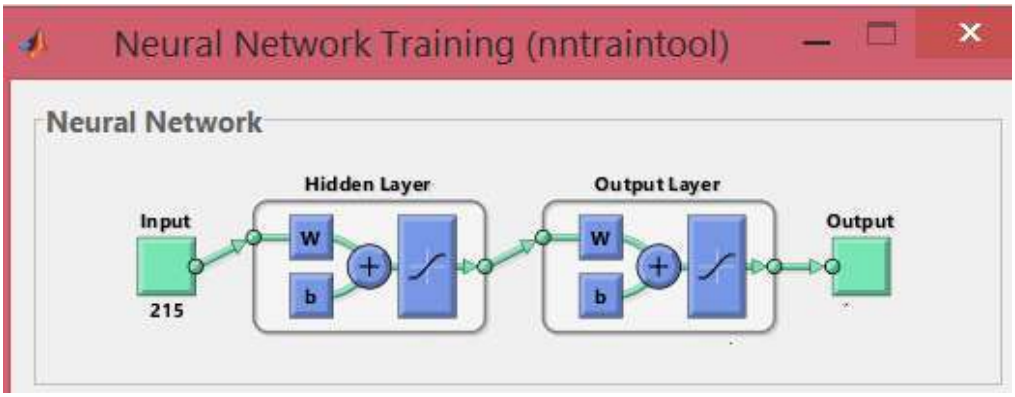


Fig: 9. Artificial Neural Network

$$f(x) = K\left(\sum_i w_i\, g_i(x)\right) \qquad [13]$$

K – *Activation function*

$f(x)$ – *Neural network function*

$g_i(x)$ – *Composition of other function*

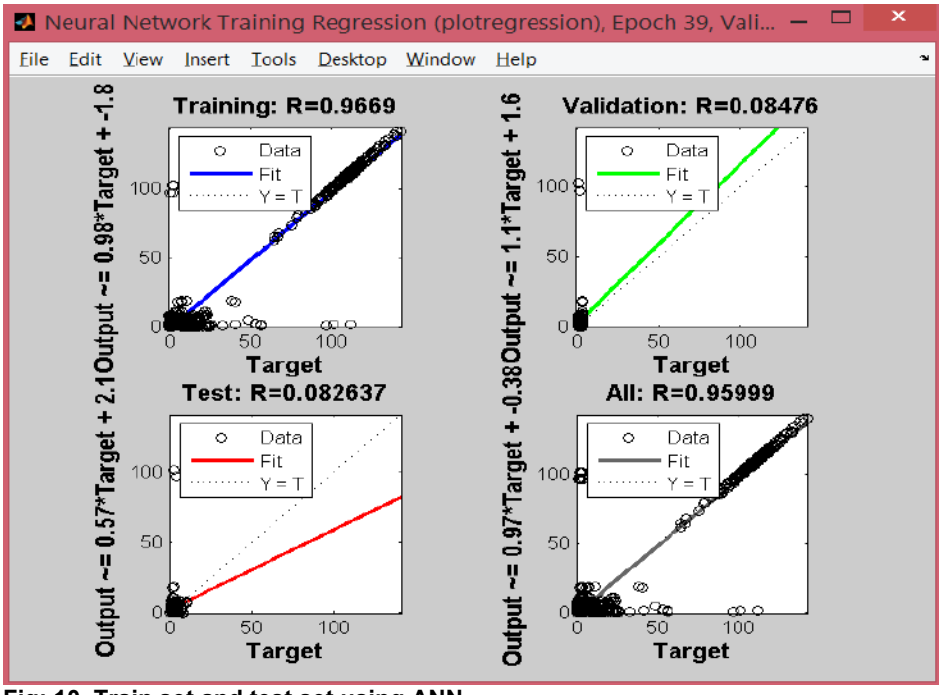


Fig: 10. Train set and test set using ANN

CLASSIFICATION ALGORITHM RESULTS AND DISCUSSION

Dataset: The data is been taken from online uci repository.*archive.ics.**uci**.edu/*

THE IIOAB JOURNAL ISSN: 0976-3104

Simulation result

Algorithm (Total Instances 215)	Correctly Classified Instance (% Values)	Incorrectly Classified Instance (%Values)	Time Taken In Secs	Kappa Statistic
Decision Tree	211(81.7757)	4(18.224)	0.23	0.9606
Naïve Bayes	152(71.0280)	62 (28.9703)	0.03	0.4321
Artificial Neural Network	170 (79.4392)	44 (20.5607)	16.67	0.5560
Support Vector Machine	173 (80.8411)	41 (19.1588)	0.67	0.5302

Training and error result:

Algorithm (Total Instances 214)	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Decision Tree	0.2318	0.3772	50.5261	78.7789
Naïve Bayes	0.3005	0.4992	65.4846	104.2402
Artificial Neural Network	0.2177	0.4279	47.4360	89.3519
Support Vector Machine	0.2018	0.4492	43.9748	93.8086

Table 1: Classification Result

Confusion matrix:

Decision Tree

		Predicted		
		Normal	**Hypo**	**Hyper**
Actual	Normal	147	0	1
	Hypo	1	35	0
	Hyper	2	0	29
		.95	0.99	0.98
		Sensitivity	Specificity	Accuracy

Table 2: Decision Tree Confusion Matrix

Naïve Bayes

		Predicted		
		Normal	**Hypo**	**Hyper**
Actual	Normal	149	0	1
	Hypo	1	34	0
	Hyper	4	0	26
		.98	0.87	0.97
		Sensitivity	Specificity	Accuracy

Table 3: Naïve Bayes Confusion Matrix

Artificial Neural Network

		Predicted		
		Normal	**Hypo**	**Hyper**
Actual	Normal	145	0	0
	Hypo	0	35	0
	Hyper	1	0	29
		.98	0.87	0.79
		Sensitivity	Specificity	Accuracy

Table 4: Artificial Neural Network Confusion Matrix

THE IIOAB JOURNAL
ISSN: 0976-3104

Support Vector Machine

		Predicted		
		Normal	**Hypo**	**Hyper**
Actual	Normal	150	1	5
	Hypo	0	34	0
	Hyper	0	0	25
		.98	0.87	0.80
		Sensitivity	Specificity	Accuracy

Table 5: Support Vector Machine Confusion Matrix

Classification Result Discussion:

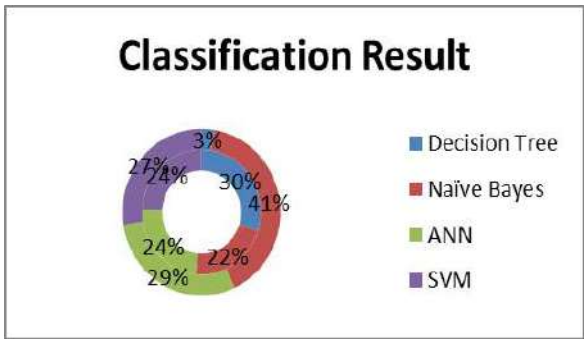


Fig 11: Classification result Graph

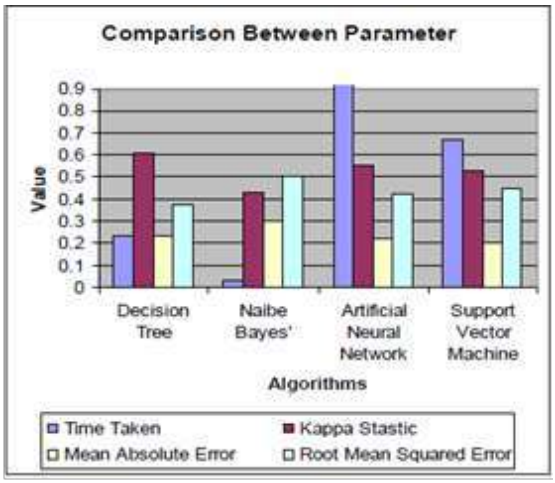


Fig 12: Comparison between parameter Graph

THE IIOAB JOURNAL
ISSN: 0976-3104

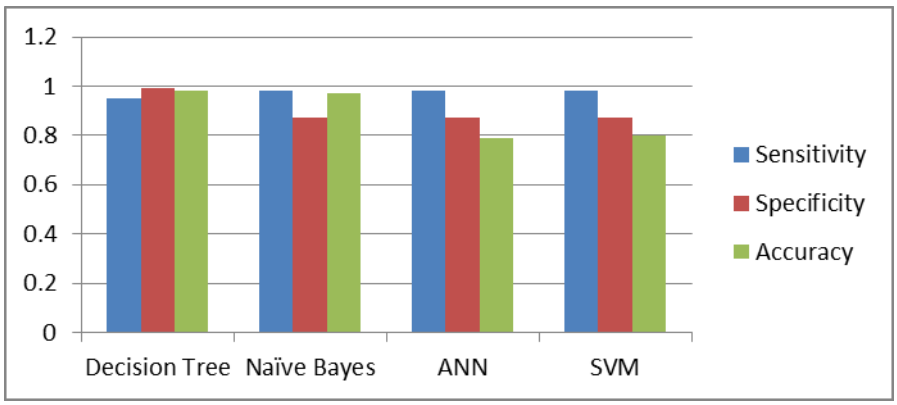


Fig 13: Confusion Matrix Graph

Classification Result Discussion

Decision tree classifier effectively grouped 214 occasions in 0.23 seconds. The erroneously arranged examples are 61 and Kappa measurement is 0.6132 for Decision Tree classifier. This examination recommends that Decision Tree classifier is the ideal calculation with maximum exactness for the charge card information. Despite the fact that, Naïve Bayes took 0.03 seconds to fabricate the model yet its precision is low. The dataset contain missing qualities. The test come about vigorously relies on upon the capacity to handle missing qualities by specific calculation. These outcomes demonstrate that among the machine learning calculations tried, the Decision Tree classifier can possibly fundamentally enhance customary characterization techniques in monetary part.

PROPOSED METHODOLOGY

Image acquisition and Data set

The algorithms where implemented based 51 real thermal image of neck consist of 21 normal and 30 abnormalities of which 14 hyperthyroidism and 16 hypothyroidism. The image captured by FLIR- E30 has temperature range of -20 to 250^{o} C with an accuracy of ±2% and thermal sensitivity of <0.10^{o}C produces thermal image of 160 x 120 resolutions. These images were collected from DITI India.

Preprocessing

Preprocessing is the first stage of medical image analysis used to reduce the noise and improve image resolution which improves the quality of image. First step of preprocessing is to convert color image to gray and remove noise from the image without affecting the edge, nonlinear filter removes the noise and preserve the edge. Median filter works by moving the window (pattern of neighbor pixels) through pixel by pixel over the image. Medial filter has special it is nonlinear for two sequence of X(m) and Y(m), Median(X(m) + Y(m)) not equal to median X(m) + median Y(m). 3.b shows the gray scale image after applying median filter. 4.a and 4.b shows the histogram equalize image .

Segmentation

We apply Otsu Thresholding to segment ROI in the Thermal Image manually. This method extract object from its background using intensity value T. **Figure- 5b** shows the segmented ROI and **Figure- 5c** shows the result of Thresholding using Otsu Algorithm. For analysis optimal threshold is taken to be 0.85 for all images. **Figure -14** shows sample thermal image before and after segmentation.

THE IIOAB JOURNAL
ISSN: 0976-3104

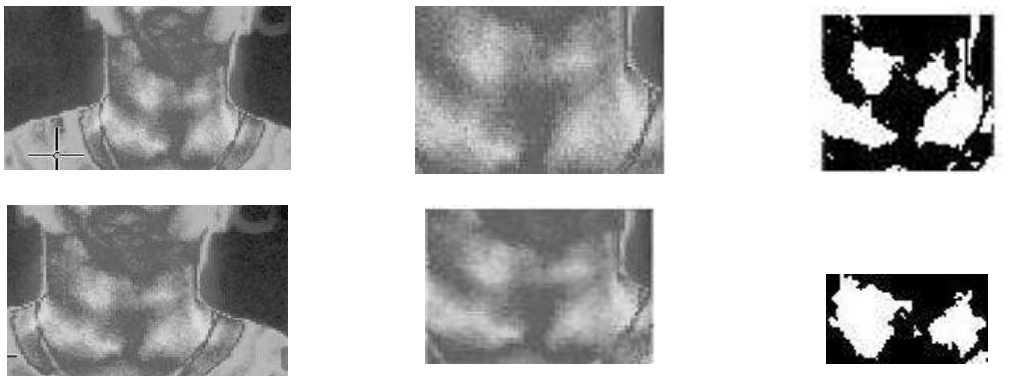


Fig: 14. Sample Thyroid Thermal Image Before and after Segmentation using Otsu Algorithm

Feature Extraction and Feature Selection

Feature like area, mean, Variance and standard deviation are calculated and shown in **Table- 6**. For our analysis we have taken the area below 43 is categorized as normal all others abnormal, variance below 5.30 are normal and above is abnormal, standard deviation below 33 are normal all other abnormal.

Sl.No/ Image	Area	Standard Deviation	Variance	Temperature range	Identified	Expert Radiologist
1.IR_0403	33	18.66	4.32	36.15°C	Normal	Normal
2.IR_0419	4	38.81	6.23	36.50°C	Hyper	Abnormal
3.IR_0437	34	41.47	6.44	35.40°C	Hypo	Abnormal
4.IR_0467	19	27.24	5.22	36.24°C	Normal	Abnormal
5.IR_0509	4	38.96	6.17	37.20°C	Hyper	Abnormal
6.IR_0527	48	44.35	6.66	37.30°C	Hyper	Abnormal
7.IR_0563	41	37.45	6.12	38.10°C	Hyper	Abnormal
8.IR_0625	33	34.69	5.89	35.25°C	Hypo	Abnormal
9.IR_0927	48	24.70	4.97	36.10°C	Normal	Normal
10.IR_1013	45	25.00	5.00	35.90°C	Normal	Normal
11.IR_1019	13	16.97	4.12	35.24°C	Normal	Normal

Table: 6. Calculate values of parameter and classified result compared with expert radiologist

THE IIOAB JOURNAL
ISSN: 0976-3104

Classification

From the abnormal it has been classified as hyper and hypo using Decision tree classification algorithm

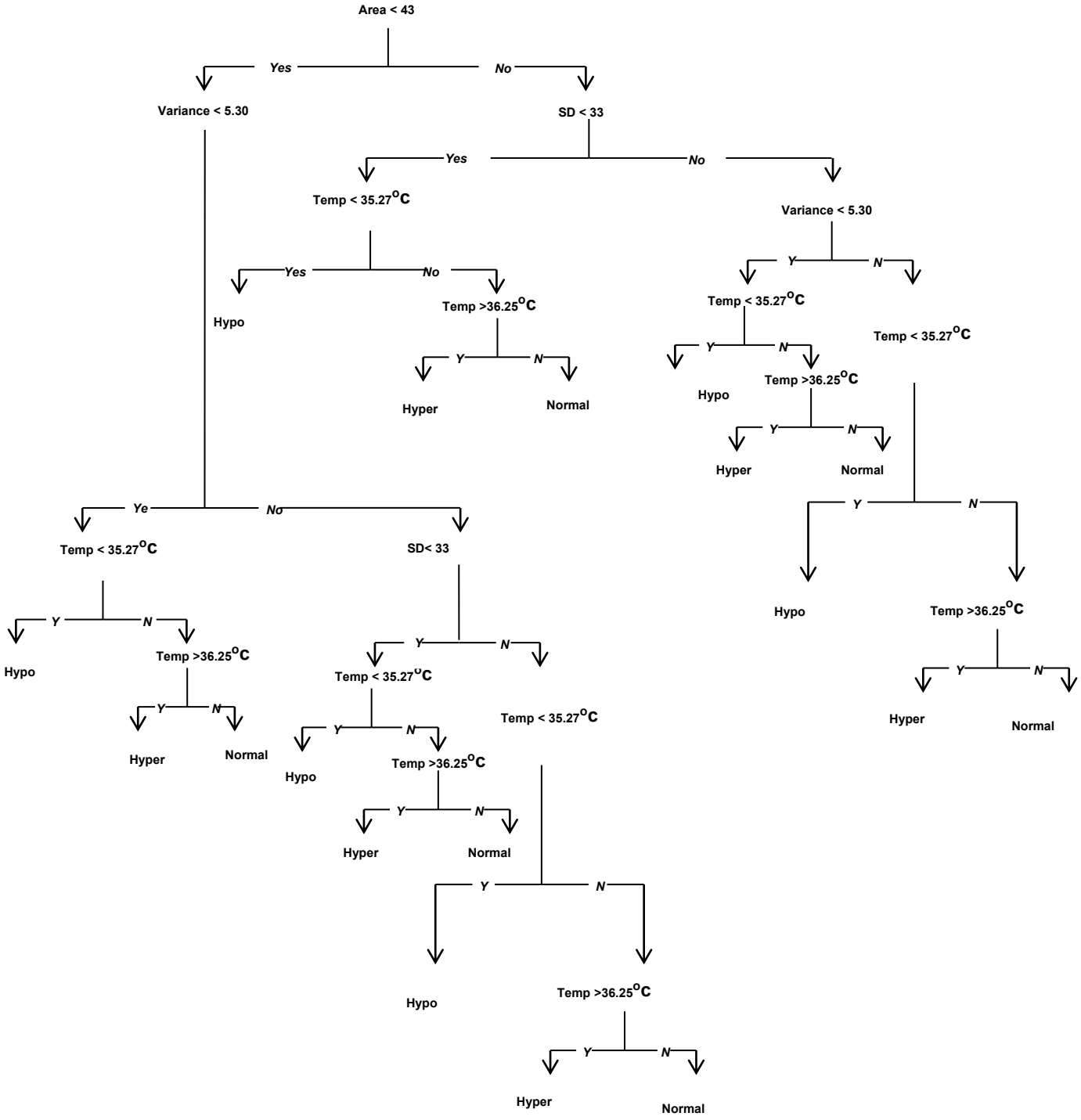


Fig: 15. Decision tree Classification for the Thyroid abnormalities

THE IIOAB JOURNAL
ISSN: 0976-3104

DISCUSSION

For the result analysis area below 43 is normal and above is abnormal, variance below 5.30 are normal and above are abnormal standard deviation below 33 are normal and above are abnormal using this parameter. For classifying abnormalities mean heat distribution value less than 35.27 are hypothyroidism, ranges from 35.27 – 36.25 are normal and greater than 36.25 are hyperthyroidism

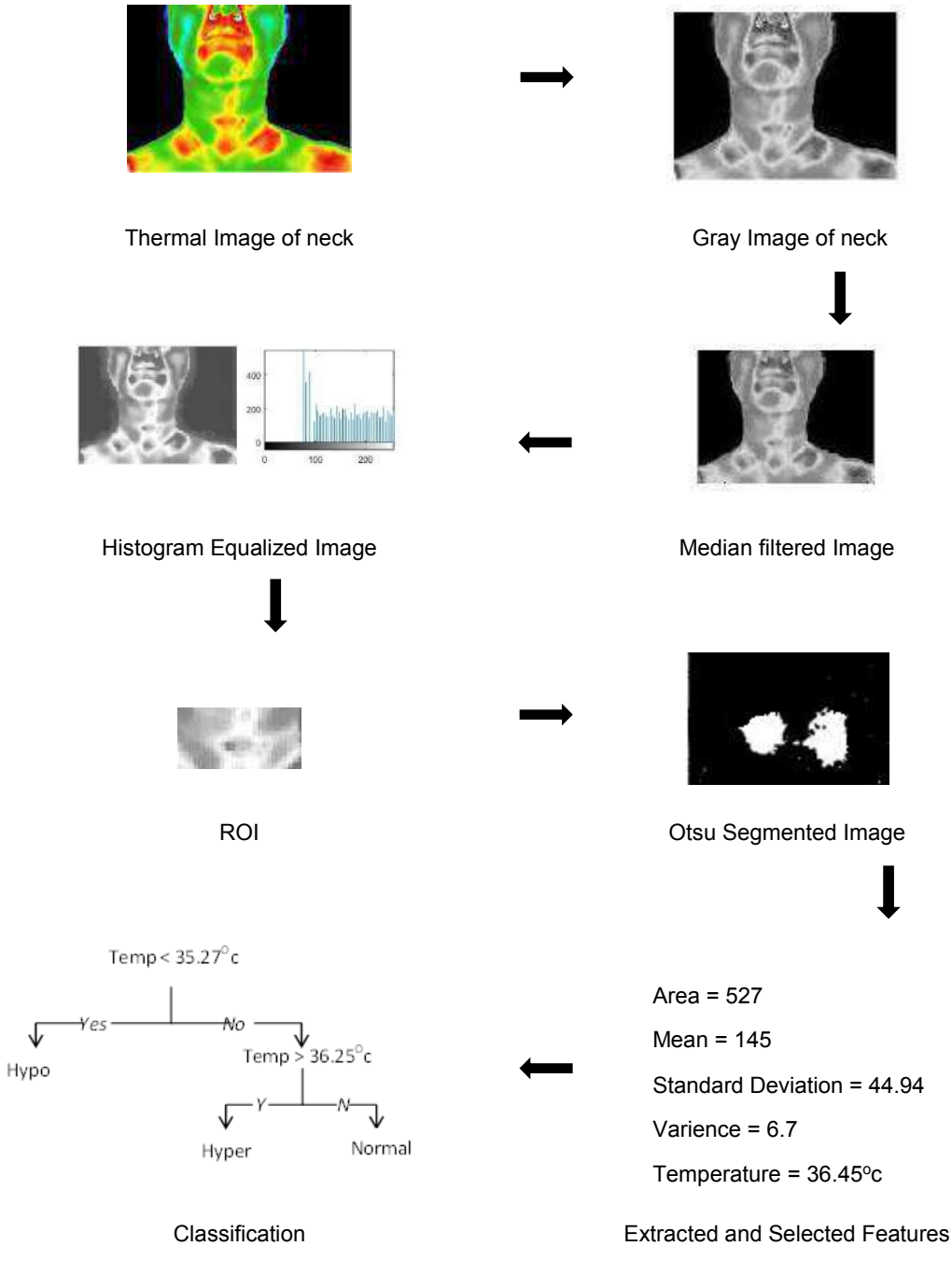


Fig 16. Proposed Mrthodology of Thyroid Thermal Image Classification

CONCLUSION

ISSN: 0976-3104

With the advancement in computational intelligence detection through thermography attracts more attention for medical analysis. Thermography is an easy non distractive technology for analyzing medical image by sensing heat emitted by human body. Thermography can be used for detecting and classifying thyroid abnormalities by simple mapping with body temperature. Thermography is one of major research subjects in medical imaging and diagnosis system. With the survey we proposed a methodology classification of thermal image. The proposed technique first apply median filter and histogram equalization for noise removal and image enhancement on converted gray scale image, then by using Otsu segmentation ROI is extracted, then employs Gabor filter and GLCM to extract features from Thermal image and by using the selected features decision tree algorithm classify as hyperthyroid, hypothyroid or normal. Result discussion on proposed method shows the robustness of proposed techniques. According to the experimental result proposed method propose classification accuracy of 95% with 96% sensitivity rate and 92% specificity rate. The challenges remain to provide generalized approach that works in all cases regardless of data base size and quality. Thermal Image Diagnosis system remains an open problem: (1) Acquisition of large database from different centers with various image qualities. (2)We have noticed segmentation plays vital role in ROI which is considered for our feature work. (3)Inclusion of machine learning techniques with hybrid model will increase the classification accuracy.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

ACKNOWLEDGEMENT

This work is part of Ph. D Research work. It is not supported by any agency
.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

REFERENCES

[1] Lahiri, B. B., Bagavathiappan, S., Jayakumar, T., & Philip, J. (2012). Medical applications of infrared thermography: a review. *Infrared Physics & Technology*, *55*(4), 221-235.

[2] Viglianti, B. L., Dewhirst, M. W., Abraham, J. P., Gorman, J. M., & Sparrow, E. M. (2014). Rationalization of thermal injury quantification methods: application to skin burns. *Burns*, *40*(5), 896-902.

[3] Calin, M. A., Mologhianu, G., Savastru, R., Calin, M. R., & Brailescu, C. M. (2015). A review of the effectiveness of thermal infrared imaging in the diagnosis and monitoring of knee diseases. *Infrared Physics & Technology*, *69*, 19-25.

[4] Rossato, M., Burei, M., & Vettor, R. (2015). Neck thermography in the differentiation between diffuse toxic goiter during methimazole treatment and normal thyroid. *Endocrine*, *48*(3), 1016-1017.

[5] Mahajan, P., & Madhe, S. (2014, August). Hypo and hyperthyroid disorder detection from thermal images using Bayesian Classifier. In *Advances in Communication and Computing Technologies (ICACACT), 2014 International Conference on* (pp. 1-4). IEEE.

[6] Nur, R., & Frize, M. (2013, March). Image processing of infrared thermal images for the detection of necrotizing enterocolitis. In *SPIE Medical Imaging* (pp. 86692M-86692M). International Society for Optics and Photonics.

[7] Tyagi, M. S., Amhia, M. H., & Tyagi, M. S. (2013). Comparative Study of Image Enhancement and Analysis of Thermal Images Using Image Processing and Wavelet Techniques. *International Journal of Computational Engineering Research*, *3*(4), 32-38.

[8] Purushotham, S., & Tripathy, B. (2014). A comparative study of RIFCM with other related algorithms from their suitability in analysis of satellite images using other supporting techniques. *Kybernetes*, *43*(1), 53-81.

[9] Etehadtavakol, M., Ng, E. Y. K., Chandran, V., & Rabbani, H. (2013). Separable and non-separable discrete wavelet transform based texture features and image classification of breast thermograms. *Infrared Physics & Technology*, *61*, 274-286.

[10] Bhattacharya, S., & Deepa, N. (2013). Non-Contact Dry Eye Syndrome Diagnosis using Thermal Imaging and GLCM Feature. In *Proc. of Int. Conf. on Advances in Computer Science and Application* (pp. 759-763).

[11] Mestha, L. K., & Venkataramani, K. (2016). *U.S. Patent No. 20,160,135,729*. Washington, DC: U.S. Patent and Trademark Office.

[12] Acharya, U. R., Tan, J. H., Vidya, S., Yeo, S., Too, C. L., Lim, W. J. E., ... & Tong, L. (2014).

Diagnosis of response and non-response to dry eye treatment using infrared thermography images. *Infrared Physics & Technology*, *67*, 497-503.

[13] Acharya, U. R., Ng, E. Y. K., Tan, J. H., & Sree, S. V. (2012). Thermography based breast cancer detection using texture features and support vector machine. *Journal of medical systems*, *36*(3), 1503-1510.

[14] Haddadnia, J., Hashemian, M., & Hassanpour, K. (2013). Diagnosis of Breast Cancer using a Combination of Genetic Algorithm and Artificial Neural Network in Medical Infrared Thermal Imaging. *Iranian Journal of Medical Physics*, *9*(4), 265-274.

[15] Mohanaiah, P., Sathyanarayana, P., & GuruKumar, L. (2013). Image texture feature extraction using GLCM approach. *International Journal of Scientific and Research Publications*, *3*(5), 1.

[16] Mitra, J., Martí, R., Oliver, A., Lladó, X., Ghose, S., Vilanova, J. C., & Meriaudeau, F. (2012). Prostate multimodality image registration based on B-splines and quadrature local energy. *International journal of computer assisted radiology and surgery*, *7*(3), 445-454.

[17] Skala, K., Lipić, T., Sović, I., Gjenero, L., & Grubišić, I. (2011). 4D thermal imaging system for medical applications. *Periodicum biologorum*, *113*(4), 407-416.

[18] Pašagić, V., Mužević, M., & Kelenc, D. (2008). Infrared thermography in marine applications. *Brodogradnja*, *59*(2), 123-130.

[19] Hildebrandt, C., Raschner, C., & Ammer, K. (2010). An overview of recent application of medical infrared thermography in sports medicine in Austria. *Sensors*, *10*(5), 4700-4715.

[20] Chang, C. Y., Chung, P. C., Hong, Y. C., & Tseng, C. H. (2011). A neural network for thyroid segmentation and volume estimation in CT images. *Computational Intelligence Magazine, IEEE*, *6*(4), 43-55.

[21] Forsberg, F., Machado, P., Segal, S., Okamura, Y., Guenette, G., Rapp, C., & Lyshchik, A. (2014, September). Microvascular blood flow in the thyroid: Preliminary results with a novel imaging technique. In *Ultrasonics Symposium (IUS), 2014 IEEE International* (pp. 2237-2240). IEEE.

[22] Luo, S., Kim, E. H., Dighe, M., & Kim, Y. (2009, September). Screening of thyroid nodules by ultrasound elastography using diastolic strain variation. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE* (pp. 4420-4423). IEEE.

[23] Lavarello, R. J., Ridgway, B., Sarwate, S., & Oelze, M. L. (2013, April). Imaging of follicular variant papillary thyroid carcinoma in a rodent model using spectral-based quantitative ultrasound techniques. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on* (pp. 732-735). IEEE.

[24] Savelonas, M. A., Iakovidis, D. K., Legakis, I., & Maroulis, D. (2009). Active contours guided by echogenicity and texture for delineation of thyroid nodules in ultrasound images. *Information Technology in Biomedicine, IEEE Transactions on*, *13*(4), 519-527.

[25] El-Dahshan, E. S. A., Mohsen, H. M., Revett, K., & Salem, A. B. M. (2014). Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm. *Expert systems with Applications*, *41*(11), 5526-5545.

ABOUT AUTHORS

Prof. M.P.Gopinath is working as Assistant Professor (Senior) in School of Computing Science and Engineering, VIT University, Vellore. His research area includes Image processing, data mining. He has published 4 journal papers in his research filed. He is life member of CSI and IEEE. He is also part of various school activity committees.

Dr. S. Prabu is working as Associate professor in the school of computing sciences and engineering, VIT University, at Vellore, India. He is the principal Investigator of Funded project from SAC-ISRO. He is life member of CSI and IEEE. He has published many technical papers in various international journals, conferences, and Springer book chapters. His research interest includes Image processing, Remote Sensing, Cloud Computing.

ISSN: 0976-3104

SPECIAL ISSUE (SCMDSA)

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A NOVEL RGB BASED STEGANOGRAPHY USING PRIME COMPONENT ALTERATION TECHNIQUE

Anil Sathyan*, Mythili Thirugnanam, Sumit Hazra

[12,3] *School of Computing Science and Engineering, VIT University, Vellore, INDIA*

ABSTRACT

Steganography is the science with the help of which secret or confidential data is hidden within any media like text, images, audio or video and protocol-based network. As privacy concerns continue to develop, it is in widespread use because it enables to hide the secret data in cover images. Steganographic techniques are best suited for digital image processing. In general, Steganography is classified into spatial and transform domain techniques. This paper presents a new RGB based algorithm in spatial domain called Prime Pixel Alternation technique. In the present scenario, many spatial domain algorithms like LSB, first component alternate, pixel indicator are usually used for steganography since they are easier to implement and less complex. Even though their hiding capacity is high, they are more prone to steganalysis. A new RGB-based algorithm is designed to store data in Random prime numbered multiple pixel locations and the encrypted key is also stored along with the data in a co-prime location (co-prime to the aforesaid 3 numbers).This algorithm requires to choose 3 random numbers to store data in R(Red), G(Green) and B(Blue) components of the cover image(24 bit image).Blue component is given the priority to store more data (lowest prime no: multiple pixel location) because a research was conducted by Hecht [14], which reveals that the blue objects ,if visually perceived , are intense and are comparatively less distinct than the red and green objects . Since the key size is fixed, it is stored in co-prime pixel locations, which will be least in number. Key bits will be stored in R, G, B components one at a time in a cyclic manner, in the above mentioned co-prime locations, in which the security is improved. A null terminator bit pattern may be used to indicate end of key or data. The reverse process of the encoding algorithm is used to decode the message.

Received on: 30th-November-2015
Revised on: 22nd-February-2016
Accepted on: 12th–February-2016
Published on: 16th–May-2016

KEY WORDS

Steganography; RGB; First Component Alternate; LSB; Steganalysis.

***Corresponding author:** **Email:** anil.sathyan2015@vit.ac.in; **Tel.:** +91-416 - 2202000; **Fax:** +91-416-2202041

INTRODUCTION

Due to rapid advancement in the fields of both computer technology and the internet one of the most essential and important factors is the communication over the internet which requires security and authenticity of information and thus that of the users as well. For exhibiting the mechanisms for the security of data transfer to and fro over internet either of the two mechanisms can be adopted which are namely Cryptography and Steganography. The art of hiding information is known as Cryptography (which is a Greek word with basically two components: kryptos (meaning hidden) and logos (meaning word).The main and ultimate goal of cryptography is user authentication, data authentication and data confidentiality. Steganography just adds another level of security to it. It basically involves the composing of messages so that only the sender and the receiver only knows that the message even exists thus avoiding any unwanted third-party attention and hence no possibilities for their intrusion. Mainly it is of three categories : Steganography in image , Steganography in audio iii) Steganography in video and in recent literature Steganography in text has also been proposed. Steganography for image and text has been worked upon. The general classification of Steganography is shown below in **Figure– 1.**

Some of the applications of our technique are being described in this paragraph. It can be used in the corporate world for transmission of confidential data without any third-party interception. It is also used in business for hiding any of the ideas or plans for a new invention. The simplest and the oldest application is that in map making where cryptographers sometimes add a tiny fictional street to their maps, which allows them to track and hence prosecute copycats. Photo collections which are sold on CD , often have hidden images in the photos which allow detection of unauthorized use. When the same technique is applied to DVDs it is even more effective and useful, as the industry builds DVD recorders to detect and at the same time disallow copying of protected DVDs and thus preventing unauthorized use. Four essential parameters for the evaluation and analysis of the quality of a Steganographic algorithm are : Imperceptibility, High data hiding capacity, Security and Robustness. In this proposed algorithm parameters such as imperceptibility, data hiding capacity and security has been focused upon mainly.Thus , through our proposed algorithm we are basically ensuring an advanced level of security by including symmetric key based

THE IIOAB JOURNAL
ISSN: 0976-3104

encryption technique and at the same time maintaining a perfect balance between other steganographic parameters as mentioned above, thus optimizing the algorithm . Some of the related works done on this topic are briefly discussed in the following sections.

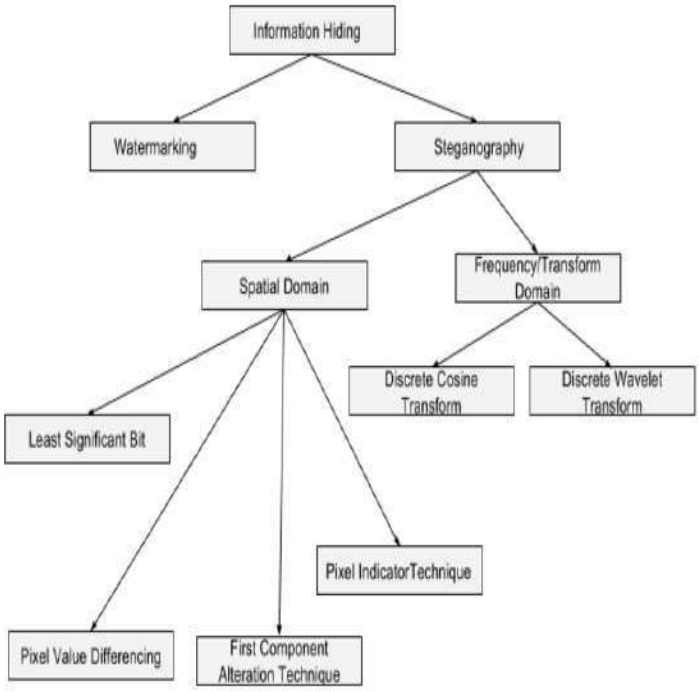


Fig: 1. Steganography Classification

MATERIALS AND METHODS

Basic Techniques and Related Works

Least Significant Bit (LSB) Data Hiding Technique

This method is the easiest way of hiding information may it be in the form of text or image in a cover image. In this very technique, the Least Significant Bits (LSB's) of the pixels are considered individually at a time which further undergoes replacement with the message that is to be sent, in which each bit of the message is hidden in each LSB. The message bits are permuted before the task of embedding is performed, which distributes the bits in an even manner, thus on an average only half of the LSB's will be altered or modified. Using this concept Darshan R., R Prabhu et al. **Error! Reference source not found.** have proposed a Steganographic technique in which though the code for Steganographic technique is easier to implement ; the statistical methods for Steganalysis makes it all the more insecure and is vulnerable to corruption.

Pixel Value Differencing (PVD) Technique

The alteration of edge areas in the visual system of human beings cannot be distinctly recognized separately or distinguished well, but if the small areas are altered such alterations can be well identified with distinguishing features. So more confidential and secret data can be hidden in an edge area than a smoother one. With this concept, Yang, Cheng-Hsing et al. [13]through a paper have put forth a Steganographic scheme in which they have used combined methodologies of Adaptive LSB and Pixel Value Differencing. Though this paper gives high adaptability, capacity, image quality and imperceptibility but then pixel value is varied from 0-255 which is a major disadvantage besides data is hidden along edges only, thus giving a lower embedding rate overall.

First Component Alteration Technique

Many Steganographic schemes are existing but a new advanced spatial domain Steganographic scheme which is the First Component Alteration Technique is being discussed here . The First Component Alteration technique is used to hide secret data or image within the cover-image which is comparatively bigger. Focus is mainly on the two bits or the four bits of a pixel in a image with

THE IIOAB JOURNAL
ISSN: 0976-3104

a maximum of five bits at the edge of an image which results in less PSNR(Peak Signal to Noise Ratio) and a high value of root mean square error. The technique uses 8 bits of blue components (as in RGB (Red, Green and Blue) where Blue is considered the First component of pixels are replaced with secret data bits one by one in the proper sequence. This scheme can hide more data than previous schemes and give a better image quality but then it cannot be applied for those images which have less blue component ,which is its major disadvantage. With this concept Kaur, Amanpreet et al. **Error! Reference source not found.** have proposed a random 8-bit alteration technique for Steganography which though provides a higher data hiding capacity, imperceptibility and better image quality but cannot be applied for those images which have a lesser blue component.

Pixel Indicator Technique

Image based Steganography basically is based upon the fact that the images are utilized as cover media to hide confidential data. The common technique that has been used replaces the LSB bits of the image pixels with the intended secret bits. Several efforts to improve the security of the LSB method have already been emphasized through earlier presentations. This paper has proposed an enhanced technique which utilizes the 24 bits in each pixel in the RGB images by using the two least significant bits of one of the channels to act as an indicator for the existence of data in the remaining two channels (R/G/B). This Steganographic method does not use a separate key to remove the overhead of the key management. Instead, it is using the secret data size as the main criteria for selection of the first indicator channel to insert security and randomness. Our proposed technique has been compared with two other similar works by analyzing the security and capacity of respective techniques. This pixel indicator technique for RGB images has showed better results compared to previous techniques.

In prior studies, one of the most popular and the oldest techniques for Steganographic Data Hiding has been the LSB mechanism in which the data bits has been embedded by substituting the Least Significant Bits of the binary representations of the RGB components of each pixel. According to the study conducted by Darshan R et al. [3] the LSB substitution technique's GPU execution is 20 times faster and even easier to implement. However, a statistical method for Steganalysis makes it insecure and thus making it vulnerable to corruption. A similar study carried out by Kaur et al. [11]focuses on the First Component Alteration Technique. Here, they embed the data bits into the blue component only thus having a higher data hiding capacity, imperceptibility and better image quality, yet it fails its application for those images which have less blue component. In another case Bharti et al. [4] adds another layer of security to the existing techniques and imperceptibility by using Vigenere Cipher; but it is having a primary weakness which is the principle of the repeating key. In another notified paper by Mahimah et al.[5] the same Steganographic technique has been utilized by using a new and different approach that is of a Zigzag Pixel Indicator Technique which gives us a high quality stego image providing better security than the existing LSB techniques. But, on the other side it is computationally complex and has high space requirements. A paper on Steganography by M.G.Gouthamanaath et.al uses the concepts of Pixel Value Differencing (PVD) and Pixel Indicator Technique (PIT)[1] which reduces the existing computational costs and provides a good image quality; yet in their paper the hidden capacity depends on Cover image pixel intensities. In another important paper by Luo et al. [12]the Steganographic analysis uses the techniques of Edge Adaptive Image Steganographic scheme in Spatial Domain LSB where data is hidden adaptively only in the specific edges. Even though the technique used, preserves the statistical and visual features its embedding rate is sufficiently low. But again the aforesaid technique can be extended to audio/video Steganography. Delving into the world of Steganalysis a related paper by Fillatre et al. [7] focuses on the Statistical Hypothesis Testing for the LSB mechanism which maximizes the probability of detection; yet the technique used is the LSB technique but with increased number of assumptions and also requires hiding of extra bits of signature with hidden message .

Most of the research carried out uses LSB (Least Significant Bit) Technique for embedding the data bits with improved security and imperceptibility yet, images of higher resolutions are required which is mostly covered by our research. With these intentions this work involves an algorithm which aims for a better data hiding capacity and improved security since it uses encryption techniques along with Steganography, thus also making the technique more robust. Focus is basically on the spatial domain as the pixels can directly be manipulated , but in contrast to it is the Transform Domain in which the process has complex transformations in the very specific frequency domain. Even though the Transform Domain technique is less susceptible to external interference e.g. Noise or Corruption but then it has less data hiding capacity when it is compared to Spatial Domain which is basically the platform under consideration. Thus the aim is to achieve a perfect balance between Imperceptibility, Data Hiding Capacity and Security through the proposed methodology that follows.

Proposed Work

The proposed Steganographic algorithm is basically a RGB-based prime-pixel alteration technique .Our Steganography is applicable for both images and texts respectively. At first a message (either an image or a text)is being taken as an input .The message is encrypted with the key and only after its proper encryption the message is embedded in the cover image. Basically to describe explicitly the process involves saving each of the data (text or input image converted into strings) bits in the prime pixel locations of the Red, Green and Blue Components' groups of 3 random numbers or prime triplets. And since the key is of a fixed size the key bits are stored in co-prime pixel locations. The key bits will be stored in the Red, Green and Blue components one at a time in a cyclic manner for enhanced security. For storing the data, the Red, Blue and Green components are basically being considered. For decryption or decoding the original message with quality fully retained and without any third party interception, the reverse process of the encryption algorithm has been followed. This can be better understood through the block diagram described below.

The basic functional block diagram of the proposed method is shown in **Figure– 2**. Firstly the payload data(image or text) is encrypted before embedding. Then, the key along with the encrypted message is embedded using the proposed technique larger

ISSN: 0976-3104

cover images are chosen for better hiding capacity. The resulting stego-file is transmitted via any communication channel to the intended receiver. At the receiving end the receiver extracts the key from stego file with the help of shared secret data(prime triplets). Using the key the extracted information is decrypted to get back the original message. The cover image is obtained as a byproduct. The complete embedding and extraction algorithms are fully described below with the help of flow diagrams in the following sections.This schematic diagram provides an outline of the basic working of the stegnographic technique.The implementation can be done in different ways depending on the types of encryption, data type, media etc.

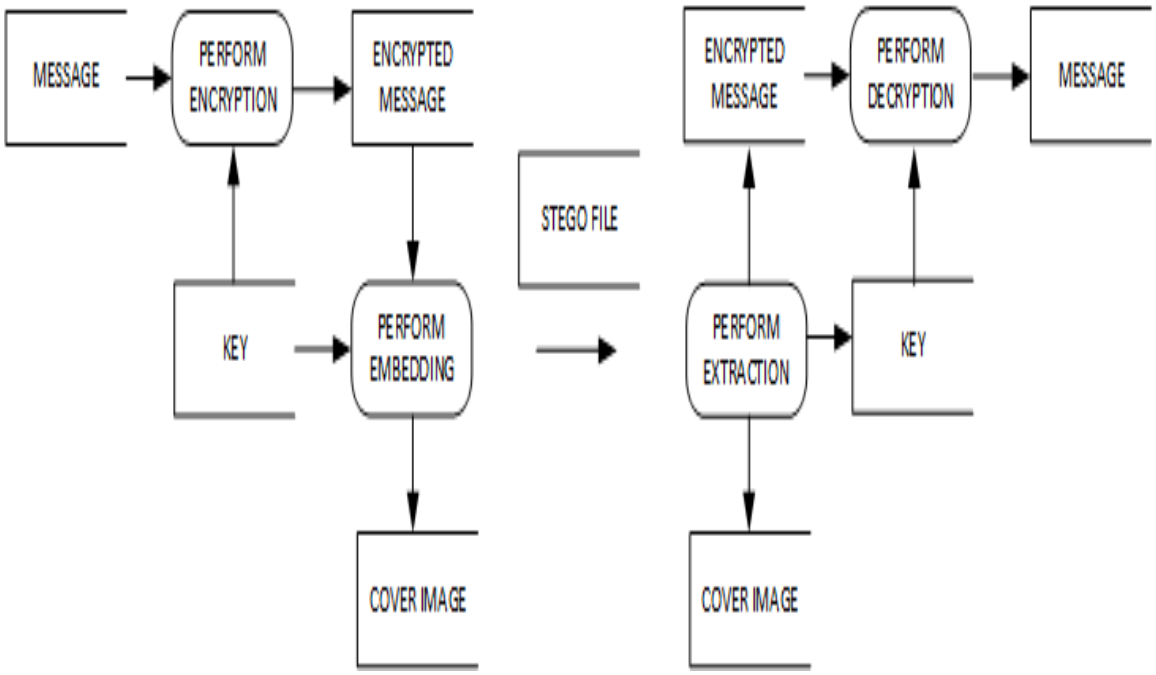


Fig: 2. Schematic Diagram

……..

The detailed flow diagram of the embedding and extraction phases of the algorithm are discussed in the following sections. The extraction algorithm is almost the reverse process of embedding. The entire process can be depicted in the form of flow diagrams as shown below.

Algorithm

The algorithm has two phases namely Embedding and Extraction. The detailed description of Embedding and the Extraction phases with the use-case scenario is described below:-

Person A wants to send a secret file (image or text file) to Person B through a insecure channel. Person C is an intruder who as access to the covert channel information. Person A and B has shared information about a prime triplet, which will be used as an input in this algorithm. Person A uses a large image to hide his secret data and a random 24 bit key.

THE IIOAB JOURNAL
ISSN: 0976-3104

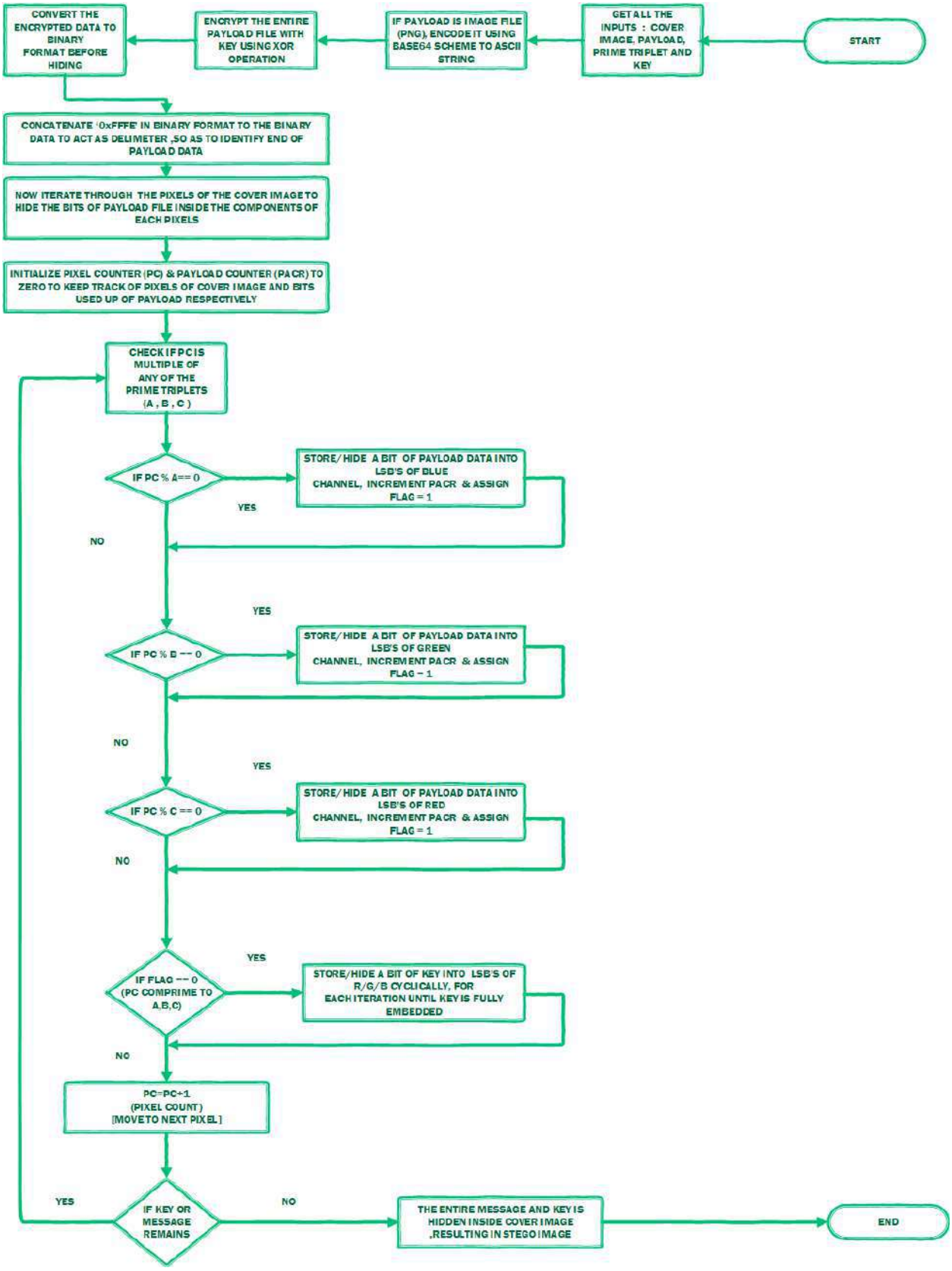


Fig: 3. Embedding Process

……

THE IIOAB JOURNAL
ISSN: 0976-3104

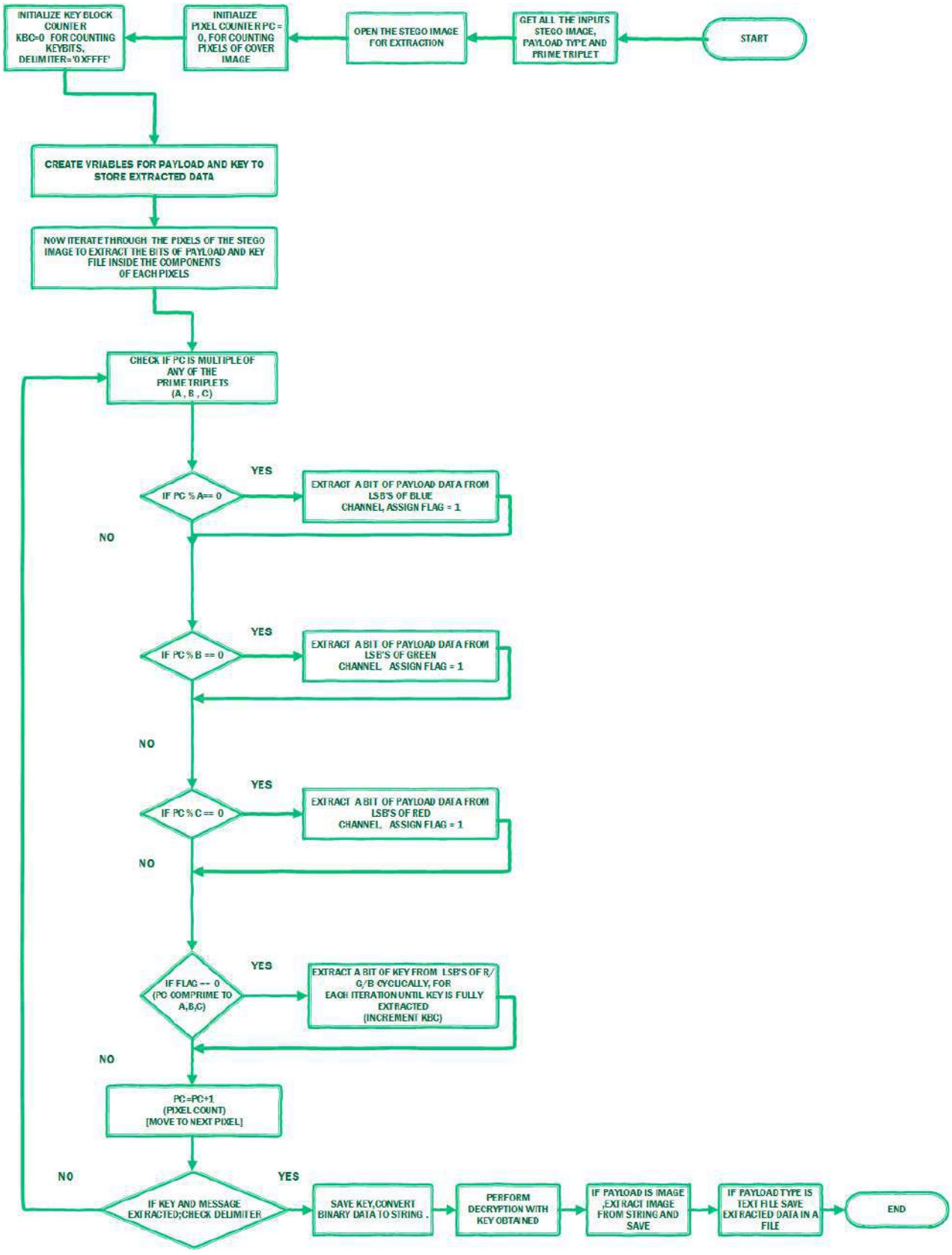


Fig: 4. Extraction Process

……

THE IIOAB JOURNAL
ISSN: 0976-3104

Steps followed for ***EMBEDDING***:

INPUT: {COVER IMAGE, SECRET FILE (PAYLOAD), KEY (24 BIT), PRIME TRIPLET (P1, P2, P3)}
OUTPUT*:* {STEGO IMAGE}

1. Get all the inputs and check the payload type, whether it is an image or text file.
2. If it is an image (PNG), encode it using base64 encoding scheme to an ASCII string else, do not encode the data.
3. Now encrypt the string using XOR with the key, with the help of standard library functions to get encrypted data.
4. Convert the encrypted message to binary string
5. Add a delimiter string '0xFFFE' to end of the binary data obtained, so as to identify the end of the payload data.
6. Now open the cover image and iterate through the pixels of the image
7. Initialize the pixel counter PC = 0 payload counter PACR = 0, and for each pixel in the cover image do the following until the entire payload and key is embedded

7.1. If PC is multiple of P1 then replace LSB's of the BLUE component by the bits of the payload binary data. Increment payload
counter accordingly.
7.2. If PC is multiple of P2, then replace LSB's of the GREEN component by the bits of the payload binary data. Increment payload
counter accordingly.
7.3. If PC is multiple of P3, then replace LSB's of the RED component by the bits of the payload binary data. Increment payload counter accor-
dingly.
7.4. If PC is NOT multiple of P1, P2, P3, then replace LSB's of the R/G/B component by the bits of binary key string ,cyclically in each
iteration (one component at a time).
8. After each pixel is saved, save this stego image in the required format (PNG).
9. If payload size exceeds capacity, throw error.

If Person B receives this cover image, he can get back the secret data from cover image, by using the extraction algorithm (reverse process), provided he has the shared prime triplet information..The key is embedded within the cover image.

Steps followed for ***EXTRACTION:***

INPUT: {STEGO IMAGE, PRIME TRIPLET (P1, P2, P3)}
OUTPUT: {COVER IMAGE, SECRET FILE (PAYLOAD), KEY (24BIT),}

1. Get all inputs, Use delimiter='0xFFFE' for checking end of message.
2. Get payload type from user that is image or text.
3. Open stego image, initialize the pixel count PC = 0, for iterating through pixels of stego image; Create variables payload,
key(empty initially for storing the corresponding bits during extraction process.
4. While the payload data is not completely extracted do the following by iterating stego pixels, use the delimiter to check for end
of data.
4.1. If PC is multiple of P1, then extract LSB's of the BLUE component and add it to the bits of the payload binary data.
4.2. If PC is multiple of P2, then extract LSB's of the GREEN component and add it to the bits
of the payload binary data.
4.3. If PC is multiple of P3, then extract LSB's of the RED component and add it to the bits of the payload binary data.
4.4. If PC is NOT multiple of P1, P2, P3, then it to extract LSB's of the R/G/B component, addthe bits of binary key string ,cyclically in each
iteration (one component at a time).
5. Save the key for decryption
6. Convert binary data to string.
7. Perform decryption using XOR , with the key obtained.
8. If payload is an image, extract the image from the string and save the image.
9. If the payload is text file, save the extracted as data text in a new file.

Thus the person B can get back the secret data (text/image) from the cover image. But here, the cover image cannot be recovered fully. If the person C (intruder) has no knowledge of the prime triplets, he won't be able to extract the message from the stego image; even though if he somehow hacks the extraction algorithm. The encryption carried out on the secret data serves as an additional layer of security.

THE IIOAB JOURNAL
ISSN: 0976-3104

Module Description

The input of the algorithm consists of a large cover image, secret data (image/text), 24 bit key and a prime triplet. The key size is selected to be 24 since an image (png) having RGB format with 24 bits per pixels has been used.

The prime triplets are ordered pairs of form (p1, p2, p3) such that p1<p2<p3 .They are selected randomly and should be very less than size of secret data to be hidden. These numbers are shared between two parties involved in communication. A **prime triplet,** in mathematics, is defined as a set of three prime numbers of the form (p, p + 2, p + 6) or (p, p + 4, p + 6).For example ordered triplets like (5, 7, 11), (7, 11, 13), (11, 13, 17), (13, 17, 19), (17, 19, 23), (37, 41, 43) etc..In our algorithm it can be any 3 random numbers or numbers of the form just mentioned above.

Here the value of each number determines the number pixel components modified. (p1, p2, p3) corresponds to (B, G, R) components respectively. The main process of the algorithm consists of iterating through the pixels of cover image. For this the pixels are numbered sequentially and then check for each pixel if it is a multiple of any of the prime triplets. If it is multiple of p1 hide the bits of secret data in LSB's of BLUE component; if it is multiple of p2 hide the data in GREEN component and finally if it is a multiple of p3 the data is hidden in the RED component of the pixel. Since p1<p2<p3, Blue component is modified most number of times, followed by green and finally Red component is modified least when considering total pixels modified. This is done because a study conducted by Hecht et al.[14] suggested reveals that that the visual perception of blue objects those are intense is comparatively less distinct than that of the perception of objects that are colored red and green.

In this algorithm ,the image is converted into base64 encoded format for easier processing. Then the specific key is used to encrypt the data using XOR operation.XOR operation can provide moderate security as long as key is not compromised. The property of XOR is that once XOR operation is done using a key, the original message is obtained back from cipher text just by performing the same XOR operation with the same key on the cipher text. For example:-

DATA: 10011100 XOR (⊻)
KEY: 01101100
CIPHER: 11110000

KEY: 01101100 XOR (⊻)
CIPHER: 11110000
DATA:10011100

The data is converted to binary format so as to easily embed the same inside the component bits of the pixels. A delimiter '0xFFFE' in binary format is used for identifying the end of message. At each iteration process described above counters are used for the pixels of the cover image and also for the data and key so that the iteration can be stopped when entire key and data is embedded inside the cover image. The key is embedded inside the pixels whose number is co prime to prime triplets. The component positions in which the key bits are hidden (of pixels of the cover image) are changed cyclically. This adds some 'confusion' to the encryption scheme. The main aim of Steganography is to conceal the existence of data inside cover-image. Now by adding additional security, even if the existence of message is found out the hacker will be unable to extract the message in unencrypted form.

The algorithm will have a maximum hiding capacity based on the size of the cover image. Also this will also depend on the selection of the prime triplets. Generally, using a smaller prime triplet one can get higher hiding capacity since data is hidden only in pixel locations which are multiple of the prime triplets. So this prime triplet values can be used for tuning the algorithms hiding capacity, keeping the cover image size constant.
Images in PNG format and text files have been used as payload (secret data) and the performance was as good as expected.
In the extraction phase only the stego-image and prime triplets are required as inputs, since the key is already embedded inside the stego-image. This removes the burden of maintaining an additional shared key. The extraction phase is almost same as the reverse process of embedding . Keep the count of pixels in stego image and also count of the key bits for identifying the end of data and key. After the extraction of data convert it back to ASCII string. Once the key is extracted use this same key to decrypt the data using XOR operation(In general any encryption scheme can be used for enforcing security).If the payload is of image type extract the image from the string and save it ;else if it is a text store the extracted text inside a new file.

The prototype application was implemented using python with the help of python image library (PIL) and Crypto cipher suite along with other encoding schemes like base64, binascii etc.. Using the proposed algorithm text files as well as images in png and tiff formats can be hidden , but the analysis that has been carried out in this paper mainly focuses on images. The performance analysis in terms of speed, quality and other parameters along with benchmarking of algorithm are discussed in the next section.

THE IIOAB JOURNAL
ISSN: 0976-3104

RESULTS

Analysis of Algorithm

Here the quality and performance of the algorithm based are being analyzed on the basis of various parameters like execution speed, data hiding capacity, imperceptibility, security, etc respectively. A sequential version of the algorithm was tested for determining the execution speed for various sizes and combinations of inputs namely the payload size and prime triplet combinations respectively. A graph was plotted with Input size vs. Execution time and the result is shown aside.

For a cover image size of 1024*1024 and prime triplet (13,17,19), it was observed that maximum possible size of image for hiding was 220*220 (22%);but with lower prime triplets higher resolution was possible.

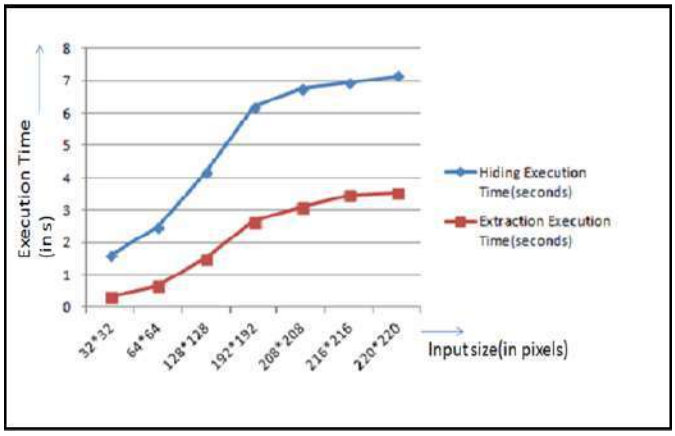


Fig: 5. Input vs. Execution Time

……..

Let m, n, p be the size of the message, key and cover image (in bits).Let (p1, p2, p3) be prime triplets. Let p>>>m, n (since cover image should be sufficiently large).Here, as observed from the flow diagram of the algorithm, for each pixel location, do '5' comparisons (p1, p2, p3, co prime-flag, end of key/message).Therefore, Total Number of comparisons = 5 * (Number of pixels used up for hiding in cover image) Number of pixels used up for hiding in cover image = At Most (m+n)/2 (worst case) (e.g. all pixel location are either exclusively multiple of any one number or a co prime) Because for each pixel store at least 2 bits of key/message in component LSB's.

i.e. $O((m+n)/2)$

So, Total no. of comparisons ,

(T.C) = $5 * O((m+n)/2)$.

If 'b' LSB bits are used then the complexity will be

T.C = $5 * O((m+n)/b)$.

Now, the Asymptotic Time Complexity for the entire algorithm is

$$T(m,n) = O(m+n).$$

[Since constant terms can be neglected, assuming encryption, conversion etc. take constant time]

The parameters for measuring the performance for Steganographic algorithms are Image quality, Data Hiding Capacity, Security, Imperceptibility etc. Our main objective is to find optimum data hiding capacity with respect to all the above parameters, for this algorithm.

The data hiding capacity is generally proportional to the number of bits hidden per pixel, which is based on number of bits hidden on each component of a pixel. At maximum level hide 4 LSB bits per component of a pixel; beyond this limit, usually artifacts will become visible reducing the quality of the image. This will directly affect the imperceptibility (which is one of the major goals of Steganography) of the stego image , and thus one will be easily able to identify the discrepancy with the image. Thus an optimum value is required to maintain good data hiding capacity and at the same time have good imperceptibility. This is like a trade-off between two parameters hiding capacity and imperceptibility. In the naive implementation 4

ISSN: 0976-3104

LSB bits have been used for hiding the data inside cover image. The maximum modification occurs in pixels whose positional number is divisible by all the prime triplets wherein hide 4 data bits in all the three components (B, G, R) LSB position. The number of such pixels will be very low (also depends on values of prime triplets) when compared to other type of pixels. The following table shows the PSNR, MSE values for various types of inputs followed by the corresponding graphs in **Figure–1** (for PSNRs) and in **Figure–2** (for MSEs).

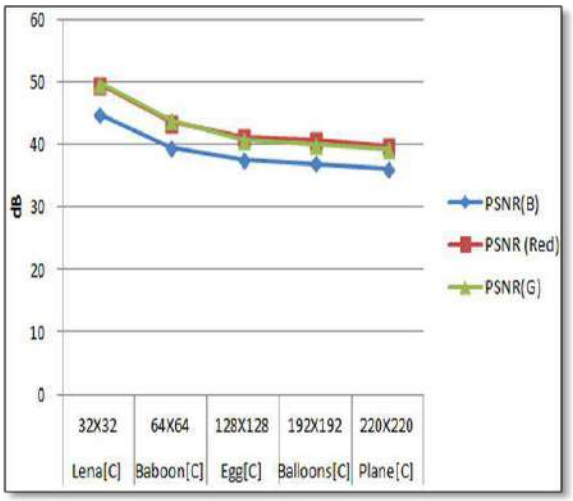


Fig: 6. PSNR - R,G,B

……..

:

Image	Size	PSNR(B)	PSNR (Red)	PSNR(G)	MSE(B)	MSE(Red)	MSE(G)
Lena[C]	32X32	44.831292	49.626305	49.876463	2.51	0.67	0.71
Baboon[C]	64X64	39.3725797	43.5099666	43.9048412	7.57	2.67	2.92
Egg[C]	128X128	37.4430374	41.2344887	40.6790346	11.81	4.93	5.61
Balloons[C]	192X192	36.8456661	40.7617169	40.0081296	13.55	5.5	6.54
Plane[C]	220X220	36.0297394	39.8449123	39.2234668	16.35	6.79	7.84

Table: 1.PSNR, MSE of R,G,B components

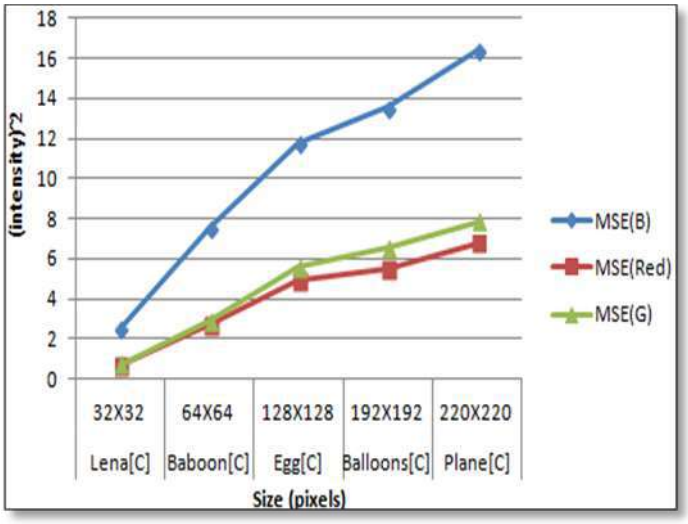


Fig: 7. MSE- R,G,B

……..

THE IIOAB JOURNAL
ISSN: 0976-3104

Different PSNR and MSE values (depicted via. the following tables):

Image	Size	PSNR(in dB)	MSE
Lena[C]	32X32	48.11	1.29
Baboon[C]	64X64	42.26	4.39
Egg[C]	128X128	39.79	7.45
Balloons[C]	192X192	39.21	8.53
Plane[C]	220X220	38.37	30.98

Table: 2. Overall PSNR,MSE (cover vs. stego)

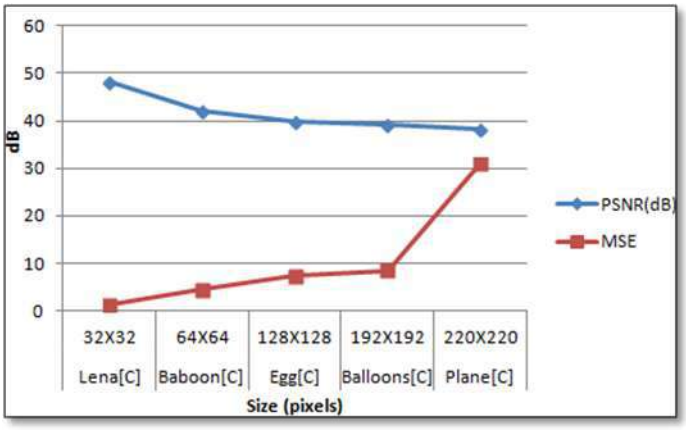


Fig: 8. Overall PSNR and MSE

Experiment and sample output

The following images were utilized for our experimentations:

Fig: 9. Images used for experimentations

As per proposed in this paper, a short implementation has been carried out and is shown below.

The image Lena.png shown above has been used as the cover image to hide payload images such as Aeroplane.jpg , Balloon.jpg,

THE IIOAB JOURNAL
ISSN: 0976-3104

Baboon.jpg, egg.jpg , etc. A sample output for hiding Baboon.jpg has been shown below.

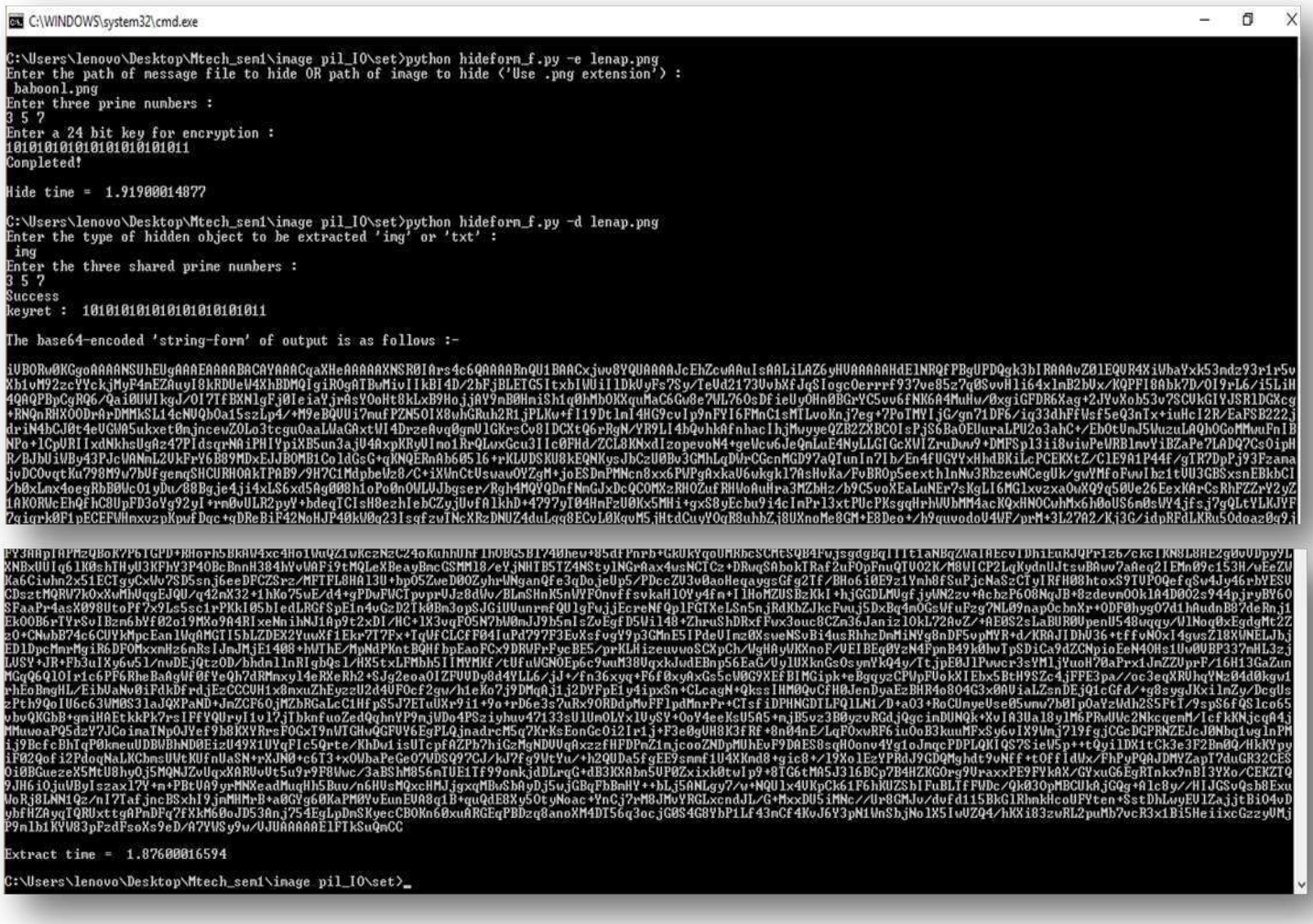


Fig: 10. Sample Outputs:

DISCUSSION

In the tables **[Tables– 1 & 2]**and graphs (depicted via. **Figures– 1, 2 & 3]** for the PSNR (Peak Signal-To-Noise Ratio) and MSE (Mean Square Error) values the following specific formulas have been used:

MSE (Mean Square Error):

It gives the mean of the pixel values of the image and by averaging the sum of squares of the error between two images.

$$MSE = \frac{\hat{}}{M * N} \rangle \quad \rangle \quad [x(m,n) - y(m,n)]^2$$

where x (m, n) and y (m, n) refers to the two images having a size of M*N. In this formula x is the original image and y is the stego image .

PSNR (Peak Signal-To-Noise Ratio):

The Peak Signal-to-Noise Ratio (PSNR) measures the estimates of the quality of the stego-image compared with an original image and is a standard (benchmark) way to measure image reliability or conformity.

THE IIOAB JOURNAL ISSN: 0976-3104

$$PSNR = 20 \log_{10} \left| \frac{\text{MAXPIX}}{RMSE} \right|$$

where, MAXPIX is the maximum value of a pixel and RMSE is the Root Mean Square Error for the input image (it gives the average sum of distortion in each pixel for the stego image i.e. the average change in pixel caused by the encryption algorithm used).

In PSNR signal is the original image and noise is the error in the stego image resulting due to encoding and decoding. PSNR is measured in decibel (dB).

Also, Peak Signal to Noise Ratio (PSNR) is inversely proportional to the Mean Square Error (MSE), which implies that lower the value of Mean Square Error (MSE), higher is its Peak Signal to Noise Ratio (PSNR). Thus higher the Peak Signal to Noise Ratio (PSNR) the more is it better as it results in lesser error.

After comparing the input image(that is to be embedded in the cover image)before embedding and after extraction it was found that the MSE was zero(0.00) and PSNR value was found to be infinite(Inf dB).This confirms that the input image embedded in the cover image and the image that has been extracted are same. In the following sections comparisons of the cover and the stego images have been carried out with the help of histograms. The histograms depict the graphical representations between the intensity values of pixels (along X axis) and their frequencies (along Y axis) of the images that are being compared.

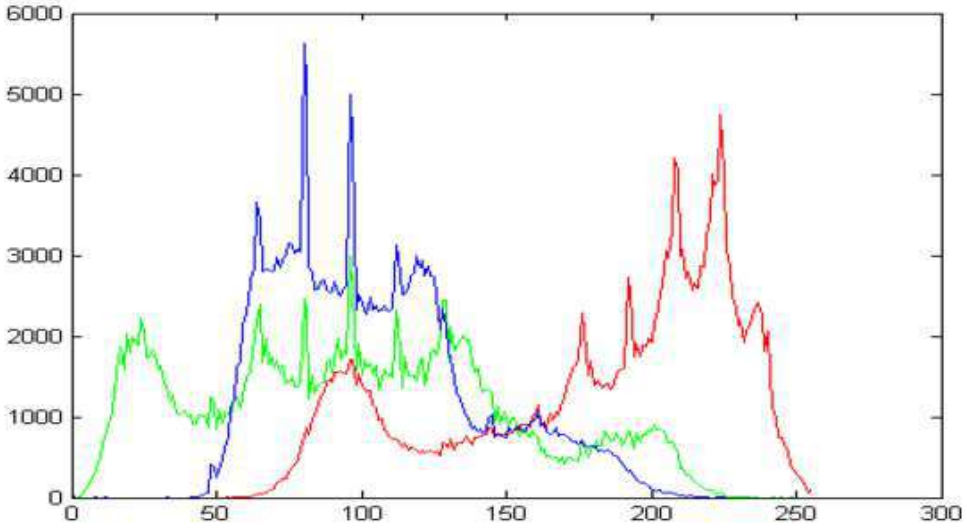


Fig: 11. Histogram before Embedding (COVER IMAGE)

Here the Cover image size was 512*512 and the payload image was 64*64.The color of the graph lines indicate the corresponding values for the R/G/B component. Histogram gives a plot of intensities of components versus. their frequencies in constituent pixels of the image. It is clear from the graph that most of the changes occur in the blue components of the pixels of the cover image followed by green and then red components are the least modified. The algorithm is a modified form of LSB substitution some extra security and also hiding capacity have been obtained. Thus it will be more difficult to crack the secret data by Steganalysis than the conventional LSB techniques. The security depends upon the encryption technique employed. Also the use of random prime triplet makes it difficult to predict the pattern of embedding. Even if the pattern is somehow decoded , still encryption provides so security against complete failure.

The prime triplet used is random and is not repeated within a particular time frame or particular number of exchanges of message using this technique. Finally the key is stored in pixel locations which are co prime to prime triplets, that too in a cyclic manner in R/G/B components for each such positions. Thus these interdependent security features, in effect provides us a high level security against the common Steganalysis techniques.

THE IIOAB JOURNAL
ISSN: 0976-3104

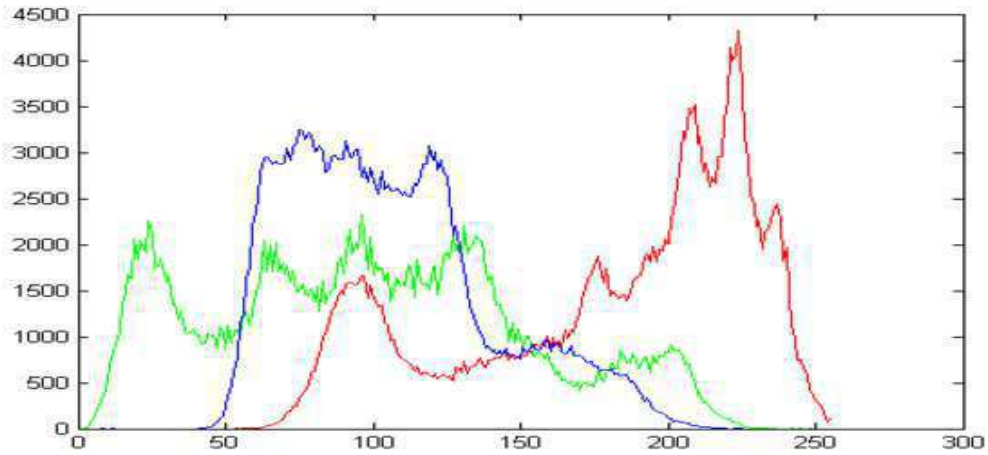


Fig: 12. Histogram after Embedding (STEGO IMAGE)

Upon comparison of some previous steganographic algorithms in [1],[11]with the proposed work, the following results have been obtained as has been described below.

COVER IMAGE	WU AND TSAI'S METHOD(PSNR in dB)	HSIEN AND HUI METHOD(PSNR in dB)	PROPOSED METHOD (PSNR in dB)
LENA	38.94	40.21	48.11
BABOON	33.43	41.35	42.26

Table: 3. Comparison of PSNR of proposed algorithm

LENA IMAGE	LSB3 METHOD	PVD METHOD	LIE CHANG'S METHOD	JAE GIL YU METHOD	FIRST COMPONENT ALTERATION TECHNIQUE	PROPOSED METHOD
PSNR	37.92	41.48	37.53	38.98	46.11	48.11

Table: 4. PSNR value comparison of lena image.

As per the comparison-based data tabulated above, the PSNR values that have been obtained by our proposed algorithm and those that of the existing ones, we infer that the proposed algorithm is superior in terms of PSNR metric. This indicates that the output images used for our experimentations have better quality and less distortions .Besides ,to make the algorithm securer ,we are using symmetric key encryption algorithms along with prime triplets which are not existing in the previous works. To verify the correctness of this proposed technique, the PSNR values of original payload image have been compared with the extracted output image. It was observed that PSNR value was infinity and the MSE value was zero. Thus successful extraction of the original image is confirmed. Likewise, the proposed technique has also been verified with text inputs. So, the proposed algorithm works well with texts and images adding flexibility and scalability to the algorithm.

CONCLUSION

In today's world, we often hear about a popular term "Hacking". This refers to an unauthorized access of data during transmission or storage. In the case of Steganography this problem is often termed as Steganalysis. This concept (basically to prevent hacking using Steganography) has been used to hide data in images, securely and efficiently. Steganography combined with Cryptography may be some of the future solutions. The general parameters for any Steganographic technique involves robustness, data-hiding capacity and security which should be optimized .A new prime-pixel alteration technique has been presented in our paper ensuring the above criteria. So future works are inclusive of extension of the

THE IIOAB JOURNAL ISSN: 0976-3104

algorithms with different image formats(like jpeg, bmp, etc) and media formats. Also this algorithm can be explored much more by using other color domains like HSI, YCBCR, etc. formats. Nowadays it is very frequent of new Steganographic techniques to be proposed and Steganalysis methods to be found. Thus emphasizing on the fact that Steganography, nowadays is an essential pre-requisite for the communications over the Internet.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

ACKNOWLEDGEMENT

We would like to thank the School of Computing Sciences and Engineering, VIT University and Special thanks to Dean SCOPE, for his kind guidance and support along with our guide Dr(Mrs.) Mythili Thirugnanam without whom it would not have been possible to complete this mammoth task . This work has been (Partially) supported by the research program in SCOPE, VIT University, India.

CONFLICT OF INTEREST

No conflict of interest

REFERENCES

[1] MG Gouthamanaath, A.Kangaiammal, Ph.D," Color Image Steganography using Combined Pixel Value Differenncing and Pixel Indicator Technique in Spatial Domain, *",International Journal of Computer Applications* (0975 – 8887) National Conference Research Issues in Image Analysis and Intelligence (NCRIIAMI-2015).

[2] Agham Vinit, and Tareek Pattewar.[2014] Data hiding technique by using RGB-LSB mechanism. Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE,.

[3] Darshan R Prabhu, and M Divya . [2014]Acce-leration of LSB Algorithm in GPU.(2014).

[4] Bharti, Deeksha, and Archana Kumar.Enh- anced Steganography Algorithm to Improve Security by using Vigenere Encryption and First Component Alteration

[5] Mahimah P, and R Kurinji. [2013] Zigzag pixel indicator based secret data hiding method. " Computational Intelligence and Computing Research (ICCIC), 2013 *IEEE International Conference on IEEE*,.

[6] JyothiUpadhya K.U Dinesh Acharya, and S Hemalatha.[2013] Speed-Up Improvement Using Parallel Approach in Image Steganography.

[7] Fillatre Lionel. [2012] Adaptive Steganalysis of least significant bit replacement in grayscale natural images. Signal Processing, *IEEE Transactions* on 60.2: 556–569.

[8] Hong Wien, and Tung-Shou Chen. [2012]A novel data embedding method using Forensics and Security, *IEEE Transactions* on 7(1) 176–184.

[9] Sharma Sonia, and Anjali Dua.[2012]Design and Implementation of an Steganography Algorithm Using Color Transformation. IJRTE) *International Journal of Recent Technology and Engineering* 1(2)

[10] Amirtharajan Rengarajan, et al.[2012] Who decides hiding capacity? I, the pixel intensity. Recent Advances in Computing and Software Systems (RACSS), 2012 *International Conference on. IEEE*,

[11] Kaur Amanpreet, Renu Dhir, and Geeta Sikka."A new image Steganography basedon first component alternation technique". arXiv preprint arX iv:1001.1972(2010).

[12] Luo Weiqi, Fangjun Huang, and Jiwu Huang. [2010] Edge adaptive image steganography based on LSB matching revisited. Information Forensics and Security, *IEEE Transactions o*n 5(2): 201–214.

[13] Yang, Cheng-Hsing, et al. [2008]Adaptive data hiding in edge areas of images with spatial LSB domain systems." Information Forensics and Security, IEEE Transactions on 3(3): 488–497.

[14] E Hecht.[1987] Optics, 2nd Ed, Addison Wesley,.

ABOUT AUTHORS

Mr. Anil Sathyan is currently pursuing his M.Tech Computer Science and Engineering at VIT University, Vellore ,Tamil Nadu, India. He has completed his B.Tech CSE from GEC Thrissur, Calicut University, Kerala, India. He is a an active follower of FOSS and has research interests in the field of image processing, cryptography, mobile application development and also latest web technologies. He has done various research projects in the fields of image processing, data mining and parallel computing.

THE IIOAB JOURNAL
ISSN: 0976-3104

Dr. Mythili Thirugnanam is an Associate Professor in the School of Computing Science and Engineering at VIT University, Vellore, India. She received a Master's in Software Engineering from VIT University. She has been awarded doctorate in Computer Science and Engineering at VIT University in 2014. She has teaching experience of around 8 years. She has an research experience of 3 years in handling sponsored projects funded by Govt. of India. Her area of specialization includes Image Processing, Software Engineering and Knowledge Engineering. She has published nine papers in international journals and presented around seven papers in various national and international conferences

Sumit Hazra is currently pursuing his M. Tech in Computer Science and Engineering at VIT University , Vellore which is a Research based course. He has completed his graduation (B.Tech in CSE) from Guru Nanak Institute Of Technology affiliated to West Bengal University Of Technology, Vellore in the year of 2015. He has received meritorious scholarship from VIT University for being one of the top scorers in the 1st Semester exams of his M. Tech. He has successfully completed the trainings on Advanced C,C++ and Java courses and hence been certified by IIT ,Bombay for the same, funded by National Mission on Education through ICT, MHRD, Govt., of India. Also he has been certified by the Telecom Sector Skill Council , Government Of India, for having cleared successfully the assessment for the role of Customer Care Executive(Relationship Centre) conforming to National Skill Qualifications Framework Level-4.

ISSN: 0976-3104

SPECIAL ISSUE (SCMDSA)
Thirugnanam et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

HYBRID TOOL FOR DIAGNOSIS OF DIABETES

Mythili Thirugnanam[1*], Tamizharasi Thirugnanam[2], Sumathy S[3] and Swarnalatha P[4]

[1,2,4]School of Computing Science and Engineering, VIT University, Vellore, INDIA
[3]School of Information Technology and Engineering, VIT University, Vellore, INDIA

ABSTRACT

This paper presents a performance of various computational approaches for diagnosis of diabetes to predict the levels of diabetes risk with better accuracy.The proposed tool comprises of all the computational techniques for first level diagnosis of diabetes. Rule based approach is applied for the results obtained from the first level diagnosis to categorize the risk level of patients. The significance of this paper is the data used in the training phase which is obtained from huge number patient's data. Based on the observation of patient details some of the influenced parameter for diabetes diagnosis was identified. The morality of the diagnosis of diabetes is also considered to reduce the percentage of inaccurate prediction. The accuracy of the prediction rate for diagnosing diabetes was found to be 95%.

Received on: 30th-Nov-2015
Revised on: 22nd-Feb-2016
Accepted on: 21st– Mach-2016
Published on: 18th–May-2016

KEY WORDS

Fuzzy Approach; Neural Approach; Case Based Reasoning; Rule Based Approach; Diagnosis; Computational techniques.

***Corresponding author:** Email: tmythili@vit.ac.in. **Tel:** +91-9042857554

INTRODUCTION

Diabetes mellitus is a metabolic disease in which a person has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. Diabetes is classified into three types; Type 1, Type 2 and Gestational diabetes. Since diabetes has become a major health problem among people of all ages, diagnoses of this disease is important. Most of the people are unaware of the symptoms caused by it and how to diagnose this disease. So there is a need to develop a tool that would help people to diagnose this disease in the early stages thereby reducing the number of people affected by it. This paper presents the performance analysis for diagnosis of diabetes which has two stages to predict the diabetes status They are divided into initial prediction stage and final prediction stage.

Fuzzy logic is a mathematical model that gives approximate values rather than fixed or exact values. It shows the truth value that ranges between 0 and 1. K.Rajeswari et al., (2011) [1] discussed fuzzy model for diabetic diagnostic decision support. Neural network is a computation model reduces the amount of computation required. Neural helps to train large amount of data very easily and has usages in artificial intelligence, image analysis, and diagnosis of diabetes. Sumathy et al., (2010) [2] proposed a method that diagnosed diabetes based on risk factors that used Artificial Neural Network (ANN) architecture for classification which had a supervised multilayer feed forward network with back propagation learning algorithm. Case Based Reasoning (CBR) is an easy approach that helps in solving new problems based on the solutions of similar past problems. It can be easily updated without altering other parts. Liping Zheng et al., (2011) [3] described a medical aided diagnosis system and maxillofacial diseases (OMD-MADS) for diagnosis of oral and maxillofacial disease through the usage of ontologies.

RELATED WORK

The survey below clearly highlights that the previously developed systems had number of issues unaddressed and also the prediction rate was not very significant. Some pioneers concentrated only on certain aspects of the occurrence of diabetes leaving few crucial factors that were important.

THE IIOAB JOURNAL ISSN: 0976-3104

S.No	Title	Author's name	Techniques used	Accuracy obtained	Pros	Cons
1	Fuzzy based modeling for diabetic diagnostic decision support using Artificial Neural Network	K.Rajeswari and V.Vaithiyanathan [1]	Fuzzy approach and Artificial neural networks	Not mentioned	Highly efficient with good accuracy support for classification and further analysis.	Works on the real-time dataset
2	Diagnosis of Diabetes Mellitus based on Risk Factors	Sumathy, Mythili Thirugnanam, Praveen Kumar, Jishnujit T M, K Ranjith Kumar [2]	Artificial Neural Networks (ANN	99%	Better results, diagnosing other diseases like coronary artery disease, hypertension.	Input values should be normalized before giving to the network
3	The Design and Implementation of Oral Disease Aided Diagnosis System	Liping Zhengl, Guangyao Li and Junqing Li[3]	VTK and uses Add-In tree	Not mentioned	Clinical diagnosis and taken as an instruction tool.	Some functions are imperfect.
4	Application of Modeling Techniques to Diabetes Diagnosis	A.M. Aibinu, M. J. E. Salami and A. A. Shafie[4]	Complex-valued neural networks (CVNN) and real-valued neural network (RVNN)	72.83%	The results produced by ANN-AR are better	Only for females
5	Diagnosis of Diabetes by using Adaptive Neuro Fuzzy Inference	AdemKarahoca, Dilek Karahoca and Ali Kara [5]	Adaptive Neuro Fuzzy Inference System , Multinomial non-linear regression	Not mentioned	Standard error of ANFIS was smaller, better system than MLR	MLR is not a good system for diabetes diagnosis
6	A Fuzzy Expert System for Heart Disease Diagnosis	Ali.Adeli, Mehdi.Neshat [6]	Fuzzy logic	94%	Results logical and efficient	Not mentioned
7	An Ontology-Based Electronic Medical Record for Chronic Disease Management	Ashraf [7]Mohammed Iqbal, Michael Shepherd and Syed SibteRazaAbidi	Electronic Medical Records (EMRs) ontology, Description Logic representation	Not mentioned	Capture clinical records, treatment of acute diseases	Medication or immunization status could not be captured
8	Decision Tree Discovery for the Diagnosis of Type II Diabetes	Asma A. AlJarullah [8]	Data Mining	78.1768%	Increases diagnostic accuracy, reduce costs and reduces human resources	Datasets themselves must be available.
9	A Fuzzy Expert System for Diabetes Decision Support Application	Chang-Shing Lee, and Mei-Hui Wang [9]	Fuzzy knowledge layer, fuzzy group relation layer, fuzzy group domain layer, fuzzy personal relation layer, and fuzzy personal domain layer fuzzy diabetes ontology (FDO)	Between 73.5%- 91.2%.	Analyse Data and further transfer the acquired information into the knowledge to Simulate the thinking process of humans.	Use of only one data set.
10	A Knowledge-based Clinical Decision Support System for the diagnosis of Alzheimer Disease	Eider Sanchez, Carlos Toro , Eduardo Carrasco, Patricia Bonachela , Carlos Parra ,Gloria Bueno and Frank Guijarro[10]	Knowledge Engineering (KE)-semantic technologies and web inspired paradigms	Not mentioned	Applied to other domains such as cardiologic diseases or autism, as well as extended to other purposes such as the treatment and monitoring of patients high adaptability, robustness and reasoning capabilities	Validation not performed, results not published.
11	A Diagnostic Fuzzy Rule-Based System for Congenital Heart Disease	Ersin Kaya, Bulent Oran and Ahmet Arslan [11]	Weighted vote method and singles winner method	Not mentioned	Weighted vote method generally increased the classification accuracy of	Not mentioned

IIOAB JOURNAL ISSN: 0976-3104

					Congenital Heart Diseases.	
12	Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease	Mostafa Fathi Ganji and Mohammad Saniee Abadeh [12]	Ant colony optimization (ACO), Fuzzy Logic	Not mentioned	Good Comprehensibility	The rules are learned for each class Independently
13	Detection of diabetic retinopathy using radial basis function	Vijayamadheswaran, Dr.M.Arthanari and M.Sivakumar [13]	Contextual clustering and Radial basis function (RBF) network.	96%	Presence of exudates is identified more clearly, effectiveness of RBF	Fundus image is taken with good quality,

Table: 1. This Literature Survey

The suggested systems had many disadvantages like some systems required dataset of very high quality, some gave accurate prediction rate only if one dataset was used, some were designed only for females, some only for females whose age was lesser than 21, some failed to differentiate certain types of diabetes with the other types and some important functions of the system were not ideal. In order to overcome these shortcomings a new system is developed that improves the prediction rate and at the same time considers various factors into account.

PROPOSED FRAMEWORK

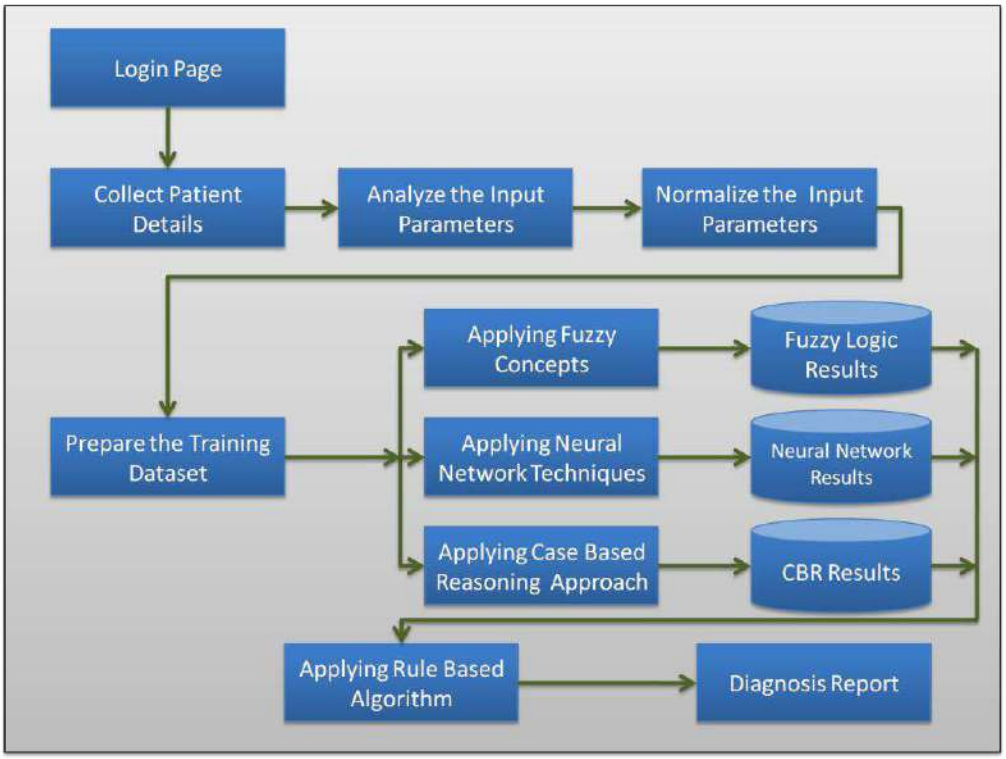


Fig: 1. Architectural Framework for FNC Tool

The architectural framework for the developed tool is shown in **Figure- 1**. The first step is the login page where the user gets access to the system. The significance of this paper is the data which is used in the training phase. The data in the training phase has 16 influenced input attributes that has been used under the expert advice of the doctor. For the normalization purpose all the values for the attributes have been assigned either '0' or '1' based on whether it is present or absent. The table below shows the normalized values for all the 16 influenced input attributes.

THE IIOAB JOURNAL
ISSN: 0976-3104

Input Field	Range	Fuzzy Sets
Age	<45 Years >45 Years	Low(0) High(1)
Gender	Female Male	Low(0) High(1)
Family background (whether your parents or brothers or sisters have been diagnosed with diabetes)	No Yes	Low(0) High(1)
Are you currently taking medicine for high blood pressure?	No Yes	Low(0) High(1)
High blood glucose during illness	No Yes	Low(0) High(1)
Smoking or using tobacco products	No Yes	Low(0) High(1)
Vegetable or fruit intake	Everyday Not Everyday	Low(0) High(1)
Body Mass Index	<24 >=24	Low(0) High(1)
Waist Hip ratio	<0.8 >=0.8	Low(0) High(1)
Increased urination, hunger, thirst	No Yes	Low(0) High(1)
Poor Wound Healing	No Yes	Low(0) High(1)
Lifestyle	Labour Class Sedentary Work and Retired People	Low(0) High(1)
Gestation Diabetes(Applicable for reproductive females)	No Yes	Low(0) High(1)
Frequent intake of non-veg.	No Yes	Low(0) High(1)
Itching	No Yes	Low(0) High(1)
Physical Activity	Everyday Not Everyday	Low(0) High(1)

Table: 2. Normalization table for influenced input parameters

This paper focuses on the performance of the computational techniques such as fuzzy logic, neural network and case based reasoning. The last step is applying rule based algorithm to the above obtained results. Fuzzy logic involves framing of rules based on the input attributes. Neural network performs training of the data set. Case based checks the similarity measures. Rule based algorithm is based on the if-then rules. Fuzzy and neural network are implemented using matlab. Case based is implemented using protégé by invoking MyCBR plugins. Rule based algorithm is implemented using java .Finally the diabetes report shows the accuracy of the techniques used and also approach is the best.

Back Propagation algorithm

ISSN: 0976-3104

Back Propagation can be considered as a generalization of the delta rule for non-linear activation functions and multi-layer networks. It is a systematic method of training multi-layer artificial neural networks. Hidden layer allows artificial neural network to develop its own internal representation of input-output mapping. The algorithm is as follows

1. First apply the inputs to the network and work out the output – this initial output could be anything, as the initial weights were random numbers.

Next is working out the error for the neuron B. The error is found using the below equation:

$$Error_B = Output_B\ (1\text{-}Output_B)(Target_B - Output_B)$$

2. The "Output (1-Output)" term is necessary in the equation because of the Sigmoid Function–if we were only using a threshold neuron it would just be (Target – Output).

3. Change the weight. Let W^+_{AB} be the new (trained) weight and WAB be the initial weight

$$W^+_{AB}=W_{AB}+ (Error_B*Output_A)$$

All the weights are updated in the output layer in this way.

4. Calculate the Errors for the hidden layer neurons.

Unlike the output layer we can't calculate these directly (because we don't have a Target), so we Back Propagate them from the output layer (hence the name of the algorithm). This is done by taking the Errors from the output neurons and running them back through the weights to get the hidden layer errors. For example if neuron A is connected as shown to B and C then we take the errors from B and C to generate an error for A.

$$Error_A = Output_A(1 - Output_A)(Error_B\ W_{AB} + Error_C\ W_{AC})$$

Again, the factor "Output (1 - Output)" is present because of the sigmoid squashing function.

5. Having obtained the Error for the hidden layer neurons now proceed as in stage 3 to change the hidden layer weights. By repeating this method we can train a network of any number of layers. **Figure–2** shows graph for trained dataset.

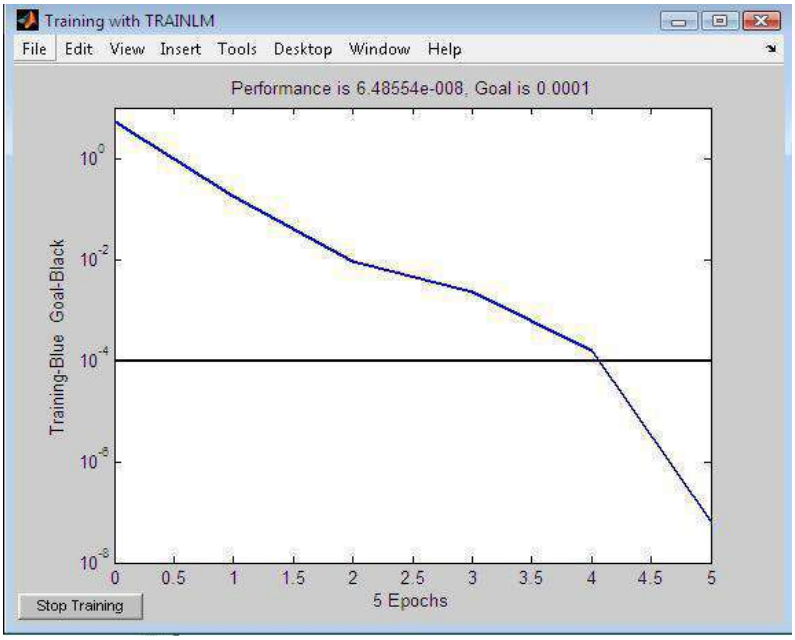


Fig: 2. Graph for Trained Dataset

Designing fuzzy expert system

THE IIOAB JOURNAL
ISSN: 0976-3104

Since fuzzy logic is a mathematical model, its concepts are implemented by preparing the fuzzy set, fuzzy expert system and defining decision rules. The concepts are briefed below.

The input attributes and their range values are specified in the fuzzy tool of Matlab. **Figure –3** shows the framed fuzzy rules. **Figure– 4** shows the fuzzy rule viewer.

Fig: 3. Fuzzy Rule Editor

Case based Reasoning Approach

In the class tab classes are created with their respective slot values. Data sets are stored in the Instance tab. MyCBR plug-in shows the output through user queries. The query results are shown below with their similarity measure values.

The algorithm for Case Base Reasoning is as follows

1. Create main class and sub-class.
2. Give Slots i.e. attributes of the class or sub-class and assign their respective values
3. Create Instance of sub-class or main class and store it into the database. Thus Ontology is created.
4. Now for the CBR, get the query from the user and store into a separate database
5. For the Similarity to be calculated, compare the instance values and user query values using Euclidean Distance Formula.

$$D= (x_2-x_1)^2+ (y_2-y_1)^2+(z_2-z_1)^2$$

THE IIOAB JOURNAL
ISSN: 0976-3104

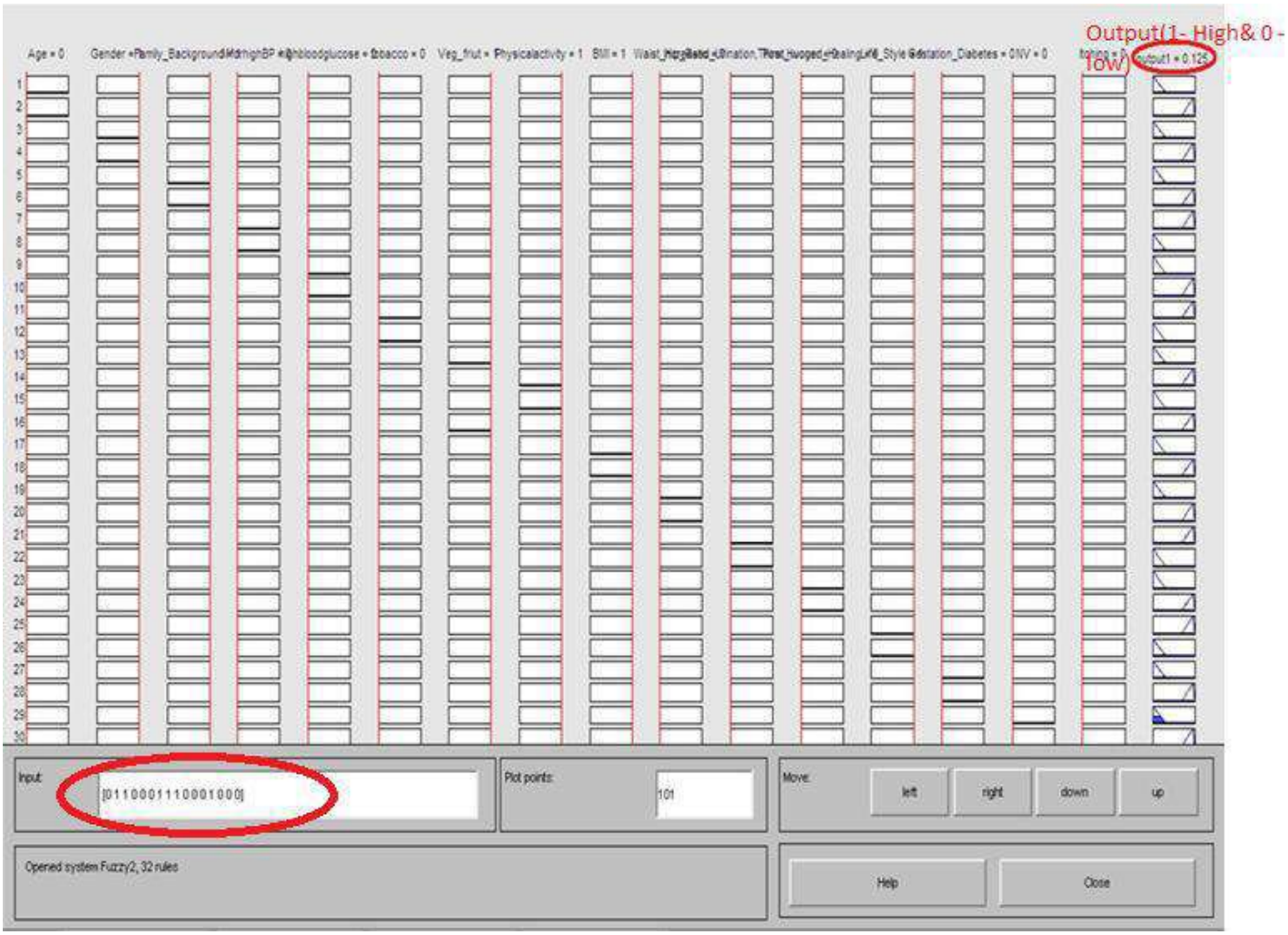


Fig: 4. Fuzzy Rule Viewer

6. Assign the weight value which is based on the result of Euclidean Distance.

If (D is 0){ Weight=1 }

Else their respective distances are shown

7. Similarity measures are calculated and the results are shown in **figure– 5**.

THE IIOAB JOURNAL
ISSN: 0976-3104

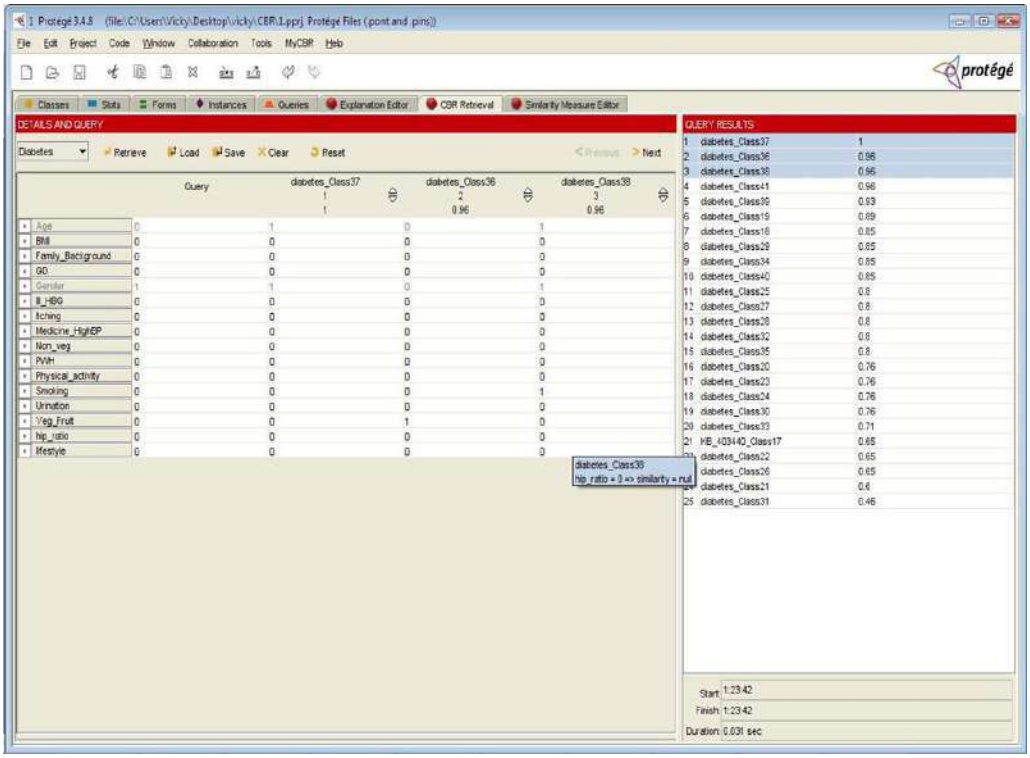


Fig: 5. Similarity measure values for CBR

RESULT ANALYSIS

For the proposed approach we have used 200 cases .The sample test cases for one patient with the output is elaborated in the table below.

Attributes	Value
Age	0
Gender	1
Family_Background	1
Taking Medicine for High Bp	0
High_Blood_Glucose illness	0
Smoking	0
Vegetable Fruit Intake	1
Physical_activity	1
BMI	1
Hip_Ratio	0
Increased_Urination	0
PWH	0
Lifestyle	1
Gestational Diabetes	0
Non_veg Intake	0
Itching	0
Fuzzy_Output	Low(0.125)
Neural_Output	Medium(0.40029)
Case_Based_Output	Medium(Class_18)
Rule_Based_Output	Medium

Table: 3. Test case for sample data

ISSN: 0976-3104

We tried to improve the prediction rate of diabetes mellitus by indicating the diabetes risk. The experimental environment was constructed to evaluate the performance of the proposed approach; in addition, Rule Based algorithm was applied to the results obtained from the FNC approach to increase the accuracy of the prediction rate.

For the neural network approach back propagation algorithm was applied. Here the data set is trained using activation function and the results are obtained. The sample test case for one patient is provided in the table above. The prediction rate obtained through this approach for the sample test case is medium (0.40029)

Next fuzzy expert system for diabetes diagnosis was designed with membership functions, input variables, output variables and rule base. Designed system has been tested with expert-doctor. Designing of this system with fuzzy base in comparison with classic designed improves results. Results have been shown from this system in compression with past time system are logical and more efficient. This system simulates the manner of expert-doctor. This system is designed in way that patient can use it himself. For this purposes the same test case is used and then results obtained is low (0.125)

The third approach is case based reasoning. For this purpose CBR approach is divided into two phases such as information gathering phase and query processing phase. The information gathering phase has the following processes knowledge acquisition, creation of ontology and trained dataset. The query processing phase has the following processes such as use, query and CBR result. In the information gathering phase we need to create classes and their hierarchies. Next values are assigned to the respective attributes, thereby creating the trained dataset. In the query processing phase user requests a query which is then compared with trained dataset of CBR. After comparing related results are obtained. The similarity measure obtained for the test cases is medium (Class_18)

In addition to the above mentioned approaches we have used another algorithm called rule based algorithm. This algorithm improves the accuracy of prediction rate to a greater level. This algorithm is applied to the results obtained from the FNC approaches. It also shows which approach provides the most accurate value. The rule based algorithm evaluates the results obtained and the final outputs for the two test cases are medium and high respectively. The best approach used is Case based approach as the accuracy of prediction is the best.

This paper presents a mixture of fuzzy logic, neural network and CBR approaches for prediction of diabetes. The developed system helps in diagnosing diabetes mellitus. With the literature survey performed and the analysis carried over, the developed system would serve as a better method for diabetes diagnosis. To conclude, we have developed a new approach called FNC approach for diagnosing diabetes by using newly designed influenced inputs parameters. The system will ease the patients undergoing medical tests for diagnosing this disease without consulting a doctor thus helping the patients to take precautionary measures well in advance.

After completing the implementation using the FNC approach, rule based algorithm is used to improve the accuracy of the prediction rate. For the fulfillment of the developed system more than 150 cases were tested. The significance of applying the three approaches is that even if one of the three approaches fail, the other two approaches would predict the occurrence of the risk level. The rule based algorithm (hybrid approach) would further provide information regarding which of the three approaches gives thebest result and it was found to be CBR.

CONCLUSION

This paper presents a mixture of fuzzy logic, neural network and CBR approaches. The developed system would serve as a better method for diabetes diagnosis. To conclude, we have designed a new tool called FNC tool for diagnosing diabetes by using newly designed influenced inputs parameters. The system eases the patients undergoing medical tests for diagnosing this disease without consulting a doctor thus helping the patients to take precautionary measures well in advance. The developed system will predict which one of the three approaches would give a better prediction rate for diagnosis of diabetes.

.

THE IIOAB JOURNAL
ISSN: 0976-3104

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

We would like to thank Dr. Praveen Kumar for providing us with the dataset and giving his expert advice on the attributes necessary for framing of the fuzzy rules.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] K.Rajeswari and V Vaithiyanathan.[2011] Fuzzy based modeling for diabetic diagnostic decision support using Artificial Neural Network, *IJCSNS International Journal of Computer S 126 science and Network Security*,.11: 126-130.

[2] Sumathy, MythiliThirugnanam, Praveen Kumar, Jishnujit TM, K Ranjith Kumar.[2010] Diagnosis of Diabetes Mellitus based on Risk Factors", *International Journal of Computer Applications*, 10(.4): 1-4.

[3] LipingZheng, Guangyao Li and Junqing Li.[2011] The Design and Implementation of Oral Disease Aided Diagnosis System, *Journal of Next Generation Information Technology*, 2: 38-44.

[4] AM Aibinu, MJE Salami and AA Shafie.[2010.] Application of Modeling Techniques to Diabetes Diagnosis, *IEEE Trans Biomed Engineering*, 194-198.

[5] AdemKarahoca, DilekKarahoca and Ali Kara.[2009] Diagnosis of Diabetes by using Adaptive Neuro Fuzzy Inference, *IEEE Trans*,

[6] Ali.Adeli, Mehdi.Neshat. [2010] A Fuzzy Expert System for Heart Disease Diagnosis, IMECS, vol 1.

[7] Ashraf Mohammed Iqbal, Michael Shepherd and Syed SibteRazaAbidi.[2011] An Ontology-Based Electronic Medical Record for Chronic Disease Management, Proceedings of the 44th Hawaii International Conference on System Sciences, pp 1-10.

[8] Asma A, AlJarullah,[2011] Decision Tree Discovery for the Diagnosis of Type II Diabetes, International Conference on Innovations in Information Technology, 303-307.

[9] Chang-Shing Lee, and Mei-Hui Wang.[2011] A Fuzzy Expert System for Diabetes Decision Support Application, *IEEE Trans.Cyberrnetics.*, 41: 139-153.

[10] Eider Sanchez, Carlos Toro , Eduardo Carrasco, Patricia Bonachela , Carlos Parra ,Gloria Bueno and Frank Guijarro.[2011],A Knowledge-based Clinical Decision Support System for the diagnosis of Alzheimer Disease", 13th International Conference on e-Health Networking, *Applications and Services*,.351-357.

[11] Ersin Kaya, Bulent Oran and AhmetArslan.[2011] A Diagnostic Fuzzy Rule-Based System for Congenital Heart Disease, World academy of science, *Engineering and technology*, 253-256.

[12] MostafaFathiGanji and Mohammad SanieeAbadeh," Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease, *Proceedings of ICEE*, 2010.

[13] R Vijayamadheswaran, .M Arthanari and MSivakumar.[2011] Detection of diabetic retinopathy using radial basis function", *International Journal of Innovative Technology & Creative Engineering*, 1: 40-47.

ABOUT AUTHORS

Dr. Mythili Thirugnanam is an Associate Professor in the School of Computing Science and Engineering at VIT University, Vellore, India. She received a Master's in Software Engineering from VIT University. She has been awarded doctorate in Computer Science and Engineering at VIT University in 2014. She has teaching experience of around 8 years. She has an research experience of 3 years in handling sponsored projects funded by Govt. of India. Her area of specialization includes Image Processing, Software Engineering and Knowledge Engineering. She has published nine papers in international journals and presented around seven papers in various national and international conferences.

THE IIOAB JOURNAL
ISSN: 0976-3104

Prof. Tamizharasi Thirugnanam is an Assistant Professor in the School of Computing Science and Engineering at VIT University, Vellore, India. She received a Master's in Networking from VIT University. She has teaching and research experience of around 3 years. Her area of specialization includes Networking and Knowledge Engineering. She has published three papers in international journals.

Dr. Sumathy S is an Associate Professor in the School of Information and Technology Engineering at VIT University, Vellore, India. She received a Master's in Computer Science and Engineering from VIT University. She has been awarded doctorate in Information Technology and Engineering at VIT University in 2015. She has teaching and research experience of around 18 years. Her area of specialization includes Networking and Operating System She has published more than 15 papers in international journals and presented around seven papers in various national and international conferences.

Dr. Swarnalatha Purushotham is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She Pursued her Ph.D in Image Processing and Intelligent Systems. She has published more than 50 papers in International Journals/International Conference Proceedings/National Conferences. She is having 14+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.

ISSN: 0976-3104

SPECIAL ISSUE (SCMDSA)
Soundrapandiyan et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

MULTIMODALITY MEDICAL IMAGE FUSION USING BLOCK BASED INTUITIONISTIC FUZZY SETS

Rajkumar Soundrapandiyan*, Rishin Haldar, Swarnalatha Purushotham and Arvind Pillai

School of Computing Science and Engineering, VIT University, Vellore, INDIA

ABSTRACT

Image fusion combines more than one image from various environments into a single image. This can be useful for subsequent processing of the image, especially in medical imaging where it can help in disease diagnosis. This paper uses the block based Intuitionistic Fuzzy Sets(IFS) to fuse the multimodality medical images. IFSs can effectively handle the inherent uncertainties of digital images. Initially, in this model, entropy is used to deduce the optimal parameter value for defining the membership and non-membership function. This, in turn generates the Intuitionistic Fuzzy Images (IFI) from the original image. Finally, the IFIs are partitioned into image blocks and then recombined by the generated membership function. This paper compares the proposed method with popular ones like Principal Component Analysis (PCA), simple averaging (AVG), Laplacian Pyramid Approach(LPA), Discrete Wavelet Transform (DWT) and MPA (Morphological Pyramid Approach) on various performance measures such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM), Universal Image Quality Index (UIQI), Mean and Standard Deviation (STD). The experimental results show better image visualization generated through the proposed method compared to the other methods, in overall.
.

Received on: 28th-Nov-2015
Revised on: 28th-Feb-2016
Accepted on: 31st– Mach-2016
Published on: 19th–May-2016

KEY WORDS

Medical image fusion; Intuitionistic fuzzy image; Entropy; Quantitative measures; Multimodality images.

***Corresponding author:** Email: rajkumars@vit.ac.in

INTRODUCTION

DNA microarray Image fusion is widely used as an effective technique for analysis of images [1]. These images are obtained from various domains like satellite images, biometrics, robotics, remote sensing etc., and there are customised image sensors for each of these domains. Consequently, the data obtained from these specialised sensors may be incompatible with each other. For example, in medical imaging, the image generated by an MRI machine gives clear details of soft tissues while a CT (X-Ray) machine gives clear details of bone structures. In this scenario, if we are required to find the clear details of both, or more, of the features, where the data is incompatible, image fusion can be an effective tool to address the issue. This gives us the motivation to apply image fusion on medical images.

Image fusion can be carried out by mainly two techniques, spatial fusion and transform fusion. Based on the unification phases, fusion can be done in three levels, namely pixel, feature and decision levels. Pixel level fusion combines the pixel values directly and creates a composite image. The simplest method just takes the average of the pixel values of source images. Laplacian pyramids [2], PCA [3] are some of the other techniques which use pixel value fusion. In order to improve upon the degraded performance of the average policy of fusion algorithm, many multi-resolution transform techniques emerged, like pyramid decomposition, wavelet transforms [4] etc. Fusion of the images by singular value decomposition (SVD) [5] works quite well on pixel basis and outperforms PCA. The MSVD [6] technique, which looks into multiple properties like sphericity, isotropy and self-similarity of signals, performs faster than SVD. Image fusion technique also generated a highly featured picture using multi-scale decomposition. The various attempts in using multi-scale transform showed that shifting of invariance is highly desirable for image fusion. In this context, NSCT [7], a complete transform, has been very effectively utilized in image fusion.

THE IIOAB JOURNAL
ISSN: 0976-3104

Image processing, however, has many uncertainties at every phase. Fuzzy sets [8] have been known to remove these uncertainties, especially in luminance and contrast of the image. In medical images, poor luminance increases the uncertainty of the image, and IFS [9], an improvement on the traditional fuzzy set, has been quite successful in removing these uncertainties. Thus, by using the multimodal properties of the image as well as using fuzzy sets, the image fusion can be extremely effective. One paper [10] uses Intuitioinstic Fuzzy Sets on multimodal images to fuse the images, and the results were very encouraging.

This paper presents a new way to fuse more than one medical image, and builds on the efforts done [10]. This paper also uses the block based Intuitionistic Fuzzy Sets (IFS) to fuse the multimodal medical images. However, a new and customised entropy function is used to deduce the optimal parameter value for defining both the membership and the non-membership function. This generates the Intuitionistic Fuzzy Images (IFI) from the original images. Finally, the IFIs are partitioned into image blocks and then recombined by our generated membership function. The reconstructed fused image has high degree of luminance and contrast. The resultant pictures are provided for subjective evaluation. This paper also objectively compares the proposed method with popular methods like simple averaging (AVG) [11], Principal Component Analysis (PCA) [3], Laplacian Pyramid Approach(LPA) [2], Discrete Wavelet Transform (DWT) [12] and MPA (Morphological Pyramid Approach) [13] by various performance measures such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM), Universal Image Quality Index (UIQI), Mean and Standard Deviation (STD). The experimental results are very encouraging and show that the proposed method, overall, has performed much better than these popular methods.

The following sections give the specific details of our work. Section II describes our proposed methodology along with the required computational models. Section III describes the performance measures through which we are evaluating our proposed technique. Section IV describes the experimental results and its subjective and objective comparison with the other popular methods. Finally, we conclude in Section V.

PROPOSED METHODOLOGY

The block diagram of our proposed method is shown in Fig. 1. The individual steps carried out in our proposed method, are as follows:

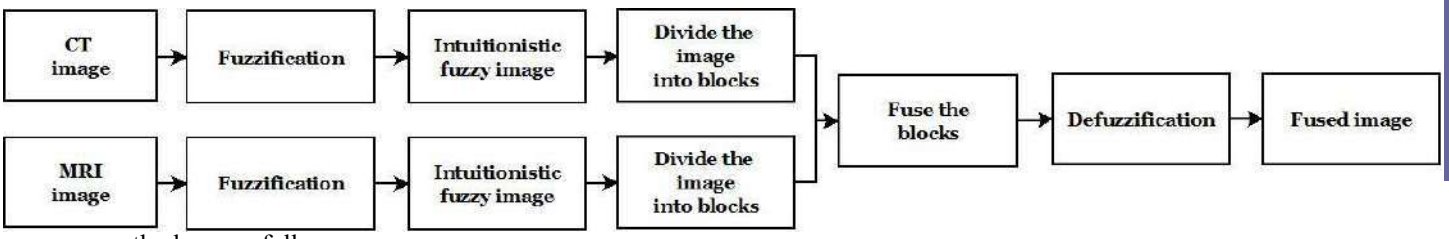


Fig: 1. Block diagram of the proposed method

1. Read/Accept the input images. There are six datasets of images, each of size 256x256 pixels.
2. Fuzzification of input images using Equation (1).
3. Generation of intuitionistic fuzzy image using Equation (8).
4. Divide the image into blocks of size 3x3.
5. Fuse the each block based on the value of entropy using Equation (9).
6. Defuzzification of the fused image using Equation (10).

Fuzzification

Fuzzification [14] is the first step of fuzzy image processing**.** It consists of converting the image from spatial domain into the fuzzy domain. It can be defined as

$$\mu_{ij} = T\left(x_{ij}\right) = \left[1 + \frac{x_{\max} - x_{ij}}{F_d}\right]^{-F_e} \quad (1)$$

THE IIOAB JOURNAL ISSN: 0976-3104

Here x_{max} is the maximum intensity level for the given input image; F_e and F_d represent the exponential and denominational fuzzifiers, respectively. When $x_{max} = x_{ij}$ then, $\mu_{ij} = 1$ indicating the maximum brightness. Fuzzifier F_d is calculated using Equation (2) and F_e *is assigned to constant value 2.*

$$F_d = \frac{x_{\max} - x_{\min}}{\left(\frac{1}{2}\right)^{-1/x_{\min}} - 1} \quad (2)$$

2.2. Intuitionistic Fuzzy Image (IFI)

In general, pixel values of images have ambiguity and uncertainty. However, some uncertainty still remains while specifying the brightness of image pixels. The main objective of the proposed method is to remove the ambiguity in those image pixels. To address this issue, the image is converted from fuzzy domain to intuitionistic fuzzy domain. The intuitionistic fuzzy domain has an additional property of degree of hesitation compared to fuzzy domain. The hesitation degree is used to align the membership function values within a range. This can effectively remove the uncertain gray level values of ambiguous image pixels [10]. An Intuitionistic Fuzzy Set (IFS) is expressed in terms of □ (membership degree), □ (non-membership degree) and □ (hesitation degree) on a finite set X [19,22], by

$$IFS = \{(x, \mu_{IFS}(x), \gamma_{IFS}(x), \pi_{IFS}(x)) | x \in X\} \quad (3)$$

Based on the Equation (1), the degree of the membership function of IFI is computed as

$$\mu_{IFS}(x_{ij};\lambda) = 1 - (1 - \mu_x(x_{ij}))^{\lambda}, \quad \lambda \geq 0 \quad (4)$$

The degree of the non-membership function is computed as

$$\gamma_{IFS}(x_{ij};\lambda) = (1 - \mu_x(x_{ij}))^{\lambda(\lambda+1)}, \lambda \geq 0 \quad (5)$$

The degree of hesitation is defined as $\pi_{IFS}(x_{ij};\lambda) = 1 - \mu_{IFS}(x_{ij};\lambda) - \gamma_{IFS}(x_{ij};\lambda)$ (6)

The parameter varies for each image. To select the single unique optimum □ value from each image, entropy (ENT) used. The entropy is defined

$$ENT(IFS;\lambda) = \frac{1}{P \times Q} \sum_{i=0}^{P-1} \sum_{j=0}^{Q-1} \frac{2\mu_{IFS}(x_{ij};\lambda)\gamma_{IFS}(x_{ij};\lambda) + \pi^2_{IFS}(x_{ij};\lambda)}{\mu^2_{IFS}(x_{ij};\lambda) + \gamma^2_{IFS}(x_{ij};\lambda) + \pi^2_{IFS}(x_{ij};\lambda)} \quad (7)$$

In Equation (7), the value of □ corresponds to the highest value of entropy. Finally, the IFI is defined as

$$x_{IFS} = \{(x_{ij}, \mu_{IFS}(x_{ij};\lambda), \gamma_{IFS}(x_{ij};\lambda), \pi_{IFS}(x_{ij};\lambda))\} \; x_{ij} \in \{0,1,....., L-1\} \quad (8)$$

Entropy based image fusion

To fuse the images, the obtained resultant image of x^{F1} and x^{F2} from Equation (8) is decompose into m X n blocks and denote the T^{th} image block of two decomposed images by x^{F1T} and x^{F2T} respectively. The entropy based fusion process is defined as

$$x_{ij}^{fuse} = \begin{cases} \min(x_{ij}^{F1T}, x_{ij}^{F2T}) \; if \; ENT(x_{ij}^{F1T}) > ENT(x_{ij}^{F2T}) \\ \max(x_{ij}^{F1T}, x_{ij}^{F2T}) \; if \; ENT(x_{ij}^{F1T}) < ENT(x_{ij}^{F2T}) \\ \frac{x_{ij}^{F1T}}{x_{ij}^{F2T}} \qquad otherwise \end{cases} \quad (9)$$

THE IIOAB JOURNAL ISSN: 0976-3104

where max and min represent the maximum and minimum operations in IFS.

Defuzzification

Equation (10) expresses the defuzzification process to convert the image from fuzzy domain [23] to the spatial domain

$$F(i,j)=T^{-1}\left(x_{ij}^{fuse}\right)=x_{\max}-F_d*\left(\left(x_{ij}^{fuse}\right)^{-1/F_e}\right)+F_d \quad (10)$$

where F(i,j) is the final fused image.

EVALUATION MEASURES

The measurement and analysis on the fused images are done both objective as well as subjective quality measures. This effectively helps in better assessment of the information in the images. For the subjective measure, pictorial representations of the images are provided. For the objective analysis, the following measures are used. In all the measures defined here, R_{ij} and F_{ij} represent the intensity value of the reference (original) image and the fused image at coordinates i, j respectively and P, Q denote the width and height of the image.

Root Mean Square Error (RMSE)

It is a [21] method to measure the differences between values predicted by an ideal (reference) image and the fused images. It is calculated as

$$RMSE = \sqrt{\frac{1}{P\times Q}\sum_{i=1}^{P}\sum_{j-1}^{Q}(R_{ij}-F_{ij})^2} \quad (11)$$

RMSE for the reference and fused images will increase with decrease in similarity, and approaches zero whenever they are similar.

Mean Absolute Error (MAE)

It is a method which measures the mean of the absolute error between the reference and fused images.

$$MAE=\frac{1}{P\times Q}\sum_{i=1}^{P}\sum_{j=1}^{Q}\left|R_{ij}-F_{ij}\right| \quad (12)$$

MAE also increases with decrease in similarity between reference and fused images and vice versa.

Peak Signal to Noise Ratio (PSNR)

PSNR [15] is a method used to measure the quality of the fused image with respect to the reference image. It is defined as:

$$PSNR=10\log_{10}\left(MAX^2/MSE\right) \quad (13)$$

$$MSE=\frac{1}{pq}\sum_{i=0}^{p-1}\sum_{j=0}^{q-1}\left[R(i,j)-F(i,j)\right]^2 \quad (14)$$

where MAX is the maximum value in an image and MSE is the mean square error value of the image.

Structural Similarity Index (SSIM)

It provides a way to measure the similarity between the two images. SSIM is an improved version of the peak signal to noise ratio [16]. It is defined as

THE IIOAB JOURNAL ISSN: 0976-3104

$$SSIM = \frac{((2\mu_F\mu_R + C_1) * (2\sigma_{FR} + C_2))}{((\mu_F^2 + \mu_R^2 + C_1) * (\sigma_F^2 + \sigma_R^2 + C_2))} \quad (15)$$

where μ_F and μ_R denote the average intensities of image F and R, σ_F and σ_R denote the variance of image F and R, σ_{FR} gives the covariance of F and R , C_1 and C_2 are constants. The SSIM index value varies from -1 to 1. When two images are identical, this value will turn out to be 1.

Universal Image Quality Index (UIQI)

It is a method to measure the quality of the images [14]. This quantifies the amount of data that has been transferred from the ideal image to the resultant fused image. UIQI defines image distortion by a combination of three factors, namely contrast distortion, loss of correlation and luminance distortion.

$$UIQI = \frac{(4 * \sigma_{RF})(\mu_R + \mu_F)}{(\mu_R^2 + \mu_F^2)(\sigma_R^2 + \sigma_F^2)} \quad (16)$$

where σ_{RF} is the covariance of RF, μ_F and μ_R denote the average intensities of image F and R, σ_F^2 and σ_R^2 denote the variance of image F and R. The UIQI index value varies from -1 to 1. Once again, a 1 indicates the identical nature of the two images.

Mean (MEAN)

The mean intensity estimates the luminance of an image. This is deduced by

$$MEAN = \frac{1}{P \times Q} \sum_{i=1}^{P} \sum_{j=1}^{Q} |F_{ij}| \quad (17)$$

Standard Deviation (SD)

It shows the extent of variation or dispersion from the average or mean [17,20]. Standard deviation takes into account the original image and the acquired transmission noise. Absence of any noise in the transmitted image increases its effectiveness and portraits the image's contrast. SD can be calculated as

$$SD = \sqrt{\frac{1}{P \times Q} \sum_{i=1}^{P} \sum_{j=1}^{Q} (F_{ij} - MEAN)^2} \quad (18)$$

RESULTS AND PERFORMANCE EVALUATION

The experimental results of the fusion techniques are analyzed with six brain images taken from CT and MRI (T2). Each CT image, combined with T2, are considered as one set for fusion. This, in turn, totally derives six combinations of input dataset. All images have the same size of 256 * 256 pixels, with 256-level gray scale.

Subjective evaluation of results

Figure-2 gives the subjective comparison of the results from average method, PCA method, Laplacian method, DWT method, MPA method and proposed method. Fig.3. evident that the proposed method generated results with good visualization (i.e. high luminance and contrast) than other existing methods.

Performance Evaluation

For the objective measures, the measures discussed in the section III are used. The results generated from the proposed method for each of the measures used to quantify the results, are compared with the average method, PCA method, laplacian method, DWT method and MPA method. The comparative analyses of each of the measures are tabulated in **[Table-1, 2, 3 and 4].** The results for the RMSE measure are tabulated in **[Table-1]**. It is evident from Table 1 that the RMSE is lower for proposed method compared to other five methods, which means that proposed method introduces very less error.

THE IIOAB JOURNAL
ISSN: 0976-3104

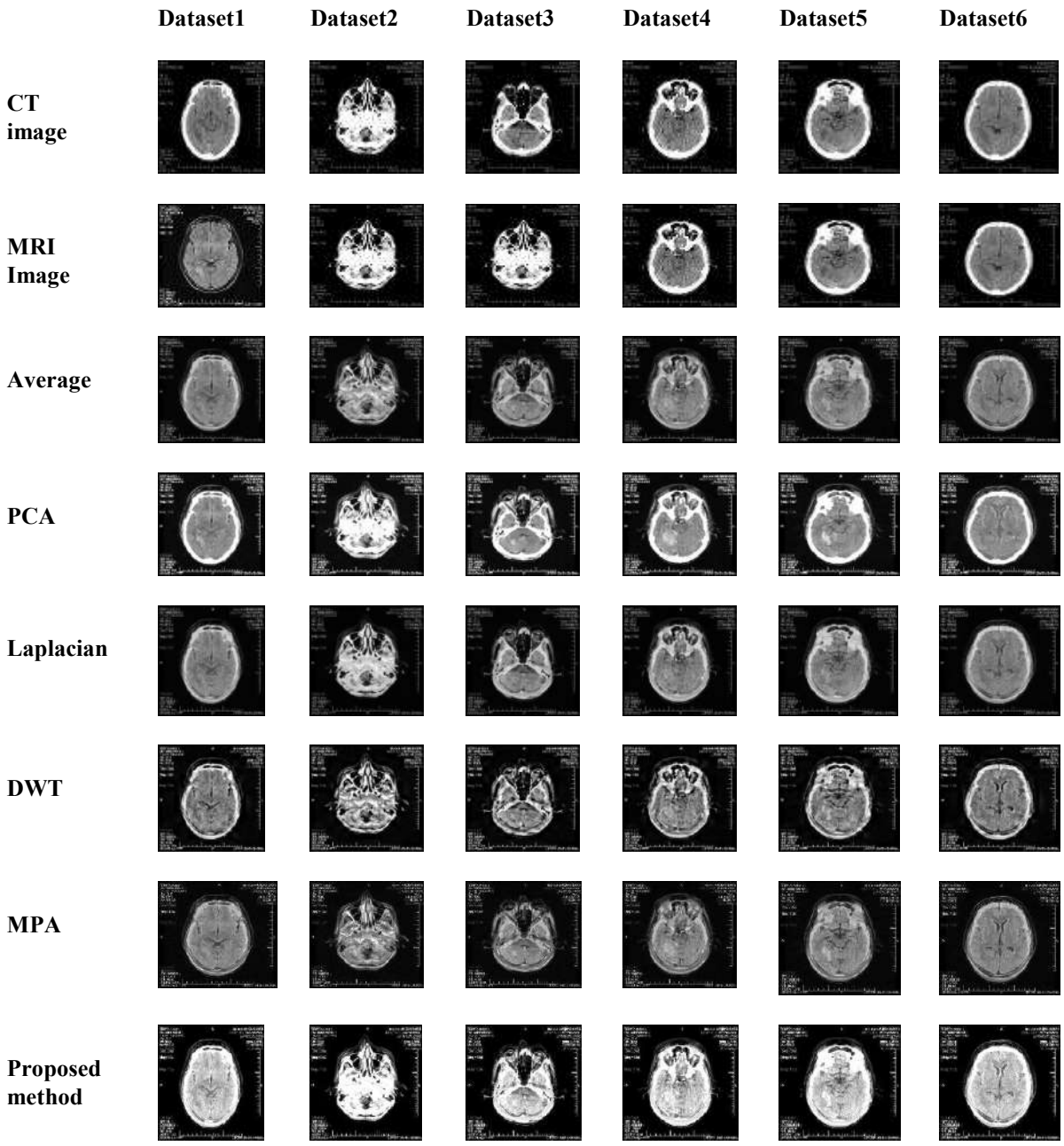


Fig: 2. Comparison (subjective) of the fusion results over 6 images

Table: 1. Comparative analysis of RMSE

Fusion method	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5	Dataset6
Average	0.1897	0.2015	0.2112	0.1987	0.1936	0.2007
PCA	0.1833	0.2216	0.2097	0.2001	0.1867	0.1972

THE IIOAB JOURNAL
ISSN: 0976-3104

Laplacian	0.1874	0.2145	0.1998	0.1972	0.1956	0.2382
DWT	0.1832	0.2127	0.1964	0.1945	0.1904	0.1823
MPA	0.2136	0.2454	0.2270	0.2273	0.2228	0.2148
Proposed	0.1827	0.2125	0.2073	0.1918	0.1860	0.1859

Table: 2. Comparative analysis of MAE

Fusion method	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5	Dataset6
Average	0.0773	0.1097	0.0927	0.0908	0.0866	0.0743
PCA	0.0663	0.0648	0.0943	0.0686	0.0674	0.0671
Laplacian	0.0837	0.1087	0.0943	0.0931	0.0920	0.1206
DWT	0.0880	0.1166	0.1004	0.0992	0.0961	0.0886
MPA	0.0996	0.1269	0.1093	0.1103	0.1078	0.0990
Proposed	0.0624	0.0703	0.0726	0.0661	0.0627	0.0655

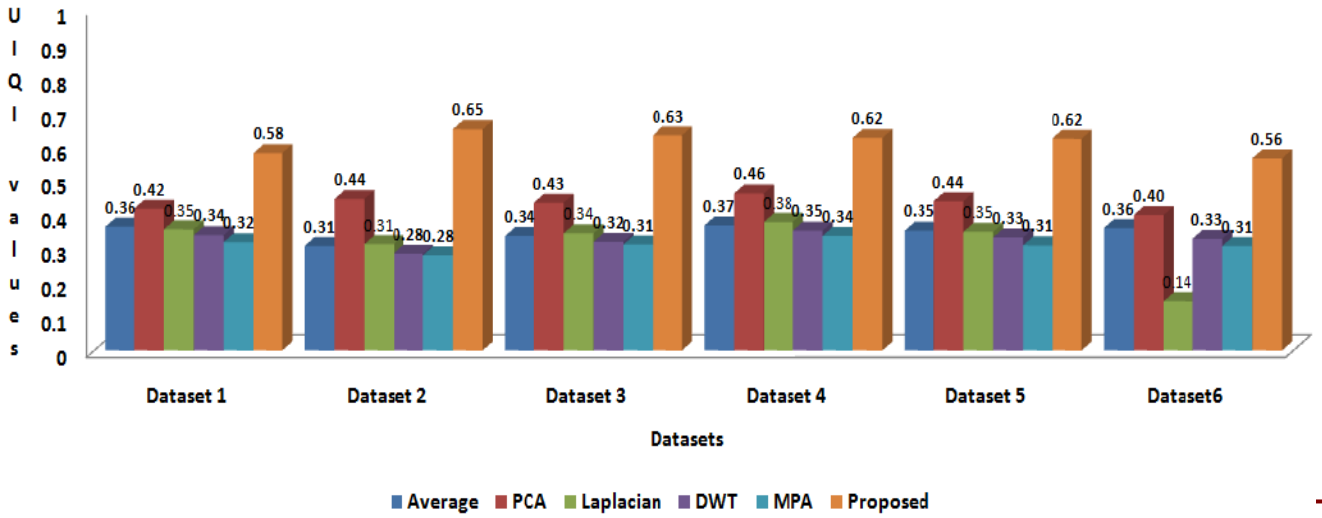
Table: 3. Comparative analysis of PSNR

Fusion method	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5	Dataset6
Average	62.3742	60.38421	59.8876	61.3645	62.0019	61.3658
PCA	62.3781	60.3652	60.2332	62.3118	62.1187	62.4722
Laplacian	62.6763	61.5024	62.1178	62.2349	62.3056	60.5905
DWT	62.8700	61.5745	62.2671	62.3508	62.5352	62.9158
MPA	61.5384	60.3347	61.0116	61.0001	61.1724	61.4890
Proposed	62.8955	61.5819	61.7968	62.4722	62.7387	62.7452

Table: 4. Comparative analysis of SSIM

Fusion method	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5	Dataset6
Average	0.9983	0.9969	0.9976	0.9977	0.9979	0.9985
PCA	0.9982	0.9966	0.9980	0.9975	0.9976	0.9987
Laplacian	0.9976	0.9962	0.9969	0.9970	0.9972	0.9956
DWT	0.9978	0.9964	0.9971	0.9971	0.9974	0.9979
MPA	0.9962	0.9945	0.9954	0.9954	0.9957	0.9963
Proposed	0.9984	0.9969	0.9981	0.9977	0.9979	0.9987

The results of the MAE measure are shown in **[Table-2].** It is observed that the proposed method introduced the least error for five of the six datasets. The results of the PSNR measure are shown in **[Table-3]**. It is apparent from Table 3 that the PSNR value of each and every dataset is superior for the proposed method, indicating a higher image quality. The results of the SSIM measure are shown in **[Table-4]**. Table 4 clearly shows that the SSIM value of every dataset is closest to 1, compared to the other five methods, indicating the maximum similarity to the original image.



ISSN: 0976-3104

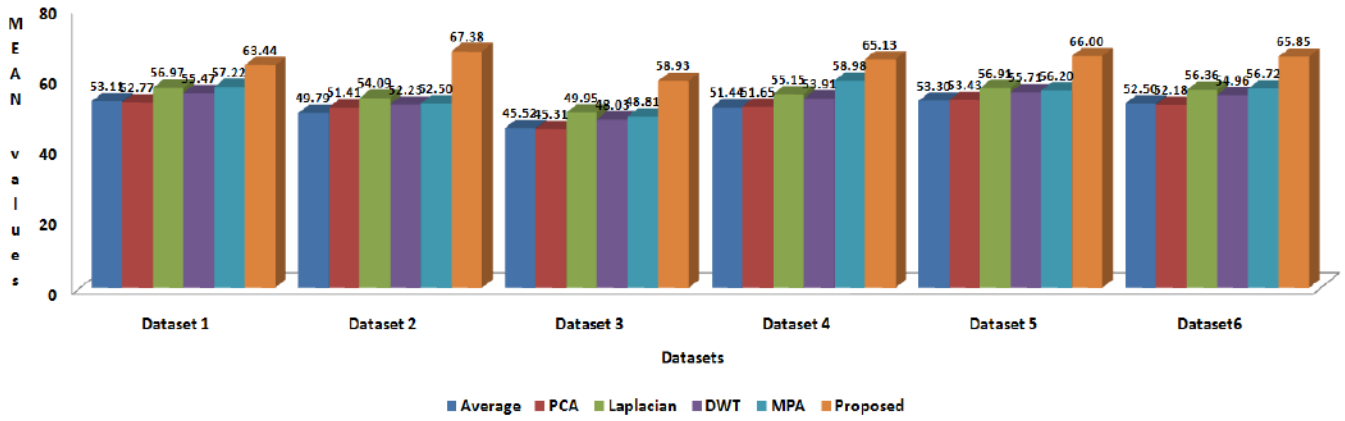


Fig: 5. Comparative analysis of MEAN

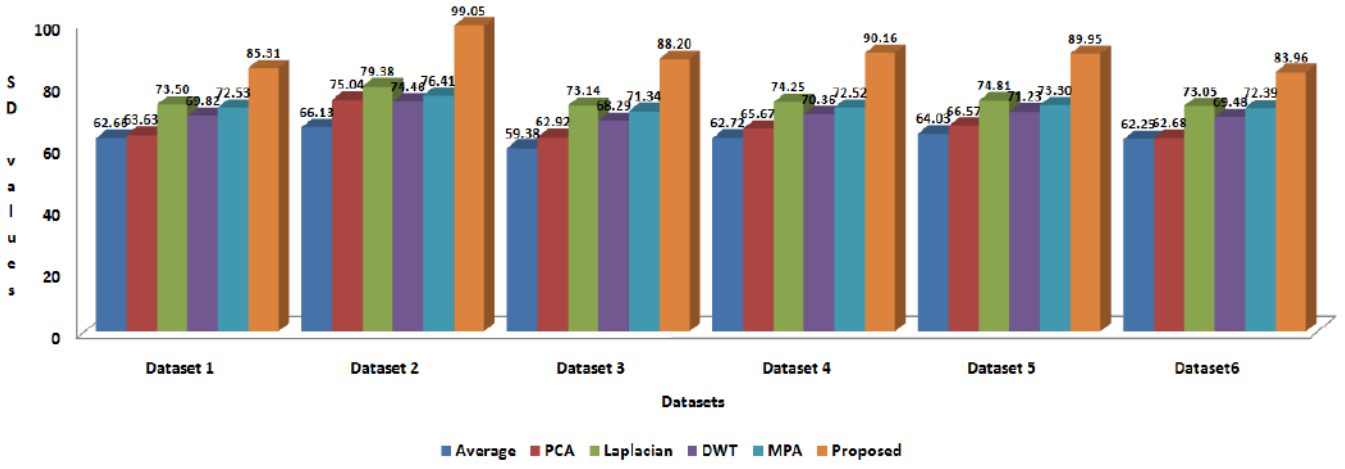


Fig: 6. Comparative analysis of SD

The UIQI values in **Figure 4** show that the results of the proposed method for every dataset is closest to 1, compared to others, thus indicating the maximum similarity.

The results in **Figure 5** show that the mean value for the proposed method is more than the other approaches, signifying more texture information on the resultant image. The impressive results are also visible in **Figure-6**, which shows the values for the standard deviation measure.

CONCLUSION

In this paper, image fusion using block based intuitionistic fuzzy sets has been proposed. Since, the entropy provides texture information of an image, the technique of block comparison with the entropy adopted in the paper

ISSN: 0976-3104

as well as the adaptive calculation of the necessary parameter for the process sums up its novelty. The experimental results show that proposed method provides better visualization than average method, PCA method, laplacian method, DWT method and MPA method. In addition, proposed method confers better result compared to the other existing methods for both the objective and quantitative measures. Furthermore, the fused image obtained from proposed method has been found to be more informative and thereby can be used for efficient disease diagnostics.

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] Stathaki, T. [2011] Image fusion: algorithms and applications. Academic Press.

[2] Burt, P. J., & Adelson, E. H. [1983] The Laplacian pyramid as a compact image code. IEEE Transactions on Communications 31: 532-540.

[3] Sun, J., Jiang, Y., & Zeng, S. [2005] A study of PCA image fusion techniques on remote sensing. International Conference on Space information Technology: 59853X-59853X.

[4] Li, H., Manjunath, B. S., & Mitra, S. K. [1995] Multisensor image fusion using the wavelet transform. Graphical models and image processing 57: 235-245.

[5] Kakarala, R., & Ogunbona, P. O. [2001] Signal analysis using a multiresolution form of the singular value decomposition. IEEE Transactions on Image Processing 10 :724-735.

[6] Lung, S. Y. [2002] Multi-resolution form of SVD for text-independent speaker recognition. Pattern recognition 35 :1637-1639.

[7] Da Cunha, A. L., Zhou, J., & Do, M. N. [2006] The nonsubsampled contourlet transform: theory, design, and applications. IEEE Transactions on Image Processing 15: 3089-3101.

[8] Ross, T. J. [2009] Fuzzy logic with engineering applications. John Wiley & Sons.

[9] Atanassov, K. T. [1986] Intuitionistic fuzzy sets. Fuzzy sets and Systems 20 : 87-96.

[10] Balasubramaniam, P., & Ananthi, V. P. [2014] Image fusion using intuitionistic fuzzy sets. Information Fusion 20: 21-30.

[11] Sharmila, K., Rajkumar, S., & Vijayarajan, V. [2013] Hybrid method for multimodality medical image fusion using Discrete Wavelet Transform and Entropy concepts with quantitative analysis. International Conference on Communications and Signal Processing (ICCSP): 489-493.

[12] Pu, T., & Ni, G. [2000] Contrast-based image fusion using the discrete wavelet transform. Optical Engineering 39: 2075-2082.

[13] Wang, Z., Ziou, D., Armenakis, C., Li, D., & Li, Q. [2005] A comparative analysis of image fusion methods. IEEE Transactions on Geoscience and Remote Sensing 43: 1391-1402.

[14] Soundrapandiyan, R., & PVSSR, C. M. [2015] Perceptual Visualization Enhancement of Infrared Images Using Fuzzy Sets. In Transactions on Computational Science XXV : 3-19.

[15] Rajkumar, S., & Mouli, P. C. [2014] Infrared and Visible Image Fusion Using Entropy and Neuro-Fuzzy Concepts. ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-I : 93-100.

[16] Prakash, C., Rajkumar, S., & Mouli, P. V. S. S. R. [2012] Medical image fusion based on redundancy DWT and Mamdani type min-sum mean-of-max techniques with quantitative analysis. International Conference on Recent Advances in Computing and Software Systems: 54-59.

[17] Rajkumar, S., & Kavitha, S. [2010] Redundancy Discrete Wavelet Transform and Contourlet Transform for multimodality medical image fusion with quantitative analysis. 3rd International Conference on Emerging Trends in Engineering and Technology : 134-139.

[18] Rajkumar, S., Bardhan, P., Akkireddy, S. K., & Munshi, C. [2014] CT and MRI image fusion based on Wavelet Transform and Neuro-Fuzzy concepts with quantitative analysis. International Conference on Electronics and Communication Systems : 1-6.

[19] Szmidt, E. and Kacprzyk, J.[2000] Distances between intuitionistic fuzzy sets. Fuzzy sets and systems 114(3): 505-518.

[20] Gupta, S., Rajkumar, S., Vijayarajan, V. & Marimuthu, K.. [2013] Quantitative Analysis of various Image Fusion techniques based on various metrics using different Multimodality Medical Images. International Journal of Engineering and Technology : 133-141.

[21] Purushotham, S. and Tripathy, B.K.. [2015] A Comparative Analysis of Depth Computation of Leukaemia Images using a Refined Bit Plane and Uncertainty Based Clustering Techniques. Cybernetics and Information Technologies 15(1):126-146.

[22] Purushotham, S. and Tripathy, B. [2014] A comparative study of RIFCM with other related algorithms

from their suitability in analysis of satellite images using other supporting techniques. Kybernetes 43(1): 53-81.

[23] Bhargava, R., Tripathy, B.K., Tripathy, A., Dhull, R., Verma, E. and Swarnalatha, P. [2013] Rough intuitionistic fuzzy C-means algorithm and a comparative analysis. In Proceedings of the 6th ACM India Computing Convention : 23-34.

ABOUT AUTHORS

Mr. Rajkumar Soundrapandiyan is currently working as Assistant Professor (Senior) in School of Computing Science and Engineering, VIT University, Tamil Nadu, India. He received his BE in Computer Science and Engineering and ME in Computer Science and Engineering from Anna University, Chennai, India in 2008 and 2010 respectively. He is pursuing his PhD at VIT University, Vellore, India. He has published more than ten research papers in reputed international conferences and journals. He is a life member of CSI. His research interest includes digital image processing, computer vision, object detection and target recognition.

Mr. Rishin Haldar is currently working as Assistant Professor (Senior) in School of Computing Science and Engineering, VIT University, Tamil Nadu, India. He received his BE in Computer Science and Engineering from Karnatak University, Dharwad and MS in Computer Science from George Mason University, Fairfax, USA. He is a life member of CSI. His research interests include artificial intelligence, soft computing, bioinformatics and semantic web.

Swarnalatha Purushotham is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Systems. She has published more than 50 papers in International Journals/International Conference Proceedings/National Conferences. She is having 14+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.

Mr. Arvind Pillai is pursuing his B.Tech in Computer Science and Engineering in VIT University, Tamil Nadu, India. His research interest includes digital image processing, computer vision and object detection

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Sevugan et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

AN ANAGLYPH APPROACH OF GEOMETRIC CORRECTION FOR DEPTH RECONSTRUCTION OF SATELLITE IMAGERY

Prabu Sevugan[1], Balakrushna Tripathy[1*],Swarnalatha Purushotham[1], Ramakrishnan R[2], Manthira Moorthi S [2]

[1] *VIT University, Vellore-Tamilnadu, INDIA*
[2]*Space Applications Centre, Ahmedabad, INDIA*

ABSTRACT

The images acquired through remote detecting frameworks are not regularly adequate for high accuracy applications because of different misrepresentations. The misrepresentations can be because of mistakes such as geometric mistakes and so forth. Likewise multi-date satellite pictures of the same range under various conditions are hard to match due to change in atmospheric propagation, sensor reaction and enlightenments. Keeping these focuses in perspective, in this paper, we manage the primary period of pre-preparing and we make the satellite pictures free from such blunders and utilize grouping systems With Geometric Correction (WGC) and Without Geometric Correction (WOGC) connected to the satellite pictures utilizing our proposed methodology. At last, the picture is remade with depth measurement/depth map era for the anaglyph picture for better understanding of satellite symbolism. We have made test investigation of our methodology utilizing suitable satellite pictures and observed the outcomes to be extremely reassuring.

Received on: 30th-Novr-2015
Revised on: 11th-March-2016
Accepted on: 26th–March-2016
Published on: 19th–May-2016

KEY WORDS

Geometric Correction, Rough Intuitionistic Fuzzy C-Means (RIFCM) clustering algorithm, Depth Reconstruction, Anaglyph Graph, Satellite Images

***Corresponding author:** Email: sprabu@vit.ac.in, tripathybk@vit.ac.in,**Tel:** +40-9001010010**; Fax:** +40-9001010012

INTRODUCTION

Remote sensing is the securing of data around an item or wonder without reaching the article. In present day use, the term for the most part alludes to the utilization of elevated sensor advances to identify and characterize objects on Earth (both at first glance, and in the environment and seas) by method for engendered signals (e.g. electromagnetic radiation) [1]. It might be part into dynamic remote detecting, when a sign is initially discharged from air ship or satellites or inactive (e.g. daylight) when data is simply recorded. Pictures acquired through remote detecting frameworks are not regularly adequate for high accuracy applications because of different bends. The falsifications can be because of geometric mistakes, radiometric blunders and barometrical mistakes. Furthermore, contrasted with a proficient bitplane which is utilized for restorative pictures may not give precise results for satellite symbolism [2].In this paper we concentrate on geometric blunders as it were.

The geometric errors can be remedied to a specific degree by considering just orbital geometry model methodology which joins worldwide changes and considers the viewpoint proportion, earth revolution, picture introduction, and so on, [1] additionally yield poor precision. Then again, with itemized data extricated from auxiliary, one can apply the rectification handling to every pixel and subsequently get much higher precision. Pixel projection methodology will consider the instrument mutilation (measured pre-flight or in-flight), the rocket demeanor, position and speed at the moment the picture is shaped in each CCD pixel. The Earth revolution impact will be ascertained by a high-exactness earth pivot model which includes customary turn, precession, and transformation [3]. In this manner the geo-area (regarding geodetic longitude and scope) of every picture pixel can be computed. The geometric mistakes can be redressed to a specific degree with definite instrument and shuttle data (subordinate information) that gives instrument alignment information, rocket state of mind, position and speed at adequately little time interim which is key for the satellite symbolism. The writing study has been completed on pre-handling of satellite symbolism with factual methods and of geometric revision utilizing non-parametric models [4][5].

THE IIOAB JOURNAL
ISSN: 0976-3104

The co-enlistment of three groups has been managed in paper [6] which is a special test because of the impact of circle and disposition in the time hole in the imaging succession. The co-enrollment issue of satellite symbolism might be utilized to have better translation of depth reproduction at the pre-preparing period of the creator's paper.

Multi-date satellite pictures under various states of the same zone are hard to analyze in light of progress in environmental spread, sensor reaction and enlightenments. Channels are generally utilized for such things as edge upgrade, clamor evacuation, and the smoothing of high recurrence information. As a result, in this paper we first evacuate these sorts of mistakes in the pre-applying so as to handle stage With Geometric Correction (WGC). Yet, we likewise continue without geometric adjustments additionally, so as to give a relative examination and get results for the WGC.

A novel methodology of fuzzy c-means implies has been proposed in the paper [7] where the Bit Plane Filter(FCMBP) separates a picture into planes utilizing different scientific systems. The outcomes demonstrated quality ordered pictures contrasted with customary edge methods with expansion in the proficiency by lessening in the epsilon estimations of bit planes. What's more, in future, the creator point is to apply other clustering techniques for recreation of the picture with the assistance of the grouped picture of the paper to show signs of improvement understanding of third measurement.

The proposition has been made in [8] utilizing rough c-means for examination of satellite pictures with/without Bit Plane Filter (RCMBP-Rough C-Means Bit Plane) which sections the satellite picture into planes utilizing different scientific procedures. Furthermore the depth calculation for all conceivable outcomes utilizing with and without bit plane procedures have been done. On comparison, utilizing using rough c-means yields better depth calculation with least time multifaceted nature. In future, the creators point is to apply other grouping procedures for further preparing of satellite pictures.

The depth map (divergence picture or Z-picture) is a picture which contains data about the depth of the photo and serves to change over a two-dimensional picture into a three-dimensional one coming about depth reproduction. In the depth delineate dim shading degree demonstrates every pixel's separation from the viewer. The lighter zone in the depth map compares to the ranges closer to the viewer, the darker one relates to more removed territories [9].

Anaglyph 3D pictures contain two diversely separated hued pictures, one for every eye which can be acquired by applying parallax calculation is utilized which is a dislodging or contrast in the evident position of an item saw along two unique observable pathways, and is measured by the edge or semi-edge of slant between those two lines [10][11].

Clustering strategies are utilized as a part of picture division and reproduction of pictures after that produces upgraded pictures for examination. Subsequent to a considerable measure of vulnerability is included in the picture investigation we discover a great deal of instability based clustering procedures which are more suited for such sort of picture grouping. In [12] an exceptionally uncertainty based hybrid algorithm has been set up and tried, called RIFCM, which joins intuitionistic fuzzy set and rough set methodologies.

The method in [13], deals with the determination of the areas that have and don't have changes. The determined area is grouped as two parts by Fuzzy C-Means Clustering method. With the method of principal component analysis, eigenvector space is gained and from here, principal components are reached. Finally, feature vector space consisting principal component is partitioned into two clusters using Fuzzy C-Means Clustering and after that change detection process has been done for clustering of images. This can also be used to extract the cloud from a satellite image which deals with R, G, and B components of a true color satellite image and convert a true color satellite image [14].

In [15], a Comparative Study has been carried out using RIFCM with Other Related Algorithms from the point of view of their suitability in analysis of satellite images with other supporting techniques which deals with proving the superiority of RIFCM with RBP in clustering with other clustering methods and other supporting metrics with and without refined which integrates judiciously RIFCM with RBP. The superiority of the RIFCM using RBP is demonstrated, along with a comparison with other related algorithms, on a satellite images with NASA.org images(Hills, Drought) and national geographic photographic images(Freshwater, Fresh water valley).

THE IIOAB JOURNAL ISSN: 0976-3104

For two depth reconstruction was stretched out to various depth reconstruction in a way like that of other [16, 17]. The utilization of various pictures gives more precise depth results by alleviating the impact of picture noise. The movement obscure estimation issue is presently spoken to by the depth estimation issue [18]. The obscure strength of the proposed calculation is checked by contrasting the depth remaking results and the ordinary variational depth reconstruction to utilize various pictures which is executed the obscure taking care of parts of the proposed technique and accordingly accomplished enhanced results [19].

A True Color Composite is connected for satellite symbolism for appropriate bunches which manages a real nature picture which is recovered when a multispectral picture comprises of the three visual essential shading groups (red, green, blue). groups might be consolidated to deliver a "genuine nature" picture coming about shading composite picture which is seen by the human eyes for better representation of satellite symbolism [19].

In the later part of this paper, the picture is recreated with depth measurement/depth map era for the anaglyph picture for better elucidation of satellite symbolism. We have made exploratory investigation of our calculation utilizing suitable satellite pictures and observed the outcomes to be extremely reassuring.

The part I of the paper manages the presentation and overview of remote detecting, geometric remedy, grouping, depth map era, anaglyph picture. Furthermore, part II examine about the technique of the paper. Results and the examination of the paper has been clarified in part III. The conclusion and future work of the paper is managed in part IV took after with affirmation and references referred to for the paper

METHODOLOGY

The proposed methodology deals with three phases of satellite imagery for depth map generation for better interpretation of images which is discussed in **Figure–1** as given below:

1. Initially, the satellite symbolism has been given as data for geometric revision which comprises of extraction of Ground Control Points(GCPs) for 35, 100 and 150 in three distinct reaches to get least rate of root mean square blunder as yield picture expressing with geometric correction(WGC) and analyzed without geometric correction (WOGC).

2. Secondly, the yield of WGC and WOGC is considered as input for clustering calculation (Rough Fuzzy Intuitionistic C-Means) coming about WOGC-C(clustered) and WGC-C(clustered) pictures.

3. Thirdly, clustered pictures are given as information for depth map era as reproduced picture with anaglyph approach for better understanding of satellite symbolism.

a. Geometric Correction

The pre-handled information is then utilized as the beginning condition for the remedy preparing utilizing GCP. The three stages must be handled for geometric remedy as

GCP extraction (using GRASS GIS)

GCPs have been separated by GRASS GIS for three classes (35,100,150 GCPs). The RMS mistake is appeared for each GCPs. The two sorts of RMS mistakes are shown over the GCP director board. They are named as forward mistake and in reverse mistake which has been shown with two boards as source image(left) and target image(right) for each GCPs classification in **Figure- 2.a, b, and c**. The execution has been tried by applying peak signal noise ratio and root mean square error measurements (figure 3) coming about least RMS for 150 GCPs contrasted with 35 GCPs and 100 GCPs

THE IIOAB JOURNAL
ISSN: 0976-3104

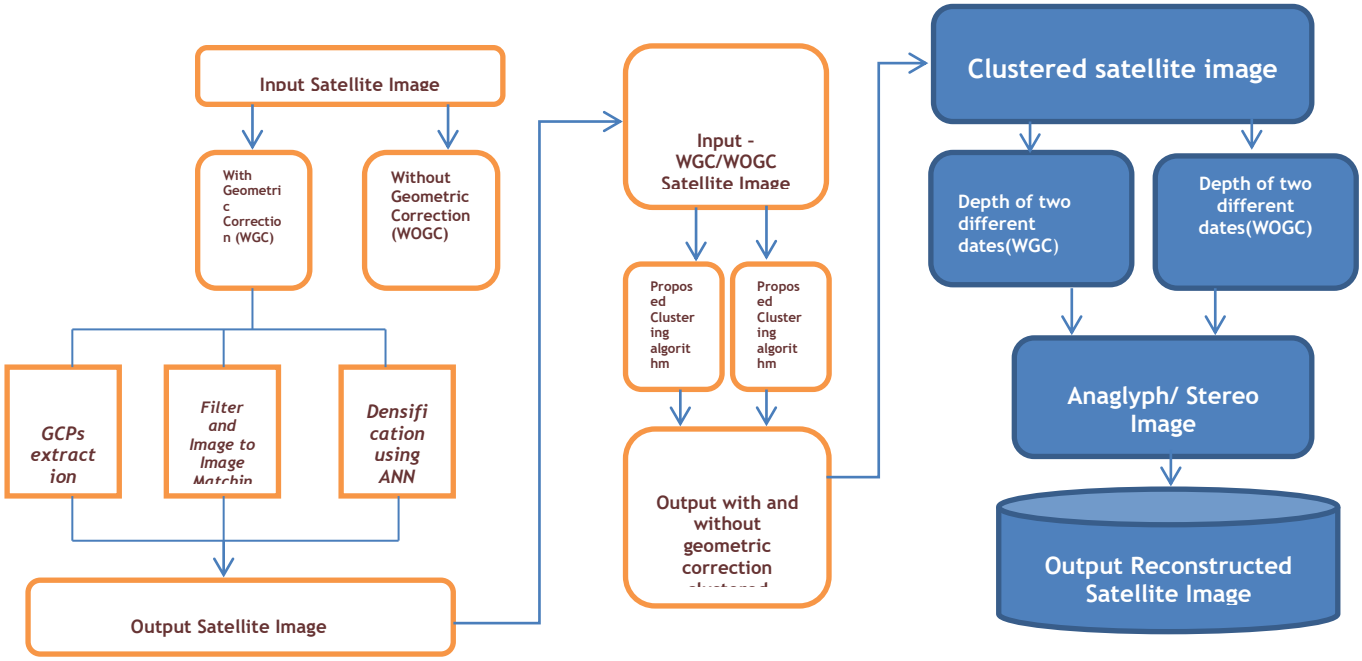


Fig: 1. Architecture of proposed methodology with anaglyph approach of WOGC – C-D and WGC-C-D

Table: 1. GCP extraction of 1-35,1-100,1-150 for 4th Jan-May

Month	GCP	RMS
a. 4th Jan-May	35	0.034272
b. 4th Jan-May	100	0.037071
c. 4th Jan-May	150	0.036559

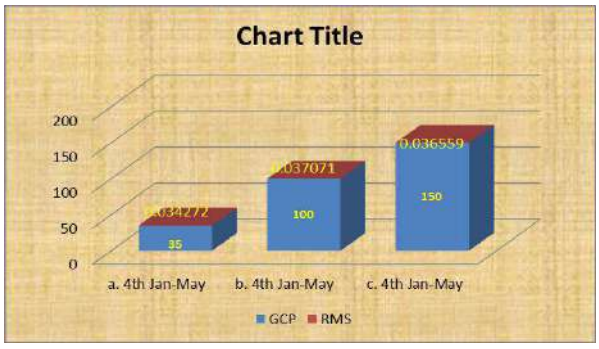


Fig: 2. Performance of PSRN & RMSE for GCPs

THE IIOAB JOURNAL
ISSN: 0976-3104

Filter and Image to Image Matching

Invert filter is applied with neighborhood analysis and Image to Image matching process is done using Cross Correlation Coefficient with Resampling Techniques (fourier transform, matrix/convolution and Neighborhood Analysis (Average)) in GRASS GIS Open Source Software. The results of resampling techniques have been given in **Figure 3(a)** fourier analysis**, 3(b)** matrix /convolution and **3(c)** neighborhood analysis (average) for further process.

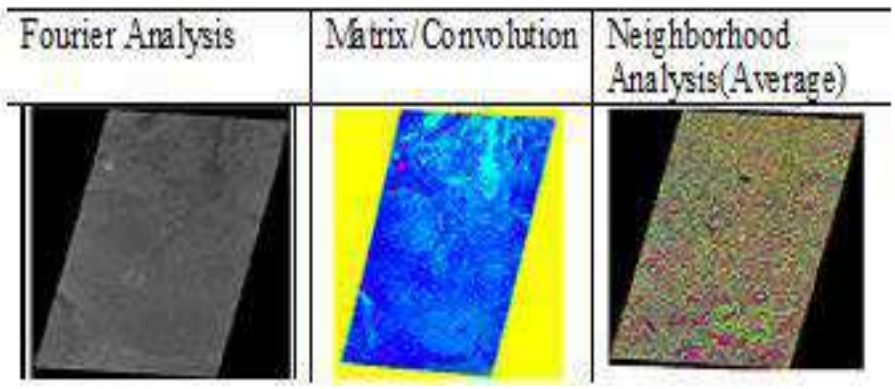


Fig: 3(a). Fourier Analysis Fig.No.4.b. Matrix/Convolution Fig..4.c. Neighborhood Analysis(Average)

Densification by ANN

Densification process is used to add vertices to a line at specified distances without altering the line's shape. This operation densifies geometries by points between existing vertices. And Densification Total Error Network Graph is applied to compute for two different months data to calculate Mean Square Error of Relative Geometric Corrected Image with least value of MSE of 0.0158**[Figure- 4]**.

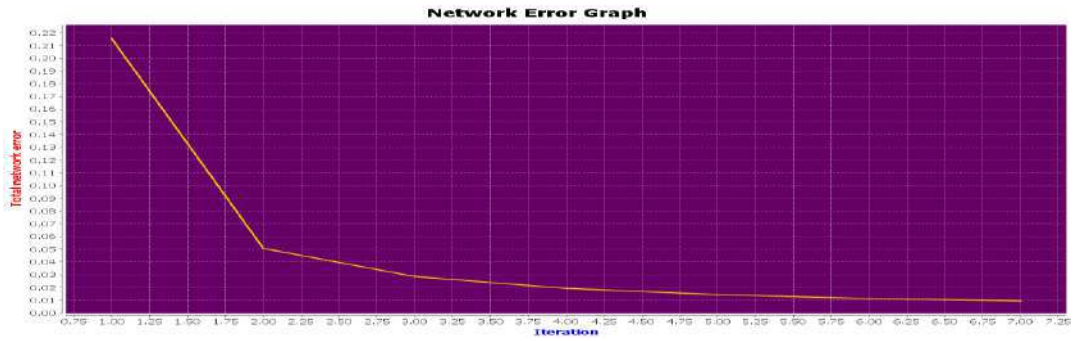


Fig: 4. January-Total Mean Square Error-0.0158

RIFCM

The Rough Intuitionistic Fuzzy C-Means clustered technique has been applied for WOGC and WGC with PSNR and RMSE metric of performance validation.

Rough Intuitionistic Fuzzy C-Means

The rough intuitionistic fuzzy c-means (RIFCM) , uses the concept of rough sets, fuzzy sets and as well as intuitionistic fuzzy sets, thereby making it a perfect combination of IFCM and RCM. It can also be considered to be RFCM with IFS, hence adding the concept of lower and upper approximation of rough set, fuzzy membership of fuzzy set, non-membership and hesitation value of intuitionistic fuzzy set. It provides a holistic approach to

THE IIOAB JOURNAL ISSN: 0976-3104

clustering of data as it deals with uncertainty, vagueness, incompleteness which, enables the efficient handling of overlapping partitions and improves accuracy.

In RIFCM, each cluster can be identified by three properties, a centroid, a crisp lower approximation and an intuitionistic fuzzy boundary. If an object belongs in the lower approximation of a cluster then its corresponding membership value is 1 and hesitation value is 0. The objects in the lower region have same influence on the corresponding cluster. If an object belongs in the boundary of one cluster then it possibly belongs to that cluster and potentially belongs to another cluster. Hence the objects in the boundary region have different influence on the cluster. Thus we can say that in RIFCM the membership value of objects in lower region is unity $(\mu_{ij}^{'} = 1)$ and for those in boundary region behave like IFCM.

The objective of this algorithm is to reduce the cost function given in [20].The parameters w_{low} and w_{up} have the standard meanings. Also $\mu ij\ '$ has the same definition as in IFCM.

The steps that are to be followed in this algorithm are as given below

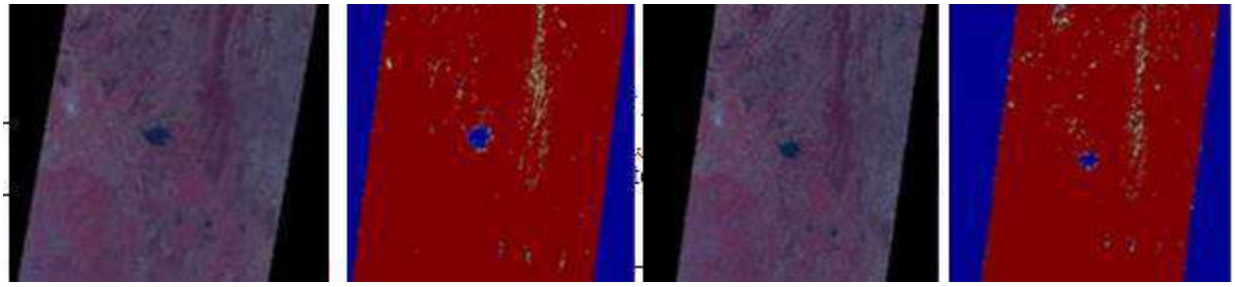
1. Assign initial means v_i for c clusters by choosing any random c objects as cluster.
2. Calculate d_{ik} using Euclidean distance formula (1)
3. Compute U matrix
   if $d_{ik} = 0$ or $x_j \in \underline{BU}_i$ then
   $$\mu_{ik} = 1$$
   else compute $\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}}$
4. Compute π_{ik}
   $$\pi_A(x) = 1 - \mu_A(x) - \frac{1-\mu_A(x)}{1+\lambda\mu_A(x)} \mid x \in X$$
5. Compute $\mu_{ik}^{'}$ and normalize $\mu_{ik}^{'} = \mu_{ik} + \pi_{ik}$
6. Let $\mu_{ik}^{'}$ and $\mu_{jk}^{'}$ be the maximum and next to maximum membership values of object x_k to cluster centroids v_i and v_j.
   if $\mu_{ik}^{'} - \mu_{jk}^{'} < \varepsilon$ then $x_k \in \overline{BU}_i$ and $x_k \in \overline{BU}_j$ and x_k cannot be a member of any lower approximation. else $x_k \in \underline{BU}_i$
7. Calculate new cluster means by using (2)
8. Repeat from step 2 until termination condition is met or until there are no more assignment of objects.

$$d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_n-y_n)^2} \ldots\ldots\ldots (1)$$

$$v_i = \begin{cases} w_{low}\dfrac{\sum_{x_{k\in\underline{BU}_i}} x_k}{|\underline{B}\,U_i|} + w_{up}\dfrac{\sum_{x_{k\in\overline{BU}_i-\underline{BU}_i}} \mu_{ik}^{m} x_k}{\sum_{x_{k\in\overline{BU}_i-\underline{BU}_i}} \mu_{ik}^{m}} & if\ |\underline{B}\,U_i| \neq \phi and\ |\overline{BU}_i - \underline{BU}_i| \neq \phi \\ \dfrac{\sum_{x_{k\in\overline{BU}_i-\underline{BU}_i}} \mu_{ik}^{m} x_k}{\sum_{x_{k\in\overline{BU}_i-\underline{BU}_i}} \mu_{ik}^{m}}, & if |\underline{B}U_i| \neq \phi and |\overline{BU}_i - \underline{BU}_i| \neq \phi \\ \dfrac{\sum_{x_k\in\underline{BU}_i} x_k}{|\underline{B}\,U_i|} & ELSE \end{cases} \ldots..(2)$$

THE IIOAB JOURNAL ISSN: 0976-3104

The **Figure 5(ai**.) deals with the Original image of Ahmedabad a.ii.Clustered image of Ahmedaba and the figure 6.b.i. Original corrected image of Ahmedabad b.ii.Clustered corrected image of Ahmedabad where the performance of satellite imagery has been calculated which is given in **Table– 2** and **Tigure– 11**.



ai bi aii bii

Fig: 5.a.i. Original image of Ahmedabad Fig: 6.b.i. Original Corrected image of Ahmedabad
a. ii.. Clustered image of Ahmedabad b.ii. Clustered corrected image of Ahmedabad

DEPTH RECONSTRUCTION

The third dimension of the affected portion [21][22] is very vital as the degree to which the component has been deteriorated which can be calculated approximately. The angle of incidence of X-rays and the thickness of the component are taken as an input from the user. The third dimension of the component can be calculated using the formula given below.
The formula for calculating third dimension is:

$$Depth = \left(\frac{L/2 + x}{Tan\,\alpha} + \delta\right) - \frac{W}{Sin\alpha} \quad \text{.....(3)}$$

L = Length of the third dimension affected region, x = 0; Assuming that the third dimension is exactly in the centre = Angle of incidence of the X-rays, W = Width of the dimension, $^{\delta}$ = Thickness of the component
If the third dimension of the affected portion is greater than the thickness of the component then the component is fully affected and it has to be replaced with another one. Based on the three parameters (length, width and depth), the papers aims to reconstruct an image which may be possible as given below:-

Depth map generation algorithm (Depth Reconstruction) and Anaglyph Technique

The third dimension map generation (Z-dimension) which is a gray-scale image should possess the resolution that is same as to the original input image for consideration of Depth Map Generation given in **Figure– 6**.

Algorithm: Generation of Depth Map (Depth Reconstruction)with Anaglyph Image

1. Start
2. Initialize images
   2.1 First image has been taken as a original img,2D image (Oi) and
   2.2 Second image is considered as a processed image (clustered image), 2D image (Ci),then
3. Two images are opened in the interface: left and right frames of a stereo pair. And
   3.1 A depth map (depth reconstruction) will be created from the two images
4. Generate frames with the parallax method which characterizes the distance of the object's projections on the plane for the left and right eyes (disparity).
5. Iterate the frames till anaglyph image is obtained from the third dimension map generated for better visualization of an image . Finally,
6. Anaglyph Image from third dimension map has been resulted
7. Stop

THE IIOAB JOURNAL
ISSN: 0976-3104

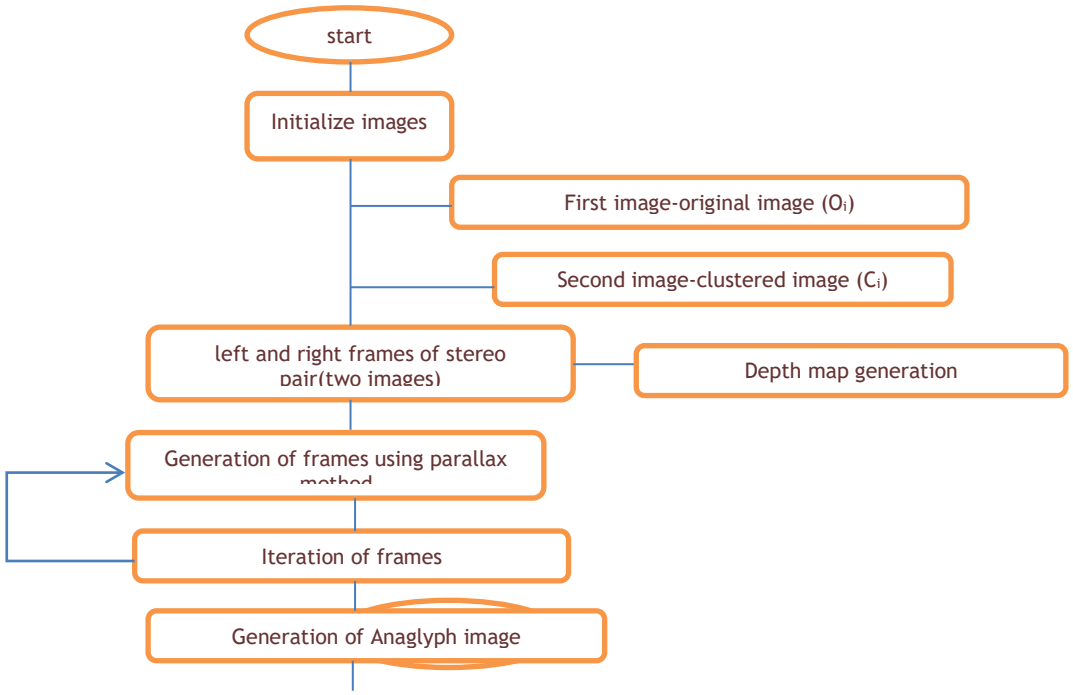


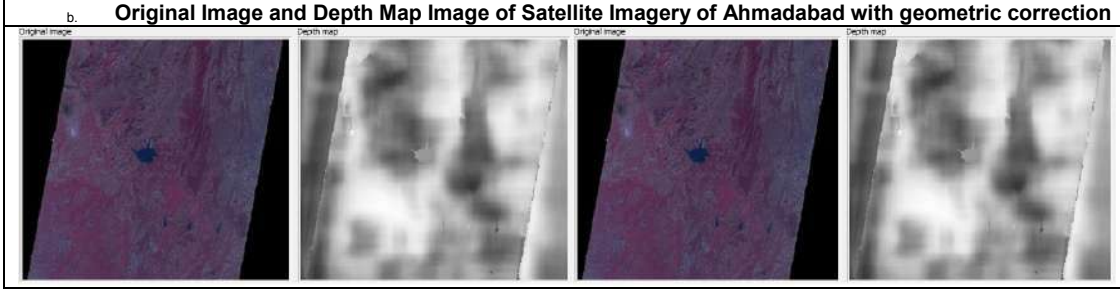
Fig: 6. Flow chart of Depth Map Generation Algorithm

To support the third dimension map generation, the parallax algorithm may be used where the results has been made in figure 8.a.WOGC-C-D and figure 8.b.WGC-C-D, which characterizes the distance of the object's projections on the plane for the left and right eyes (disparity). And the method, Parallax [23] which is a dislocation or dissimilarity in the apparent position of an entity viewed along two dissimilar lines of sight, is calculated by the point of view or semi-angle of leaning between those two lines.

A third dimension (Z) map generation serves to exchange the original image into a depth map one. Intensity of a pixel in a third dimension map shows the space from the similar pixel in the original image to the watcher. The lighter areas in depth map match to the regions nearer to the viewer, the darker ones match to more distant areas. A white pixel in a depth map deals with the the pixel of the original image that has the smallest distance to the viewer (foreground),**Figure 7(a), 8(a)** and a black pixel with the pixel of the original that has the biggest distance to the viewer (background),**Figure 9(b) 10(b)**

a. riginal Image and Depth Map Image of Satellite Imagery of Ahmadabad without geometric correction

b. Original Image and Depth Map Image of Satellite Imagery of Ahmadabad with geometric correction



THE IIOAB JOURNAL ISSN: 0976-3104

Fig:7 (a):Original Image and Depth Map of Satellite Imagery without geometric correction (WOGC-C-D) b. with geometric correction (WGC-C-D)

……

And frames of images will be generated from the third dimension map, **Figure 9(c) & 10c** resulting anaglyph 3D image, from a set of generated frames. The Anaglyph 3D images can be viewed through the "color-coded" "anaglyph glasses", each of the two images reaches one eye, revealing an integrated stereoscopic image. The visual cortex of the brain fuses this into perception of a three dimensional scene or composition as given in **Figure 9(d)** (WOGC-C-D) and **Figure 10(d)** (WGC-C-D).

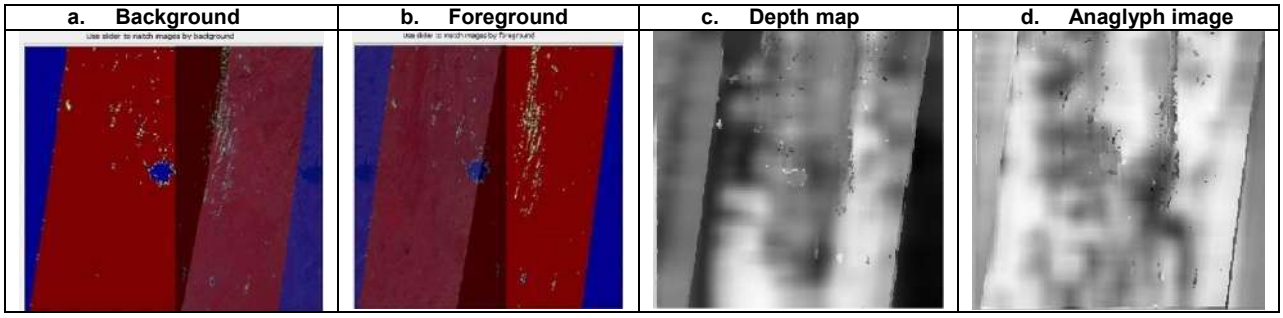


Fig: 8. Depth images of a .background b. foreground c. depth map d. anaglyph images of without geometric correction (WOGC-C-D)

……..

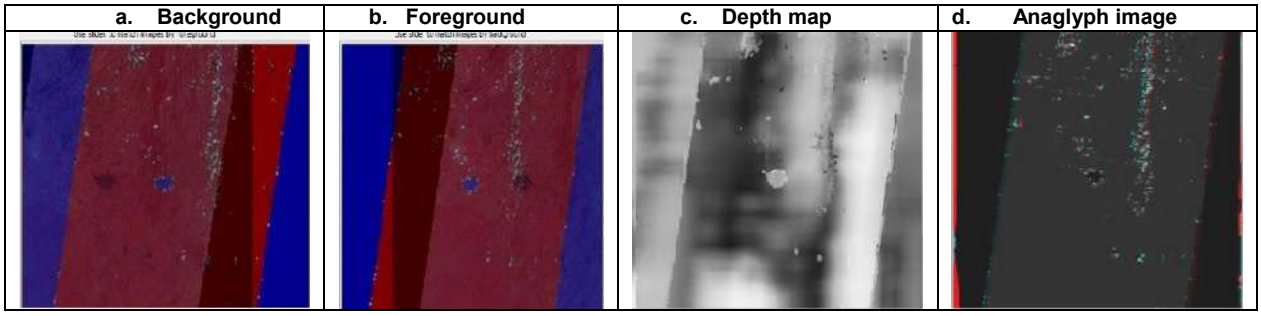


Fig: 9. Depth images of a. background b. foreground c. depth map and d. anaglyph images of with geometric correction (WGC-C-D)

……..

RESULTS AND DISCUSSION

Performance The proposed paper resulted overall method of WOGC-C –D and WGC-C-D resulting depth map with geometric correction (WGC)and without geometric correction(WOGC) which involves three steps as first phase deals without using geometric correction (WOGC) and with using geometric correction (WGC) has been dealt in **Figure- 2(a)(b)(c)** and tabular representation**[Table-1]** of performance analysis of metric PSNR and RMSE as per **Table- 2 and Figure- 10** has been carried on which yields improvised value (8.3992 for WOGC to 7.5597dB for WGC).

THE IIOAB JOURNAL
ISSN: 0976-3104

The clustering of WGC and WOGC has resulted in proper cluster of satellite imagery which yielded again better results as per **Table- 2** and **Figure- 10** (8.3992-11.0707dB for WOGC) and (7.5597-10.8370dB for WGC) and clustered images also been tested which gives minimum PSNR and RMSE using WOGC.

The clustered imagery has been considered for depth map generation as reconstruction with the help of stereo pair images resulting in anaglyph 3d image for better visualization of satellite imagery with an efficient unsupervised true colour composite results as per **Figure-10**.The metric PSNR and RMSE also applied to check the performance of satellite imagery (8.3992-6.9112dB for WOGC) and (7.5597-6.5697dB for WGC) as per **Figure - 11** with related **Table- 2**.

TABLE: 2. PERFORMANCE OF PSNR AND RMSE VALUES

Input Images	PSNR Value-WOGC	RMSE Value-WGC	PSNR Value-clustered	RMSE Value-clustered	PSNR Value-depth map	RMSE Value-depth map
WOGC	8.3992	13.7734	11.0707	13.7734	6.9112	10.3854
WGC	7.5597	12.9627	10.8370	12.9627	6.5697	8.9570

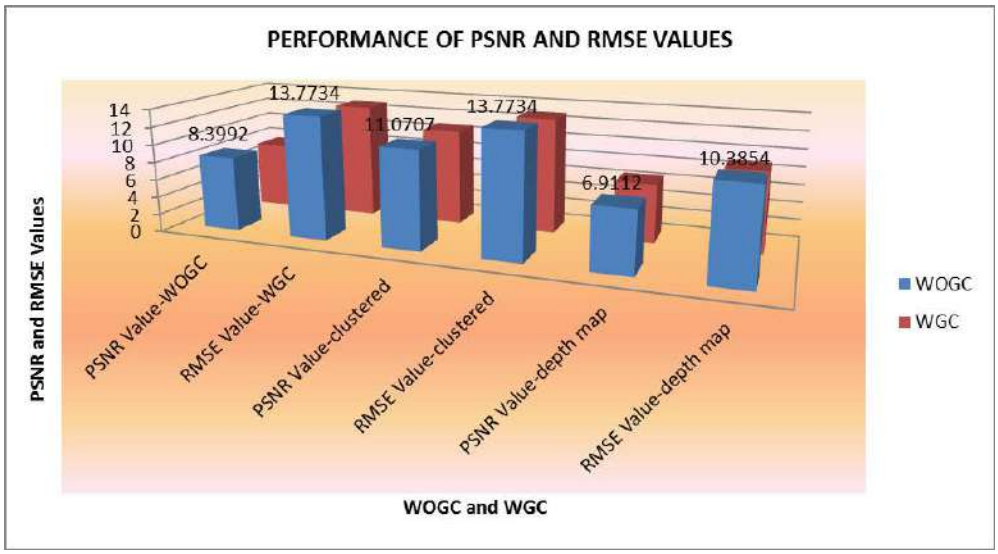
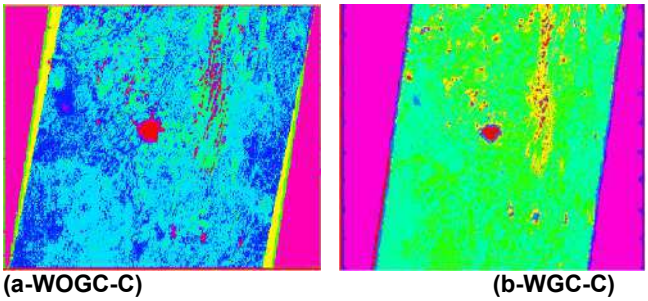


Fig: 10. Overall performance of PSNR and RMSE of WOGC,WGC for original, clustered images and Depth Maps



(a-WOGC-C) **(b-WGC-C)**

Fig: 11.(a)(b) Accuracy of True Color Composite of Clustered with Original Satellite Imagery

THE IIOAB JOURNAL ISSN: 0976-3104

CONCLUSION

The proposed work of the paper deals with the Depth Reconstruction using Geometric Correction with Anaglyph approach for Satellite imagery. It discuss about without geometric correction (WOGC) and with geometric correction (WGC) as two ways for clustering using Rough Intuitionistic Fuzzy C-Means algorithm as WOGC-C and WGC-C. The resampling techniques have been used at WOGC and WGC level with PSNR and RMSE [0.036559 rate] at WOGC-C and WGC-C level with PSNR and RMSE[7.5597-10.8370dB]and true color composite at WOGC-C and WGC-C level with PSNR and RMSE [7.5597-6.5697dB] for reconstruction of depth map generation. At all levels, the performance of satellite imagery yielded improvised results of WGC-C-D in comparison with WOGC-C-D. As a future work, author aim to propose a method that can be used which extracts the cloud from a satellite image. This will extract R, G, and B components of a true color satellite image and converted to a true color satellite image to a 256 gray image (from one image to four images). The R-components" is used to extract the cloud from a satellite image, instead of the converted gray image to reduce time complexity.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers and the editor-in-chief of the journal for their valuable guidance which has improved the quality and presentation of the paper.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Lillesand TM and Ralph W Kiefer [2003] Remote sensing and image interpretation *Fourth Edition, John Wiley and sons inc, Singapore.*

[2] Swarnalatha P and P.[Venkata Krishna 2013] An efficient method of Bitplane filtering algorithm using convex hull of medical images *Anale.Seria Informatică*. 1(1): 9–16.

[3] Arif M. Akbar and A. M. Wu (2006) Most favorable automatic georeferencing based on GCPs selection using least square method *proceedings of 4th International Conference on Digital Earth*: 25–26.

[4] Ramakrishnan.R, Manthira Moorthi.S, Prabu.S and P Swarnalatha [2013] Comparative Analysis of Various Methods for Preprocessing of Satellite Imagery *International Journal of Engineering and Technology (IJET),* 5(1):431–437.

[5] Prabu Sevugan, Ramakrishnan.R, Manthira Moorthi.S and Swarnalatha.P (2013) Non Parametric Model Re-Sampling Techniques for Geometric Correction of Remote Sensing Satellite Images *International Journal of Applied Engineering Research (IJAER),* 8(14):1611-1621.

[6] Manthira Moorthi S., Kayal Raja,.Ramakrishnan R and P.K.Srivastava [2008] RESOURCESAT-1 LISS-4 MX bands on ground co registration by in-flight calibration and attitude refinement *International Journal of Applied Earth Observation and Geoinformation,* 10 (2):140–146.

[7] Tripathy B.K.and P.Swarnalatha [2013] A Novel Fuzzy C-Means Approach with Bit Plane Algorithm for Classification of Medical Images *IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN 2013)*, 360 – 365.

[8] Swarnalatha P. and B.K.Tripathy [2013] Depth Computation using bit plane with clustering techniques for satellite images: *accepted for publication in International journal of earth sciences and engineering,"in press",* ISSN: 0974—5904.

[9] Triaxes® StereoTracer [2013], *version 7.0, Triaxes Lab LLC. Russia, Tomsk.*

[10] Shorter Oxford English Dictionary [1968] *Mutual inclination of two lines meeting in an angle.*

[11] Jump upJump up "Parallax". Oxford English Dictionary (Second Edition ed.).[1989] stron. Apparent displacement, or difference in the apparent position, of an object, caused by actual change (or difference) of position of the point of observation; spec. the angular amount of such displacement or difference of position, being the angle contained between the two straight lines drawn to the object from the two different points of view, and constituting a measure of the distance of the object.

[12] Tripathy.B.K., Rohan Bhargava, Anurag Tripathy, Rajkamal Dhull, Ekta Verma and P.Swarnalatha [2013] Rough Intuitionistic Fuzzy C-Means Algorithm and a Comparative Analysis *in proceedings of ACM Compute-2013, ,International Conference, SITE, VIT University.*

[13] Kesikoğlu, M. H. Atasever, Ü. H., C Özkan. [2013] Unsupervised change detection in satellite images using fuzzy c-means clustering and principal component analysis *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 7(2) : 129–132.

[14] Li XL, Tian YC, Xing KZ, GQ Ma [2015] Cloud Extraction of Satellite Image Using Fuzzy C-Means Clustering Approach *Applied Mechanics and Materials,*743:289–292.

[15] Tripathy B.K., P.Swarnalatha [2013] A Comparative Study of RIFCM with Other Related Algorithms from Their Suitability in Analysis of Satellite Images using Other Supporting Techniques *Kybernetes,* 43(1): 53–79.

[16] Newcombe RA, Lovegrove SJ and AJ Davison Dtam. [2011] Dense tracking and mapping in real-time *in Proc. ICCV.*

[17] Stuhmer J, Gumhold S, D Cremers. [2010] Real-time dense geometry from a handheld camera *proceedings of the 32nd DAGM conference on Pattern recognition*:11—20.

[18] Seok Lee Hee, Kyoung Mu Lee. [2012] Dense 3D Reconstruction from Severely Blurred Images using a Single Moving Camera *CVPR.*

[19] http://www.crisp.nus.edu.sg/~research/tutorial/opt_int.htm

[20] Maji P, SK Pal. [2007] Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices *IEEE Transaction on System, Man and Cybernetics, Part B: Cybernetics*, 37(6):1529—1540.

[21] Swarnalatha P, Madhuri Kota, Nagarjuna Reddy Resu , G Srivasanth. [2009] Automated Assessment Tool for the Depth of Pipe Deterioration *SCOPUS-IEEE International Advance Computing Conference, IACC 2009*:721—724.

[22] Swarnalatha P and BK Tripathy [2012] Centroid Model for the Depth Assessment of Images using Rough Fuzzy Set Techniques *International Journal of Intelligent Systems and Applications,* 1(2), ISSN: 2249-9954: 20—26.

[23] http://www.en.wikipedia.org/wiki/Parallax#cite_.

ABOUT AUTHORS

Dr Prabu Sevugan have completed Bachelor of Engineering in Computer Science and Engineering from Sona College of Technology (Autonomous) and Master of Technology in Remote Sensing from College of Engineering Guindy, Anna University Chennai and one more Master of Technology in Information Technology at School of Computer Science and Engineering, Bharathidasan University Trichy. Did his Doctoral studies on Integration of GIS and Artificial Neural Networks to Map the Landslide Susceptibility from College of Engineering Guindy, Anna University, Chennai. He was a Post-Doctoral Fellow at GISE Advanced research lab, Department of Computer Science and Engineering, Indian Institute of Technology Bombay. He has more than 45 publications in national and international journals and conferences. He organized 3 International Conferences which includes one IEEE Conference as chair and also participated in many workshops and seminars. He is a member of many professional bodies and senior member of IACSIT, UACEE and IEEE. He is having more than ten years of experience in teaching and research. Currently I am working as a Division Chair for Parallel and Distributed Computing, School of Computing Science and Engineering, VIT University Vellore.

B.K.Tripathy is a senior professor in the school of computing sciences and engineering, VIT University, Vellore, since 2007. He has produced 26 PhDs, 13 M.Phils and 03 M.S students so far. He has published more than 340 papers in different international journals, conference proceedings and edited research volumes. He has edited two research volumes for the IGI publications and has written 02 books on Soft Computing and Computer Graphics. He is in the editorial board or review panel of over 60 international journals including Springer, Science Direct, IEEE and World Scientific publications. He is a life member/ senior member/member of 21 international forums including ACM, IEEE, ACEEE and CSI.His current interest includes Fuzzy Sets and Systems, Rough sets and Knowledge Engineering, Multiset Theory, List Theory, Data clustering and Database Anonymization, Content Based Learning, Neighbourhood Systems, Soft Set Analysis, Image Processing, Cloud Computing, Social Internet of Things, Big Data Analytics, Multi Criteria Decision Making and Social Network Analysis.

Swarnalatha Purushotham is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Systems. She has published more than 50 papers in International Journals/International Conference Proceedings/National Conferences. She is having 14+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.

THE IIOAB JOURNAL
ISSN: 0976-3104

S. Manthira Moorthi received his M.sc in Applied Physics from Gandhigram University, Tamil Nadu, in 1989. He has been working in Signal and Image Processing Group, Space Applications Centre (ISRO), Ahmedabad as a scientist since 1991. He is involved in the development of methods, approaches, algorithms and software systems for earth and planetary remote sensing data pre and post processing. His areas of interests are geometric and radiometric processing, automatic image registration techniques, image analysis, geodetic procedures, information models, space science data management, and planetary data models.

He obtained his M.Sc. (Applied Mathematics) from University of Madras (Madras Institute of Technology) in 1977 and Ph.D in Mathematics form U.P.Technical University , Lucknow in 2004. Joined as scientist at Space Applications Centre in 1978 and leading Data Products Software Group. His contributions are in the area of development of software systems for Indian Remote Sensing data processing as well as Meteorological Data Processing such as INSAT for operational needs. His research area includes distributed computing, software architectures, Software Quality assurance and Image Processing of optical remote sensing data.

ISSN: 0976-3104
SPECIAL ISSUE (SCMSA)
Swarnalatha and Tripathy

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

ANALYSIS OF DEPTH USING CLUSTERING TECHNIQUES WITH BIT PLANE FILTER FOR SATELLITE IMAGERY

Swarnalatha Purushotham and Balakrushna Tripathy*

VIT University, Vellore-Tamilnadu, INDIA

ABSTRACT

Calculation of measurements has noteworthy inspiration in different applications now a day. The count of sizes from the profundity of satellite pictures is fundamental for better representation, for recognizable proof of ranges, climate gauging, and so on. Fitting grouping strategies will give better yield for profundity calculation of satellite pictures with least time many-sided quality. The utilization of bit plane channel technique produces improved pictures with least epsilon esteem which helps in showing signs of improvement grouping of pictures. In this paper we apply the strategies for bit plane channel strategy with fuzzy c-means (BPFCM) and bit plane unpleasant c means strategy (BPRCM) on satellite pictures and it is watched that both of these systems give upgraded groups for satellite pictures contrasted with traditional techniques. Likewise, we set up that BPRCM gives a superior upgraded grouped picture in contrast with BPFCM. Furthermore we perform the profundity calculation of satellite pictures, which demonstrates that our methodology (BPRCM) needs least time unpredictability and edge esteem as for the customary techniques and in addition BPFCM for the bended satellite pictures.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26th–March-2016
Published on: 19th–May-2016

KEY WORDS

Bit Plane Filter; Otsu Thresholding; Mathematical Methods; Clustering; Depth Computation; Satellite Images

***Corresponding author: Email:** tripathybk@vit.ac.in, **Tel:** +40-9001010010; **Fax**: +40-9001010012

INTRODUCTION

The satellite pictures acquired by remote detecting frameworks are not frequently adequate for high exactness applications because of different mutilations. The twists can be ordered as geometric, radiometric and atmospheric corrections. The satellite pictures manage three errors/mutilations, and the objective of this paper is to correct satellite pictures from such mistakes by suitable grouping of layers. The wellsprings of geometric mutilation incorporate sensor impacts, stage impacts, object impacts, and atmospheric impacts. We will quickly say a few wellsprings of mutilations. In a raw picture the pixel size on the ground is not steady since it varies with the scanning angle. On the off chance that the speed of the satellite in respect to the ground is not steady, the separation between two neighboring sweep lines is not consistent. Further contortions are brought on by the way that the elevation of the satellite changes with time.

What's more, atmospheric refraction might bring about geometrical distortions. A portion of the mistakes are deliberate, while a percentage of the blunders happen pretty much arbitrarily. By geometric correction, a picture is gotten from the raw picture such that it has the same geometry as a source of perspective picture or a very much characterized map projection.

For every pixel in the amended picture, we need to decide a mapping to the raw picture. Satellite pictures are helpful for checking changes in land use and land cover. Yet, significant issue with these pictures is that locales underneath mists are not secured by sensor. The picture distortion because of overcast spread is an established issue of unmistakable band of remote detecting symbolism. Particularly, for non-stationary satellite, it is normally found in the earth asset perception application. Expelling overcast spread from satellite symbolism is exceptionally helpful for helping picture understanding. Consequently cloud identification and evacuation is extremely essential in the preparing of satellite symbolism.

THE IIOAB JOURNAL ISSN: 0976-3104

Further it is more hard to measure and decipher changes on multi-fleeting pictures under various brightening, climatic or sensor conditions without radiometric adjustment. The relative way to deal with radiometric remedy, known as relative radiometric standardization is favored.

Then again, atmospheric particles diffuse the daylight into the sensor's field of perspective specifically, bringing about a radiation that does not contain any surface data by any means. The consolidated climatic impacts because of disseminating and retention are wavelength subordinate, shift in time and space, and rely on upon the surface reflectance and its spatial variety [1].

With a specific end goal to uproot mutilations, the bit plane filter calculation is utilized for pictures which improves the picture by isolating a picture into an arrangement of bits relating to a given piece position in each of the 0's and 1's speaking to a satellite picture. Furthermore satellite pictures can be isolated into cuts to locate an upgraded suitable piece planes to decide the trifling data of the satellite picture.

The edge location strategy is utilized for portioning a satellite picture into edges. There are four distinctive ordinary strategies i.e., Canny, Sobel, Prewitt, Zero-Cross and Robert-Cross Methods. The issue of false edge location gives dainty or thick lines which was vexed because of commotion and so on., has been utilized by diminishing the aggregate of information and channels out futile data, by safeguarding the critical auxiliary properties in a satellite picture. This kind of satellite picture is useful for applications inside and out (three dimensional spaces) calculation, picture recreation, and so on.

What's more, the calculation of depth was prepared by finding the third measurement of a clustered satellite picture utilizing traditional strategies. The clustered satellite picture is important for further preparing of picture reproduction which is fundamental for climate gauging or recognizable proof of locales, and so on of societal purposes. In [Swarnalatha et al, 2009], the misshaped pictures which are upgraded by utilizing a specific force appraisal, that exists underneath the past quality utilizing edge has been found to not to give precise results

As the routine strategies don't give legitimate division, fuzzy c-means and rough c-means clustering algorithms have been utilized as a part of this paper. To be more exact, in this paper, we utilize three grouping procedures (conventional, fuzzy c-means and rough c-means) utilizing with and without bit plane methodology for a satellite picture. One of the proposed strategies in this paper is the bit plane rough c-means (BPRCM) which yields better grouped pictures contrasted with traditional and bit plane channel technique with fuzzy c-means (BPFCM). As the depth calculation is crucial to locate the mutilated segment of a satellite picture or the control of various regions, it turns into a period expending process which is essential for further remaking of a satellite picture. We have processed the time complexity of changing a satellite picture utilizing the proposed BPRCM technique.

The skeleton of the paper is as per the following. Segment 2 manages writing overview identified with the bit plane strategy, conventional methods and clustering techniques of a satellite picture. This dialog additionally incorporates fuzzy c-means, rough c-means and time complexity for depth of a satellite picture. In segment 3 we portray our proposed strategy with bit plane channel for conventional, fuzzy c-means and rough c-means. In segment 4, we give a similar investigation of the diverse procedures such as the traditional strategies, fuzzy c-means, rough c-means, where for every situation we consider two adaptations of utilizing bit plane strategy and without utilizing it for better grouping of satellite pictures. Additionally an examination has been made on their execution basing upon test results. Segment 5 manages depth calculation and time complexity nature. Likewise, in this area we compress our commitment in this paper under conclusions. We propose some future work in this bearing under area 6. We introduce the source materials referred amid the readiness of this work in the references segment.

LITERATURE SURVEY

In the review, it has been found that distinctive picture upgrade methods have been actualized and have effectively removed the components of the affected/distortion areas/regions which is not precise.

Nowadays, satellite imaging sensors use multispectral or even hyper spectral devices, which results in acquiring multiple images require advanced techniques and experimental methods for processing [2].

THE IIOAB JOURNAL ISSN: 0976-3104

There are many edge detection methods, which are used to visualize different layers existing in satellite images obtained through satellites. Edge detection techniques (Sobel detector, Robert Cross detector, Canny edge detector, Zero-based detection) are used to extract boundaries[13]. As a result of an abrupt change in brightness levels in the satellite images, we cannot obtain the correct smooth edges. That is why satellite images can be segmented using some methods handling imprecision. Some of these methods are the fuzzy c-means method introduced in [3] and the rough c-means [4]. In this paper, we shall be using these two methods for segmentation of satellite images.

Edge detection significantly reduces the amount of data and filters out useless information when storing the significant structural properties in an image. A comparative analysis of various Image Edge Detection techniques can be found in [5].

From a different prospective a vague spatial phenomenon has been dealt in [6] for the improvement of characterisation and quantification of vegetative drought. Vegetative drought is characterised using a membership function to model the gradual transition between drought and non-drought classes. A Crisp approach, using the median of the transition range as the threshold value, does not quantify the vagueness of vegetative drought. A membership function is used to represent the quantification of vegetative drought in order to form the steady transition between drought and non-drought classes The procedure has been implemented using fuzzy set to quantify the areas those have vagueness of vegetative drought.

The auto-searching and matching algorithm is introduced and miss-matching elimination is used in dealing with ortho-rectification of images aided by GCP (Ground Control Point) image databases method [7].

A hybrid lossless algorithm based on simple selective scan order with bit plane slicing method [8] is used for lossless Image compression of limited bits/pixel images, such as medical images, satellite images and other still images common in the world.

In [9] discussions have been made about the updating the map in GIS, environmental inspection, transportation and urban planning, etc for the purpose of speedy retrieval of road network that is useful in. And using fuzzy theory, for urban area from multi-spectral IKONOS imagery, an automatic road extraction algorithm was developed.

The paper [16] deals with the calculation of depth of the defect, which is very vital as the measure to which the image has been depreciated, can be estimated from this. The angle of incidence of X-rays and the thickness of the image are taken as an input from the user and thereby the dimensions of the affected areas such as the length, breadth and depth have been found which is not accurate in results for further reconstruction [11][13].

In [14] a three step method is presented, which is a simple, robust and efficient one to detect defects in the underground concrete pipes. It identifies and extracts defect-like structures from pipe images whose contrast has been enhanced. The objective of the paper is to reduce the effort and the labour of a person in detecting the defects in underground pipes.

The above issues have been solved individually so far. But in this paper we deal with a method to detect edges of distortions using conventional methods, fuzzy c-means and rough c-means techniques which cluster an enhanced image using bit plane filter method with minimum thresholding. Clustering can be used to segment an enhanced image using bit plane method into several clusters with Otsu thresholding and statistical techniques we are able to remove unwanted clusters. Finally, image is edge detected, where a clear boundary is obtained. As a result, the bit plane rough c-means performs better than existing edge detection methods which are useful for the assessment of time complexity of depth computation with minimum value compared to the existing tools [11][13] and bit plane fuzzy c-means algorithm.

Depth computation is essential for better visualization of images for satellite applications. The disorder portion of satellite images may lead to many natural disasters of rocket launching due to inaccurate classification and clustering of satellite images.

THE IIOAB JOURNAL ISSN: 0976-3104

To have proper analysis of the above applications, using statistical moments, the third dimension calculation was implemented using centroid model gives less accurate results. Hence in the paper [11][13], we aimed at developing an approach for better classification and clustering of the affected portion which may be helpful for depth computation and further process.

METHODOLOGY

The architecture of the paper provided in fig.1 provides a clear idea about the methodology. In this section we discuss on some of the key steps involved in the process.

Bit Plane Slice Method

This is a preprocessing technique which removes distortions from an image, yielding a better image which can be used for further processing.

Input Image: Ordinary Satellite image

Step 1: Declare an array of variables for matrix and to read pixel values and an integer of variables of n, m.
Step 2: Exercise loop for increment to the 'j' variable and count the occurrence of 1's in the first 3 most significant bit slices.
Step 3: Then use the method which will compute mean(standard deviation, variance) value for the declared array by applying binary
Addition operator. And to the calculated mean value, we have to replace the center pixel by iterating for the whole satellite image in the variable of an img.
step 4: By preserving edges of a satellite images, we can remove noise thereby.
for every row and column, till < than n and m Method
for (row to n){ for (col to m) {} }
Step 4.1: And to arrive the bit slices, declare the bit 1 to bit 8 array variables,
Step 4.2: And do instantiation as set pixels values ,
pixels[i] ∈ {1,2,…8}
initially, updated by bit planes loop number ,
i ∈ {1,2,….n*m},
initially 0, will be instantiated to a integer value.
Step 5: End
Step 6: Assign cbp=zeros(size(img));
Step 7: Compare the planes to display the correct bit plane using iteration
cbp=bs(cbp,1,pixels[15])||(cbp,2,pixels[16]) ……
Step 8: Display the cbp
Step 9: End
Output Image: Bit Plane Satellite image

The bit plane algorithm is used to partition an image to 0-7 slices. As for illustration the 8 bit value 10011011 will become 155 in decimal. Pre-processing the satellite image should be carried out for the detection and extraction of the significant features [17].

Thereby reducing the first bit slice/plane of a satellite image, plane by plane till gives the final slice as a value of $2^{(m-n)}$ with better approximation having trivial information [18]. The cropped bit plane (cbp) can be applied for a particular portion of a satellite image for further interpretation.

Conventional Edge Detection Techniques

The problem of false edge detection [12] gives thin or thick lines which was troubled due to noise etc., was used by reducing the total of data and filters out useless information, by preserving the significant structural properties

ISSN: 0976-3104

in a satellite image. We can have four conventional edge detection techniques (Sobel, Canny, Robert Cross and Zero-Cross). The "Sobel operator" edge detection method is simple enough to detect the edges and their orientation negatives which are sensitive to noise.

The Zero Cross edge detection method aims in detection of edges and their orientations having fixed characteristics in all directions that can respond to some of the existing edges, sensitivity to noise. Canny edge detection method can use probability for finding error rate, localization and response which improves signal to noise ratio with better detection specially in noise conditions. But it involves complex computations, false zero crossing and also it is time consuming.

The Robert Cross edge detection method deals with the properties of the produced edges which should be well-defined, the background should contribute as little noise as possible and the intensity of edges should match as close as likely to what a human would recognize [17]. The Canny Edge Detection technique is used for the detection of high range of edges in satellite images that may lead to detection and localization as good. Filters as horizontal, vertical and diagonal edges can be detected using canny edge detection algorithm to determine the intensity gradient of the satellite image by way of as given below:

$$\Theta = \arctan\left(\frac{G_y}{G_x}\right) \quad (3.2.1)$$

$$G = \sqrt{G_x^2 + G_y^2} \quad (3.2.2)$$

Otsu Thresholding, Statistical Methods and Histogram Analysis

Thresholding is computationally inexpensive and fast. It is one of the oldest segmentation methods and is still widely used in simple applications. Using range values or threshold values, pixels are classified and clustered using either of the thresholding techniques like global and local thresholding. Global thresholding method selects only one threshold value for the entire satellite image. Local thresholding selects different threshold values for different regions. Structuring elements are applied to the pixels of the satellite image. That is, using the structuring elements the pixels in the satellite image can be clustered and classified into different classes and then by performing the set difference operation the features of the affected area can be extracted from the satellite image for which the horizontal structuring element must be varied. A particular intensity value is considered and all the pixels whose intensity values lie below that value are obtained.

Statistical Methods

The different mathematical methods can be applied on the bit-plane to get the enhanced satellite image [11-13] In the paper, root mean square error and peak signal noise ratio has to be computed for better interpretation of a satellite image.

RMSE and PSNR Value

Peak Signal-to-Noise Ratio can be characterized as **PSNR,** which is the relation with the majority, likely power of a signal and the power of corrupting distortions that influence the fidelity of its demonstration to validate the performance of the clustering techniques.

The PSNR value can be computed through mean squared error(MSE). For an example distortion-free mxn monochrome satellite image 'I' with its noisy approximation 'K', MSE can be represented as given below:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \quad (3.3.2.1)$$

ISSN: 0976-3104

Hence, the PSNR is defined as:

$$PSNR = 10.\log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

(3.3.2.2)

$$= 20.\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

(3.3.2.3)

$$= 20.\log_{10}\left(MAX_I\right) - 10.\log_{10}(MSE)$$

(3.3.2.3)

where MAX_I is the most possible 0's and 1's values of a satellite image. And it will be replaced with 255, as and when the 0's and 1's are given using 8 bits per model. And MSE will become '0', when the distortion is null indicating that the two input satellite images are same.

Here, MAX_I is the maximum possible pixel value of the satellite image. When the pixels are represented using 8 bits per sample, this is 255. In the absence of noise, the two satellite images I and K are identical, and thus the MSE is zero. In this case the PSNR is undefined**.**

Histogram Analysis

Histogram can be applied to review graphically the satellite images have resulted in terms of allocation and deviation. Let the variable 'u' represent the gray levels of the satellite image to be enhanced. We assume that 'u' has been normalized to the interval [0, 1], with u = 0 representing black and u = 1 representing white.

Later, we consider a discrete formulation and allow pixel values to be in the interval [0, G-1] where G is the highest gray level value.

For any 'u' satisfying the aforementioned conditions, we focus attention on transformations of the form

$s = T(u) \quad 0 \le u \le 1$

(3.3.4.1)

that produces a level s for every pixel value r in the original satellite image. We assume that the transformation function T(u) satisfies the following conditions:

(a) T(u) is single - valued and monotonically increasing in the interval $0 \le u \le 1$; and

(b) $0 \le T(u) \le 1$ for $0 \le u \le 1$.

(3.3.4.2)

Let $p_u(u)$ and $p_v(v)$ denote the probability density functions of random variables u and v, respectively, where the subscripts on p are used to denote that p_u and p_v are different functions.

For discrete values we deal with probabilities and summations instead of probability density functions and integrals. The probability of occurrence of gray level u_k in a satellite image is approximated by $P_u(u_k) = n_k/n \quad k = 0,1,2,\ldots, G-1$ (3.3.4.3)

THE IIOAB JOURNAL ISSN: 0976-3104

Where, n is the total number of pixels in the satellite image, n_k is the number of pixels that have gray level u_k, and G is the total number of possible gray levels in the satellite image. The discrete version of the transformation function given in Eq. (3.3.4.2) is

$v_k = T(u_k) = \sum p_u(u_j) = \sum n_j/n \quad k = 0,1,2,...G-1$ (3.3.4.4)

Thus, a processed (output) satellite image is obtained by mapping each pixel with level u_k in the input satellite image into a corresponding pixel with level v_k in the output satellite image via Eq. (3.3.4.4).

Fuzzy C Means (FCM)

Here, we are presenting the FCM which is used for data clustering. For this, first we need to define the concept of a Fuzzy Set.

Definition

A fuzzy set A is determined by its membership function μ_A, where $\mu_A : X \rightarrow [0.1]$ such that every x ∈ X is associated with its grade of membership in A, $\mu_A(x)$. If an element does not belong to A then $\mu_A(x)$=0. The closer the membership value $\mu_A(x)$ to 1, the more x belongs to A. The grade 1 represents full membership, [19].

Fuzzy C-Means Algorithm

Let X={ x_1,….,x_j,,……., x_n } be the set of n objects and V={ v_1 ….,. v_i ,…..,v_c } be the set of c centroids (means), where x_j ∈to $\Re^m$, v_i ∈ $\Re^m$ and v_i ∈ X. The FCM provides a fuzzification of the HCM by [3]. It partitions X into c clusters by minimizing the objective function.

$$J = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^{m_1} \| x_j - v_i \|^2$$

(3.3.5.1.1)

Where 1<= m_1 < ∞ is the fuzzifier, v_i is the ith centroid corresponding to cluster β_i, μ_{ij} ∈[0,1] is the probabilistic membership of the pattern x_j to cluster β_i , and ||.|| is the distance norm, such that

$$v_i = \frac{1}{n_i}\sum_{j=1}^{n}(\mu_{ij})^{m_1} x_j, where \qquad n_i = \sum_{j=1}^{n}(\mu_{ij})^{m_1} x_j,$$

(3.3.5.1.2)

$$\mu_{ij} = (\sum_{k=1}^{c}(\frac{d_{ij}}{d_{kj}})^{\frac{2}{m_1-1}})^{-1}, where \qquad d_{ij} = \| x_j - v_{j|} \|^2$$

(3.3.5.1.3)

Subject to $\sum_{i=1}^{c}\mu_{ij} = 1, \forall_j, and\ 0 < \sum_{j=1}^{n}\mu_{ij} < n, \forall_i.$ The process begins by randomly choosing c objects as the centroids of the c clusters. The memberships are calculated based on the relative distance of the object x_j to the centroids { v_i } by (3.3.5.1.3). After computing the memberships of all the objects, the new centroids of the clusters are calculated as per (3.3.5.1.2).

The process stops when the centroids stabilize. That is, the centroids from the previous iteration are identical to those generated in the current iteration.

In the FCM, the memberships of an object are inversely related to the relative distance of the object to the cluster centroids. In effects, it is very sensitive to noise and outliers. In addition, from the standpoint of "compatibility

ISSN: 0976-3104

with the centroid," the membership of an object x_j in a cluster β_i should be determined solely by how close it is to the mean(centroid) v_i of the class and should not be coupled with its similarity with respect to other classes.

Rough C-Means (RCM) Algorithm

The notion of rough sets was introduced in [18] as an extension of the concept of crisp sets. We provide below the definition of a rough set.

Let U (≠∅) be a finite set of objects, called the universe and R be an equivalence relation over U. By U/R we denote the family of all equivalence classes of R (or classification of U) referred to as categories or concepts of R and $[x]_R$ denotes the class of x in R containing elements of U related to x by the relation R. By a Knowledge base, we understand a relational system K = (U, *R*), where U is as above and *R* is a family of equivalence relations over U.

For any subset P ($\neq \varnothing$) $\subseteq R$, the intersection of all equivalence relations in P is denoted by IND (P) and is called the indiscernibility relation over P. The equivalence classes of IND (P) are called P- basic knowledge about U in K. For any Q∈*R*, Q is called a Q-elementary knowledge about U in K and equivalence classes of Q are called Q-elementary concepts of knowledge *R*. The family of P-basic categories for all $\varnothing \neq P \subseteq R$ will be called the family of basic categories in knowledge base K. By IND (K), we denote the family of all equivalence relations defined in k. Symbolically, IND (K) = {IND (P): $\varnothing \neq P \subseteq R$}.
For any X⊆U and an equivalence relation R∈ IND(K), we associate two subsets, $\underline{R}X = \bigcup\{Y \in U/R : Y \subseteq X\}$ and $\overline{R}X = \bigcup\{Y \in U/R : Y \cap X \neq \varnothing\}$, called the R-lower and R-upper approximations of X respectively. The R-boundary of X is denoted by $BN_R(X)$ and is given by $BN_R(X) = \overline{R}X - \underline{R}X$. The elements of $\underline{R}X$ are those elements of U which can be certainly classified as elements of X employing knowledge of R. The borderline region is the undecidable area of the universe. We say X is rough with respect to R if and only if $\underline{R}X \neq \overline{R}X$, equivalently $BN_R(X) \neq \varnothing$. X is said to be R- definable if and only if $\underline{R}X = \overline{R}X$, or $BN_R(X) = \varnothing$. So, a set is rough with respect to R if and only if it is not R-definable.

$$\mathbf{v}_i = \begin{cases} w_{\text{low}} \dfrac{\sum_{\mathbf{x}_k \in \underline{B}U_i} \mathbf{x}_k}{|\underline{B}U_i|} + w_{\text{up}} \dfrac{\sum_{\mathbf{x}_k \in (\overline{B}U_i - \underline{B}U_i)} \mathbf{x}_k}{|\overline{B}U_i - \underline{B}U_i|}, & \text{if } \underline{B}U_i \neq \emptyset \wedge \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ \dfrac{\sum_{\mathbf{x}_k \in (\overline{B}U_i - \underline{B}U_i)} \mathbf{x}_k}{|\overline{B}U_i - \underline{B}U_i|}, & \text{if } \underline{B}U_i = \emptyset \wedge \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ \dfrac{\sum_{\mathbf{x}_k \in \underline{B}U_i} \mathbf{x}_k}{|\underline{B}U_i|}, & \text{otherwise} \end{cases} \quad \textbf{(3.3.7.1)}$$

RCM Algorithm

Rough c-means algorithm was introduced by [4], which describes a cluster by its centroid and its lower and upper approximation. In rough c-means an object can belong completely in one cluster or in between two clusters. The lower and upper approximations are weighted differently. Since the objects in the lower approximation completely belong to the cluster, therefore they are assigned a greater weight denoted by w_{low}. The objects in the upper approximation as assigned a relatively lower weight denoted by w_{up}. The algorithm is given as follows:

1. Assign initial means V_i for c clusters.
2. Let $d_{i,k}$ be the minimum and $d_{j,k}$ be the next to minimum distance of x_k from clusters U_i and U_j. Assign each data object to the lower or upper approximation by computing $d_{j,k} - d_{i,k}$.
3. If $d_{j,k} - d_{i,k}$ is less than threshold (ε) then
   $x_k \in \overline{B}U_i$ and $x_k \in \overline{B}U_j$ and is not the member of any lower approximation.
   else $x_k \in \underline{B}U_i$

THE IIOAB JOURNAL ISSN: 0976-3104

4. Calculate new centroids for each cluster using equation 3.3.7.1
5. Repeat from step 2 until there are no more assignment.

Our approach is based on fuzzy c-means (FCM) and rough c-means (RCM) with bit plane filter method. In fact we shall be using the Bit Plane Conventional Methods (BPCM), Bit Plane Fuzzy C- Means (BPFCM) and Bit Plane Rough C-Means (BPRCM), which are obtained as follows:

1. Bit Plane Filter Algorithm where the satellite images are divided into slices
2. Otsu Thresholding and statistical Methods with conventional edge detection techniques.
3. Otsu Thresholding and statistical methods with Fuzzy C-Means
4. Otsu Thresholding and statistical methods with Rough C-Means for better clustering and classification for depth computation and future analysis.

As the first method, Otsu thresholding (OT) is applied for a original satellite image that can be used for the bit plane filter, which divides the satellite images into slices to have better visuality and methodology as follows: a satellite image is divided into a set of bits corresponding to a given bit position in each of the 0's and 1's which represents a satellite images. And that satellite image can be divided into slices to determine the trivial information with updated Otsu thresholding of the Bit Plane (BOT).

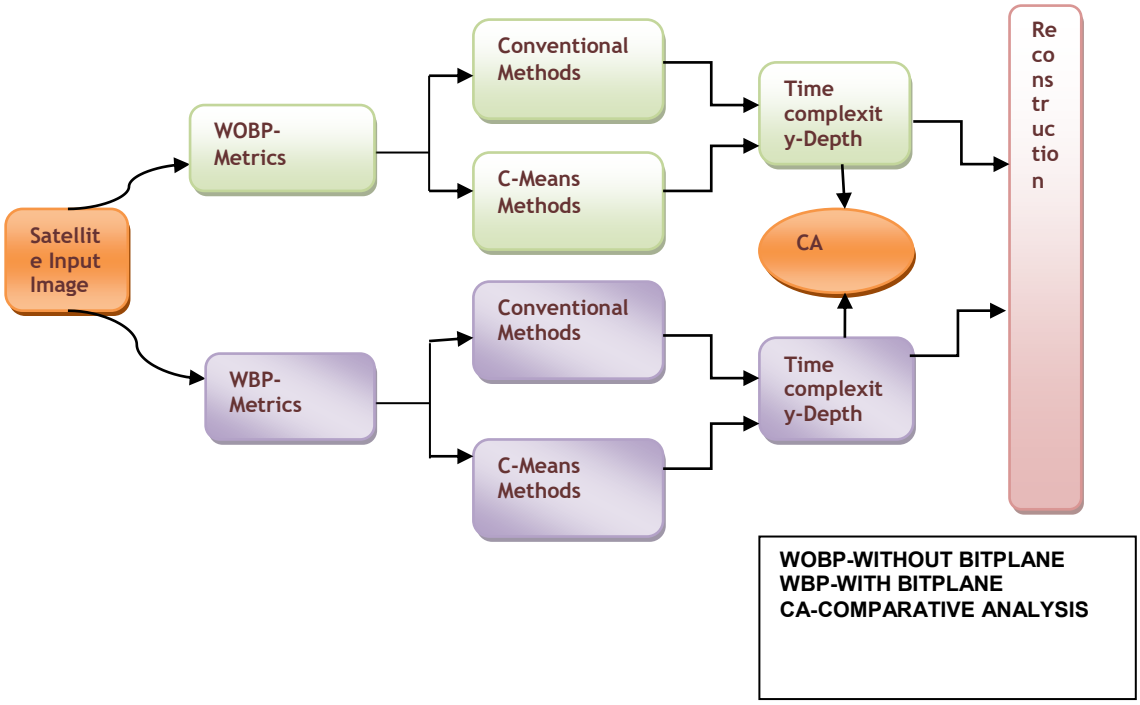


Fig: 1. Block Diagram of Depth Computation of Time Complexity using Without Bit Plane Fuzzy C-Means, Rough C-Means, Conventional Methods

……..

And as the second method, for the bit plane sliced satellite image can be used for conventional edge detection techniques (Sobel, canny, zero-based, Robert cross etc,) with Conventional Otsu thresholding (COT) and Statistical Methods.

In the third method, for the bit plane satellite image, we have to apply fuzzy c-means with Fuzzy Otsu Thresholding (FOT). In the fourth method, for the bit plane satellite image, we have to apply rough c-means with Rough Otsu Thresholding (ROT).

THE IIOAB JOURNAL ISSN: 0976-3104

Hence, obviously there is a necessity to generate results using the conventional edge detection techniques, fuzzy c-means and rough c-means with bit plane and without bit plane for better clustering of a satellite image. And comparison of three clustering techniques (conventional, fuzzy c-means, rough c-means) has been applied to a satellite image to extract control points for depth computation and further reconstruction of an image.

Figure– 1 deals with four levels as 1) with gray scale satellite image as an input with Original Otsu Thresholding (OOT) for two categories in level 2) of Without using Otsu Thresholding (WOOT) /Bit Plane and using With using Otsu Thresholding/Bit Plane (WOT). In comparison, WOT/BOT gives an efficient performance to OOT and WOOT. In level 3) the three methods of conventional, fuzzy c-means and rough c-means should be applied for without using and with using bit plane, thereby getting Conventional Otsu Thresholding (COT), Fuzzy Otsu Thresholding (FOT) and Rough Otsu Thresholding (ROT).

Table: 1. Bit Planes with Statistical Values and Thresholds

Bit Slices with statistical values and thresholds				
Bit Slice	**Mean**	**Standard Deviation**	**Variance**	**Threshold**
Original satellite image	138.1219	122.9636	125.0	62
0	103.6880	121.5076	129.0	42
1	103.1821	121.4028	129.5	46
2	107.6164	122.2095	129.5	46
3	115.1320	123.1899	126.5	40
4	99.6205	120.6428	126.5	42
5	90.5032	118.1787	127.0	46
6	104.6774	121.8552	128.0	41
7	60.1268	104.7469	126.5	40

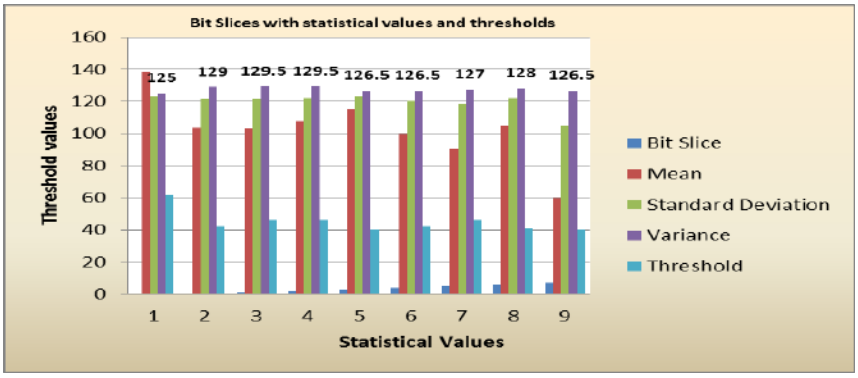


Fig: 2. Performance of Bit Planes with Statistical Values and Thresholds

……

Figure– 2, discuss about the bit planes that have been justified based on the statistical values of mean, standard deviation , variance with their thresholds.

THE IIOAB JOURNAL ISSN: 0976-3104

The **Table –1** states that satellite image **(Figure 4.a)** has been sliced into 8 planes(i.e., 0 to 7 and 8 as original satellite image for comparison) for better visualization been used as one of the filter algorithm. Thereby, the mathematical values for the 7th bit slice like mean (60.1268),standard deviation(104.7469),variance(126.5) with Otsu threshold (40) shows improved result compared to the 0th bit slice of original satellite image of mean(103.6880),standard deviation(121.5076), variance(129.0) with Otsu threshold (42).

Table: 2 Overall performance of Original satellite image using conventional methods with epsilon values of without and with bit planes

Conventional Otsu	Without bitplane	With bitplane
Original	60	51
Canny	49	102
Robert	36	102
Sobel	124	71
Zero-cross	88	123

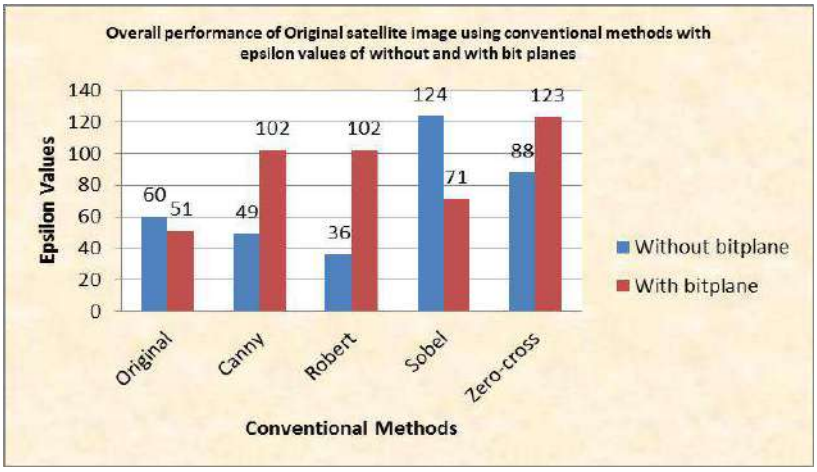


Fig: 3. Overall performance of Original satellite image using conventional methods with epsilon values of without and with bit planes

Table– 2 and **Figure– 3** deal with overall performance in figures and graphical form respectively of the original satellite image using conventional methods with epsilon values, by using bit plane technique and without using it. The epsilon values for traditional methods without bit plane give segmented satellite image with different performance values that is for one image, canny method has minimum value whereas for another Robert method does so. But the epsilon value is unique with minimum value for one method for all images Thus establishing that using bit plane, a stable reduction is possible in one method for all images in a field (satellite, medical, etc.,). The original satellite image with conventional methods without bit plane and with bit plane is given in **Figure 4.a.b.c.d. & 5.a.b.c.d.** after references of the paper.

Original Satellite image	Conventional Methods			
	Canny(a)	Sobel(b)	Zero-Based(c)	Robert-Based(d)

THE IIOAB JOURNAL
ISSN: 0976-3104

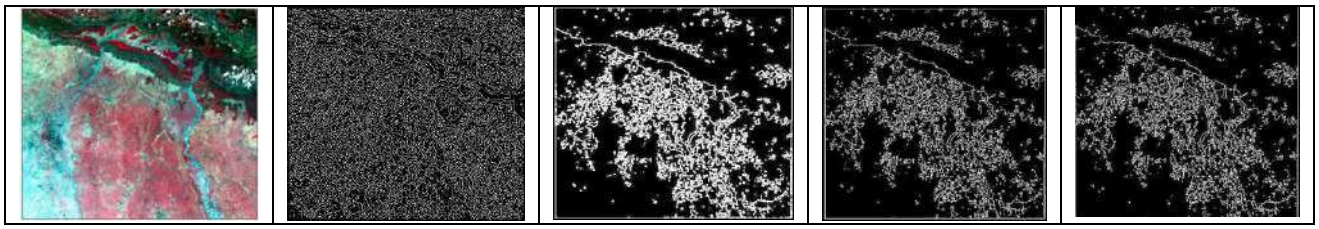


Fig: 4.a.b.c.d. Original satellite image with conventional Methods

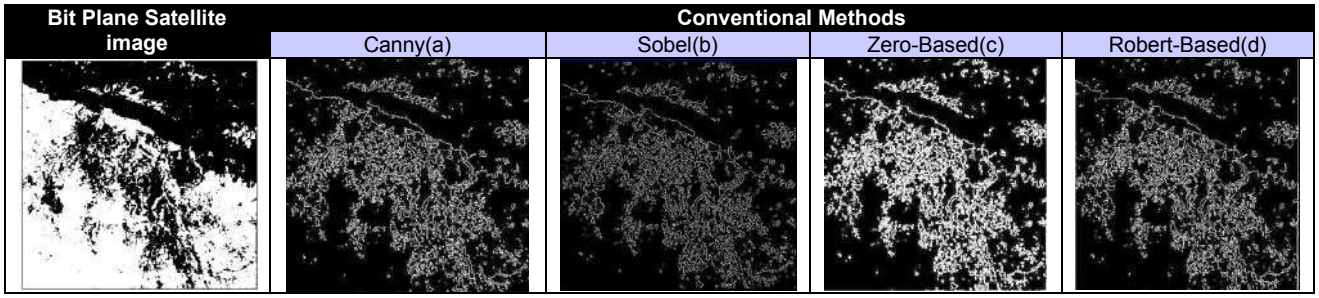


Fig: 5.a.b.c.d. Bit Plane satellite image with conventional Methods

Table: 4. Original , fuzzy c-means and rough c-means satellite images with epsilon values of without and with bit planes

	Without bitplane	With bitplane
Input Satellite image	130	122
Fuzzy c-means	126	120
Rough C-means	109	102

As a result, we can get a blurred satellite image from the given input satellite image by removing the noise present on raw unprocessed data . The **table– 3(a)(b)** discuss about satellite images using without bit plane(WOBP) and with bit plane(WBP) for traditional/conventional, fuzzy c-means and rough c-means. With histogram analysis, a graphical representation of satellite images that results in terms of allocation and deviation have been processed giving better visualization of three methods.

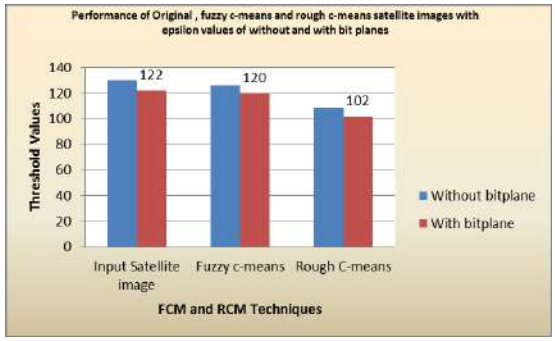
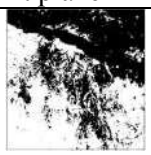
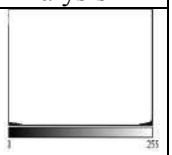
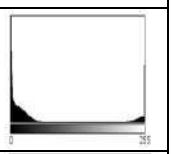
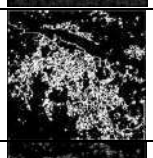
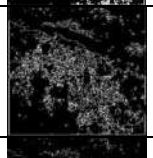
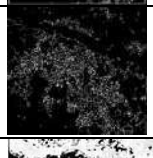
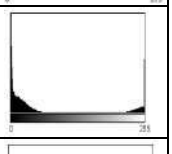
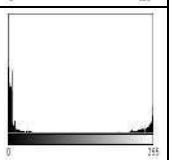
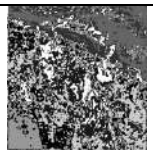
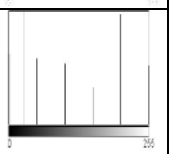
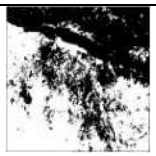
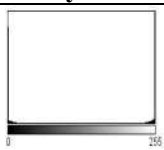
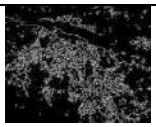
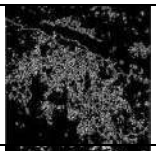
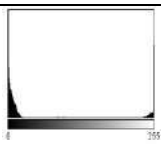
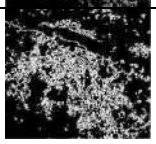
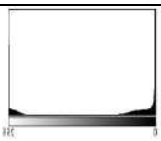
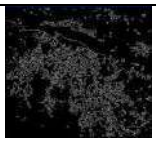
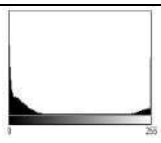
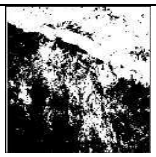
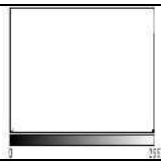
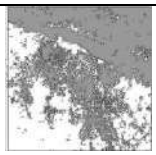
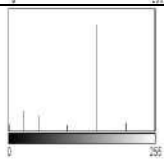


Fig: 6. Performance of Original , fuzzy c-means and rough c-means satellite images with epsilon values of without and with bit planes

THE IIOAB JOURNAL ISSN: 0976-3104

Table : 3(a_b Satellite images using without bitplane and with bitplane for traditional, fuzzy c-means and rough c-means with Histogram Analysis

Original Satellite image (a)-Without Bit plane		Histogram Analysis
Traditional Methods-Original Satellite image		
Canny Edge Detection		
Sobel Operator		
Zero-Cross Method		
Robert-Cross Method		
Fuzzy C-Means		
Rough C-Means		

Bit Plane Satellite image (b)- With Bit Plane		**Histogram Analysis**
Traditional Methods-Bit Plane Satellite image		
Canny Edge Detection		
Sobel Operator		
Zero-Cross Method		
Robert-Cross Method		
Fuzzy C-Means		
Rough C-Means		

The **Table– 4** and **Figure -6** gives the performance of threshold values for fuzzy c-means and rough c-means. The clustered techniques have been used without bitplane and with bitplane procedures. **Figure– 3** deals with using four (canny, sobel, zero-based, robert-based) conventional methods, any one of the method can be applied to one satellite image and differs for mages in one field. Thereby, Rough C-Means without bit plane and with bit plane (154, 71) gives better clustered satellite image compared to fuzzy c-means method (180,142) and conventional methods(96, 80) . Original Satellite image with fuzzy c-means and rough c-means and with that of bit plane satellite image is given in **Figure– 7, 8** for better visualization.

IIOAB JOURNAL ISSN: 0976-3104

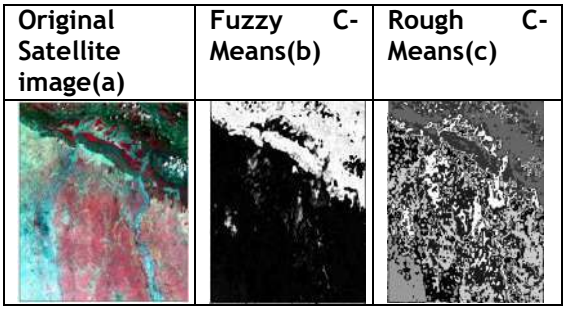


Fig: 7.a.b.c Original Satellite image with Fuzzy C-Means and Rough C-Means Methods

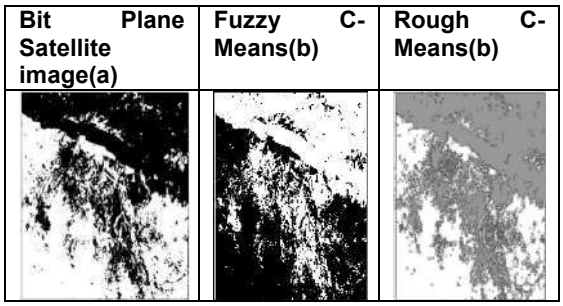


Fig: 8.a.b.c. Bit Plane Satellite image with Fuzzy C-Means and Rough C-Means Methods

In the paper, we experimented with the above edge detection techniques which may miss true edges. With the application of clustering techniques i.e., conventional, fuzzy c-means and rough c-means methods, better segmented satellite image of rough c-means. Hence rough c-means method can be used for depth computation which carries minimum time consumption of a satellite image.

COMPARATIVE STUDY AND RESULTS

In this section we provide a comparative study of the experimental results obtained by applying the conventional methods as well as the fuzzy c-means and rough c-means with bit plane technique applied to a satellite image.

For comparison, sliced satellite image with Otsu thresholding has been used for conventional edge detection methods. And a comparative analysis of conventional, fuzzy c –means and rough c-means has been used with their root mean square error(RMSE) and peak signal noise ratio(PSNR) values for without bit plane /original satellite image (1347,38.7689,1352,38.7318) and with bit plane (477, 37.2323,473,37.1501) as given in table 5 and **Figure– 9**. And a proposed Bit plane with conventional, fuzzy c means, rough c-means clustering methods (three) with Otsu thresholding is applied to a satellite image for better clustering as given in table 6. Performance of rough c-means yields better results compared to fuzzy c-means better than conventional methods [20].

As per **Figure– 10**, comparison has been made by taking Otsu thresholding as the epsilon value resulted in an improved clustering of the rough c-means Otsu thresholding. And on comparison, rough c-means (RCM) yields good segmentation compared to that of old detection methods and fuzzy c-means (FCM) [20].

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 5. PSNR and RMSE of three clustering methods for original and bit plane satellite images

Images with various clustering techniques	Original Satellite Image		Bit Plane Satellite Image	
	RMSE-Original	PSNR-Original	RMSE-Bit Plane	PSNR-Bit Plane
Input Satellite image	1347	38.7689	1352	38.7318
Canny	2180	33.9544	2163	34.0327
Robert	2060	34.5206	2280	33.5059
Sobel	1411	38.3047	2340	35.0733
Zero-cross	2357	33.1738	1482	37.8137
Fuzzy c-means	2489	32.644	520	48.286
Rough c-means	1311	37.3047	1238	39.612

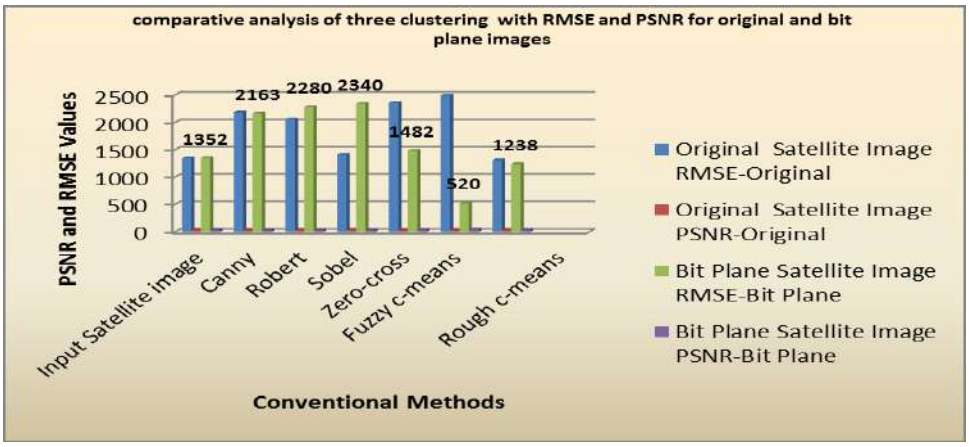


Fig: 9. Comparative Analysis of three methods with RMSE and PSNR values for original and bit plane satellite images

Table: 6. Original, fuzzy c-means and rough c-means satellite images with epsilon values of without and with bit planes

Old/fuzzy-otsu/rough otsu	Withoutbitplane	With bitplane
Original	130	122
Canny	99	102
Robert	130	102
Sobel	124	71
Zero-cross	88	123
Fuzzy c-means	126	120
Rough c-means	109	102

THE IIOAB JOURNAL
ISSN: 0976-3104

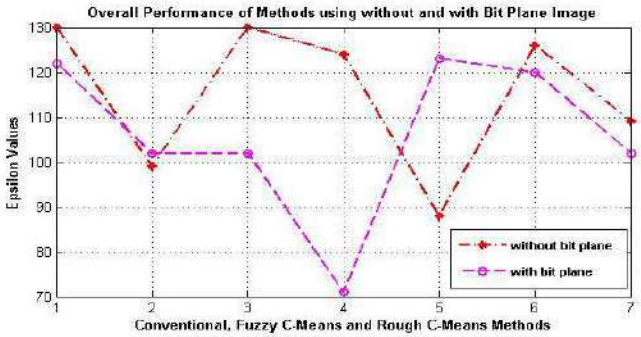


Fig: 10. Overall Performance of Original, fuzzy c-means and rough c-means satellite images with epsilon values of without and with bit planes

The **Table –6** and the corresponding graph in **Figure– 10** deals with overall performance of Otsu thresholding with bit plane and without bit plane, performance for conventional methods using with bit plane and without using bit plane, performance of fuzzy methods using with bit plane and without bit plane and performance of rough c-means using with and without bit plane resulting PSNR of all performances for interpretation of control points.

DEPTH COMPUTATION

We computed results of time complexity in terms of seconds using conventional, fuzzy and rough c-means with bit plane of depth computation for further reconstruction of an image. The depth computation of various clustering techniques using bit plane and without using bit plane is shown in **Table– 7** and **Figure– 11**. First, we divided the satellite images into planes using bit plane and reducing the noise by applying on satellite images. The different mathematical statistical methods like mean, standard deviation, variance and their PSNR values were applied on the bit-plane to calculate the efficiency of the Bit Plane Method. The output of conventional and fuzzy c-means techniques gave clustering results of a satellite image. And still, rough c-means yields improved clustering of a satellite image compared to other methods. The depth image of a satellite image is shown in **Figure –11**.

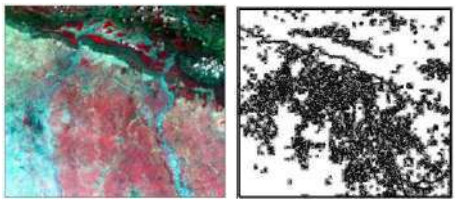


Fig. 11.a.b.Depth of a Original Satellite Image

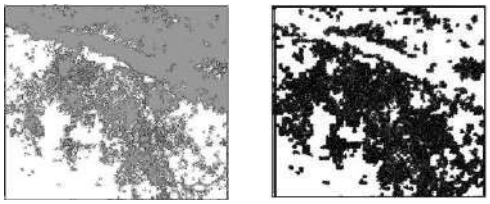


Fig. 12.a.b.Depth of a Rough C-Means Satellite Image

Table: 7. Time complexity of Depth Computation of three methods using without and with bit planes

Old/fuzzy-otsu/rough otsu	Withoutbitplane-Time in Secs	With bitplane-Time in Secs
Input Satellite image	0.35	0.17
Canny	0.9	0.76
Robert	4.87	0.87
Sobel	10.63	5.99
Zero-cross	0.28	0.2
Fuzzy c-means	0.26	0.38
Rough c-means	0.24	0.20

ISSN: 0976-3104

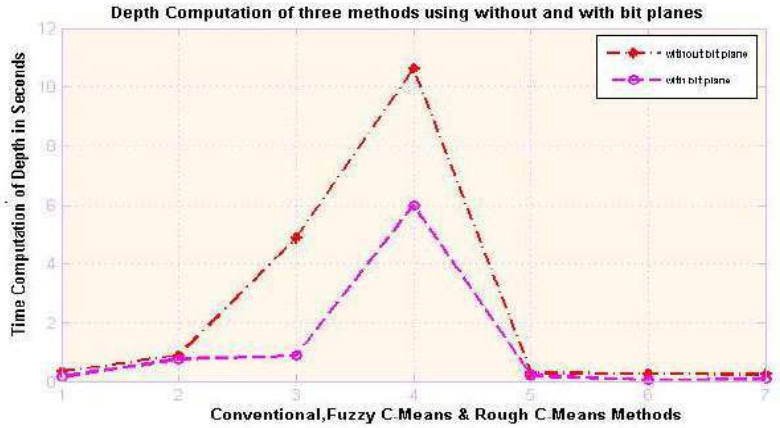


Fig: 13. Time complexity of Depth Computation of three methods using without and with 4.bit planes

Figure-10. discuss about the graph that deals with an overall performance of a satellite image using with bit planes and without bit planes for conventional, fuzzy and rough c-means methods. **Figure 11(a), (b).** deals with depth of a satellite image, **figure 12.a.b**.deals with depth of a fuzzy c-means satellite image and figure 13.a.b. deals with depth of a rough c-means satellite image. **Figure 13** deals with the proposed methodology (BPRCM-Bit Plane Rough C-Means), has proved to yield better performance with minimum time in computing the depth. The time complexity of original satellite image is 48.57%, conventional satellite image takes 56.34%fuzzy c-means takes 53.23% and rough c-means is 25.51%. And as a result, the performance of rough c-means with bit plane takes minimum time complexity compared to fuzzy c-means and conventional methods.

Thereby, a better cluster of satellite images has been processed after preprocessing using bit plane filter method. The enhanced satellite image has been used with proper edge detection techniques using conventional edge methods, fuzzy c-means method and rough c-means method. The distortions or errors or noise has been removed for further process of image reconstruction. And time complexity for the computation of depth has been carried out. The computation is needed for interpretation of degree of seriousness of any distorted portion or identification of any region for cultivation purpose or any weather forecast essential for the society. In this paper, satellite images for Himalayas have been considered which deals with geographical analysis. The proposed method may help the societies who are willing to proceed to Himalayas for the completion of task which involves efficient clustering.

CONCLUSION AND FUTURE WORK

In this paper, we proposed two new approaches; the FCMBP and the RCMBP, which segment the satellite image into planes. We justified the bit plane image using PSNR values to conclude that the satellite images are distortion free. These satellite images have been put under edge detection process involving better clustering algorithms. In all the phases, histogram analysis has been carried out to provide easy interpretation of satellite images. In conclusion, we can say that using rough c-means with bit plane technique results in better clustered satellite images compared to other techniques (conventional and fuzzy c-means) with increased efficiency and reduction in the epsilon (Otsu) and PSNR values. The depth computation for all possibilities using with and without bit plane techniques have been carried out. But compared to all three methods using without bit plane and with bit plane, the depth computation time is minimum for rough c-means method with bit plane than conventional methods and fuzzy c-means method. Thus clustering using rough c-means, using bit plane filter method at preprocessing technique yields better clustered image with minimum epsilon value and depth computation with minimum time complexity. In future, the authors aim is to apply other clustering techniques for further processing of satellite images.

THE IIOAB JOURNAL ISSN: 0976-3104

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers and the editor-in-chief of the journal for their valuable guidance which has improved the quality and presentation of the paper.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Kaufman YJ.[1989] The atmospheric effect on remote sensing and its correction *Chapter 9 in Optical Remote Sensing*, *technology and application*, *G. Asrar, Ed., Wiley.*

[2] Ersin G.[2011] Adaptive Wiener-turbo systems with JPEG & bit plane compressions in image transmission.*Turk J Electrical Engineering & Comp Science*, 19:141-155.

[3] Bezdek JC.[1981] Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, *New York.*

[4] Lingras P. and C West [2002] Interval set clustering of web users with rough k – means *Dept. Math.Comput. Sci., St. Mary's Univ., Halifax, NS, Canada, Tech.* Rep. No. 2002–002.

[5] Raman Maini and Aggarwal Himanshu.[2002] Study and Comparison of Various Image Edge Detection Techniques *Int..J Img. Proc.,* 3(1):1–12.

[6] Rulinda CM, Dilo A, Bijker W, A Stein [2012] Characterising and quantifying vegetative drought in East Africa using fuzzy modeling and NDVI data. *Journal of Arid Environments*, 78:69-178

[7] Guoyuan Li, Wang H, Li Canhai.[2012] Ortho-Rectification Of HJ-1A/1B ,Multi-Spectral Image Based On the GCP Image Database *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 39(22):25.

[8] Pandian A.P and S.N.B Sivanandam.[2012] Hybrid algorithm for lossless image compression using simple selective scan order with bit plane slicing *Journal of Computer Science*, 8(8): 1338–1345.

[9] Luo Y, Xue Y, S Zhong. [2005] Road extraction from IKONOS image using Grid computing platform *IEEE International Geosciences and Remote Sensing Symposium*, 3895-3898

[10] Sinha SK. and Fieguth W Paul.[2006] Automated detection of cracks in buried concrete images *Automation in Construction*, 15: 58-72.

[11] Mitra S., Banka H. and W. Pedrycz [2006] Rough–Fuzzy Collaborative Clustering *IEEE Transactions on Systems, Man, And Cybernetics-Part B: Cybernetics,* 36:795-805.

[12] Raman M. and Himanshu A.[2002] Study and Comparison of Various Image Edge Detection Technique *International Journal of Image Processing (IJIP*), 3 (1).

[13] Swarnalatha P and BK. Tripathy [2012] A Centroid Model for the Depth Assessment of Images using Rough Fuzzy Set Techniques *International Journal of Intelligent Systems and Applications*, 1:20-26.

[14] Swarnalatha P., Kota M., Resu N.R. G. Srivasanth [2009] Automated Assessment Tool for the Depth of Pipe Deterioration. Advance Computing Conference, 2009. IACC 2009 *IEEE International,* ISBN.978-1-4244-2928-8:721,724.

[15] Aguiar Pesso. AS, Stephany. S, L.M Garcia Fonseca.. [2011] Feature selection and image classification using rough sets theory. *International Geoscience and Remote Sensing Symposium (IGARSS),* 2904-2907.

[16] Ashok Kumar T, Priya S, V. Paul [2012] Automatic Feature Detection in Human Retinal Imagery Using Bitplane Slicing and Mathematical Morphology", *European Journal of Scientific Research*, 80:57-67.

[17] Fr´ias-Vel´azquez A. and W Philips.[2010] Bit-Plane Stack Filter algorithm for Focal Plane Processor *IEEE Proceedings of 2010, IEEE 17th International Conference on Image Processing*, 1-5.

[18] Pawlak Z. [1981] Rough Sets -Theoretical Aspects of Reasoning About Data *Kluwer Academic publishers, The Netherlands*.

[19] Zadeh L. A. [1965] Fuzzy Sets. Information and Control, 8:338-353.

[20] Swarnalatha P and B.K. Tripathy [2013] A Novel Fuzzy C-Means Approach with Bit Plane Algorithm for Classification of Medical Images *IEEE – International Conference on Emerging Trends in Computing, Communication and Nanotechnology(ICE-CCN),* 978-1-4673-5036-5:360-365.

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Rajamani et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

AN INVESTIGATION ON THE TECHNIQUES USED FOR ENCRYPTION AND AUTHENTICATION FOR DATA SECURITY IN CLOUD COMPUTING

Tamilarasi Rajamani*, PrabuSevugan, Swarnalatha Purushotham
VIT University, Vellore-Tamilnadu, INDIA

ABSTRACT

The paper deals with data security of cloud and their authentication techniques. Now-a-days, the cloud data security method uses the symmetric encryption and asymmetric encryption algorithms with their strong authentication techniques. The use of relevant algorithm deals with the level of data safety in cloud because data security in cloud computing is a serious issue as the data centers are located worldwide. Authentication is the most essential procedure to ensure the cloud data in a secured manner. However, strong user authentication is the main requirement for cloud computing that reduces the unauthorized user access of data on cloud. Data security is a more important issue of cloud computing. The survey is completely based upon the estimation for the cloud data security and authentication resolution. Almost, the inventors use the symmetric and asymmetric encryption algorithms with other authentication methods. Symmetric algorithms are AES, DES, Blowfish, RC2, 3DES and asymmetric algorithm are RSA, DSA, Diffie-Hellman and ELGamal. The Authentication techniques are one time password, Digital signature, and Biometric method. So a hybrid technique which is a combination of these encryption techniques and authentication method gives a more excellent and strong security on cloud data.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26th – March-2016
Published on: 18th–May-2016

KEY WORDS

Authentication; Cloud security; symmetric encryption; asymmetric encryption.

***Corresponding author: Email:** tamilarasi.r2014@vit.ac.in; **Tel:** +40-9001010010; **Fax:** +40-9001010012

INTRODUCTION

Cloud computing supports distributed service oriented architecture, multi-users and multi-domain administrative infrastructure, it is more prone to security threats and vulnerabilities. At present, a major concern in cloud adoption is its security and Privacy. Cloud computing nowadays is the precondition and essential part of the computing globe using whole day developing in its usages and popularity. Huge estimate of users is currently depending on cloud computing application for their everyday work of authority and produce services over the computer internet. Cloud represent as data centre. A client makes use of cloud resources, applications, storage and different services and is charged accordingly [1].

CLOUD SERVICE MODEL

Cloud computing providers' offers their service according to several fundamental models.

- Software as a Service(SaaS)
- Platform as a Service (PaaS) and
- Infrastructure as a Service (IaaS).

Software as a service

- In SaaS model, cloud suppliers introduce and work application, programming in the cloud and cloud clients get to the product from cloud customers.Cloud users do not manage the cloud infrastructure and platform where the application runs. This eliminate the need to install and run the application on the cloud users own computers which simplifies maintenance and support.

THE IIOAB JOURNAL
ISSN: 0976-3104

Platform as a service

- In platform as a service, the cloud providers deliver a computing platform including operating system, programming language execution environment database and web server. The resources scale automatically to match application demand so that the cloud user does not have to allocate resources manually.

Infrastructure as a service

- In cloud service model, providers of IaaS offers computer – physical/virtual machines, other resources include hypervisor, Virtual Machine (VM) disk image library, raw/file based storage, firewalls, load balancers, virtual LAN and internet Protocol (IP) address.IaaS cloud providers supply these resources on demand from their data centers. To deploy their applications, cloud users install OS and application, software on the cloud infrastructure.Cloud providers bill IaaS service on a utility computing basis (i.e.) cost which reflects the amount of resources allocated and consumed.This part is basically belongs to admin part or about service provider [2].

CLOUD DEPLOYMENT MODELS

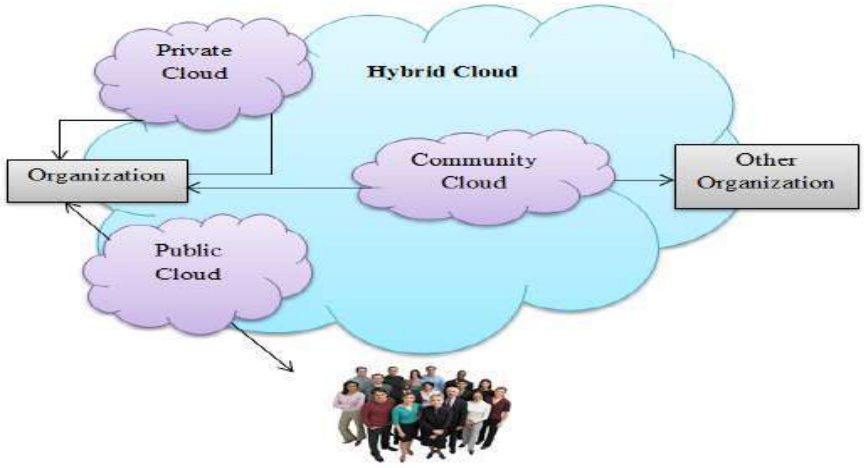


Fig.1: Cloud Deployment Model

Private cloud

- Private cloud computing model it's operating solely for a single organization within the boundary of the organization.
- It maintains internally or externally to support business operation.
- Cloud providers share the accessible resources and applications, so that in private cloud users can flexibly share and use it. They work similar to an intranet within an organization. Unauthorized persons can't access the data and share the resources. Using this security the private cloud is more secured when compared to the public cloud [3].
- To create a private network restructures the existing infrastructure by adding virtualization and cloud line interface.

Public cloud

- Public cloud is a cloud infrastructure that is made available to the general public as pay per use concept-able model. The resources are hosted on the service provider's premises.
- Resources are dynamically provisioned through publically accessible web application or web services (SOAP) from a third party premises.
- It is cost effective because all the computing resources are shared worldwide.
- It is fully customer self services; the customer can access the public cloud through internet.

THE IIOAB JOURNAL ISSN: 0976-3104

Hybrid cloud

- This cloud infrastructure is a combination of two or more different cloud infrastructures (community, private, or public) bound together by some standardized or using some technology and allow migradition data and application between them [4].

Community cloud

- The cloud infrastructure is combined by several organisms [5]. (or) In community cloud is a multi-tenant cloud service model that is shared among several or organizations and that is governed, managed and secured commonly by all the participating organizations or a third party managed service provider [6].

CLOUD COMPUTING SECURITY

It refers to a broad set of policies, technologies and controls deployed to protect data, applications and associated infrastructures of cloud computing. Cloud computing can be implemented with different service models (Saas, Paas, Iaas) and deployment models (private, public, hybrid, community) so we need to provide security for cloud data. Security issues for cloud computing is classified as two levels they are.

Security issues faced by cloud providers (organization providing software platform (Infrastructure -as-a-service)) then cloud.

The cloud provider assures that the infrastructure is protected and secure. The cloud provider must protect the data and application of the cloud users.

Security issues faced by their customers

The provider must ensure that their infrastructure is secure and their clients, data and application are protected while the customer must ensure that the provider has takes the proper security measures to protect their infrastructure. The main security issue is virtualization. Virtualization must be properly configured managed and secure.

CLOUD SECURITY CONTROLS

Cloud security architecture is effective only if the right protective execution is set up. Efficient cloud security design ought to perceive the issues that will emerge with security administration. Security administration addresses these issues with security controls. These controls are placed set up to shield/secure any short coming in the framework and reduce the effects of attack.

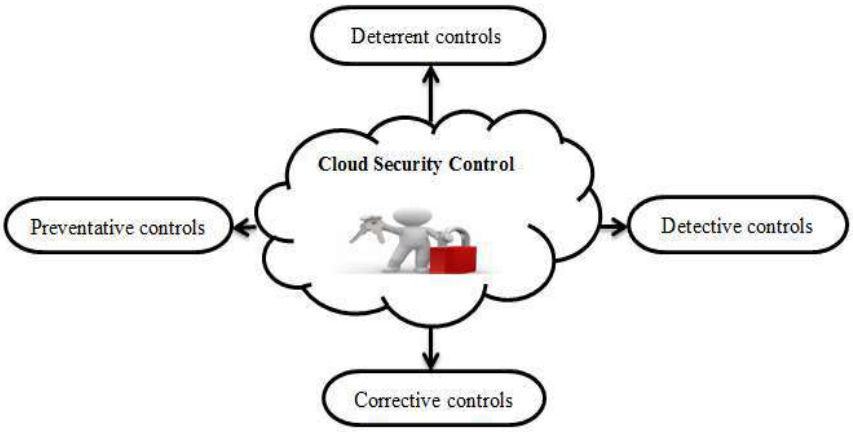


Fig: 2. Cloud Security Control

THE IIOAB JOURNAL ISSN: 0976-3104

Deterrent controls

- These controls are set in place to prevent any purposeful attack on a cloud system. It is like a warning sign on a fence or a property. They do not reduce the actual vulnerability of a system.

Preventative controls

- These controls upgrade the strength of the system by managing the vulnerabilities. It safeguards vulnerabilities of the system. It attacks the users from the preventative controls in place to cover the attacks and reduce the damage and violation to the system security.

Corrective controls

- The corrective controls are used to reduce the effect of an attack. Unlike the preventative controls. The corrective controls takes action as an attack is occurred.

Detective controls

- Detective controls are used to detect any attacks that may be occurring to the system. In the event of an attack, the detective controls will signal the preventative or corrective controls to address the issues.

RELATED WORK

In proposed a data security in cloud computing. There are additionally existing algorithms. That use various encryption and decoding techniques to make safe client data on a cloud. This also makes user be able to load their data at cloud except for any worry and can retrieve their data as per use. Here many algorithms are provided to security for users data on cloud with secure the user information from malicious activity [7].

It presents the review which totally the estimation for the cloud assurance solution. The creators nearly advantage the encryption code strategy. The accepted encryption techniques were discussed the overview such as AES, DES, 3DES, RC2, RC6 and Blowfish. The cloud security approach has proposed the use of AES algorithms, without any gaps in AES [4].

The proposed a hybrid (combination) encryption method compresses of using asymmetric and symmetric cryptographic algorithms. The symmetric key distribution among cloud provider and legitimate users is executed utilizing asymmetric techniques. At the last moment, we compare hybrid (mixed) technique applying the connection of the AES and RSA algorithms. The present summary shows that the new method gives merits of symmetric algorithm by efficient processing time and the forcefulness of asymmetric algorithm in their key size. Actually the present lightweight algorithm is faster than other cryptographic methods in processing data [8].

In studied in user authentication to protect data of encryption techniques within cloud computing. Cloud computing permit client to access the internet browser without application installation and approach their information at any system utilizing internet browser. This framework approves the security of the data in cloud server. Here user is using AES algorithm for encryption and decryption structure to create cloud users data secured and also ensure the information privacy in the cloud. This paper presents the protected data mechanism to clarify the issues of data security and privacy in cloud computing [2].

Inproposed techniques that deal with secure of static data as well as dynamic data in cloud. AES is faster than RSA. RSA presentation depends on prime number and complication depends on that value of primes and also this paper provides linking viewpoint of security, protection, detection of threads [9].

In displayed symmetric key encryption strategies, for example, DES and AES encryption quality. The study tells that the symmetric key encryption strategies are utilized for the majority well-organized than asymmetric key encryption techniques. Data security in cloud can be improved by applying a grouping of separate methods at the same time [10].

Inproposed to store secure data in various cloud, Data encryption and separation (splitting) are the two techniques used. Then AES-256 also used to secure with data encryption purpose. And also data splitting is used which

THE IIOAB JOURNAL ISSN: 0976-3104

separates file into subparts to store into different clouds. These techniques benefit to secure files from unauthorized users in looking all files [11].

In proposed analytical study of different data security methods on cloud computing as it produce a huge number of security problems like RASP It issues secure and effective series request and ken query services for secured data, TRSE, CP-ABE, KP-TSABE, KP-ABE, RSA, AROCrpyt, AES, Blowfish, DES, RSA, ASIF. Different data security algorithms are referred in the following section [3].

In proposed an algorithm called RSA technique. It is a new method which assembles the essentially of public key systems. While using this method data security will increase with minimum execution time and cost. But RSA also have disadvantages like imitation public key techniques, complication of key production, security requirements and with low speed [12].

In proposed 512 bit AES encryption algorithm that is used to provide security on medical data present moment in cloud in other side.Only the authorized user can decrypt the data. It is very securable and highly powerful. It has only one drawback that is AES 512 is the requirement for huge design area [13].

Inpresented the issue of data security in cloud data storage is considered, which a distributed storage system is basically. Using cryptography method is used to transformthe secure data and storing between user and cloud storage services. A hybrid method of symmetric and asymmetric encryption methods like AES and RSA are used. These techniques mainly achieve the confidences of data security in cloud. RSA encryption will provide complexity for attackers for decreasing the time of data communication by using AES encryption techniques [14].

In proposed an Elliptic curve cryptography that provides the higher level ofsecurity and also presentation is very much improves than another encryption techniques. The paper proposed ECC algorithm which deals with one views smallest amount of time to encrypt the data. Here there are three security frameworksnamely Authentication, encryption and separation of customs for the security that has been fulfilled which benefits to succeed the greater level of safe and secure. The solution displays that ECC is more secured and has the well execution than other encryption techniques [5].

In proposed Architecture is based on block based symmetric cryptography Algorithms, Block based symmetric is efficient & secured. Using Symmetric encryption techniques but it has only one drawback the authors have not proved with real time implementation of the architecture [15].

In proposed Encryption and obfuscation techniques increase the confidentiality of data, obfuscation as two different systems to ensure the data in the cloud storage. Encryption is the procedure of changing over the discernable (readable) content into incomprehensible structure utilizing an algorithm and a key. Obfuscation is same like encryption. Obfuscation is a procedure which masks illegal clients by executing a specific scientific capacity or utilizing programming methods. Based on the kind of data, encryption and Obfuscation can be applied. Encryption can be applied to letters in order and alphanumeric sort of data and obscurity can be applied to a numeric sort of data. Applying encryption and Obfuscation methods on the cloud data will give more provide against unapproved use. Classification could be accomplished with a mix of encryption and obscurity [16].

In presented Rijndael EncryptionAlgorithm and EAP (Extensible Authentication Protocol) the proposed design and architecture that can help to encrypt and decrypt the document at the client side that gives security to data at rest as well as while moving. The authentication is utilized for the transport and utilization of keying material and parameters produced by EAP strategies [17].

In studied conceptual paper Comparison of Symmetric and Asymmetric algorithms, shows the superiority of symmetric algorithm and AES shows superior algorithm performance among the various algorithms [18].

In proposed group based authentication and key agreements protocols. This framework was executed in the cloud environment with a test bed of 20 systems. This usage results in appropriate authentication with access control for query signature based cloud services.This plan likewise acknowledges with important control for authentication for multi-group. Additionally, the user can get the cloud benefit once they are enrolled as an individual from a group, also ensured the exactness of the authentication and data sharing for the user in the dynamic group.Hence

THE IIOAB JOURNAL ISSN: 0976-3104

by dealing with this authentication time and movement in the dynamic group. At long last, the capacity and computational cost are stable [19].

In presented this paper induced brief analysis on data security in cloud environment, It is well-known that cloud computing has numerous potential focal points, there are still many actual issues that should be tackled and the data is moving to the public or hybrid cloud.As indicated by the analysis of data security, it is required to have an incorporated and extensive security solution for addressing the issues of safeguard top to bottom [20].

In studied symmetric and asymmetric algorithms, the paper draws the inference proxy re-encryption and hierarchical attribute-based encryption is most proposed approaches to ensure data security but nearly 33% is not validated. But this paper considered 15 research papers to draw conclusion [21].

In presented a hybrid encryption algorithm using RSA and AES algorithms for providing data security to the user in the Cloud. The greatest purpose of interest it gives us is that the keys are delivered on the reason of structure time in this way no intruder can even figure them there by giving us extended security close by convenience. Private Key and a Secret key are just known to the client and in this manner client's private data is not accessible to anybody not even the Cloud's Administrator. [22].

In proposed AES Algorithms three way mechanism techniques. Firstly Diffie-Hellman algorithm is utilized to create keys for key exchange step. Then digital signature is utilized for authentication, from that point, AES encryption algorithm is utilized to encrypt or decrypt client's data file. This is executed to give trusted processing environment with a specific end goal to avoid data change (modification) at the server end. The AES (Advanced Encryption Standard) encryption algorithm to protect confidentiality of data stored in cloud [23].

In proposed Elliptic curve Diffie-Hellman (ECDH) and vicariate polynomial secret sharing. The main goal of these schemes is to ensure the data protection and security in the cloud serverandinclude the symmetric property in secret sharing to effectively decrease the expense to share the shares the customer and the server [24].

In proposed RSA algorithm,it only authorized person can use the data. Since Cloud Computing stores the data and dispersed assets in the open environment, security has become the fundamental obstruction which is hampering the organization of Cloud environments. Even though the Cloud Computing is promising and productive, there are many difficulties for data security as there is no region of the data for the Cloud client. [25].

In presented Symmetric algorithms (AES, Triple DES and DES). Transferred of data is encrypted in the upper layer on top of the transport layer instead of using IPS and SSL. In this manner, the plan for the execution change can be connected without altering the execution of IP layer and efficient secure communication by pre-handling of encryption in the upper-layer is realized [26].

In proposed Fragmentation techniques. To applies the fragmentation techniques in the cloud environment with minimal encryption to prevent data exposurethe fragmentation method which is efficiently stores the data on CSP servers utilizing the base possible measure of encryption. The fragmentation procedure is applied to a relational database where the tables are treated as independent fragments. This fragmentation and distribution the methodology reduces the trust expectancies towards the outer administration suppliers and subsequently enhances security and privacy [27].

In studied symmetric encryption algorithm, AES is found suitable and the most secured algorithm for Amazon EC2, the model utilized three-layer framework structure, in which every floor performs its own obligation to guarantee that the information security of cloud layers. The main layer: in charge of user authentication, nearly this is two component confirmations; however, free cloud suppliers utilize one element as samples eyeos, cloudo, and freezoh. The second layer: responsible for user's data encryption, and ensure the protection of user through a specific path by utilizing one symmetric encryption algorithms. Likewise, permit security from a user.The third layer: The user data for fast recovery this depends on the speed of decryption [28].
In studies in issues relevant of cloud data storage techniques and security in virtual environment. The paper proposed RSA public key cryptosystem that provides data storage and security in cloud computing. RSA algorithm provides more security in high possibility data encryption organization [29].

THE IIOAB JOURNAL
ISSN: 0976-3104

In this paper encryption algorithms have been proposed to make cloud data secure, defenseless (vulnerable) and offered worry to security issues, challenges furthermore algorithms have been made between AES, DES, and RSA calculations to locate the best one security algorithm, which must be utilized as a part of cloud computing for making cloud data secure and not to be hacked by attackers [30].

In proposed another patient-centric Personal Health Record framework, which would be packaged with a two stage validation instrument. The framework likewise utilizes a standout amongst the best encryption methods of AES encryption for securing the Personal Health Record (PHR) which are put away in the Third party semi-trusted cloud server and also proposed OTP authentication for using secure data in a cloud [31].

In presented the requirement for data integrity through Third Party Storing so as to inspect and the data heterogeneously over multiple servers or databases and thus provides gives additional security to cloud clients. Furthermore, utilize One Time Password for client authentication like the banking system. This distributed data split vertically when consolidated with encryption gives an additional layer of security and ensures clients information access and capacity storage in a simple and cost effective manner as well [32].

In proposed data security model gives client authentication and data assurance. This makes certain safe correspondence framework and concealing data from others. In this model message digest based document encryption framework and secure open key encryption framework utilizing RSA for trading data is incorporated. This model additionally incorporates one-time password (OTP) framework for client validation process [33].

EXISTING ALGORITHMS FOR SECURITY

To provide data security in cloud computing there are all the more existing methods. Which utilizes encryption and decoding strategies for well-being and security with client information on the cloud? Here we are utilizing symmetric and asymmetric encryption algorithms. The symmetric encryption method is having only one key that is used to encrypt and decrypt the data. Another method is asymmetric encryption that is having two key one is private key and another one public key. Private Key is used for decryption and public key is used for encryption. [34].

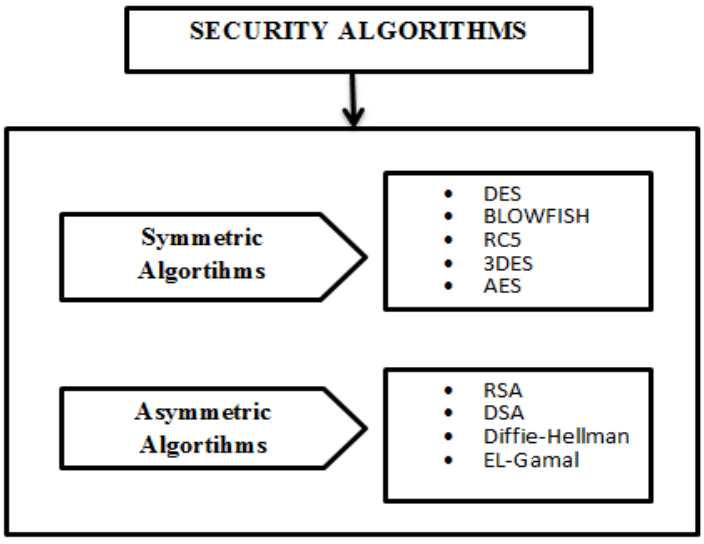


Fig: 3. Security Algorithms

THE IIOAB JOURNAL ISSN: 0976-3104

SYMMETRIC TECHNIQUES

DES

DES stands for Data Encryption Standard established in 1977. It applies a 56-bit key to each 64-bit block of data. It was the first encryption standard to be approved by NIST. This Method can run in number of modes and requires 16 rounds or controls, even though this is designed with "strong" encryption. We have used DES algorithm with destruction-editing approach for providing data security with integrity [35]. Each round in the deals with uses a separate 48-bit round key which is produced from the consistent cipher key according to the DES techniques [10].The Data Encryption Standard (DES) is a formerly transcendent symmetric-key algorithm for the encryption of electronic data. It was highly influential in the advancement of present day cryptographic systems.DES is the block cipher an algorithm that takes a fixed length string of plaintext bits and changes into a series of muddled operations into another cipher-text series of bits with the same length. On account of DES, for the most part, the block size is 64 bits. DES additionally utilizes a key to altering the change, so that decryption must be performed by the individuals who know the specific key used to encrypt. At the present DES issued to be unconfident for multiple applications, and therefore it has been replaced by the Advanced Encryption Standard (AES)[36].

BLOWFISH

It is symmetric encryption algorithm. It have 64 bit block cipher developed by Bruce Schneider; enhanced for 32-bit mainframes with huge data stores, it is greatly faster than DES on a Power PC-class machine. Key lengths can differ from 32 to 448 bits in range. Also it's have 16 rounds. Blowfish, accessible easily and developed as an alternate for DES or IDEA which is in use in a large number of production [7].

RC5

It is symmetric encryption algorithm that deals with 128 bit block cipher based upon, and a development done, RC5. Also its have 12 rounds. The utilization of RC5 algorithm for encryption, cloud computing can be connected to the data transmission security.Transmission of data will be encrypted, regardless of the fact that the data is stolen, there is no relating key can't be restored[37].

3DES

In Triple DES (3DES) Triple Data Encryption Algorithm (TDEA or Triple DEA) symmetric-key block cipher encryption is discussed with the development ofthe Data Encryption Standard (DES) cipher techniques. TDES uses a block size of 64 bits and operate 48 processing round corresponding to DES. In 3DES three times iteration is produced to improve the encryption and security level [38]. It makes three encryption and decryption permits done the block using DES 56 bit keys [7].

AES

Advanced Encryption Standard (AES) uses a symmetric key encryption design known as called Rijnadael, a block cipher proposed by Belgian cryptographers Joan Daemen and Vincent Rijmen [7]. The key size can have variable lengths such as 128,192 or 256 bits. The default keysize is 256 bits. AES encryption standards are very fast, it is flexible and effective than DES. It has had total 14 rounds contingent on the key sizes which are used in [10]. It is one of the very regularly used and isavailable for utilizing the data secure purposes; the algorithm depends on a few substitutions, permutations and direct changes. It said that up until today, no functional assault against AES exists. In this manner, governments, banks and high-security frameworks around the globe are favored utilizing AES for the encryption standard. [14]. It is highly securable and more efficient algorithm and it secure all types of data that deals with medical information's but only thing need for more design area. Here AES algorithm is used that consists of 22 rounds it may minimize to 18 rounds which reduce time consuming and cost [13]. The use of AES encryption algorithm is highly securable with no loopholes and AES encryption and decryption is highly

THE IIOAB JOURNAL ISSN: 0976-3104

secured and fastest method.AES is the main algorithm which is not inclined to any of the cryptanalysis assaults (attacks).[4].

ASYMMETRIC TECHNIQUES

RSA

Ronald Rivest Adi Shamir and Leonard Adlemandesigned the RSA algorithm 1977 cryptosystem uses the properties of the generative homomorphism encryption. RSA key size is having1024 bit. Then its have one rounds [10]. RSA is generally use public key techniques and RSA is accomplished to maintain encryption and digital signatures. RSA provides the best security plan by encrypting the data that is confidential; this is the motivation behind why the enormous administration suppliers like Google mail, Yahoo mail and so on are utilizing this algorithm to give their clients the protection of secrecy in utilizing their administrations.[14].RSA today is utilized as a part of a few programming items and it can be utilized for digital signatures, key exchange, or encryption of a little block of data. RSA uses a changeable size key and a variable size encryption block but RSA encrypts and decrypts data that consumes more time [7].

DSA

In presented an autonomous investigation of security algorithms in cloud computing. Which provides the particular technique to secure data on cloud computing.The DSA techniquegives digital signature possibilities for the authentication of messages [7] DSA (Digital Signature Algorithm) is a Federal data processing Standard for digital signatures. DSA was introduced by the NIST (National Institute of Standards and Technology) it is used to detect the unauthorized alterations to the data send by the source to the receiver [10].

Diffie-Hellman

Diffie-Hellman introduces secret-key exchange protocol only. It is not for authentication or digital signatures and it is public key exchange methods, it uses of the discrete logarithm problem. Actually the sender and receiver set the secret key [7]. These techniques protect the data confidentiality and safe and security, Diffie Hellman Key Exchange method tolink organization and Elliptic curve cryptography for data encryption [39].

El-Gamal

El-gamalalgorithm is also public key cryptographic techniques. The private key will be secret. It is not capable to expose the information. So encryption and decryption of message will gives more security for the data, ELGamal's cryptosystems have numerous helpful applications, with its strong properties. It is an exceptional sec. This is most certainly not restrictively difficult to encrypt the message in the cloud also [40].The bit operation of encryption or decryption in El-Gamal cryptosystem is polynomial used in the paper. The ElGamal algorithm is utilized as a part of this Paper for homomorphic encryption. The unique data is then acquired by the user with the cipher Keys. This must be reached out for a various number of clouds and with various operations. [41].

AUTHENTICATION TECHNIQUES

Authentication is an important part of an environment. The cloud is no exclusion. Actually, in some aspects, authentication is still more important in a public cloud environment than in a traditional environment. Authentication use primary techniques of exclusive of access of the applications and data. This paper discusses the three levels of authentication techniques. The paper discusses given below.

ONE TIME PASSWORD

In this paper we have proposed to create of factor one time password with two factor authentication as a powerful authentication method that is necessary mobile phone as an authentication device. In this technique mobile phones are in control to produce OTP which is valid only for 3 minutes [42]. This is oldest techniques but it also provides

THE IIOAB JOURNAL ISSN: 0976-3104

secure authentication. In order to secure the system, the produced OTP must be strong to find, recover, or trace by hackers. Therefore, it's very important to develop a secure OTP generating algorithm. Users appear to be willing to utilize straight forward variables, such as their mobile number and a PIN for administrations, such as approving mobile micropayments [43]. One time password can be produced in any of the two ways, HMAC based One Time Passwords (HOTP) and Time based One Time Passwords (TOTP) The user creates a one-time password and submits it to the server. Server additionally creates a one-time password for that inhabitant for that instance of time and confirms it with the password received from a user[44].

DIGITAL SIGNATURE

The digital signature gives a dynamic solution to implement services that assure data protection and data integrity. RSA Digital Signature Scheme guarantees legitimacy and respectability of data [45]**.** The digital signature authentication method is better than all techniques. Digital signature is a method supports authenticity and integrity of information. Digital signature is public key cryptography. Digital signature is used with any kind of data whether data is encrypted or not [46]**.** Digitally signed messages may be everything that represents strong bit string: examples include electronic mail, contracts,or a message sent by any other cryptographic protocol. Digital signature is a mathematical structure for establishing the authenticity of digital information or document [47]. The reasonable digital signature provides a receiver cause to believe that the message was produced by an identified user, which was not chanced in transit. It is generally used for software distribution, financial transactions, and in other instance to detect the forgery which was important to use [48].In proposed another security design which executes RSA for both encryption and secure correspondence purposes though MD5 hashing is utilized for digital signature and hiding key data. This model gives security to the whole cloud computing environment.Both RSA encryption and Digital Signatures algorithms subsequently a capable security and information respectability administration framework is acquired [52].

BIOMETRIC AUTHENTICATION

This proposes a new approach based on biometric encryption for to increase the security of data sharing in public cloud. The biometric based authentication to make sure that the user is unauthorized person. For the authentication purpose we usethe physiological assessment as the encrypted image, here it is analysis graph of heartbeat [49].In this proposal thumbnail expression of user for Authentication is used. When register with the new user, it take the thumbnail expression of user using thumb recognition device and stored in image format in System Database. Whenever the user logs in, user should give the thumbnail expression using thumb detection device then system checks that image is same or not. If wrong then provide the error and if it is accurate then gives approval for other authentication scheme [50]. Generally traits used for biometric recognition are: faces, fingerprints, irises, palm-prints, speech etc. Designing biometric services in cloud highlight with result that has to be built with value to that mechanism of the biometric system that should be transferred to the cloud. Biometric is the authentication process which is used for the security purpose. In proposed system we are using biometrics like figure print and iris images are used. It also uses the minute matching algorithm to compare the images [51].

Here correlation of symmetric block cipher and unbalanced algorithm talked about, DES (Data encryption standard) algorithms have a key size of 56 bit key, it's called as private key type and block size 64 bit furthermore it's have 16 rounds. Blowfish is likewise symmetric algorithm its have 32 to 448-bit level, it has 16 round to handling and execution time is quick. At that point RC5 is a symmetric algorithm it has 0 to 2040 key sizes and square size 64 bit, it's similar to a private key type yet execution speed is low. 3DES key size is 32, 64 or 256 bit and block size 64, its have 48 rounds. AES (Advanced encryption standard algorithm) it is quick and secure technique and it have 128,192 or 256 pieces then the number rounds are 10, 12, and 14. It has quick execution speed too. In last RSA is have a 1024-bit level and square size is variation number of rounds is 1 and it's called as public key algorithm.

THE IIOAB JOURNAL ISSN: 0976-3104

COMPARISON OF SYMMETRIC AND ASYMMETRIC ALGORITHMS [6][8][9][13]

Table: 1. Comparison of symmetric and asymmetric techniques

Algorithms	DES	Blowfish	RC5	3DES	AES	RSA
Key Size	56 (+8 parity bits)	32-448(default 128)	Max 2040	112,168	128,192 or 256	1024 to 4096
Block Size	64	64	32,64 or 256	64	128,192 or 256	Variant.
Number of Rounds	16	16	1-255(12 suggested)	48	10(128),12(192),14 (256)	1
Cipher type	Symmetric block cipher	Symmetric block cipher	Symmetric block cipher	Symmetric block cipher	Symmetric block cipher	Asymmetric block cipher
Key type	Private Key	Private Key	Private Key	Private Key	Private Key	Public Key
Speed	Very slow	Fast	Slow	Slow	Very fast	Slow

CONCLUSION AND FUTURE WORK

The paper concludes with an independent study of security algorithms and authentication methods in cloud computing such as symmetric, asymmetric and authentication techniques. The symmetric and asymmetric techniques such as DES, Blowfish, RC5, 3DES, AES,RSA, DES, Diffie-Hellman and El-Gamal and the authentication methods includes Onetime password, Digital Signature and Bio-metric. The study says that on comparison with many secured algorithms available till date, blowfish is faster than other encryption algorithms. Also the survey shows that the authentication method, an oldest method, restricts the user to access easily for incorrect password can be handled by digital signature and biometric methods.

The future work of the paper closes with a usage of information security and verification in cloud computing with respect to the AES and Blowfish algorithms. Both of them can be best compared.We are going to have correlation examination of AES and Blowfish strategy and then we will analyze and consider which one gives more security and we can enhance the blowfish systems.

FINANCIAL DISCLOSURE

No financial support was received to carry out this research.

ACKNOWLEDGEMENT

Authors would like to thank School of Computing Science and Engineering, VIT University for providing resources and support to carry out this research work.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

REFERENCES

[1] RamandeepKaurBhinder el al. [2015] A Review on Using Cryptography Techniques for Securing User Data in Cloud Computing Environment. *International Journal of Computer Science & Communication (*IJCSC),.6:83–86.

[2] NiteenSurv et al. Framework for Client Side AES Encryption Techniques in Cloud Computing. *International Advance Computing Conference (IACC)*, 525– 528.

[3] Periyanatchi S, Chitra.K. [2015] Analysis on Data Security in Cloud Computing-A Survey. *International Conference on Computing and Intelligence Systems* 04:1281 – 1284.

[4] LovepreetKaur et al. [2015] A Survey on the Encryption Algorithms in the Cloud Security Applications. International journal of Science Technology & Management (IJSTM), pp.1–9.

[5] Neha A Puri et al. [2014] Deployment of Application on Cloud and Enhanced Data Security in Cloud Computing using ECC Algorithm. pp. 1667– 1671.

[6] Thiyagarajan B, Kamalakannan R. [2014] Data Integrity and Security in Cloud Environment Using AES Algorithm. Information communication and Embedded Systems. 1– 5.

[7] CharanjeetKaur et al. [2015] Data Security Algorithms In Cloud Computing: A Review. *International Journal For Technological Research In Engineering* 2:372– 375.

[8] Sana Belguith et al. [2015] Enhancing Data Security in Cloud Computing Using a Lightweight Cryptographic Algorithm. *The Eleventh International Conference On Autonomic and Systems*. 98– 103.

[9] Tembhurne S et al. [2015] An Improvement In Cloud Data Security That Uses Data Mining. *International Journal of Advanced Research in Computer Engineering & Technology* 4: 2044– 2049.

[10] Nikhitha K, Navin K S. [2015] A Survey On Various Encryption Techniques For Enhancing Data Security In Cloud. *International Journal of Advanced Research Trends in Engineering and Technology* 194– 197.

[11] Rashmi S et al. [2015] Architecture for Data Security In Multi-cloud Using AES-256 Encryption Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication* 157-161.

[12] Masthanamma V et al. [2015] An Efficient Data Security in Cloud Computing Using the RSA Encryption Process Algorithm. *International Journal of Innovation Research in Science, Engineering and Technology* 4: 1441– 1445.

[13] SaiSindhuTheja R et al. [2015] Data Security in Cloud for Medical Sciences using AES 512-bit Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication* 1746– 1749.

[14] Nasrin K, ZurinaMohd. [2014] A Framework Based on RSA and AES Encryption Algorithms for Cloud Computing Services. *IEEE Conference on Systems, process and Control* pp. 58-62.

[15] Sugumaran M et al. [2014] An Architecture for Data Security in Cloud Computing. 2014 World Congress on Computing and Communication Technologies. pp. 252– 255.

[16] Arockiam L, Monikandan. [2014] Efficient Cloud Storage Confidentiality to Ensure Data Security. International Conference on Computing Communication and Information. 5:1– 5.

[17] Anuj Kumar et al. [2014] Cloud Data Security using Authentication and Encryption Technique. *International Journal of Innovative Research In Technology* 1: 388– 391.

[18] Vanya divan et al [2014] Cloud security solution: comparison among various cryptographic Algorithms. *International journal of advanced research in computer science and software Engineering* 4:1146– 1148.

[19] pradeep Kumar et al [2014] An authentication approach for data sharing in cloud environment for dynamic group. International conferences on issues and challenges in intelligent computing techniques(ICICT)9:262–267.

[20] Meenakshi et al. [2014] Data security analysis in cloud environment. *International journal of innovations & advancement in computer science* 2:14– 19.

[21] Aized Amin Soofi et al [2014] Encryption Techniques for cloud data confidentiality. International Journal of Grid Distribution computing. 7:11–20. .

[22] Vishwanth, S.Mahalle et al [2014] Enhanced the data security in cloud by implementing hybrid(RSA&AES)encryption algorithms. *International conference on automation and communication*. pp.146–149.

[23] Prashantrewagad et al. [2013] Use of digital signature with Diffie Hellman key Exchange and AES encryption algorithm to enhanced data security in cloud computing. *International conference on communication systems and network technologies*. 3:437-439.

[24] Ching-Nung yang, jia-bin lai. [2013] Protecting data privacy and security for cloud computing based on secret sharing. *International symposium on biometrics and security technologies* 7:259–266.

[25] ParsiKaplana ,sudha. [2012] Data security in cloud computing using RSA algorithm. *International journal of research in computer and communication technology,* vol.1.

[26] Rohit ,sunil [2012] A proposed secure framework for safe data transmission in private cloud. *International journal of recent technology and engineering*, vol.1

[27] Aleksandar et al. [2012] Data confidentiality using fragmentation in cloud computing. *International journal of networks and distributed system*, 1:85–90.

[28] Mohand M et al [2012] Enhanced data security model for cloud computing. International conference on Informatics and systems. vol.36, pp.cc-12.

[29] PachipalaYellamma et al. [2013] Data Security In Cloud Using RSA. 4'th International Conference on Computing Communication and Networking Technologies, pp. 1-6.

[30] Sonia sindhu. [2015] A survey of security algorithms in cloud computing. International journal of Advanced Research in Computer Engineering & Technology,.4(5):2368–2371.

[31] Ramesh K, Ramesh S. [2014] Implementing One time password based security mechanism for securing personal health records in cloud. International Conference on control, Instrumentation, Communication and computation technologies. pp. 968–972.

[32] Subbhiah S, Selva S [2015] Distributed data security for data prevention in cloud computing using One time password for user authentication. Journal of Environmental Science, *Computer Science and Engineering & Technology* 4:752–758.

[33] Priyanka Nema [2014] An Innovative Approach for dynamic Authentication in Public cloud: Using RSA, Improved OTP and MD5. *International Journal of Innovative Research in Computer and Communication Engineering*. 1(11):6697–6702.

[34] RandeepKaur, SupriyaKinger. [2014] Analysis of Security Algorithms in Cloud Computing. *International Journal of Application or Innovation in Engineering & Management* 3: 171–176

[35] SunithaSharma et al. [2013] Enhancing Data Security In Cloud Storage. *International Journal of Advanced Research in Computer and Communication Engineering* , 2: 2132–2134

[36] Vijendra et al. [2014] Data Storage Security in Cloud Environment with Encryption and Cryptographic Techniques. *International Journal of Application or Innovation in Engineering & Management,* 3: 209–213

[37] Jay Singh et al. [2012] Improving Stored Data Security In Cloud Using RC5 Algorithm. Nirma University International Conference on Engineering. pp. 1–5.

[38] DeepikaVerma, Karan Mahajan. [2014] To Enhance Data Security in Cloud Computing Using Combination of Encryption Algorithms, 2: 41–44.

[39] Honey Patel, JasminJha. [2012] Securing Data in Cloud Using Homomorphic Encryption. *International Journal of Science and Research*. 4, :1892–1895.

[40] Jayanthi M et al. [2014] Analysis on Secure Data Sharing using ELGamal's Cryptosystem in Cloud. *International Journal of Computer Science and Electronics Engineering*, 4:50–55.

[41] Raghul et al. [2015] Data Security in Federated Cloud Environment using Homomorphic Encryption Technique. *International Journal of Emerging Technology and Advanced Engineering*, 5:137–141.

[42] Vishal Paranjape, VimmiPandey [2013] An Approach towards Security in Private Cloud Using OTP. *International Journal of Emerging Technology and Advanced Engineering* 3:.683–687.
[43] Abhishektripathy, TarunGoyal. [2014] Cloud Data Security Using Encrypted Digital Signature & 3D Framework. *International Academic of Science, Engineering and Technology* 3:114–121.
[44] ShikhaChoksi. [2014] Comparative Study on Authentication Schemes for Cloud Computing. *International Journal of Engineering Development and Research*, 2: 2785–2788.
[45] HanumanthaRao et al. [2013] Data Security in Cloud using Hybrid Encryption and Decryption. *International Journal of Advanced Research in Computer Science and Software Engineering*. 3: 494–497.
[46] Roshani et al. [2015] Data Security in Cloud through Confidentiality and Authentication. *International Journal for Scientific Research & Development* 3: 1735–1738.
[47] Dimpi Rani, Rajiv. [2014] Enhanced Data Security of Private Cloud Using Encryption Scheme with RBAC. *International Journal of Advanced Research in Computer and Communication Engineering* 3: 7330–7337.
[48] Pradeep et al. [2012] Enhancing Data Security in Cloud Computing Using 3D Framework & Digital Signature with Encryption. International Journal of Engineering Research & Technology. 1:1–8.
[49] Ranu S, Hasna. [2015] Biometric Based Approach for Data Sharing in Public Cloud. *International Journal of Advanced Research in Computer and Communication Engineering*. 4:95-97.
[50] SuchitaKolhe et al. [2015] Five-Level Authentication Security in Cloud Computing. *International Journal for Research in Emerging Science and Technology*, 2:116–118.
[51] Sasi E, Saranyapriyadharshini.[2015] Secured Biometric Authentication In Cloud Sharing System. *International Journal of Computer Science and Mobile Computing*, 4: 572–577.
[52] Sudhansu & Biswaranjan [2014] Enhanced data security in cloud computing using RSA encryption and MD5 Algorithm. International *Journal of Computer Science Trends and Technology* 2(2):60–64.

ABOUT AUTHORS

Ms. Tamilarasi R is a Research Scholar in School of Computing Science and Engineering, VIT University, Vellore. She graduated M.S c in Computer Science. She is doing research in Cloud Computing. She published two papers in reputed Journals. Her area of specialization is Cloud Computing.

***Dr. Prabu Sevugan** completed Bachelor of Engineering in Computer Science and Engineering from Sona College of Technology (Autonomous) and Master of Technology in Remote Sensing from College of Engineering Guindy, Anna University Chennai and one more Master of Technology in Information Technology at School of Computer Science and Engineering, Bharathidasan University Trichy. Did his Doctoral studies on Integration of GIS and Artificial Neural Networks to Map the Landslide Susceptibility from College of Engineering Guindy, Anna University, Chennai. He was a Post-Doctoral Fellow at GISE Advanced research lab, Department of Computer Science and Engineering, Indian Institute of Technology Bombay. He has more than 45 publications in national and international journals and conferences. He organized 3 International Conferences which includes one IEEE Conference as chair and also participated in many workshops and seminars. He is a member of many professional bodies and senior member of IACSIT, UACEE and IEEE. He is having more than ten years of experience in teaching and research. Currently I am working as a Division Chair for Parallel and Distributed Computing, School of Computing Science and Engineering, VIT University Vellore.*

***Swarnalatha Purushotham** is an Associate Professor, in the School of Computing Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Systems. She has published more than 57 papers in International Journals/International Conference Proceedings/National Conferences. She is having 15+ years of teaching experiences. She is a senior member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE.She is an Editorial board member/reviewer of reputed International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.*

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Sankar and Sevugan

ARTICLE OPEN ACCESS

AN INVESTIGATION ON HYBRID COMPUTING FOR COMPETENT DATA STORAGE AND SECURE ACCESS FOR GEO-SPATIAL APPLICATIONS

Karthi Sankar* and Prabu Sevugan

School of Computing Science and Engineering, VIT University Vellore, Tamil Nadu, INDIA

ABSTRACT

The GIS and cloud computing which is a hybrid method, a proposed technique is made on par with existing stenography and cryptographic algorithms. In modern years different steganography methods has been used to construct data in a more secured way along with special steganalysis methods. There are various tools that may be used for detecting hidden information which is easily available over the internet in order to secure the data from steganalyst. This is considered to be a gap from the review of the paper. Internet based computing provides security to cloud users as encrypted data in the cloud that protects the data from many attacks. Cryptography using GIS is collective geographic information for geographic or locational component purposes. It discharges Infrastructure as Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS) to computers. The existing papers meet the services individually, but as a whole, this will be done by GIS and CC which is a proposed hybrid model. Many open source tools will be used for hybrid model which is secured for cryptography. Also MapReduce is used to minimize the storage area which is mapped to GIS database in a cloud. The author proposed hybrid model to deal with information in a secured way that will store and hide the information. This happens along with some normalization techniques. Finally, the proposed model yields results in an efficient way of delivery of information without any loss of data and with minimum time through effective load balancing

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26– March-2016
Published on: 20th–May-2016

KEY WORDS

Steganalysis, Geo-Spatial, Cloud Computing, GeoDB, GeoCloud

***Corresponding author: Email:** cskarthi11@gmail.com **Tel:** +91 9943369964

INTRODUCTION

Nowadays, Geographic Information System (GIS) is having enormous growth in all kinds of industries. It enables users to visualize, issue, analyze, process store and interpret spatial data to understand relationships, patterns, and trends. A GIS provides an electronic representation about the Earth's natural and man-made features which in turn deal with real-world spatial data elements to a coordinate system through high computing resources. It involves large size of spatial repositories and the complicacy of the geospatial models with an increase in time complexity and high computing resources.

Cloud Computing is evolving as a key computing platform for sharing resources that include infrastructures, software, applications, and business processes. Cloud computing is an Internet based computing which provide servers, storage applications and resources to many organizations. Cloud Computing is a model to enable expedient, on-demand network access to a shared pool of configurable computing resources which can be rapidly provisioned and released with minimal management effort or service provider interaction. Security in cloud computing is very necessary so that it would be more effective and useful. The users do not have any idea where their data is placed. Hence, triple Security in Cloud Computing is made using related algorithms to provide availability, confidentiality, integrity to the data.

Cloud computing technology includes Hadoop platform and MapReduce parallel computing model into the domain of geographical information system to solve issues like spatial data storage, spatial data index and spatial operation in various applications of GIS. MapReduce is applicable for computing-intensive spatial applications and the lead in scalability and data storage efficiency is done by Hadoop. Thus large-scale distributed data management provides an efficient way for big data storage based on cloud computing technology.

THE IIOAB JOURNAL
ISSN: 0976-3104

GIS is an Integrated System of Computer Hardware, Software and Spatial Data that performs controlling and analytical operations on this data to produce reports, graphics, statistics and controls geographic data processing workflows. In turn, Cloud Computing is a parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements. Geographic Information Systems applications have been moving into the cloud with improved drive, Global organizations like ESRI, GIS Cloud Ltd etc have already taken the quantum leap and taken a technical shift to Cloud Computing Paradigm. GIS Cloud infrastructure providers are Amazon, Microsoft and IBM which provides reliability and security towards cloud technology to the end-users. GIS cloud is a service where large datacan be stored and accessed securely over the Internet through computer's hard drive

RELATED WORK

The author [1] describes that GIS and spatial data techniques lead to distributed data method and computing environment. This made the cloud computing technology to focus on extensive resource sharing and low cost for large data storage technology. They proposed new architecture for spatial data processing based on cloud computing technology. in particular, the problems of large volume of spatial data in some of the disaster cases which make the needs for stretch way for storing and analysis and computing all of these resources. Proposed architecture for spatial data processing based on cloud computing technology then results showed better presentation in assessment to previous works. To enhancing the proposed model to support raster data of GIS system also using full use of computing services in the cloud computing infrastructure to swarm large volume of data as possible using Amazon storage services such as Amazon S3.

The paper [2] deals with the cloud computing which is the most admired information technology for data processing. This enables users to process large amount of information without having their own computing power and this also apt for geospatial data processing. To examine how the cloud computing can improve geospatial data processing with comparing its advantages and disadvantages and discussing cloud computing features. The cloud computing technique's growth in the field of geospatial data seem to become a hot spot and has to be more extensive in the further tasks as the valuable consequences produced and seen from various organizations.
The author [3] deals with the CC which is the internet based technique provide various resources. The accessibility of these resources is very flexible in nature and few are obtainable to customers free of cost but some are on a pay-as-use basis. Then the customers also allowed to access information and utilize the computer resources from anywhere having internet access. This cloud computing technology uses some security technology like SHA 512, AES, and stenography to overcome the various security issues (stealing, hacking, unauthorized access etc.). The author planned SHA-512 is used for verify integrity of data. For encryption, use AES and then apply Steganography which can change presentation of data so that unauthorized person could not access the hidden data. Hence this is essential for providing maximum security to the data. Receiver can get original plain text by reversing the, SHA-512, AES, steganography. Outlook implemented SHA-512, AES and STEGANOGRAPHY to provide maximum security in cloud computing. By implementing these three algorithms we are intended to provide maximum security to the data. The technique used also increased the time complexity which should be reduced and hence we will try to improve the same

The author [4] said the enlargement of spatial technology and the spatial data have been collected through different approaches. Raising the anxiety is being put forward on use of spatial information in emergence system. This GIS technique and spatial information have lead to more distributed process and computing environment. Then the cloud computing technology is focused on extensive resource sharing and low cost for large data storage technology. Cloud based service to compute the large spatial data technique in emergence management and provides efficient spatial service to the spatial platform. Cloud means infrastructures that provide resources and services over the Internet. Frequently the services are layered to create a stack of cloud services that serves as a platform for developing cloud-based appearance management applications. Face new challenges as try to manage our spatial data that might be stored in a variety of devices.

This author [5] says about the Orfeo Toolbox (OTB) which is the tool for the operational development of the future sub-metric optic and radar image. The OTB is useful for all public working in the Remote sensing imagery community. Using a open source license, CNES hopes to gain from charity of many specialists to develop the particular use of satellite imagery.

THE IIOAB JOURNAL ISSN: 0976-3104

The paper [6] deals with the cloud computing technology which is an important issue of scientific community to the spatial data. It refers the computing infrastructure for an illustration of network. Then the main individuality of cloud computing is being dynamic, high power in computing and storage. It is cost effective for web based spatial data and for difficult analysis. This cloud technology is an easiest way of distributed computing that handles varied data. One of the most important feature of cloud computing is capability in powerful computing and dynamic storage. Also they discussed that cloud computing is intended for data infrastructure which in turn be able to develop the relation for spatial data infrastructure in the upcoming.

The author [7] told Geospatial data is an essential element in decision making processes and planning efforts across a variety of industries and information sectors. This quantity and variety of data is rapidly increasing and where as more of this data is risk oriented of being lost or becoming unusable. There is a increasing recognition of the significance of being able to access historical geospatial data.

The author [8] describes about the enlargement of new generation of spatial databases, where the DBMS is able to manage spatial and non-spatial data types together. This spatial database deal with vector geometries but it has limited facilities for managing image data. It is important to make database capable of dealing with image together with other spatial and non-spatial data types. To describe a solution for efficient handling of large image data sets in a standard object-relational database management system. By means of sufficient indexing, density and retrieval techniques, acceptable performances can be achieved using a standard DBMS, even for very large satellite images. The part of the development of which aims to provide a complete environment for the development of GIS applications.

The author [9] said geo information agencies are managed, maintained and build the geospatial information platform. They combine the various geo-spatial data to give data analysis services for supporting government decision making. This big data is difficult to address the data and computing intensive issues by established platform. This uses HDFS for managing image data and MapReduce-based computing service. Further optimization of platform architecture. Especially for IaaS layer, need to improve the efficiency of DC2's resource schedule, and optimize workload balancing and auto-scaling algorithm, as well as increase platform stability and reliability. For PaaS layer, it will expand GIS service functions for service chaining with thematic application services, such as land use and planning, government emergency response, and geographical condition analysis, etc. On the other hand, need to consider the unified framework of spatial data cloud storage, which implements integrated management based on the spatial data access interface for spatiotemporal data, such as managing spatial vector data and integrating HDFS with geodatabase.

The paper [10] says about Hadoop raised area for image processing rather than for its unique purpose of text processing. It has never been proved that Hadoop can be efficiently utilized for high-volume image files. The purpose of Hadoop for image processing has been researched using eight different practical image processing algorithms. It expand the file loom in Hadoop to deem the entire TIFF image files as a unit by increasing the file format that Hadoop uses. This technique is scalable and resourceful in processing various large images; used frequently for remote sensing applications. This creates variation between the single PC runtime and the Hadoop runtime which is evidently noticeable.

The author [11] described about evaluation of multi-datacenter Hadoop deployment with single-datacenter Hadoop deployment to identify the performance issues that inherent in a geographically distributed cloud. A simplification of the problem description in the context of geographically distributed cloud datacenters is also provided with deliberations on general optimization strategies. It describes about the design and realization of a suite of system-level optimizations for improving performance of Hadoop service provisioning in a geo-distributed cloud. This also deals with the prediction-based job localization that configures HDFS data placement and data perfecting.

The author [12] discuss about the Hadoop which is a java based programming framework that supports the storage and process of large data sets in a distributed computing environment. It is suitable for high volume of data. It is using with HDFS for data storage and MapReduce to process the data. The main aim of Mapreduce programming model is to parallelize the job implementation across multiple nodes for execution. Multiple nodes for execution. , all center of the researchers and companies toward to Hadoop. due this, many scheduling algorithms have been proposed in the past decades. There are three important scheduling issues in MapReduce such as locality, harmonization and equality. The most common objective of scheduling algorithms is to minimize the completion

THE IIOAB JOURNAL ISSN: 0976-3104

time of a parallel application and also achieve to these issues. In describe the overview of Hadoop MapReduce and their scheduling issues and problems. then, it have studies of most popular scheduling algorithms in this field. finally, highlighting the implementation Idea, advantages and disadvantage of these algorithms.

This paper [13] deals with the MapReduce systems that gives suitable due to their superior scalability, fault tolerance, and flexibility to handle unstructured data. It discovers the viability of building a hybrid system that takes the best features from both technologies. Hadoop and Hive are relatively young open-source projects. HadoopDB will automatically benefit from these improvements. HadoopDB is therefore a hybrid of the parallel DBMS and Hadoop approaches to data analysis, achieving the performance and efficiency of parallel databases, yet still yielding the scalability, fault tolerance, and flexibility of MapReduce-based systems. The ability of HadoopDB to directly incorporate Hadoop and open source DBMS software makes HadoopDB particularly flexible and extensible for performing data analysis at the large scales expected of future workloads

The author [14] discuss about Hadoop data where users face with sensible advice on how to protect these environments. This method of data storage is increasing extremely in lots of folds. There has been increasing data security and isolation concerns for people who outsource data on this Hadoop clusters. The spatial data analysis of big datasets using distributed method. Experimented the cloud computing technology in spatial fields over single spatial databases and proved the performance and efficiency of operations on spatial data in Hadoop environment. The main feature of hadoop is to partition the data by calculating 1000's of hosts in similar to the remaining data. Geospatial data analysis has been done through Hadoop and MapReduce in cloud computing technology. The author concludes that the execution of Spatial query in mapreduce involves join transactions on small scale yields efficient results with minimum time on comparison with hadoop and distributed DB systems.

The paper [15] discuss about Apaches Hadoop- HDFS which is used to store the streaming data that is too big in size most organization. Using Hadoop Map Reduce for computing and HDFS for storage. This Hadoop technique is most trendy for analysis, storage and to process very large data which does not require lots of changes in hadoop system. This Hadoop application requires streaming access to data files. Data storage and data processing try to solve which helps hadoop system to improve processing speed and reduce time to execute the task. Hadoop application requires stream access to data files. During placement of data files default assignment of Hadoop does not consider any data characteristics. If the related set of files is stored in the same set of nodes, the efficiency and access latency can be increased. Hadoop uses Map Reduce framework for implementing large-scale distributed computing on unexpected data sets. There are possible duplicate computations being performed in this process. No mechanism is to identify such duplicate computations which increase processing time. Solution for above problem is to co-locate related files by considering content and using locality sensitive hashing algorithm which is a clustering based algorithm will try to locate related file streams to the same set of nodes without affecting the default scalability and fault tolerance properties of Hadoop and for avoiding duplicate computation processing mechanism is developed which store executed task with result and before execution of any task stored executed tasks are compared if task find then direct result will be provided . By storing related files in same cluster which improve data locality mechanism and avoiding repeated execution of task improves processing time, both helps to speed up execution of Hadoop.

The author [16] describes the cloud data base that is based on hadoop technologies that is hadoop distributed file system (HDFS). Data is simulated in different data nodes which can be accessed by name of the node using logs that are present in them. They use Mapreduce method to process the data on cloud with various types of systematic to perform using map reduce codes. As storage mechanism on cloud to the sender subscriber to a cloud Daas Hadoop enables surplus data to be streamlined for any distributed processing system across clusters of computers using simple programming models. It truly is made to scale up from single servers to a large number of machines, each and every offering local computation, and storage space. Instead of depending on hardware to provide high-availability, the library itself is built to detect and handle breakdowns at the application layer, so providing an extremely available service along with a cluster of computers, as both versions might be vulnerable to failures. This module described the Map Reduce execution platform at the heart of the Hadoop system. By using Map Reduce, a high degree of parallelism can be achieved by applications. The Map Reduce framework provides a high degree of fault tolerance for applications running on it by limiting the communication which can occur between nodes, and requiring applications to be written in a "dataflow-centric" manner

The author [17] discuss about the development of a scalable spatial data supervision system and their various spatial queries with MapReduce which cannot be supported using MapReduce without difficulty. Then they focus on problems of difficult spatial applications and current methods correspondingly. It implements two distinctive

THE IIOAB JOURNAL ISSN: 0976-3104

spatial applications, all nearest neighbor and astronomical cross-match which face the same complex problem where distance computing is essential. MapReduce is a key-value based programming model and an associated implementation for processing large data sets. It has been adopted in various scenarios and seems promising. However, when spatial computation is expressed straightforward by this key-value based model, difficulties arise due to unfit features and performance degradation. The present methods as splitting method for balancing workload, pending file structure and redundant data partition dealing with relation between spatial objects, a strip-based two-direction plane sweeping algorithm for computation accelerating. Based on these methods, ANN(All nearest neighbors) query and astronomical cross-certification are developed. Performance evaluation shows that the MapReduce-based spatial applications outperform the traditional one on DBMS. We implement two typical spatial applications, all nearest neighbor and astronomical cross-match, which faces the same difficult problem that distance computing is required

The author [18] says that the Hadoop Distributed File System (HDFS) is used to store the huge amount of data set consistently to stream the data sets at high bandwidth to user applications. In this type of cluster, thousands of servers mutually deals with storage and implement user application tasks. Through distribution of storage and computation across many servers, the resource can develop at an economical growth at each size. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size. We describe the architecture of HDFS and report on experience using HDFS to manage 25 peta bytes of enterprise data at Yahoo.

The paper [19] deals that the MapReduce framework and Hadoop are temporarily introduced in this segment. It describes about the gradual execution process of this programming model, as well as its core modules. This also analyzes the distributed file system of Hadoop (HDFS) and its specialized nodes. The outlined the importance of spatial data, since they are used in almost every research field that needs to depict multidimensional data. The illustrated how they can be efficiently processed in a parallel manner. Furthermore, due to ever-growing data sets, the need arises for technological advancements towards information management. It is clearly understandable that parallel computing is more beneficial over sequential methods. As a result, the way towards parallelization – MapReduce and its implementation, Hadoop – will be the solution to a variety of difficult computational problems The author [20] describes about the SpatialHadoop and a full-fledged MapReduce framework with native support for spatial data. SpatialHadoop is a comprehensive expansion to Hadoop that injects spatial data awareness in each Hadoop layer, namely, the language, storage, MapReduce, and operations layers. This method used to improve the cloud architecture which is hybrid for object storage based for geospatial data processing. Due to increase in retrieval of the geo-spatial data by various fields, the necessity for delivery of the geospatial process is also increased. This has led to hybrid cloud architecture to handle geospatial data compared to object based data storage. The hybrid model allows retrieval of geospatial data from public cloud. The execution is possible over a distributed large scale devices which are virtually reserved by the organization of cloud

The author [21] says that Encryption techniques are the most secured method to transfer the perceptive information from sender to the intended receiver. The main Objective of this method is to create a perceptive information illegible to all other except the receiver. The secure transmission of information AES encryption has been used which provides most secure way to transfer the sensitive information from sender to the intended receiver. The main purpose of using this technique is to make sensitive information unreadable to all other except the receiver. The data thus compressed enables utilization of storage space in cloud environment. It has been augmented with Hadoop's map-reduce paradigm which works in a parallel mode. The experimental results clearly reflect the effectiveness of the methodology to improve the security of data in cloud environment.

The paper [22] deals with the cloud computing technology which is the tremendous next generation information technology. This technology is used to access remote data and provide a large amount of storage space .It provides a secured cloud storage system and also maintains privacy between Third party auditor and Cloud service providers. In cloud computing the Third Party Auditor guarantee that the cloud service provider & also itself TPA would not learn any knowledge about the data that is stored on the cloud server. During the efficient auditing process, it not only eliminates the burden of cloud user from the tedious and possibly expensive auditing task. The partitioning of data will enables storing of the data in easy and effective manner. It also gives way for flexible access and there is less cost in data storage. The space and time will also effectively reduce during storage. Also the remote data integrity checking detects the threats and misbehaving server while storing the data in cloud ensuring data security. Calculating digital signature may secure file more efficiently.

THE IIOAB JOURNAL
ISSN: 0976-3104

The author [23] discuss about the Steganography technique to secure the data dealing with number of attacks. Tools have been considered for detection of information over Internet. Instead of secured algorithms of the existing algorithms to hind the image with data , still public key like Deffie-Hellman and RSA is susceptible in mathematics. Hence the paper [] proposed quantum cryptography with steganography which provides security in an efficient way. They use of quantum cryptography has added another layer of defense to our data. Even if we use most secure encryption algorithm and best stegno technique to hide our data, if the keys gets compromised these security will be of no use. Quantum cryptography addresses current as well as emerging threats and it definitely has “competitive advantage” over other public key cryptosystems. They used quantum cryptography only for key distribution rather than for entire messages because of limitations of transmission speeds and hardware expenses. The representation of bits through polarized photon is the foundation of quantum cryptography that serves as the underlying principle of quantum key distribution. They are concentrates on theory of quantum cryptography and the use of quantum cryptography for key distribution and how the use of quantum cryptography contributes to the field of steganography. They using quantum mechanics and photon polarization we can provide perfect security

The paper [24] deals with the Advanced Encryption Standard (AES) which is a secured authentication technique for cloud computing environment. In Advanced Encryption Standard, the key deliberation dealt in this application is the encryption schema for protected data by making it incoherent for all. . Implementing AES for security over data provides less memory utilization and less computation time as compared to other algorithms. It also provides security to cloud users as encrypted data in the cloud that protects the data from many attacks

The paper [25] deals with cloud computing which is an internet based computing technology. It is used to share the large amount of software information and resources to the world. In this cloud environment, resources are shared to all servers to separate users. This computing system supports distributed services, multi-domain Infrastructure, and multi-users to secure data efficiently. It also guarantees the secured data storage system by using several encryption algorithms with digital signature. The security model in a cloud computing environment, here file one encrypted with RSA algorithm in which keys are created sequence one by one to the system. This ensuring a major secures and also solves the main security issues like a new login user data hacking to the attacker. Login into the main system is compulsory and download, store the files. The encrypted a file is hide from unauthorized users. The files already store the main system server. It only single user multiple servers. The user forgets a password and not able to access same user name have key value to identify unique values. Once login the entry detail is cannot access the same user name login. The new user security for cloud computing platform includes RSA and encryption algorithm. The user login execution period is not a part of higher.(ie) implementation of each algorithm is perform different servers, and download, upload a files to take overall system is stop difficult. The RSA algorithm and digital signature with encryption model high secured and light encryption system information. We want to work ensure safe communication computers between systems to user.
The paper [26] deals with cloud computing that is used to resolve the day to day computing problems like Hardware, Software and Resource accessibility unhurried by Computer users. It is used to provide unchallenging and non-ineffectual solution for computational issues. They implement the digital signature with RSA algorithm to provide the Cloud Storage Methodology and Data Security in cloud.

The author [27] discussed three algorithms to secure the communication on the network. These algorithms provides authentication of the document by the sender through digital signatures. Data encryption algorithm provides encrypted data to secure data from outsider attacks. Steganography can change the presentation of data to make the data secured and to avoid the interpretation of data by users in order to provide data security on the network. This security related issues are the greatest obstacle in the popularity of cloud. Therefore we are going to use the combinations of three different algorithms- DSA, DES and Steganograohy. These algorithms help to reduce the problems of security on cloud. Digital signature Algorithm, Data Encryption Standard and Steganography to improve the security in cloud computing. We find that the Time complexity is high because it is a one by one process but in future this time complexity could be reduced. We try to improve the time complexity by using other security algorithms.

The author [28] describes integration of GIS and cloud computing solves many issues based on large-scale scalable server cluster. The author discusses about issues like storage of spatial data, index of spatial data and its operations. They evaluated the presentation and effectiveness of operations of spatial data in the platform of Hadoop with the original data available. Normally spatial objects are bigger and more complex structure with

THE IIOAB JOURNAL ISSN: 0976-3104

various properties. Every object not only has spatial properties such as points, lines and polygons in itself, but also group of similar objects with their relationship in clustering of spatial data. These clusters give easy access to process the data where the access does not use the Hadoop Distributed File System (HDFS) data management interface directly to access and operate the spatial data. Data files are divided into a lot of small discrete objects which are stored in different processing nodes in HDFS file system using MapReduce. This achieves storage efficiency and minimum computing power in accordance with scope and the space size. Author proposed a hybrid index process of data preprocessing which is spatial search and spatial operation of Hadoop a distributed search method. Hadoop has done dynamic spatial data index. The index data split to basic unit handled by each node and Map/Reduce node can concurrently handle the data file and its equivalent index file. The author done preliminary research on storing and indexing of spatial data. For this, work on spatial data analysis, processing and mass spatial data storage is required to attain the objective of geospatial cloud service.

The author [29] explained that the Hadoop – GIS is a high performance system that deals with big data by modifying query output (execution of query on MapReduce, etc).This paper proved the high scalability and performance of Hadoop-GIS with global partition indexing system on comparison with parallel SDBMS which performs efficient SDBMS to calculate intensive queries. Hadoop-GIS is available in terms of package in Hive, a software package, a set of library to process queries of spatial data. The author proposed a technique on index of building which partition the data to execute queries. The paper concludes with efficient query processing of spatial data. They have proposed an effective solution for analytical spatial queries over big datasets but they did not concentrate much on complicated spatial queries in their proposed system.

The author [30] developed a spatial Hadoop in MapReduce frame for spatial data which overcomes Hadoop-GIS drawbacks. This paper yields efficient throughput than Hadoop for k-nearest-neighbor queries and triple presentation for spatial joins.

The author [31] describes that Hadoop is a programming platform used to support the processing of large data sets in a distributed computing environment. Authors described the evolution of Hadoop which is a framework that divides an application into different parts. The current Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper.

The author [32] discussed about geospatial data in a distributed environment for Big Data analysis in WebGIS. This involves cloud-based storage and calculation of commodity cluster that lead to increase in performance of Input and Output. Also, parallelized performance is achieved simultaneously. This paper concludes with performance improvement of 14.6 % and 57%-153% better than traditional database management system.
The author [33] proposed techniques for handling the large size of images acquired by airborne and remote sensing satellite sensors. Also explained various techniques used for effective management and process to handle large data sets.
The author [34] described the HDBS based Distributed cache system in their work. The cache services are designed with three access layers an in-memory cache, a snapshot of the local disk, and the actual disk view provided by HDFS.
The author [35] discussed three algorithms to secure the communication on the network. These algorithms provides authentication of the document by the sender through digital signatures. Data encryption algorithm provides encrypted data to secure data from outsider attacks. Steganography can change the presentation of data to make the data secured and to avoid the interpretation of data by users in order to provide data security on the network.

The author [36] describes the security related problems like hacking, stealing, misusing etc. and also deals with the security related problems in cloud computing. The author used combinations of three different algorithms such as DSA, DES, and Steganography. These algorithms are used to rectify the issues on cloud to store the data. Some unwanted activities can damage the data that is cybercrime. The security issue in cloud is that, it has a single point of failure. One mistake or failure can impact the whole group. The hacker not only hack the cloud data it also hack the user account. The main intent of security is to provide availability, confidentiality, integrity to the data. The author used triple security in cloud computing by using three different security algorithms. These algorithms are based on signature of digital based encryption of data and data hiding beyond the file of audio. The author implement Digital signature Algorithm, Data Encryption Standard and Steganography to improve the security in cloud computing with the disadvantage of time complexity.

THE IIOAB JOURNAL ISSN: 0976-3104

The author [37] describes the method to pass information with unknown existence to repel attention of the potential attacker. This method of information hiding can be given to copyright protection for digital media along with hiding data for confidentiality. The paper focuses on the Least Significant Bit (LSB) method to hide messages in an image. The author enhanced the LSB technique by randomization to disperse the messages in an image to make harder for unauthorized users to get the original messages.

The author [38] describe data security initiate the threats and the unpleasant trend that has necessitated the need for an advanced approach to data security. Serpent encryption algorithm and distributed steganography are already proven techniques for securing the data. This paper describes an developed mechanism to ensure data security by strategically combining serpent cryptographic algorithm and distributed steganography. The unified approach is aimed at leveraging the strength of these two proven techniques to achieve a robust mechanism for ensuring confidentiality and integrity of data in the cloud. In this paper, they focus on securing the data in the cloud to ensure the data confidentiality and integrity as a proposed work.

The author [39] described data storage and management technology because of cloud computing are getting more extensive attention in various applications. Security will be a serious challenge to store the data in the cloud environment. This paper describes the advantages of spatial data management in cloud computing environment. The paper sets a spatial data security management model under the cloud environment, and reviews the key security technologies of the spatial data storage and management. But the application of cloud computing in spatial data storage and management is favorable to many IT enterprise and users. It has a lot of advantages in cloud computing technology which is still to be improvised further.

The author [40] describes the collection of thousands of servers host directly that are connected to store and execute the client request tasks with distribution of storage of data and computation across many servers which the resource can produce with demand which are economical at every size. The author explains the architecture of HDFS and manages 25 peta bytes of enterprise data at yahoo.

The author [41] describes the Hadoop Distributed File System (HDFS) storage files. HDFS gives scalable, fault-tolerant storage at low cost. HDFS stores files across a collection of servers in a cluster. Files are splitted into a blocks and each block is written to more than one servers. This copying provides both fault-tolerance and performance. HDFS make sure data availability by continually monitoring the servers in a cluster and the blocks that they manage.

The author [42] describes the fault occurs in large scale distributed systems such as Hadoop clusters. Native Hadoop provides basic support for Fault tolerance. However, simply re-processing the whole task decreases the efficiency of job execution, especially when the task is almost done is discussed in this paper.

The author [43] describes the inspects Hadoop cluster and security for Hadoop clusters using Kerberos. Then security enhancement using role based access control, reviewing built in protections and weaknesses of these systems. The goal is to explore security problems in which Hadoop data users face with pragmatic advice on how to secure these environments. Since this new technology of data storage is increasing tremendously in many folds, there have been increasing data security and privacy concerns for people who outsource data on this Hadoop clusters.

The author [44] deals with confidentiality for encrypting large data. As it is a time consuming process, this is controlled by an efficient application of the process in parallel mode. This paper yields efficient cost solution to process large scale data by encryption which undergoes compression to save the space. The persistent storage compared to network storage that gives high performance which is cost-effective than other traditional alternatives. And the cloud infrastructures should support the use of persistency which is attached locally for efficient support of Big Data and other I/O intensive applications. As there are existing methods which supports networked storage for availability of data may not be applicable to big data systems.

THE IIOAB JOURNAL
ISSN: 0976-3104

COMPARISON OF SPATIAL DATABASE AND SPATIAL CLOUD DATABASE

Table: 1. shows the comparison of spatial database and spatial cloud database

Sno		Spatial Db	Spatial Cloud Db
1	security	Normal security	High security
2	user access	Limited	Global access
3	Storage capacity	limited	unlimited

PROPOSED METHODOLOGY

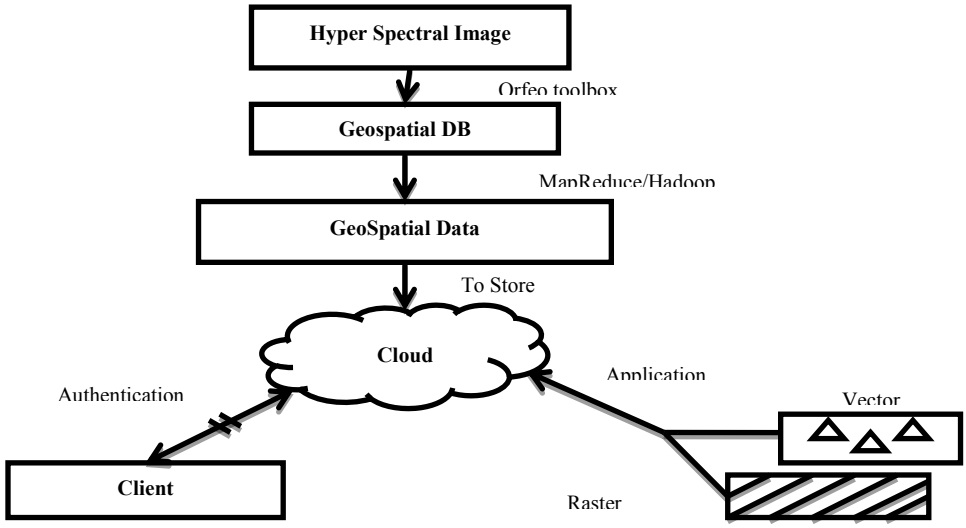


Fig: 1. proposed Architecture

For the proposed architecture, the hyperspectral images acquired by remote sensing satellites are given as input. The input images are already preprocessed. The input images are processing by using Orfeo toolbox to create Geo-Spatial Database.

Secondly, the geo spatial database is normalized and reduced using MapReduce and Hadoop to store it into the cloud.

Thirdly, the Map Reduced Geospatial data are processed for various application in the cloud environment. A simply authentication mechanism is used for providing access to the data which is stored in cloud. Based on the application and query given by the client the output will be given in either raster or in vector form

CONCLUSION

The paper is concluded by proposing a hybrid model through survey of papers on other existing stenography and cryptographic algorithms. Many tools also have been surveyed to detect hidden information over the internet. Also, MapReduce is used to minimize the storage area which is mapped to GIS database in a cloud using normalization techniques. The paper concludes with the detection of hidden information available over the internet in order to secure the data from steganalyst .Thereby the proposed model may result with delivery of

THE IIOAB JOURNAL ISSN: 0976-3104

information without any loss and with minimum time consumption. Finally the integration of GIS and cloud gives efficient data retrieval and other E-Commerce services that reduce any manual work by outsourcing to the cloud GIS Effective data balancing is made by applying load balancing strategy.

FINANCIAL DISCLOSURE

No financial support was received to carry out this research.

ACKNOWLEDGEMENT

Authors would like to thank school of computing science and engineering, vit university vellore for providing resource and support to carry out this research work.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

REFERENCES

[1] Eman Mahmoud, Osman Hegazy, Mohamed Nour El-Dien. [2013] Integration of GIS and Cloud Computing for Emergency System. International *Journal Of Engineering And Computer Science*.ISSN:2319-7242

[2] Yaser Khalilizangelani , Saman Ghaffarian.[2013] A Study of Geospatial Data Processing Based on Cloud Computing. Proceedings of ACRS 2013

[3] Ms. Manisha, Dr. Kamlesh sahrma, Ridhika Sharma Manisha.[2015] A Technique To Increase Integrity Of Cloud Data Using Hybridalgorithms. ISSN: 2277-9655

[4] Ge, X., & Wang, H. [2009] Cloud-based service for big spatial data technology in emergence management. In Proceedings of the ISPRS 38(7):C4.

[5] Christophe, E., Inglada, J., & Giros, A. (2008). Orfeo toolbox: a complete solution for mapping from high resolution satellite images. International Archives of the Photogrammetry, *Remote Sensing and Spatial Information Sciences*, 37: 1263-1268.

[6] Mohammad Naghavi.[2012]Cloud Computing as an Innovation in GIS & SDI: Methodologies, Services, Issues and Deployment Techniques. JGIS: 597-607

[7] McGavra G, Morris S, Janée G. [2009] Technology Watch Report: Preserving Geospatial Data.

[8] Vinhas L, De Souza RCM, Câmara G. [2003] November). Image Data Handling in Spatial Databases. In GeoInfo.

[9] Song WW, Jin BX, Li SH, Wei XY, Li D, Hu F. [2015] Building Spatiotemporal Cloud Platform for Supporting GIS Application.ISPRS Annals of Photogrammetry, *Remote Sensing and Spatial Information Sciences*, 1:55-62..

[10] Almeer MH. [2012] Hadoop mapreduce for remote sensing image analysis.*International Journal of Emerging Technology and Advanced Engineering*,2(4): 443-451.

[11] Zhang Q, Liu L, Lee K, Zhou Y, Singh A, Mandagere N, Alatorre G. [2014] Improving Hadoop Service Provisioning in A Geographically Distributed Cloud. In Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on (pp. 432-439). *IEEE*..

[12] Seyed Reza Pakize.[2014] A Comprehensive View of Hadoop MapReduce Scheduling Algorithms. *International Journal of Computer Networks and Communications Security* ISSN 2308-9830

[13] Abouzeid A, Bajda-Pawlikowski K, Abadi D, Silberschatz A, Rasin A. [2009] HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. *Proceedings of the VLDB Endowment,*2(1):922-933..

[14] Gaikwad, R. L., Dakhane, D. M., & Pardhi, R. L. (2013). Network Security Enhancement in Hadoop Clusters. International Journal of Application or Innovation in Engineering And Management (IJAIEM), 2:151-157.

[15] Sayali Ashok Shivarkar.[2014] Speed-up Extension to Hadoop System. *International Journal of Engineering Trends and Technology (IJETT)* ISSN: 2231-5381

[16] Brahmanaidu N, Riaz S. [2012]. Distributed Data Storage and Retrieval on Cloud by using Hadoop. *International Journal of Science and Research*..

[17] Wang K, Han J, Tu B, Dai J, Zhou, W, Song, X. [2010]. Accelerating spatial data processing with mapreduce. In Parallel and Distributed Systems (ICPADS), 2010 IEEE 16th International Conference on(pp. 229-236). *IEEE*..

[18] Shvachko K, Kuang H, Radia S, Chansler R. [2010] The hadoop distributed file system. In Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on (pp. 1-10). *IEEE*..

[19] Economides G, Piskas G, Siozos-Drosos S.[2013] Spatial Data and Hadoop Utilization..

[20] Eldawy A, Mokbel MF. [2015]. SpatialHadoop: A MapReduce framework for spatial data. In Proceedings of the IEEE International Conference on Data Engineering (ICDE'15*). IEEE*..

[21] Mehak , Gagandeep .[2014] Improving Data Storage Security in Cloud using Hadoop. Mehak Int. Journal of Engineering Research and Applications ISSN : 2248-9622.

[22] Yogesh V Bhapkar, Rakesh S Gaikwad, Milind R. Hegade Yogesh V. Bhapkar et al,[2015] Providing Security And Privacy To Cloud Data Storag. *International Journal of Computer Science and Information Technologies*, ISSN:2231-5381

[23] Kumar M, Gupta A, Shah K, Saurabh A, Saxena P, Tiwari VK. [2012] Data Security Using Stegnography and Quantum Cryptography.Network and Complex Systems, 2(2):46-55.

[24] Abha M, & Bhansali, M. (2013). Enhancing cloud computing security using AES algorithm. *International Journal of Computer Applications*, 67(9).

[25] T. Sivasakthi,Dr. N Prabakaran[2007] Applying Digital Signature with Encryption Algorithm of User Authentication for Data Security in Cloud Computing. *International Journal of Innovative Research in Computerand Communication EngineeringISSN(Online):* 2320-9801 ISSN (Print): 2320-9798

[26] Somani U, Lakhani K, Mundra, M. [2010] Implementing digital signature with RSA encryption algorithm to enhance the Data

Security of cloud in Cloud Computing. In Parallel Distributed and Grid Computing (PDGC), 2010 1st International Conference on (pp. 211-216). *IEEE*.

[27] Khushboo Gupta, Neha Goyal, Puneet Rani.[2014] Study of Security Algorithm to Provide Triple Security in Cloud Computing. *International Journal of Scientific and Research Publications* ISSN 2250-3153.

[28] Arya, Richa. [2014] Triple Security of File System for Cloud Computing. *International Journal of Computer Science and Engineering*, 2(3):148-154.

[29] Bhosale Harshawardhan S., and Devendra P. Gadekar. [2014] Review Paper on Big Data and Hadoop. *International Journal of Scientific and Research Publications*,4(10):1-7.

[30] Hu P, Dai W.[2014] Enhancing Fault Tolerance based on Hadoop Cluster, *International Journal of Database Theory and Application*, 7(1):37-48

[31] Lethby Toby, and A Jagan. [2015] Spatial Data Analysis using Map-Reduce Technique. International Journal of Electronics Communication and Computer Technology, 5(1):94-97.

[32] Izevbizua PO. [2015] Data Security In The Cloud Using Serpent Encryption and Distributed Steganography, *European Scientific Journal*, ESJ, 11(18):347-359.

[33] Mukhi, Parul, and Bhawna Chauhan. [2014] Triple System Security in Cloud Computing. *International Journal Of Engineering And Computer Science* 3(7):7364-7374.

[34] Rama Naga Durga Rao Khaja, Venkateswara Rao Kota. [2014] Hybrid Cloud framework for Object Storage based Geo Spatial Remote Sensing Data Processing. International Journal of Engineering Sciences & Research Technology, 449-454.

[35] Olson, Mike. [2010] Hadoop: Scalable, flexible data storage and analysis. IQT Quart 1(3):14-18.

[36] C Wanga,b, F Hub, X Hua, S Zhaoc, W Wena, C Yangb. [2015] A Hadoop-Based Distributed Framework for Efficient Managing and Processing Big Remote Sensing Images. ISPRS Annals of Photogrammetry, *Remote Sensing and Spatial Information Sciences*, 1:63-66.

[37] Eldawy Ahmed, and Mohamed F Mokbel. [2015] SpatialHadoop: A MapReduce framework for spatial data. Proceedings of the IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea.

[38] Li Xiang, Zi Fan Liu, Wen Bing Liu, An Xu, and Ling Ma. [2013] A spatial data security model under the cloud environment. *In Advanced Materials Research*, 765:1267-1270.

[39] Akshay MS, Suhas Mohan, Vincent Kuri, Dinkar Sitaram, HL. Phalachandra. [2014] Efficient Support of Big Data Storage Systems on the Cloud. arXiv preprint arXiv, 1411-7507.

[40] Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. [2010]. The hadoop distributed file system. 2010 IEEE 26th Symposium Mass Storage Systems and Technologies (MSST), 1-10.

[41] Wang, Yonggang, and Sheng Wang. [2010] Research and implementation on spatial data storage and operation based on Hadoop platform. Geoscience and Remote Sensing (IITA-GRS), 2010 Second IITA International Conference on IEEE, 2: 275 – 278.

[42] Zhang J, Wu G, Hu X, Wu X. [2012] A distributed cache for hadoop distributed file system in real-time cloud services. 2012 ACM/IEEE 13th International Conference on Grid Computing (GRID), 12-21).

[43] Zhong, Yunqin, Jizhong Han, Tieying Zhang, and Jinyun Fang. [2012] A distributed geospatial data storage and processing framework for large-scale WebGIS. 20th International Conference on Geoinformatics (GEOINFORMATICS), 1-7.

[44] Aji A, Wang F, Vo H, Lee R, Liu Q, Zhang X, Saltz J. [2013] Hadoop GIS: a high performance spatial data warehousing system over mapreduce. *Proceedings of the VLDB Endowment*, 6(11):1009-1020

ABOUT AUTHORS

Karthi Sankar *has completed his BE in CSE from Muthayammal Engineering College Rasipuram. And then he completed his ME in CSE from Jerusalem College of Engineering Chennai. Currently he is doing his PhD in School of Computing Sciences and Engineering, VIT University, Vellore, India. His area of research includes Geo-Spatial Application and Cloud computing.*

Dr. Prabu Sevugan *have completed Bachelor of Engineering in Computer Science and Engineering from Sona College of Technology (Autonomous) and Master of Technology in Remote Sensing from College of Engineering Guindy, Anna University Chennai and one more Master of Technology in Information Technology at School of Computer Science and Engineering, Bharathidasan University Trichy. Did his Doctoral studies on Integration of GIS and Artificial Neural Networks to Map the Landslide Susceptibility from College of Engineering Guindy, Anna University, Chennai. He was a Post-Doctoral Fellow at GISE Advanced research lab, Department of Computer Science and Engineering, Indian Institute of Technology Bombay. He has more than 50 publications in national and international journals and conferences. He organized 3 International Conferences which includes one IEEE Conference as chair and also participated in many workshops and seminars. He is a member of many professional bodies and senior member of IACSIT, UACEE and IEEE. He is having more than eleven years of experience in teaching and research. Currently I am working as a Division Chair for Parallel and Distributed Computing, School of Computing Science and Engineering, VIT University Vellore.*

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Hudedagaddi and Tripathy

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

KERNEL BASED SPATIAL FUZZY C-MEANS FOR IMAGE SEGMENTATION

Deepthi P Hudedagaddi[1*] and Balakrushna Tripathy[2]

[1,2]SCOPE, VIT University,Vellore,Tamil Nadu-632014, INDIA

ABSTRACT

An extension of various available clustering algorithms has been serving as a solution to serve many current problems by the researchers. The Fuzzy C Means (FCM) algorithm that has been in use all these days is extremely noise sensitive. Hence it fails to provide the desired results. This was solved to an extent with the introduction of spatial fuzzy c means. This included a spatial function which was the summation of all the membership values of the neighbors of the pixel considered for study. This paper proposes a new and better modification of the spatial fuzzy c means(sFCM) by introducing kernel distance metric. This groups the objects into clusters which are not separable linearly.Here radial basis kernel function is applied for sFCM clustering. The proposed clustering algorithm is tested on MRI image and noise induced MRI image. The results reveal that kernel based spatial fuzzy c means (sKFCM) is better than Euclidean based spatial fuzzy c means

.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26– March-2016
Published on: 20th–May-2016

KEY WORDS

Clustering, spatial,kernel, fuzzy sets, DB and D index, image segmentation

***Corresponding author:** Email: deepthiph@gmail.com **Tel:** +91-9986387435

INTRODUCTION

Image processing is a rapidly growing field of which image segmentation forms a major part. It has diverse field of applications. Some of them are object recognition, machine vision and medical imaging. Development of segmentation algorithms which are efficient and insensitive to noise has become a challenge. Hence, it has become the need of the hour to develop better algorithms in the field of image segmentation. They must also be capable of solving real world applications and which are sensitive to noise. Noise in real world images are inevitable[1].

Conventional FCM algorithm fails to provide appropriate results on images which have noise. Spatial FCM (sFCM) [2], is a two step procedure. It includes the neighborhood information of pixel taken for study. Though it fails in completely removing the distortion of noise, the algorithm proves to handle noise more efficiently than conventional FCM[3].

The sFCM uses Euclidean distance formula for finding the spatial data point distances. It is found in literature that Euclidean distance fails to provide good results in situations where clustering algorithms are distance based. However, kernel methods provide better results as compared to Euclidean. This paper uses kernel distance formula and compares it with the results from Euclidean. This is hence an extension to the existing spatial FCM.

DISTANCE METHODS

EUCLIDEAN DISTANCE

Suppose $a = (a_1, a_2, \ldots a_n)$ and $b = (b_1, b_2, \ldots b_n)$ are two points in the n-dimensional Euclidean space. Then the Euclidean distance d(a, b) between a and b is given by

THE IIOAB JOURNAL ISSN: 0976-3104

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (1)$$

KERNEL DISTANCE

Let 'a' denote a data point. Then transformation of 'a' to the feature plane which possess higher dimensionality be denoted by $\phi(a)$. Description of inner product space is given by $K(a,b) = \langle \phi(a), \phi(b) \rangle$. Let $a = (a_1, a_2, ...a_n)$ and $b = (b_1, b_2, ...b_n)$ are two points in the n-dimensional space. Kernel functions used in this paper are stated as follows

- Radial Basis Kernel

$$R(a,b) = \exp\left(- \frac{\sum_{i=1}^{n}(a_i^{\,p} - b_i^{\,p})^q}{2\sigma^2} \right) \quad (2)$$

Implementations of all the algorithms corresponding to Radial Basis Kernel have been done using p=2 and q=2 in equation (2).

- Gaussian Kernel (RBF with p=1 and q=2)

$$G(a,b) = \exp\left(- \frac{\sum_{i=1}^{n}(a_i - b_i)^2}{2\sigma^2} \right) \quad (3)$$

- Hyper-tangent Kernel

$$H(a,b) = 1 - \tanh\left(- \frac{\sum_{i=1}^{n}(a_i - b_i)^2}{2\sigma^2} \right) \quad (4)$$

where $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} \| a_i - a' \|^2$ and $a' = \frac{1}{N}\sum_{i=1}^{N} a_i$

For all kernels functions, N denotes total number of existing data points and ||x-y||denotes Euclidean distance between points x and y which pertain to Euclidean metric space. By D(a, b) denotes the complete form of kernel distance function where D(a, b) = K(a, a) + K(b, b) – 2K(a, b) and when similarity property (i.e. K(a, a) =1) is applied, the following is obtained

$$D(a,b) = 2(1 - K(a,b)) \quad (5)$$

EXISTING METHODS

Fuzzy models are incorporated in analysing spatial data.

Fuzzy Clustering

James C Bezdek developed fuzzy set based Fuzzy c-mean algorithm[4,5]. In this clustering method, each element can belong to more than one cluster. Each element is also associated with a set of membership values. Fuzzy clustering process invloves assigning every data element to one or more than one cluster by taking into account their membership values[6,7].

THE IIOAB JOURNAL ISSN: 0976-3104

1. Assign initial centers for c clusters.

2. Calculate distance d_{ik} between data objects x_k and centroids v_i using Euclidean formula

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

3. Generate the fuzzy partition matrix or membership matrix U:

If $d_{ij} > 0$ then

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \quad (6)$$

Else

$$\mu_{ik} = 1$$

4. The cluster centroids are calculated using the formula

$$V_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m x_j}{\sum_{j=1}^{N} (\mu_{ij})^m} \quad (7)$$

5. Calculate new partition matrix by using step 2 and 3

6. If $\|U^{(r)} - U^{(r+1)}\| < \varepsilon$ then stop else repeat from step 4.

Usually, for all experimental purpose, m is considered to be 2 and ε to be 0.02.

SPATIAL KERNEL FUZZY C MEANS (sKFCM)

Chuang. et al[2]developed spatial Fuzzy C means(sFCM) which incorporated spatial data of the image.It was a modification to the conventional FCM. The spatial function is given by the summation of the membership values in the neighboring of every pixel under consideration. The advantages were that they yielded more homogeneous clusters, diminished spurios blobs and boisterous spots and is insensitive to noise when compared to other systems. On comparable lines, spatial IFCM was additionally developed by Tripathy et.al [8] by presenting the intuitionsitic feature.

Kernel functions add an added advantage to clustering[9,10].In general, when two beside pixels are considered, correlation between them is relatively high. Since the neighboring pixels share similar intensity, the probability of them getting grouped into a same cluster is extremely high. The spatial FCM algorithm exploits this criteria. A spatial function is defined as:

$$h_{ij} = \sum_{k \in NB(x_j)} u_{ik} \quad (8)$$

where $NB(x_j)$ gives neighborhood pixels of x_j. A mask of 5x5 which is equally weighted is used having pixel x_j as it's center. Spatial function h_{ij} portrays the likeliness degree of x_j is in i^{th} cluster. The spatial function values is usually high if most of the pixels in the neighborhood of a particular pixel belong to the same cluster. It is included in the membership function as:

$$u'_{ij} = \frac{u_{ij}^{p} h_{ij}^{q}}{\sum_{k=1}^{c} u_{kj}^{p} h_{kj}^{q}} \quad (9)$$

Here p and q denote the relative weightage of the initial membership and the spatial function respectively. The spatial kernel FCM with parameters p and q is denoted by $sKFCM_{p,q}$. In noisy image conditions, the spatial function reduces the number of misclassified pixels by taking the neighboring pixels into account.

IIOAB JOURNAL ISSN: 0976-3104

The sKFCM clustering algorithm has two-steps. For every iteration, the conventional FCM algorithm is followed in the first step. Here the distance formula being used is radial-based kernel distance. Later, the centroid and the membership functions are updated. In the second step, the spatial function h_{ij} is calculated and then the new membership function (6) is computed.

RESULTS

A 225x225 dimension brain MRI image has been considered for implementation of sKFCM. We have considered the number of clusters, c=3. **Figure- 1** denotes the original image and **figure- 2** is the image induced with speckle noise with mean 0 and variance 0.04.

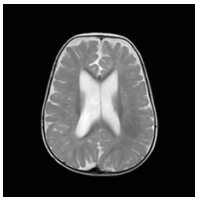


Fig:1. Original MRI image **Fig:2. MRI image with speckle noise**

V_{pc} and V_{pe} indicate fuzzy partition coefficient and partition entropy respectively. Maximum V_{pc} and minimum V_{pe} indicate better clustering[2]. DB and D indices are used to measure the cluster quality[11,12].Higher V_{pc} and lower V_{pe} indicate a good clustering. FCM, sFCM and sKFCM have been applied to both the images. V_{pc} and V_{pe} is given by

$$V_{pc} = \frac{\sum_j^N \sum_i^c u_{ij}^2}{N} \quad (10)$$

and

$$V_{pe} = -\frac{\sum_j^N \sum_i^c [u_{ij} \log u_{ij}]}{N} (11)$$

DAVIS-BOULDIN (DB) INDEX

The DB index is defined as the ratio of sum of within-cluster distance to between-cluster distance. It is formulated as given.

$$DB = \frac{1}{c}\sum_{i=1}^{c} \max_{k \neq i}\left\{\frac{S(v_i)+S(v_k)}{d(v_i,v_k)}\right\} \quad \text{for } 1 < i,k < c \ (12)$$

This index tries to minimize the within cluster distance and maximize the between cluster separation. Therefore a good clustering procedure should give value of DB index as minimum as possible[11].

DUNN(D) INDEX

Similar to the DB index the DD index is used for the identification of clusters that are compact and separated. It is computed by using

$$Dunn = \min_i\left\{\min_{k \neq i}\left\{\frac{d(v_i,v_k)}{\max_l S(v_l)}\right\}\right\} \quad \text{for } 1 < k,i,l < c \ (13)$$

This tries in maximizing the between-cluster distance and minimizing the within-cluster distance. Hence a larger value for the D index proves clustering to be more efficient[13].

The results of the validity measures on original image are shown in **Table- 1**.

THE IIOAB JOURNAL ISSN: 0976-3104

Table:1.Cluster Evaluation Results On Normal Image

METHOD	RESULTS ON ORIGINAL IMAGE			
	V_{pc}	V_{pe}	DB	D
FCM	0.7107	1.5350×10^{-4}	0.2581	5.0521
$sFCM_{1,1}$	0.7151	3.4602×10^{-09}	0.2553	5.3562
$sFCM_{2,1}$	0.7191	1.7579×10^{-13}	0.2603	5.1039
$sFCM_{1,2}$	0.7159	6.4532×10^{-14}	0.2592	5.399
$sKFCM_{1,1}$	0.727	5.2001×10^{-32}	0.2492	5.6926
$sKFCM_{2,1}$	0.7013	4.3964×10^{-43}	0.2537	5.3698
$SKFCM_{1,2}$	0.7276	4.7799×10^{-48}	0.2501	5.7399

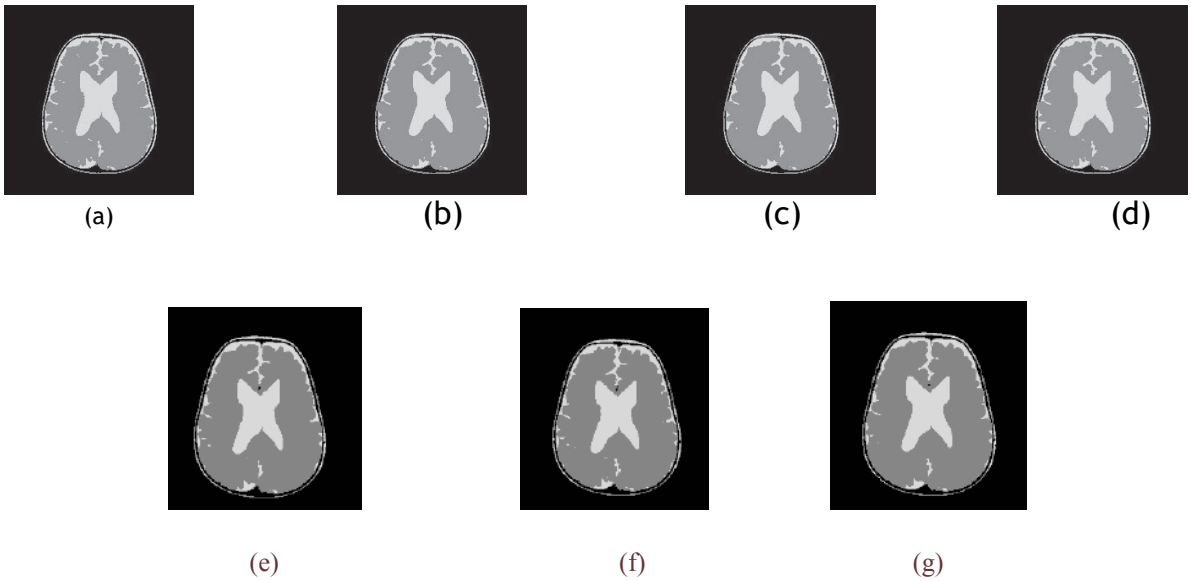


Fig: 3. Image segmentation of original image. (a) FCM. (b) $sFCM_{1,1}$(c)$FCM_{1,2}$.(d)$sFCM_{2,1}$.(e)$sKFCM_{1,1}$.(f)$sKFCM_{1,2}$.(g)$sKFCM_{2,1}$

From the above table and images, it can be seen that sKFCM has better partition coefficient and also possesses less partition entropy. sKFCM also has lower DB and higher D value when compared to conventional FCM and sFCM, thereby, sKFCM provides better clustering.

In the scenario of image with noise, the results are proved to be much better. Conventional FCM does not cluster the image to the expected level in presence of noise. Hence, leads to misclassification. The table below shows the performance of the sKFCM with other techniques implemented on the noisy image.

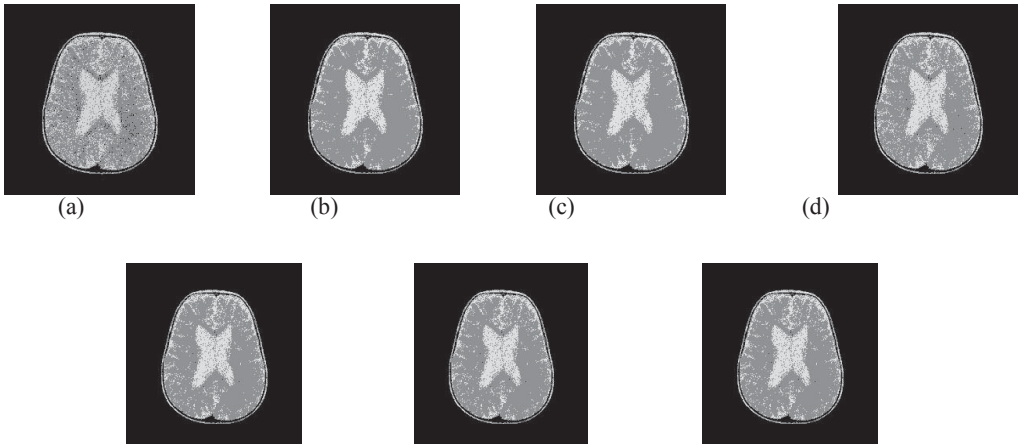
Table: 2. Cluster Evaluation Results On Noisy Image

METHOD	RESULTS ON NOISY IMAGE			
	V_{pc}	V_{pe}	DB	D
FCM	0.6975	2.8195×10^{-4}	0.4517	3.4183
$sFCM_{1,1}$	0.7101	5.9541×10^{-9}	0.4239	3.6734
$sFCM_{2,1}$	0.6922	7.7585×10^{-12}	0.4326	3.4607

THE IIOAB JOURNAL ISSN: 0976-3104

$sFCM_{1,2}$	0.6874	$4.2711x10^{-12}$	0.4412	3.6144
$sKFCM_{1,1}$	0.7432	$1.3488x10^{-25}$	0.3708	4.0343
$sKFCM_{2,1}$	0.7309	$3.7309x10^{-31}$	0.3839	4.1718
$sKFCM_{1,2}$	0.7086	$7.3218x10^{-37}$	0.3832	3.9454

From the **Table- 2** and **Figure- 4**, it can be seen that sKFCM produces better results. For noisy image, sKFCM overpowers FCM and all other forms of sFCM.sKFCM reduces the number of spurious spots and blobs to a large extent. It produces a segmented image with a good homogeneity. Smoother segmentation is achieved by taking a high value of *q*.But disadvantage is that, it may blur some of the finer details. The below figures show the segmentation results of sKFCM on the image induced with specklenoise.



(a) (b) (c) (d)

(e) (f) (g)

Fig: 4.Image segmentation of noisy image (a)FCM(b)$sFCM_{1,1}$(c)$sFCM_{1,2}$(d)$sFCM_{2,1}$(e)$sKFCM_{1,1}$(f)$sKFCM_{1,2}$(g)$sKFCM_{2,1}$

CONCLUSION

The proposed method adds a kernel approach to the conventional sFCM algorithm. It can be seen that the results of image segmentation and clustering by this approach has brought in better results. The Euclidean distance fails when clustering is to be done where distance is the major parameter. It is when kernel distance provides better results. Proposed method provides a novel way of clustering. The other kernel techniques like Gaussian and Hyper Tangent were also applied during the study. But only radial basis kernel's results were significantly different. However, the reason for unsatisfactory results of Gaussian and hyper tangent kernels is an area open for study. Likeways, this area calls research for developing hybrid clustering algorithms based on uncertainty.

FINANCIAL DISCLOSURE

No financial support was received to carry out this research.

ACKNOWLEDGEMENT

None.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

THE IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] EH Ruspini.[1969] A new approach to clustering, *Information and control*, 15(1): 22-32.

[2] Eman KS Chuang et al.[2006] Fuzzy c-means clustering with spatial information for image segmentation, *Computerized Medical Imaging and Graphics*, 30.

[3] P Swarnalatha, BK Tripathy, PL Nithin and D Ghosh.[2014] Cluster Analysis Using Hybrid Soft Computing Techniques, CNC- 2014International Conference of Network and Power Engineering ,Proceedings of Fifth CNC-2014, 516-524.

[4] BKTripathy, A.Ghosh, GK Panda.[2012] Kernel Based K-Means Clustering Using Rough Set, Proceedings of 2012 International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, (2012), Coimbatore, INDIA, pp.1 -5.

[5] B K Tripathy, A Tripathy, K Govindarajulu, R Bhargav.[2014] On Kernel Based Rough Intuitionistic Fuzzy C-means Algorithm and a Comparative Analysis. In Advanced Computing, *Networking and Informatics*- 1: 349-359). Springer International Publishing.

[6] R Bhargav, BK Tripathy, A Tripathy, R Dhull, E Verma, and P Swarnalatha.[2013] Rough Intuitionistic Fuzzy C-Means Algorithm and a Comparative Analysis, ACM conference, Compute 2013, (2013), ACM 978-1-4503-2545-5/13/08.

[7] S Mitra, H Banka, and W Pedrycz. [2006] Rough-Fuzzy Collaborative Clustering, IEEE Transactions on System, Man, and Cybernetics, Part B: *Cybernetics*, 36(4):795-805.

[8] Tripathy BK, Avik Basu, and Sahil Govel.[2014]] Image segmentation using spatial intuitionistic fuzzy C means clustering." Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE,

[9] BKTripathy and P Swarnalatha.[2014] A Comparative Study of RIFCM with Other Related Algorithms from Their Suitability in Analysis of Satellite Images using Other Supporting Techniques, Kybernetes,.43(1):53-81.

[10] B.K.Tripathy and R. Bhargav: Kernel Based Rough-Fuzzy C-Means, PReMI, ISI Calcutta, December, LNCS 8251, (2013), pp.148-157.

[11] D L Davis, and DW Bouldin. [1979] A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.PAMI-1,.2:224 – 227.

[12] D Zhang, and S Chen.[2002]Fuzzy Clustering Using Kernel Method, Proceedings of the international conference on control and automation, Xiamen, China, 123 – 127

[13] JC Dunn.[1973] A fuzzy relative of the ISODATA , Process and its use in detecting compact well-separated clusters, 32-57.

ABOUT AUTHORS

Deepthi P Hudedagaddi is pursuing her Masters at Vellore Institute of Technology. She is working on fuzzy clustering techniques on spatial data.

Dr. B K Tripathy, a triple gold medalist, is a senior professor in VIT University. He has supervised 19 PhD s, 13 M. Phil s and 02 M.S degrees. He is a senior member of IEEE, ACM, ACEEE and CSI. and is associated with over 60 international journals,published 320 articles, two research volumes and two books.He is working on Rough sets, Fuzzy sets, Social networks, Data mining, Soft Computing, E-Learning, Granular computing, Multi criteria decision making, Neighbourhood systems, SIoT and Soft Sets.

ISSN: 0976-3104
SPECIALISSUE (SCMDSA)
Aarthy and Prabu

ARTICLE OPEN ACCESS

A COMPUTERIZED APPROACH ON BREAST CANCER DETECTION AND CLASSIFICATION

SL. Aarthy[1*], S. Prabu [2]

[1]*School of Information Technology and Engineering, VIT University Chennai, Tamil Nadu, INDIA*
[2]*School of Computing Science and Engineering, VIT University Chennai, Tamil Nadu, INDIA*

ABSTRACT

Breast Cancer has become the world largest disease among women. Cancer can be prevented or cured only when it has been detected at the earliest that can save many lives. Digital mammograms are the one and only digital image which can be used with image processing techniques to detect breast cancer at the earliest. The various image processing techniques have been applied to images and the best features are extracted from the image which is further classified into benign and malignant. The classifier is used to find out the different stages in the cancerous patients. The CAD system developed with various image processing technique will assist the radiologist for further investigation of the affected person. The CAD system will easily identify breast malignancy and group all the malignancy which will help the radiologist to proceed with biopsy and chemotherapy so that malignancy is not in the final stage of the affected person. The CAD system has four different stages like preprocessing, segmentation, feature extraction and classification. The classifier used will be a binary classifier stating the class to be benign or malignant. In this paper, all the basic image processing stages have been applied on the digital mammogram and it's further classified to be cancerous or non-cancerous.
.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26th – March-2016
Published on: 20th–May-2016

KEY WORDS

Mass, feature extraction, segmentation, malignant, benign.

***Corresponding author: Email:** aarthy.sl@vit.ac.in **Tel:**+91-9600033077

INTRODUCTION

Breast cancer is the common cancer which affects many women. Breast cancer is developed from the breast tissues. The breast cancer can be seen from the lumps in the breast. Breast cancer is divided into four different stages like micro classification, masses, architectural distortion and bilateral asymmetry. Some of the imaging modalities like magnetic resonance image (MRI), Ultrasound and digital mammograms can be used to capture the breast of the suspected person .Breast cancer need not affect only women it can also present in some man also so the screening need to be done for both but the higher percentage of cancer get affected is women so the world is mostly addressing the women health care. When comparing with all the image modalities digital mammogram is inspected to improve the specificity and sensitivity of the breast cancer diagnosis.The cancer which is from the milk ducts lining is called ductal carcinomas and the one which is from the lobules are called lobular carcinomas. The breast cancer which undergoes the diagnosis process will be verified by taking biopsy of the concerning lumps. The classification which is done will be tested and verified.

A main objective of the pre-processing is a development of image data that suppresses unwanted distortions or enhances some important image features for further processing. The image pre-processing methods are categorized according to the size of the pixel neighborhood. The pre-processing images removes low-frequency background noise, normalizing the intensity of the individual particles images, reflection and masking portions of the images. A feature extraction accurately simplifies the volume of resources needed to define a large set of data. It is used in the image processing area that detects and isolates different desired portion or shapes of a digitized image. The features such as shape, texture, color, etc. are used to define the image content. These features can be further divided into pixel-level features, local features and global features. The image features are classified into primitives. The color feature is a majority used visual feature. The main advantages of the color features are robustness, effectiveness, implementation simplicity, computational simplicity and low storage requirements. In content-based image retrieval, the images are automatically indexed by generating a feature vector for describing the image content. The similarity

THE IIOAB JOURNAL ISSN: 0976-3104

of the feature vectors of the query and database images is evaluated to retrieve the image. The image content feature extraction technology proves the professional applications in industrial automation, biomedicine, social security, biometric authentication and crime prevention.

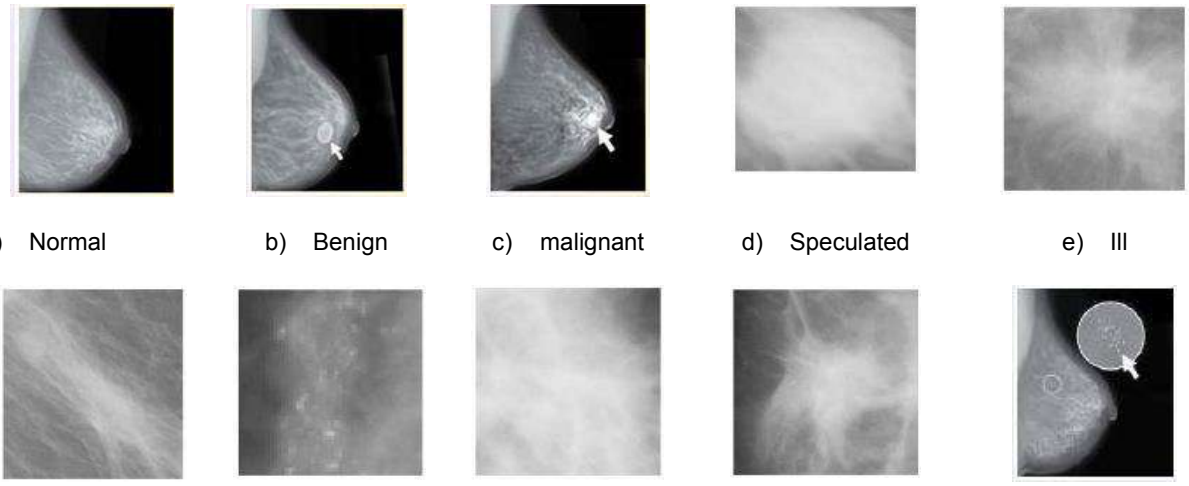


Fig: 1. Types of breast Image

This paper is organized as follows. Section 2 describes recent work in cancer detection Section 3 describe the General Methodology of CAD which explains, pre-processing, Feature Extraction, Image Classification, Image Segmentation and Feature Selection techniques. Section 4 describe the proposed methodology which uses median filter and histogram equalization for preprocessing, GLCM and PSO for Feature extraction and Selection, Selected features are classified as normal or abnormal using SVM followed by result discussion. Conclusion and feature work in Section 5

LITERATURE SURVEY

The median filter is a non-linear filtering technique used for smoothing. It preserve the edges of the image so as to remove speckle noise and salt and pepper noise in the image [1]. This filtering technique is needed to remove the effect of poor contrast due to Glare, noise and effects. This poor contrast is caused by poor lighting conditions during capturing of the image. The pixel value is replaced with a median pixel value for generating a low frequency image [2]. The histogram equalization technique is simple and effective types of well-known image enhancement technique for image contrast enhancement. A conventional histogram equalization is mainly applied to each sub-histogram [3]. In the output image histogram is authorized to enclose within the allocated gray levels range. In the image processing, the histogram of an image obviously mention the pixel intensity values of the histogram. This histogram equalization is mainly used in the prospect of the image enhancement. Compared with the existing techniques this technique has very less amount of calculations with lesser complexity. This technique is not only efficient in image enhancement and also produces rapid computations [4]. The GLCM is a famous statistical method of extracting the second order statistical textural features from the image [5]. This GLCM is adapted with a Gabor filter that has been used in many applications such as image identification and image representation [6] In this technique, a wavelet provides multi-scale representation of the image. These wavelets evaluate the contrast levels of the image according to the various resolution levels. This GLCM have important features such as Co-occurrence and Covariance. Particle Swarm Optimization (PSO) is swarm-intelligence-based, approximate, non-deterministic optimization techniques. This PSO algorithm maintains multiple potential solutions at one time. A population based

THE IIOAB JOURNAL
ISSN: 0976-3104

evolutionary algorithm is inspired in the social behavior [7].SVM classifier is mainly used for classification and regression analysis. This classification has been performed well as a computer-aided diagnostic classification mechanism for breast cancer screening in the mammography [8] The SVM is act as a mechanism for designing a SVM classifier to separate the image into normal and abnormal images. After the separation, the abnormal image is taken for further process i.e. segmentation. An ability of the Minkowski functionals is used to detect the changes in the image heterogeneity post-treatment are tested by using this classifier [9].

GENERIC METHODOLOGY OF MAMMOGRAM (CAD) SCHEME

This overview concentrates on outlining the methodologies for breast cancer detection. By and large, the Digital Mammogram CAD frameworks for breast cancer identification include four stages as indicated in **Figure 2**

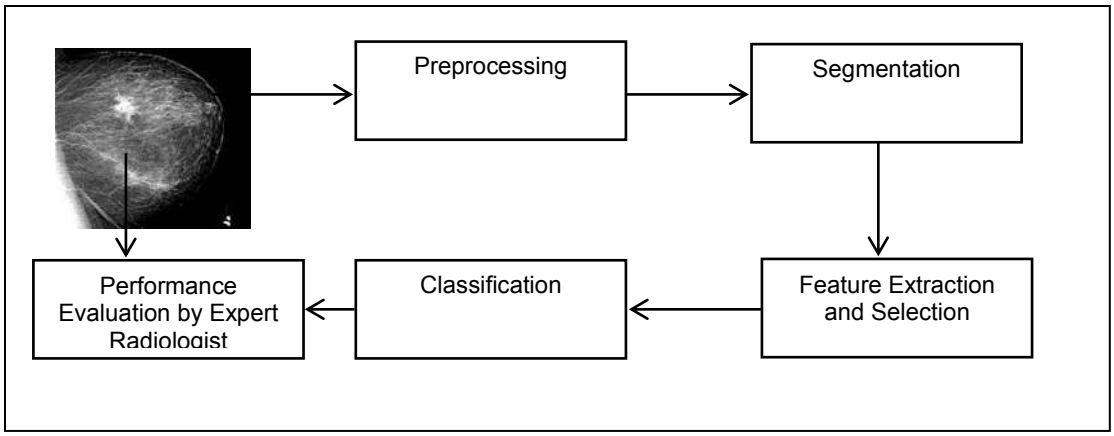


Fig: 2. Typical Methodology of Computer Aided Diagnosis System for Breast Cancer Detection

- *Image Pre-processing* –The purpose of pre-processing is to enhance the visual appearance of the images. Image pre-processing will increase the reliability of an image. Filters are used to remove the background noise which supress the quality of the image. The image which is pre-processed will be cleared from distortion and can be used as an input for the next image processing stage.
- *Image Segmentation* – Segmentation are needed to improve the analysis of an image when there is no correspondence between the pixels of an image type of tissues. This generally separates the image into various segments based on the region of interest (ROI).
- Feature Extraction and Selection – From the segmented image the features are extracted from each image which is trained. Among the features extracted the best few features have been selected and given to the classifiers.
- *Classification* – The extracted features along with his class will be given as the input to the binary classifier which finally says benign or malignant for our tested image.

All the above techniques will also be applied to the test image and the result of the image will be addressed by the classifier by stating cancerous or non-cancerous.

Pre - Processing

The main task of preprocessing an image is to suppress the unwanted distortion from the original image. It is also used to enhance some image features for better clarity in detecting the abnormalities. This section represents some preprocessing techniques applied for mammogram. One of the best techniques for preprocessing is median filter and histogram equalization for image contrast enhancement.

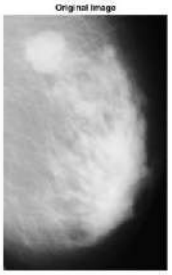
Median Filter

The median filter is one type of smoothing technique and it is also known as nonlinear digital filtering technique. The main idea behind using median filter is to preserve the edges of the images so as o remove speckle noise and

THE IIOAB JOURNAL
ISSN: 0976-3104

salt and pepper noise in the image [10]. When median filtering technique is used, the edges of the images are preserved and it is not blurred since it is not a linear filter.

$$f\hat{}(x,y) = median\{g(s,t)$$
$$where$$
$$(s,t) - belongs\ to\ S_{xy} \qquad [1]$$



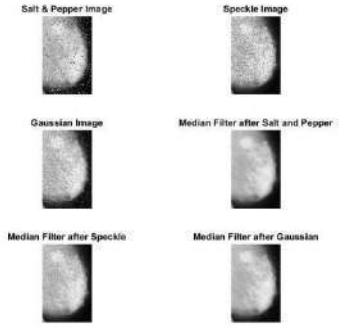


Fig: 3a. Original Image

Fig: 3b. Median Filter

In the median filtering operation, the pixel values in the neighborhood window are ranked according to intensity and the middle value (median) becomes the output value for the pixel.

The advantages of median filtering techniques are

- Easy to understand
- The brightness difference in minimal blurring of regional boundaries is preserved.
- The positions of the boundaries are preserved.

Histogram Equalization Technique

It is a techniquefor adjusting the intensity of an image.It is a simple and effective image enhancement technique for image contrast enhancement. It represents the pixel intensity values of the histogram. Compared with other existing techniques, this technique has very less complexity. It is an efficient image enhancement algorithm and also produces rapid computations. This technique improves the contrast characteristics by mapping the pixels from the histogram. It is used to enhance the de-noised image [11].

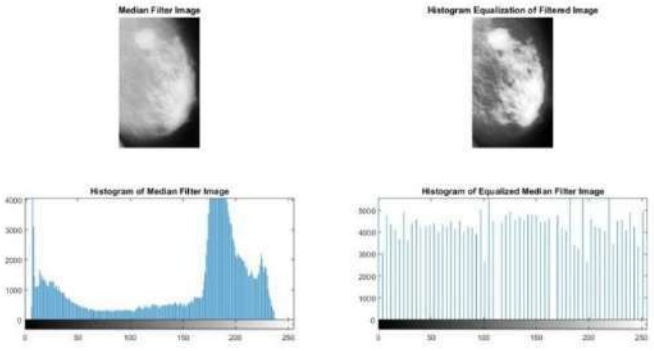


Fig :4.Histogram Equalized Image

THE IIOAB JOURNAL ISSN: 0976-3104

Segmentation

Image Segmentation is a basic and fundamental segment and is a standout amongst the most troublesome patterns in image transforming and example acknowledgment, and decides the nature of the last examination. Division is a part of the picture I into non covering districts.

$$UI_i = I \text{ and } I_i \cap I_j = \emptyset \; i \neq j \quad [2]$$

Computer-aided diagnosis framework will help radiologists in perusing and deciphering sonography. The objective for the division is to find the suspicious regions to support radiologists in conclusions.

Histogram Thresholding

Histogram thresholding is one of the generally utilized methods for monochrome picture division. Histogram thresholding was proposed for fragmenting breast ultrasound pictures. The calculations proposed for dividing masses in US pictures included the accompanying steps: (1) pre-processing utilizing editing and middle separating, (2) duplicating the pre-processed picture with a Gaussian oblige capacity, (3) deciding the potential sore edges through dim quality thresholding [12], and (4) augmenting an utility capacity for potential injury edges. Then again, the middle, width and tallness of the sores expected to be chosen physically or semi-physically.Another thresholding calculation had four stages: First, the regions of interest (ROIs) were pre-processed with a 4×4 middle channel to decrease the dot clamor and to improve the highlights. Second, a 3×3 unshar channel was built utilizing the negative of a two-dimensional Laplacian channel to underline the components with significant sign level and to upgrade the difference in the middle of article and foundation. Third, the ROIs were changed over to a twofold picture by thresholding. The limit was dictated by the histogram of ROIs. On the off chance that a valley of histogram somewhere around 33% and 66% of the pixel populace could be discovered, this force quality was chosen as the edge. In the event that there was no such valley in that range, the force of 50% pixel populace was chosen as the limit esteem. At long last, the chose knob's limit pixels were gotten utilizing morphologic operation.

Feature Extraction and Selection

Highlight extraction and determination are imperative ventures in breast disease identification and order. An ideal list of capabilities ought to have successful and separating highlights, while basically decrease the excess of highlight space to evade "condemnation of dimensionality" issue. The "condemnation of dimensionality" recommends that the testing thickness of the preparation information is so low it couldn't be possible guarantee a significant estimation of a high dimensional characterization capacity with the accessible limited number of preparing information. For some best in class arrangement routines, for example, counterfeit neural system and bolster vector machine, the measurement of highlight vectors not just exceedingly influences the execution of the characterization, additionally decides the preparation time of the calculation. Along these lines, how to concentrate helpful highlights and make a decent determination of the highlights is a pivotal assignment for CAD frameworks.
The highlights of breast US pictures can be separated into four classes: composition, morphologic, model-based and descriptor highlights. We condense and list the common and adequacy demonstrated highlights in Table 4. Unquestionably, one can't utilize every one of them in the meantime. Extraction and determination of compelling highlights is a fundamental step. The general rules for selecting huge highlights principally incorporate four contemplations: segregation, dependability, freedom and optimality. On the other hand, basically consolidating the best performed highlights won't most likely make the frameworks function admirably and successfully. The objective of highlight extraction and choice is to expand the separating execution of the highlight bunch.

Texture Features

A large portion of the surface highlights are computed from the whole picture or ROIs utilizing the dim level qualities. FT1 (auto-covariance coefficient) is an essential and conventional surface highlight which can mirror the internal pixel relationship inside a picture. FT2 (Block difference of inverse probabilities) – FT3 (Block variation of local correlation coefficients) measure the variety of intensities and surface smoothness, individually. The higher estimation of BDIP is, the bigger the change of intensities in a square is, and the bigger Block variation of local correlation coefficients quality shows that the fixings in the piece are harsh. Both the first and second request of FT2 and FT3 can be utilized as the highlights as well. FT4 is characterized as the proportion of the difference, auto-relationship coefficients or power normal inside the sore to that outside the injury. The bigger the proportion is, the bring down the likelihood of the cancer being dangerous is. FT5 is characterized as the summation of contrasts

THE IIOAB JOURNAL ISSN: 0976-3104

among the genuine appropriation of wavelet coefficients in every high-recurrence sub-band and dispersion of the normal Laplacian conveyance. This highlight can mirror the edge smoothness. FT6 is a request measurements based highlight vector removed from wavelet disintegration sub-groups. After 3rd level wavelet decay, the length (length=20) of request insights channel is picked taking into account Monte Carlo reenactment and Akaike's last expectation measure. Twenty mean qualities and 20 difference estimations of request insights parameters for the 12 wavelet coefficient groups were ascertained and framed 480-D highlight vectors. The measurement of the highlight vector was diminished from 480-D to 7-D by utilizing highlight examination. The stepwise highlight choice system or PCA could be a superior decision for lessening the highlight dimensionality. FT7 and FT8 are characterized as:

$$CON = \sum_{ij}(i-j)^2\, p(i,j) \quad and \quad COR = \sum \frac{ijp(i,j) - m_x m_y}{\sqrt{s_x^2 s_y^2}} \qquad [3]$$

where *p(i,j)* is the probability that two pixels with gray value *i* and gray value *j* are in a fixed distance apart, and

$$m_x = \sum_i i \sum_j p(i,j), m_y = \sum_j j \sum_i p(i,j) \qquad [4]$$

$$S_x^2 = \sum_i i \sum_j p(i,j) - m_x^2, S_y^2 = \sum_j j \sum_i p(i,j) - m_y^2 \qquad [5]$$

In view of comprehension of the back acoustic conduct or back shadow, distinctive numeric interpretations are proposed to compute FT12. Three ROIs were characterized whose width and profundity were the same as the ROI contains the injury itself. As Figure demonstrates, the post ROI speaks to the back locale of the injury and the privilege ROI and left ROI are nearby tissues at the same profundity of the post ROI. The thin clear limits are utilized to dodge the edge shadows. The skewness picture is sifted with an edge to get the location focuses, i.e., the shadow. The back shadow was characterized as the distinction between the dark scale histograms of the districts inside the sore and back to the injury. For the same normal for breast injuries, we can utilize distinctive approaches to characterize the numeric interpretations. To discover more exact and productive outflows ought to be one without bounds meets expectations. FT13 is the Boltzmann/ Gibbs entropy over the dark scale histogram with respect to the greatest entropy. The higher the entropy is, the more homogeneous the sore is. FT15 – FT16 are surely understood surface highlights which have as of now been all around characterized. On the other hand, they are not often utilized as a part of late US picture portrayal. This may be because of their high processing expense. The meaning of the fractal measurement (FT17) is like the Hausdorff measurement.

Classification

After the highlights have been extricated and chosen, they are info into a classifier to order the pictures into injury/ non-sore or amiable/ harmful classes [13]. Lion's share of the distributions concentrates on ordering threatening and considerate sores (normally called injury arrangement), and a portion of the articles concentrate on arranging injuries and non-injuries (typically called sore identification), and just a couple of them concentrate on both. Sore discovery is essential before injury grouping.

Linear Classifiers

Often utilized direct classifiers for breast malignancy location and grouping are straight discriminant examination and logistic relapse (LOGREG). The primary thought of LDA is to locate the direct blend of the highlights which best separate two or more classes of the information [14]. In the event that there are n classes, and LDA groups the information by the accompanying straight capacities:

$$f_i = \mu_i C^{-1} X_K^T - \frac{1}{2}\mu_i C^{-1} \mu_i^T \; In(P_i), 1 \le i \le n \qquad [6]$$

where

$$c = \frac{1}{N}\sum_{i=1}^{n} n_i C_i, \; P_i = \frac{n_i}{N} \qquad [7]$$

n_i is the quantity of tests in the *ith* class, N is the quantity of aggregate examples, i is the mean of class i, and C_i is the covariance lattice of class i. The above parameters can be acquired from the preparation information. At the point when another information x_k is in, it is doled out to class i with the most astounding f_i. Logistic relapse is a

THE IIOAB JOURNAL ISSN: 0976-3104

model for foreseeing the likelihood of an occasion happening as an element of different variables. The likelihood of $X=x_1, x_2,...,x_n$ is detailed as:

$$\text{logit}(P) = \log\frac{P}{1-P} = b0 + \sum_{i=1}^{n} b_i x_i \quad [8]$$

Where $b_0,...,b_n$ are model parameters which could be assessed from the preparation information. At the point when LOGREG is utilized to group two-class issue, for every highlight vector x_i, threshold=0.5 is utilized to choose which class X has a place with. LDA was connected to the information set of 400 cases with four naturally extricated highlights. The normal A_z under ROC bend was 0.87 more than eleven autonomous trials. LOGREG was utilized to focus the likelihood of danger in a database of 58 cases. Three edge based highlights were assessed and the territory under the ROC bend with the best highlight blend of age, edge echogenicity and rakish variety was 0.87±0.05.Here, we can see that the exhibitions of LDA and LOGREG are not high on the grounds that the classifiers are straight, and for nonlinear divisible information, the systems have natural breaking points.

Evaluation

We consider a few oftentimes utilized assessment criteria. A Receiver Operating Characteristic (ROC) curve is most commonly used because of its far reaching and reasonable assessment capacity. AROC bend is a plotting of genuine positive part (TPF) as a component of false positive portion (FPF). The range (A_z) under the ROC bend can be utilized as a paradigm. This demonstrates a case of ROC bend assessment of the execution of CAD frameworks utilizing three diverse information sets [15, 16].

$$overall\ accyracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Positive\ perdictive\ value\ (PPV) = \frac{TP}{TP+FP}$$

$$Negative\ predictive\ value\ (NPV) = \frac{TN}{TN+FN}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Where

TP is the quantity of genuine positives,

TN is the quantity of genuine negatives

FP is the quantity of false positives

FN is the quantity of false negatives.

The last recipe is Matthew's connection coefficient (MCC), which has from time to time been utilized for breast disease CAD execution assessment. In any case, MCC is a capable precision assessment foundation of machine learning routines. Particularly, when the quantity of negative examples and positive specimens are clearly unequal, MCC gives a superior assessment than general exactness. As more breast malignancy CAD frameworks utilized machine learning techniques, for example, SVM, ANN and BNN, MCC ought to be utilized as an extra assessment paradigm [17].

THE IIOAB JOURNAL
ISSN: 0976-3104

THE PROPOSED METHODOLOGY

Proposed Methodology works on mammogram image which has train set of images processed for extracting features which used for classifying features extracted from test image shown in Figure-5.

Image acquisition and Data set

The methodology presented in this work was tested on the complete mini-MIAS database. It is freely available for scientific purposes. The images of the database originate from a film-screen mammographic imaging process in the United Kingdom National Breast Screening Program. The algorithms where implemented based 90 trained image consist of 40 normal and 50 abnormal.

Preprocessing

We have seen that smoothing (low pass) filters reduce noise. However, the underlying assumption is that the neighboring pixels represent additional samples of the same value as the reference pixel, i.e. they represent the same feature. At edges, this is clearly not true, and blurring of features results. You have used convolution techniques to implement weighting kernels as a neighborhood function, which represented a linear process. There are also nonlinear neighborhood operations that can be performed for the purpose of noise reduction that can do a better job of preserving edges than simple smoothing filters. In the median filtering operation, the pixel values in the neighborhood window are ranked according to intensity, and the middle value (the median) becomes the output value for the pixel under evaluation.Median filtering does not shift boundaries, as can happen with conventional smoothing filters. Since the median is less sensitive than the mean to extreme values (outliers), those extreme values are more effectively removed. Median filtering preserves the edges.

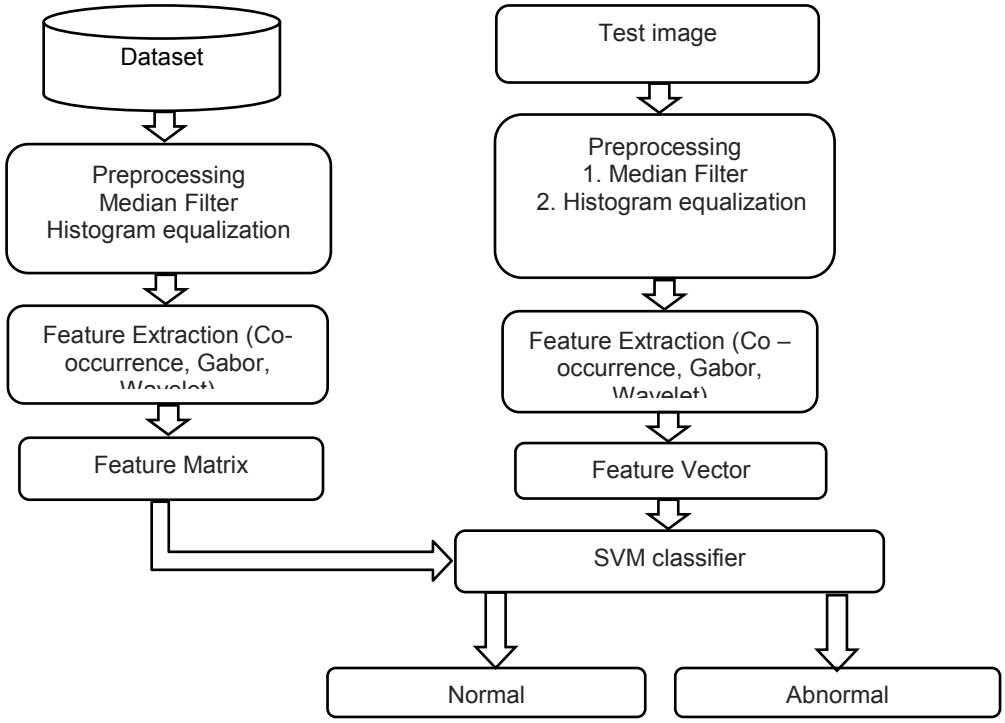
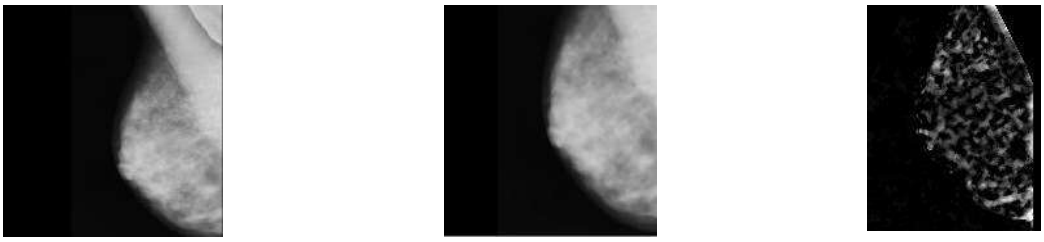


Fig: 5. proposed architecture

THE IIOAB JOURNAL ISSN: 0976-3104



Original Image | Median Filtered Image | Muscle removal

Fig: 6. Preprocessed Image

.....

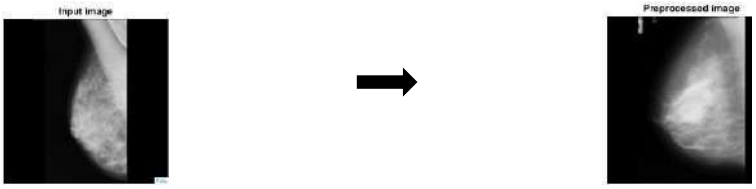
Feature Extraction and Feature Selection

Feature extraction is a process of capturing a visual content of images for originating and recovering. The primitive or low level image features are the extraction of color, texture and shape or domain specific features from the image. The feature selection is mainly used to select the needed features available in the image after the segmentation process is completed. A data in the feature selection contains multiple redundant or irrelevant features

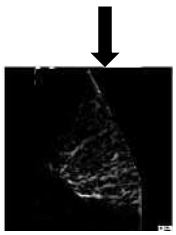
Table: 1. Extracted features of Mammogram image

Image Name	Auto Correlation	Correlation	Energy	Entropy	Homogeneity
mdb123	5.4707	9.7127	5.1009	1.3519	9.5548
mdb153	3.7491	9.6309	6.9143	8.6214	9.6629
mdb304	2.8537	9.4924	6.9211	8.3604	9.6986
mdb319	1.0021	9.6628	1.9198	2.1124	9.3583
mdb322	3.9797	9.6073	5.5252	1.1732	9.601

Classification



Original Image | Preprocessed Image



Auto Correlation 5.4707
Correlation 9.7127
Energy 5.1009
Entropy 1.3519
Homogeneity 9.5548

ISSN: 0976-3104

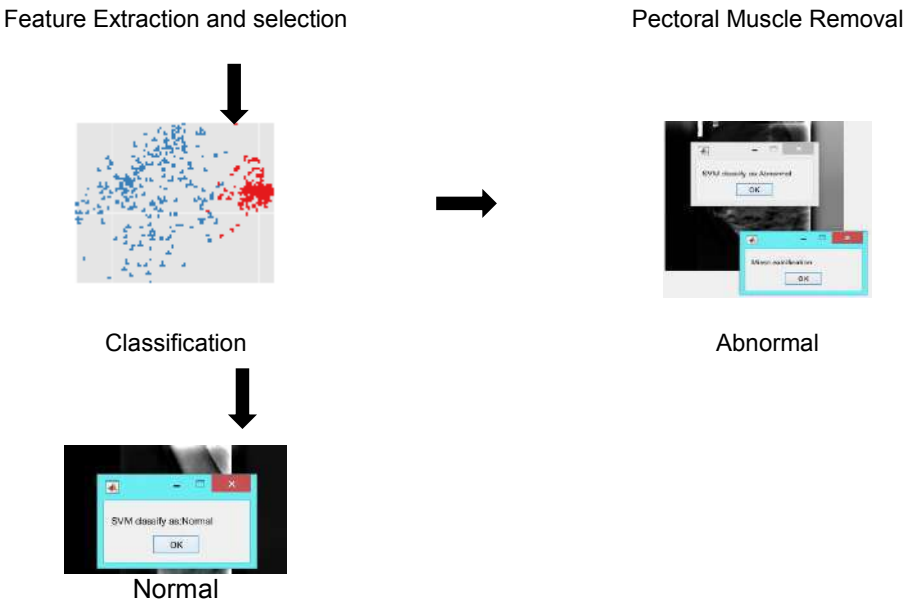


Fig: 7. Result Discussion of Proposed CAD Classification on mammogram image

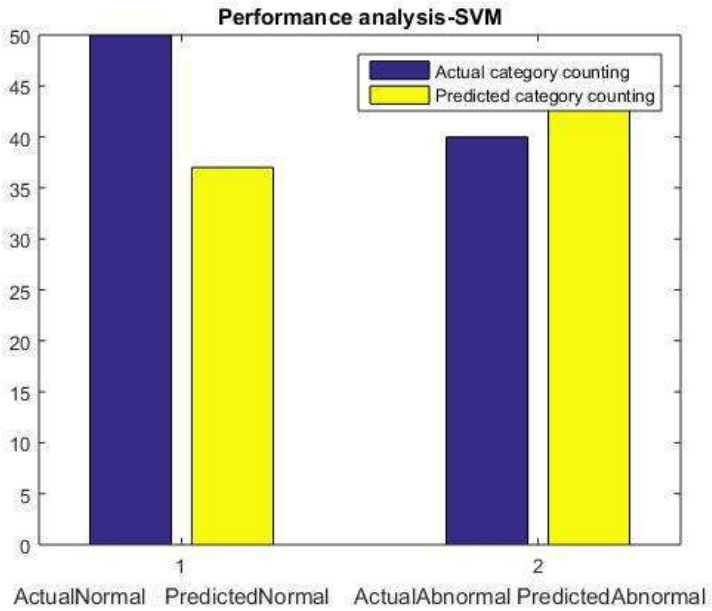


Fig: 8. Performance of SVM Classifier

Finally, extracted features are passed through SVM classifier. SVM is one of the best classifier. SVM is a binary classifier to classify whether the given mammogram image is normal or abnormal.

CONCLUSION

With the advancement in CADthrough mammogram attracts more attention for Brest cancer analysis. Mammogram is one of major research subjects in medical imaging and Cancer diagnosis system. With the survey we proposed a methodology for classifying of Mammogram image. The proposed technique first apply median filter and histogram equalization for noise removal and image enhancement, then employs Gabor filter co-occurrence matrix to extract features from Mammogram and by using the selected features SVM algorithm

THE IIOAB JOURNAL
ISSN: 0976-3104

classify as normal and abnormal. Result discussion on proposed method shows the robustness of proposed techniques. According to the experimental result proposed method propose classification accuracy of 93% with 94% sensitivity rate and 92% specificity rate. Proposed method classify input image as normal and abnormal feature work is to apply segmentation algorithm on abnormal image for extracting ROI and selection Features from FOI and classify using Feed Forward Neural Network classification Algorithm.

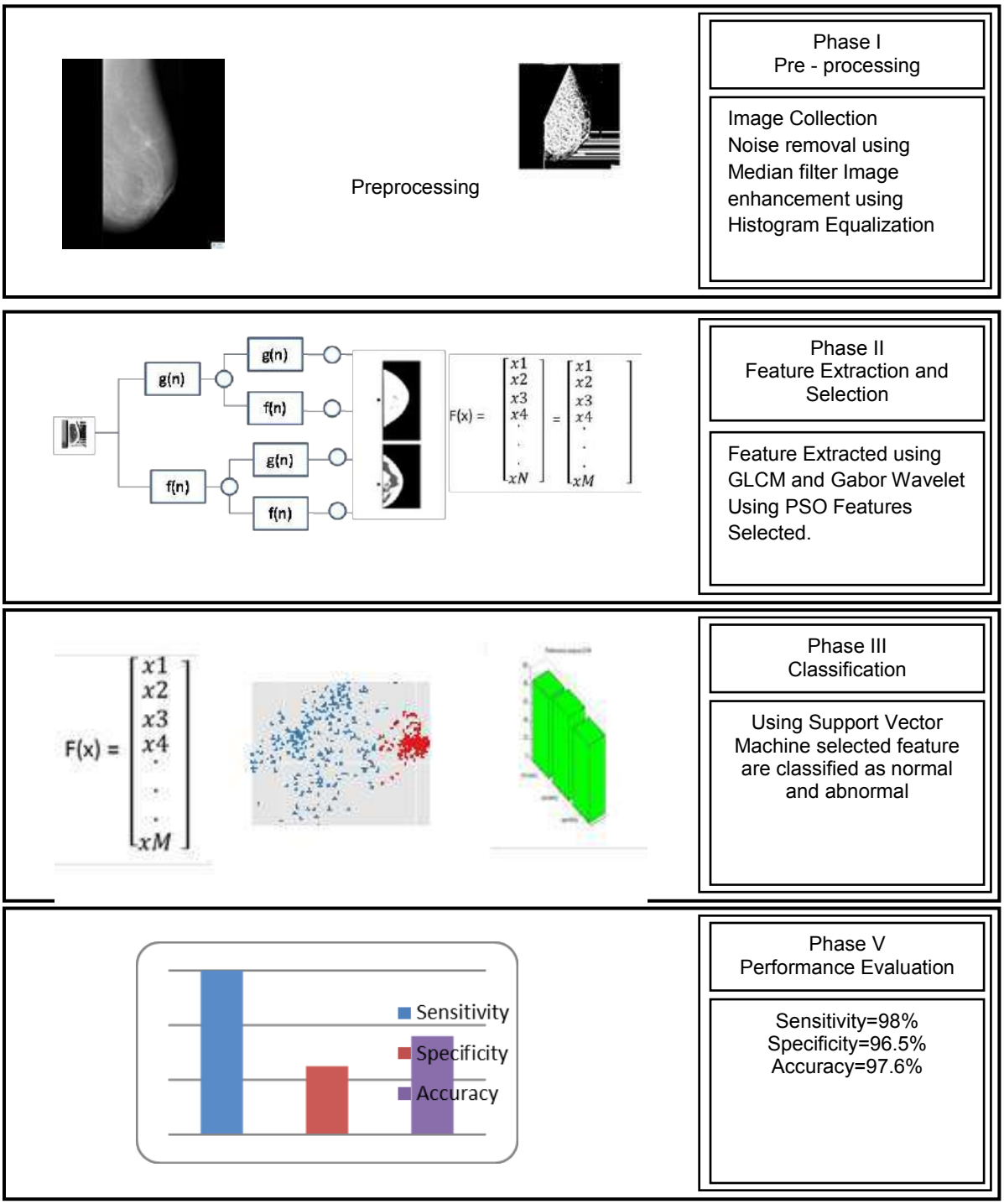


Fig: 9. Proposed Mrthodology of CAD

THE IIOAB JOURNAL
ISSN: 0976-3104

FINANCIAL DISCLOSURE
No financial support was received to carry out this project.

ACKNOWLEDGEMENT
This work is part of Ph. D Research work. It is not supported by any agency

CONFLICT OF INTERESTS
Authors declare no conflict of interest.

REFERENCES

[1] Sopharak A, Uyyanonvara,B, Barman S. [2009] Automatic exudate detection from non-dilated diabetic retinopathy retinal images using fuzzy c-means clustering. *Sensors*, *9*(3):2148-2161.

[2] Nivetha P, Manickavasagama MR. [2014] Lung Cancer Detection at Early Stage Using PET/CT Imaging Technique. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(3).

[3] Yu Z, Bajaj C. [2004] A fast and adaptive method for image contrast enhancement. In *Image Processing, 2004. ICIP'04. 2004 International Conference on* 2: 1001-1004 *IEEE*.

[4] Tuteja M, Kaur B, Gujral S. [2014] Quality Enhancement of Various Diagnosed Medical Images Using Different Signal Processing Methods. *International Journal of Emerging Trends in Science and Technology*, *1*(04).

[5] Mohanaiah P, Sathyanarayana P, GuruKumar L. [2013] Image texture feature extraction using GLCM approach. *International Journal of Scientific and Research Publications*, *3*(5): 1.

[6] Zheng Y. [2010] Breast cancer detection with Gabor features from digital mammograms. *algorithms*, *3*(1): 44-62.

[7] Alba E, García-Nieto J, Jourdan L, Talbi EG. [2007] Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In *2007 IEEE Congress on Evolutionary Computation* , 284-290.

[8] Choi JY, Kim D H, Plataniotis KN, Ro YM. [2014] Computer-aided detection (CAD) of breast masses in mammography: combined detection and ensemble classification. *Physics in medicine and biology*, *59*(14): 3697.

[9] Larkin TJ, Canuto HC, Kettunen MI, Booth TC, Hu DE, Krishnan AS, Brindle KM. [2014]Analysis of image heterogeneity using 2D Minkowski functionals detects tumor responses to treatment. *Magnetic resonance in medicine*, *71*(1): 402-410.

[10] Ergin S, Kilinc O.[2014] A new feature extraction framework based on wavelets for breast cancer diagnosis. *Computers in biology and medicine*, *51:* 171-182.

[11] Srivastava S, Sharma N, Singh SK, Srivastava R. [2013] Design, analysis and classifier evaluation for a CAD tool for breast cancer detection from digital mammograms. *International Journal of Biomedical Engineering and Technology*, *13*(3): 270-300.

[12] Purushotham S, Tripathy B. [2014] A comparative study of RIFCM with other related algorithms from their suitability in analysis of satellite images using other supporting techniques. *Kybernetes*, *43*(1): 53-81.

[13] Mohamed H, Mabrouk MS, Sharawy A. [2014] Computer aided detection system for micro calcifications in digital mammograms. *Computer methods and programs in biomedicine*, 116(3):226-235.

[14] Acharya UR, Ng EY K, Tan JH, Sree SV. [2012] Thermography based breast cancer detection using texture features and support vector machine. *Journal of medical systems*, 36(3): 1503-1510.

[15] Dheeba J, Singh NA, Selvi ST. [2014] Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of biomedical informatics*, 49: 45-52.

[16] Gøtzsche PC, Jørgensen KJ. [2013] Screening for breast cancer with mammography. *The Cochrane Library*.

[17] Dromain C, Boyer B, Ferre R, Canale S, Delaloge S, Balleyguier C.[2013] Computer-aided diagnosis (CAD) in the detection of breast cancer. *European journal of radiology*, 82(3), 417-423.

ABOUT AUTHORS

Prof. SL Aarthy *is working as Assistant Professor (Senior) in School of Information Technology and Engineering, VIT University, Vellore. Her research area includes Image processing, soft computing and data mining. She has published a good number of journal papers in her research filed. She is life member of CSI and IEEE. She is also part of various school activity committees.*

Dr. S Prabu *is working as Associate professor in the school of computing sciences and engineering, VIT University, at Vellore, India. He is the principal Investigator of Funded project from SAC-ISRO. He is life member of CSI and IEEE. He has published many technical papers in various international journals, conferences, and Springer book chapters. His research interest includes Image processing, Remote Sensing, Cloud Computing.*

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Kamalakannan and Babu

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

EARLY DETECTION OF BREAST CANCER USING GLCM FEATURE EXTRACTION IN MAMMOGRAMS

Kamalakannan J[1]* and Rajasekhara Babu. M[2]

[1]*School of Information Technology and Engineering, VIT University Vellore, Tamil Nadu, INDIA*

[2] *School of Computing Science and Engineering, VIT University, Tamil Nadu, INDIA*

ABSTRACT

Breast cancer is the one of the most common invasive cancer type among women. Early detection and diagnosis of breast cancer can be facilitating the chance of better treatment for the cancer affected people with mammography image analysis, since mammograms are cost effective and the world standard for screening of breast. Extracting the features from mammograms will help in identifying and classifying the breast abnormalities. There are many ways to extract the features; in this paper we have used GLCM to extract features from the mammographic images. GLCM is a statistical method of examining texture that uses the spatial relationship of pixels [6] . the features which are extracted can be given to a classifier to classify the abnormalities as benign and malignant. The mammograms from mini MIAS database is used for extracting the features. The radiologist uses the CAD system for differentiating benign and malignant abnormalities from the mammograms in a better way. The technique which is adapted in this paper can be helpful in improving the performance of the CAD system which can assist the radiologist for better diagnosis of breast cancer.

Received on: 30th-Nov-2015
Revised on: 25th-March-2016
Accepted on: 29th – March-2016
Published on: 23rd–May-2016

KEY WORDS

Mammogram; GLCM; Benign; Malignant; CAD; Screening; Feature

***Corresponding author: Email:** jkamalakannan@vit.ac.in **Tel:** +91-9944730423

INTRODUCTION

Today, Breast cancer is one of the common among cancer both men and women. Breast cancer is the most common invasive cancer in females worldwide. It accounts for 16% of all female cancers and 22.9% of invasive cancers in women.18.2% of all cancer deaths worldwide, including both males and females, are from breast cancer. According to the National Cancer Institute, 232,340 female breast cancers and 2240 male breast cancers are reported in the USA each year and 39,620 caused by the disease as well [1]. The breast cancer is the most affecting cancer in women compared to other types of cancer. The risks of the breast cancer increases with the factors such as female gender, obesity, lack of physical exercise , having children late or not at all etc .It has been found that the 80% of women are above age of 50[1]. Breast cancer can be easily diagnosed with various techniques. Imaging tests use x-rays, magnetic fields, sound waves, or radioactive substances to create pictures of the inside of your body. The process of examination of breast to identify the abnormality is called mammography. It is recommended that women of age 40 and older have regular mammogram to detect the breast cancer at early stage. The gold standard and cost effective way of screening the breast cancer is through mammograms [2].

Mammogram

A mammogram is one of the best radiographic methods to detect the breast cancer at early stage. It detects the tumors which are tiny and it is very difficult to identify by the radiologist . Mammography gives us the X-ray image as an output [3]. Image Processing techniques that provides a sufficient assessment to category the abnormalities[3] such as calcification(a),circumscribes masses (b),speculate masses(c),ill-defines masses (d),Architectural distortion(e), asymmetry (f) to make a clear diagnosis of the images[3]. The Current usage of early detection of breast cancer is done through mammography screening [4]. Mammogram is a medicinal practice for distinguish the breast growth which was initially coined by Bob Eagan in 1950. Mammogram is the radiology tool which gives better accuracy than clinical breast examination [4]. It not only identifies the abnormalities but also identify the normal breast among women [4]. This Detection strategy is termed as

THE IIOAB JOURNAL ISSN: 0976-3104

mammography, in which X-beams of low vitality will be anticipated on an emulsion film that gives a white washed duplicate which symbolizes the tissue in the bosom [4]. Basically, there are two sorts of perspectives in a mammography namely crania-caudal view (CC) and Mediolateral Oblique (MLO).Earlier view is normally recognized in both diagnostics test using mammogram and the clinical breast examination [4]. In this viewpoint, we can see maximum conceivable vicinity of a granular tissue, the adjoining greasy tissue and edge of the midsection divider muscle [4]. Later view is considered for the routine mammogram. Cumbersome region is additionally given by CC view than by MLO view which are shown in **Figure- 1**.

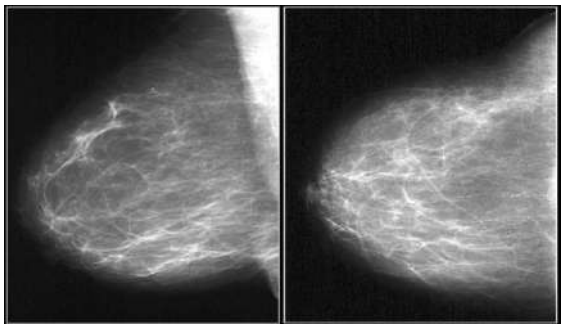


Fig: 1. MLO and CC views of the same breast

When we narrow down the perspective of mammogram, Later medial perspective are viewed from the outside towards the focal point whereas mediolateral perspective are viewed from the inside portion of breast [4]. There are different kinds of tumor may present in the mammogram. The tumor with speculated shape will be the cancerous tumor (Malignant) and the tumor with circular shape will be the noncancerous tumor (Benign). The masses with different shape and margin are depicted in the **Figure- 2**.

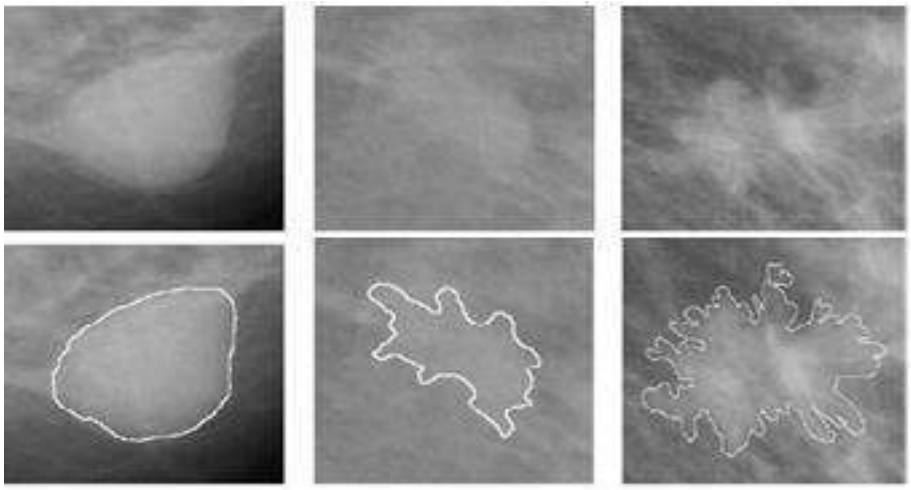


Fig: 2. Three mass examples with different shape and margin: (a) circular shape and circumscribed margin, (b) lobular shape and well defined margin, and (c) speculated shape and ill-defined margin. The last of the three has a higher malignancy probability.

The Grey Level Co-occurrence Matrix (GLCM) Features

Grey-Level Co-occurrence Matrix (GLCM) texture measurements is the one of the way to extract features for image texture since they were proposed by Haralick [5], and 14 statistical features were introduced. GLCM is a statistical method of examining texture that uses the spatial relationship of pixels [6]. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image. These features are generated by calculating the features for each one of the co-occurrence matrices obtained by using the directions 0°, 45°, 90°, and 135°[7] , then averaging these four

THE IIOAB JOURNAL ISSN: 0976-3104

values .The GLCM is a intensity change histogram as a function of distance and direction. It is an estimate of the second order joint probability[7], which is the probability of pixel going gray level i to gray level j with the given distance and direction.

The basic GLCM algorithm

1. Count all pairs of pixels in which the first pixel has a value i, and its matching pair displaced from the first pixel by 'd' has a value of j.
2. This count is entered in the ith row and jth column of the matrix Pd[i,j]
3. Note that Pd[i,j] is not symmetric, since the number of pairs of pixels having gray levels[i,j]does not necessarily equal the number of pixel pairs having gray levels [j,i].
4.The elements of Pd[i,j]can be normalized by dividing each entry by the total number of pixel pairs.
5. Normalized GLCM N[i,j], defined by:

$$N[i,j] = \frac{\quad}{\sum_{i}\sum_{j} P[i,j]} \qquad [1]$$

For a window of size wxw, we get one GLCM matrix and the dimension of the co-occurrence matrix is GxG. If we have G gray-levels in the image. Considering distance d and a direction θ Check all pixel pairs with distance d and direction inside the window. Q(i,j|d,θ) is the number of pixel pairs where pixel 1 in the pair has pixel value i and pixel 2 has pixel value j. This has been illustrated in the **Figure- 3(a) and 3(b).**

0	1	1	2	3
0	0	2	3	3
0	1	2	2	3
1	2	3	2	2
2	2	3	3	2

Level j

Gray

1	2	1	0
0	1	3	0
0	0	3	5
0	0	2	2

Gray level i

$Q(i,j|d,\theta)$.

Fig: 3(a). Image **Fig: 3(b): Pixel Pairs**

GLCM features can be used directly to measure statistical measures between the pixels. GLCM extracts the structural information about the texture pattern [8] which has to be analyzed at different orientation and scale. The features which are extracted from GLCM are listed in the **Table- 1**.

Table: 1 . List of GLCM Features

Feature Number	Feature Name	Feature Number	Feature Name
f1	Angular Second Moment (Energy)	f11 [9]	Difference Entropy [9]
f2	Contrast	f12 [9]	Information Measure of Correlation 1[9]
f3	Correlation	f13 [9]	Information Measure of Correlation 2[9]
f4	Sum of Squares: Variance [9]	f14	Autocorrelation
f5	Inverse Difference Moment (Homogeneity) [9]	f15	Dissimilarity
f6	Sum Average [9]	f16	Cluster Shade
f7	Sum Variance [9]	f17	Cluster Prominence
f8	Sum Entropy [9]	f18	Maximum Probability
f9	Entropy	f19	Inverse Difference Normalized
f10	Difference Variance	f20	Inverse Difference Moment Normalized

THE IIOAB JOURNAL ISSN: 0976-3104

Features f1-f13 are features proposed by Haralick [9], Soh proposed features f14-f18 [10] and features f19 and f20 are proposed by Clausi [9].The formula for calculating the some of the features are given below

Energy (angular second moment (asm)

$$f1 = \sum_{i,j=0}^{N-1} P_{i,j}^2 \quad [2]$$

Contrast

$$f2 = \sum_{i,j=0}^{N-1} P(i,j) * (i-j)^2 \quad [3]$$

Inverse Difference Moment (IDM) / Homogeneity.

$$f5 = \sum_{i,j=0}^{N-1} \frac{P(i,j)}{1+(i-j)^2} \quad [4]$$

Entropy

$$f9 = \sum_{i,j=0}^{N-1} P(i,j) * [-\ln(P(i,j))] \quad [5]$$

Dissimilarity

$$f15 = \sum_{i,j=0}^{N-1} P(i,j) * |(i-j)| \quad [6]$$

Maximum Probability

$$f18 = \max(i,j) P(i,j) \quad [7]$$

BI-RADS (Breast Imaging Reporting and Data System)

This is developed by American college of Radiology which provides standards for mammographic findings. The standard has been followed by researchers for assessment of different categories of abnormalities present in the mammogram. It specifies the final assessment categories into six categories. It helps in categorising abnormalities as negative, Benign, probably benign, suspicious, malignancy.

METHODOLOGY

For this method, We have taken the image database of digital mammographic images for creating a database of feature extraction from the mini MIAS and it is stored in an excel sheet and loading in matlab. We used the samples of 322 images specified in the Mammographic Image Analysis Society Mini-mammographic Database as our references. In the proposed methodology, the mammogram images are given as input and then noise is

ISSN: 0976-3104

removed from the given input image. After removing the noise ,the Otsu method is used for thresholding and then the GLCM features are extracted. The steps which are involved in the methodology is depicted in the **Figure- 3.**

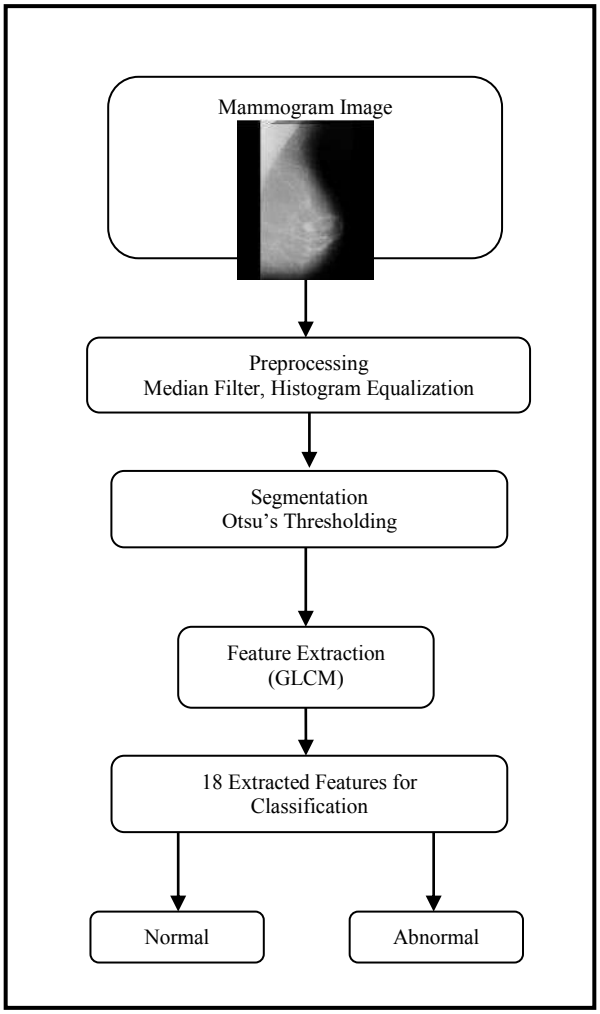


Fig:3. Architecture of the proposed method

Pre-processing

Median Filter is used to remove the noise from the image and improves the quality of the image. Median filter is a well-known order-statistics filter that replaces the original gray value of a pixel by the median of gray values of pixels in the specified neighborhood. A median filter for a smoothed image *f(x,y)* computed from the acquired image *g(x,y)* is defined as

$$f(x,y) = Median\{g(x,y)\}$$

(i,j) € N [8]

where N is the pre-specified neighborhood of the pixel(x,y) [8].

THE IIOAB JOURNAL
ISSN: 0976-3104

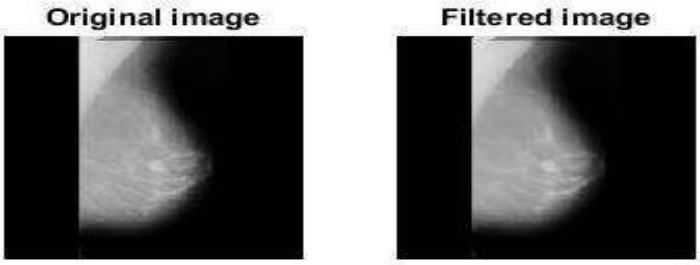


Fig: 4. Original image and output of median filter

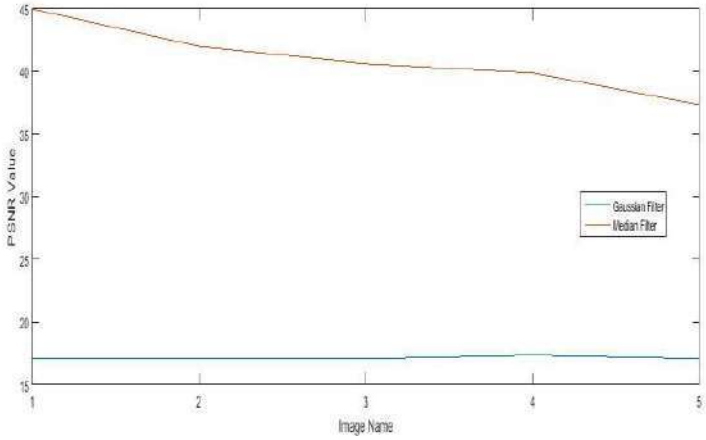


Fig: 5.PSNR value of Gaussian and Median filter

Image Enhancement

The popular technique for enhancing an image is histogram equalization. It is used to reduce the overhead darkness or brightness. It improves the distinct features and visual appearance of the images. The fig.6 shows the histogram of the original image and histogram of the gaussian filtered image.

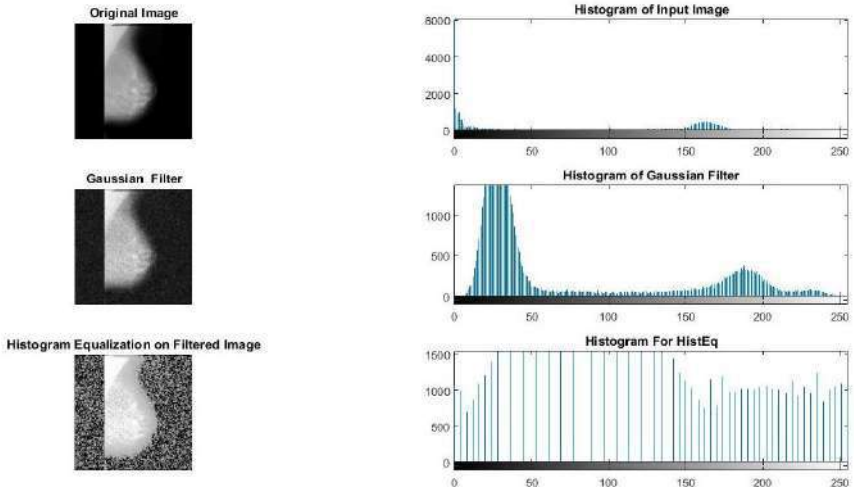


Fig: 6. Histogram of original, Gaussian and histogram equalization

THE IIOAB JOURNAL
ISSN: 0976-3104

Median filter gives better result which is shown in the **Figure-,4** .The plot which has been made by considering the PSNR value of different input images shows that the median filters suits well.Median filter removes the noise present in the mammogram and histogram equalization applied for enhancing the input image. It is very clearly observed that the histogram of histogram equalization produces better result.

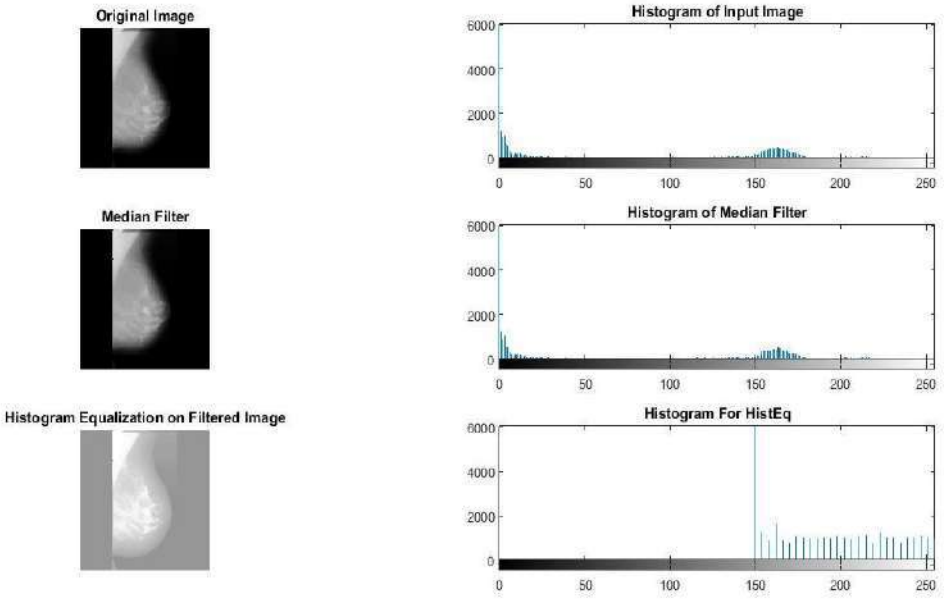


Fig:7. Histogram of median filter and histogram equalization

……

Otsu's method is used for thresholding the image, which uses global image threshold . graythresh uses Otsu's method, which chooses the threshold to minimize the intraclass variance of the threshold black and white pixels.

Feature Extraction

The GLCM feature algorithm is used to extract features from the result image. Resulting, 18 features of GLCM with each min and max value in the Array. In the **Figure- 5** the output of the input images given to the system which is applied through the median filter is shown in the **Figure- 5**.The output of the filter is further applied with thresholding and the result of this is used for extracting features from GLCM. The features which are extracted from GLCM are tabulated in the **Tables- 2,3,4,5,6 and 7**.

Table: 2. GLCM-Features(From 1-3)

Name of file	Autocorr_1	Autocorr_2	Contrast_1	Contrast_2	corrm_1	corrm_2
216	12.94617293	12.98222935	0.706981897	0.646053619	0.931484	0.937414
10	3.393753944	3.436094222	0.677105823	0.581530948	0.678767	0.723921
141	3.843938419	3.904111704	0.686293956	0.56925649	0.62367	0.687866
32	14.45759369	14.54969316	1.736042078	1.518416825	0.701693	0.739658
248	3.865650149	3.916060406	0.761368647	0.652453979	0.561842	0.624272

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 3. GLCM- Features (From 4-6)

Name of file	corrp_1	corrp_2	cprom_1	cprom_2	cshad_1	cshad_2
216	0.931483795	0.937414114	673.8898109	679.6966109	56.38696	56.85415
10	0.678767483	0.723921425	53.87342239	58.64084553	9.695674	10.35827
141	0.623670315	0.687865538	26.75795502	29.80665505	4.511493	4.999365
32	0.701692519	0.739657659	194.0906175	200.3930544	-4.23991	-3.9579
248	0.561842143	0.624272177	20.49201804	23.07003556	3.262367	3.708918

Table: 4. GLCM-Features(From 7-9)

Name of file	dissi_1	dissi_2	energ_1	energ_2	entro_1	entro_2
216	0.385958567	0.3556594	0.273878337	0.277087251	2.195317	2.159219
10	0.413554157	0.37036945	0.343597444	0.353335353	1.848073	1.806057
141	0.471271691	0.41134447	0.197939702	0.208040654	2.146046	2.08653
32	0.835936884	0.74381716	0.059125974	0.064984659	3.168805	3.083147
248	0.52394995	0.46956983	0.174258516	0.18237613	2.20637	2.159627

Table : 5. GLCM- Features(From 10-12)

Name of file	homom_1	homom_2	homop_1	homop_2	maxpr_1	maxpr_2
216	0.847930617	0.85909676	0.837456827	0.849757499	0.506535	0.50854
10	0.830112416	0.84493377	0.818897739	0.835503907	0.572579	0.58112
141	0.796612119	0.81825086	0.78564461	0.809973817	0.394088	0.403445
32	0.689862293	0.71939011	0.664798715	0.698683188	0.136923	0.14237
248	0.773026178	0.79233607	0.761385847	0.783207694	0.353372	0.361869

Table: 6. GLCM- Features(From 13-15)

Name of file	sosvh_1	sosvh_2	savgh_1	savgh_2	svarh_1	svarh_2
216	13.21142782	13.2196865	5.70628882	5.707517566	34.60879	34.74806
10	3.687030929	3.68405151	3.273163218	3.270283066	6.813117	6.925269
141	4.137673514	4.12850937	3.619543657	3.620424506	6.724814	6.852368
32	15.24063245	15.2085565	7.047245233	7.040659239	31.49719	31.67999
248	4.200287118	4.19682665	3.675614596	3.67372823	6.631532	6.714841

Table: 7. GLCM- Features(From 16-18)

Name of file	senth_1	senth_2	dvarh_1	dvarh_2	denth_1	denth_2
216	1.86709554	1.874989983	0.646053619	0.706981897	0.777347	0.818135
10	1.45531407	1.463579527	0.581530948	0.677105823	0.797139	0.851445
141	1.67771022	1.679475856	0.56925649	0.686293956	0.828385	0.896013
32	2.40021975	2.400263073	1.518416825	1.736042078	1.160029	1.225974
248	1.70029865	1.69629512	0.652453979	0.761368647	0.880817	0.937263

In the **Tables** from **2 to 7** ,the extracted 18 features are tabulated and each feature has two ranges which are numbered as 1 and 2.1 indicates the lower range and the 2 indicates the higher range. The name of the features given in the tables represented as

THE IIOAB JOURNAL ISSN: 0976-3104

Autocorr_one- Autocorrelation , Contrast_one- Contrast , corrm_1- Correlation , corrp_1- Correlation, cprom_1- Correlation, cshad_1- Cluster Shade , dissi_1- Dissimilarity , energ_1- Energy , entro_1- Entropy , homom_1- Homogeneity , homop_1- Homogeneity , maxpr_1- Maximum probability , sosvh_1- Sum of sqaures , savgh_1- Sum average , svarh_1- Sum variance , senth_1- Sum entropy , dvarh_1- Difference variance , denth_1- Difference entropy .The features which are extracted from GLCM can be reduced and then the reduced number of features can be given to classifier to classify the abnormalities as benign and malignant.

CONCLUSION AND FUTURE WORK

In this paper, we have extracted GLCM features from the mammogram which has to be done after preprocessing and the segmentation process. The features which are extracted can be used for the further classification technique. The preprocessing is done using median filter and enhancement done through histogram equalization to make the image suitable for segmentation [12] The different features which are extracted based on GLCM can be tried with different classifiers to categorize more precisely the abnormality as normal (benign) and cancerous (malignant).

FINANCIAL DISCLOSURE
No financial support was received to carry out this project.

ACKNOWLEDGEMENT
This work is part of Ph. D Research work. It is not supported by any agency

CONFLICT OF INTERESTS
Authors declare no conflict of interest.

REFERENCES

[1] NCI Cancer Fact Sheets. [Online].Available:http://www.cancer.gov/cancertopics/types/breast

[2] http://www.breastcancer.org/symptoms/testing/types/mammograms/benefits_risks

[3] Kamalakannan J, Tamilarasi Thirumal, Abinaya Vaidhyanathan, and Kansagara Deep MukeshBhai.[2015.] Study on different classification technique for mammogram image", 2015International Conference on Circuits Power and Computing Technologies [ICCPCT-2015],

[4] Kamalakannan J, P Venkata Krishna, M. Rajashekara Babu, and Kansagra DeepMukeshbhai.[2015] Identification of abnormility from digital mammogram to detect breast cancerInternational Conference on Circuits Power and Computing Technologies [ICCPCT- 2015], 2015

[5] RM Haralick, K Shanmugam, and I Dinstein.[1973] Textural features for image classification, IEEE Transactions on systems, man and cybernetics, 3(6): 610-621.

[6] DA Clausi.[2002] An analysis of co-occurrence texture statistics as a function of grey level quantization, *Canadian Journal of Remote Sensing*, 28(1): 45–62.

[7] Kulkarni, Nilambari, and Vanita Mane. [2015] Sourcecamera identification using GLCM", 2015 IEEE International Advance Computing Conference (IACC), 2015.Kamalakannan, J., Tamilarasi Thirumal, Abinaya Vaidhyanathan, and Kansagara Deep MukeshBhai. "Study on different classification technique for mammogram image", 2015International Conference on Circuits Power and Computing Technologies [ICCPCT-2015].

[8] Saroja G Arockia Selva, C Helen Sulochana.[2013] Texture analysis of non-uniform images using GLCM, 2013 IEEE Conference on Information and Communication Technologies.

[9] Radovic, Milos, Marina Djokovic, Aleksandar Peulic, and Nenad Filipovic. "Application ofdata mining algorithms for mammogram classification", 13th IEEE International Conference on BioInformatics and BioEngineering, 2013.

[10] LK Soh, and C Tsatsoulis.[1999] Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices*," IEEE Transactions on geoscience and remote sensing*, 37(2): 780-795,

[11] Sameer Singh, and Keir Bovis.[2005] An Evaluation of Contrast Enhancement For breast Techniques for Mammographic Breast Masses, *IEEE Transactions On Information Technology In Biomedicine*, 9(1): 109-119

[12] Jawad Nagi, Sameem Abdul Kareem, Farrukh Nagi, Syed Khaleel Ahmed.[2010] Automated Breast Profile Segmentation for ROI Detection Using Digital Mammograms, *IEEE EMBS Conference on Biomedical Engineering & Sciences*

[13] H Abdellatif, TE Taha, OF Zahran, W Al-Nauimy, FE. Abd El-Samie.[2013] K9. Automatic Segmentation of Digital Mammograms to Detect Masses, *IEEE* .

[14] Tiago T Wirtti, Evandro OT. Salles. Segmentation of Masses in Digital Mammograms, *IEEE* pp 1-6.

[15] A Oliver, J Freixenet ,J Marti,Elsa Perez, J Pont and RE.Denton.[2010] A review of Automatic mass Detection and Segmentation in Mammographic Images, Medical Image Analysis, 14-2 :.87-110

[16] D.Brzakovic et.al.[1990] An approach to automated detection of tumors in mammograms, *Medical Imaging, IEEE Transactions on*, 9:233-241.

[17] RM Haralick, K Shanmugam, I Dinstein.[1973]Textural features for image classification, *IEEE Transactions on systems, man and cybernetics*, 3(6):610-621.

[18] DA Clausi.[2002] An analysis of co-occurrence texture statistics as a function of grey level quantization, *Canadian Journal of Remote Sensing*, 28(1): 45–62.

[19] LK Soh, and C Tsatsoulis.[1999] Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices," *IEEE Transactions on geoscience and remote sensing*, 37(2): 780-795.

[20] Kamalakannan J, P Venkata Krishna, M Rajashekara Babu, and Kansagra DeepMukeshbhai.[2015] Identification of abnormility from digital mammogram to detect breast cancer",2015 International Conference on Circuits Power and Computing Technologies [ICCPCT- 2015]

[21] RM Haralick, K. Shanmugam, I Dinstein. [1973] Textural Features for Image Classification, IEEETransactions on Systems, *Man and Cybernetics*, 3(6): 610–621.

[22] FI Alam, RU Faruqui. [2011] Optimized Calculations of Haralick Texture Features, European Journal of Scientific Research, 50 (4): 543-553.

[23] D A. Clausi, (2002) An analysis of co-occurrence texture statistics as a function of grey level quantization, Can. J. Remote Sensing, Vol. 28, No. 1, pp 45-62.

[24] Radovic, Milos, Marina Djokovic, Aleksandar Peulic, and Nenad Filipovic.[2013] Application ofdata mining algorithms for mammogram classification", 13th IEEE International Conference on BioInformatics and BioEngineering.

[25] Kulkarni, Nilambari, and Vanita Mane.[2015] Sourcecamera identification using GLCM", 2015 IEEE International Advance Computing Conference (IACC),.

[26] elena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic."A Survey of Image Processing Algorithms in Digital Mammography, *J Bozek* et al. 635.

[27] Jain AK, Duin RPW, Mao J.[2000] Statistical Pattern Recognition, A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(1), 4–37

ABOUT AUTHORS

Prof.Kamalakannan.J, He is a faculty member at School of Information Technology and Engineering, VIT University, Vellore, India. He has completed his Bachelors in Electronics and Communication Engineering and Masters in Computer Science and Engineering from Madras University, Currently pursuing research at School of Computing Science and Engineering,VIT University, India

Dr.M.Rajasekhara Babu, He is a Senior faculty member at School of Computer Science and Engineering, VIT University, Vellore, India. He has completed his Bachelors in Electronics and Communication Engineering from Sri Venkateswara University, Tirupathi, India and took his Masters in Computer Science and Engineering from Regional Engineering College(NIT), Calicut and he has completed his Ph.D from VIT University,India.

ISSN: 0976-3104
SPECIAL ISSUE (SCMDSA)
Mohan et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

F-TEM: A FUZZY BASED TRUST EVALUATION MODEL FOR HEALTHCARE APPLICATIONS IN CLOUD

K Mohan[1*], M Aramudhan[2], Sasikala Ramasamy[3], Swarnalatha P[3]
[1,] *Department of Computer Science and Engineering, Sathyabama University, Chennai, INDIA*
[2]*PKIET, Karaikal, INDIA*
[3]*School of Computer Science and Engineering, VIT University, Vellore, INDIA*

ABSTRACT

Healthcare organizations are beginning to move cloud services in recent years to enhance services and quality without spending much investment for IT infrastructure. Medical records are very sensitive and private to any individuals. Hence, Trust in the relationship among Cloud Service Providers (CSP) and Cloud Service Users (CSU) is very important prerequisite before any service interaction. Hence effective mechanisms are required to identify trustworthy CSP. It is very difficult to evaluate trust level in an open and dynamic environment like cloud. In this paper we have proposed F-TEM: A Fuzzy based Trust Evaluation Model, which filters suitable CSP for any CSU based on their input parameters. Experiments were conducted with several parameters like security, capability and behavior of CSP and the results were compared with conventional method.

Received on: 30th-Nov-2015
Revised on: 22nd-March -2016
Accepted on: 13th– April-2016
Published on: 10th –June-2016

KEY WORDS

Fuzzy; trust; cloud computing; healthcare system

***Corresponding author: Email:** meetmohan.k@gmail.com

INTRODUCTION

In the past few years, Health care is moving to digital and particularly Cloud services are becoming patient focused and data driven. By using cloud computing facility, medical data sharing among doctors and patients, accessing and storage becomes easier. Responsibilities of health care service providers and choosing the best service provider by the user becomes the important part of digital health care systems [1,2].

Health Care System

Healthcare and the quality of Health care service are playing significant roles in any country. The responsibilities of Health care include prevention of disease, diagnosis, treatment, report preparation and analysis. Accessing Health care by common people across countries varies by economic conditions of individuals [3]. Technologies can create major impact in health care includes Robotic surgery, Health care web service, Decision support system. As like above technologies, Cloud computing is also emerging technology, which can create major impact on health care. In future, Cloud Computing will play major role in modernizing Health care service. The proposed system integrates the decision making system, Health care web services and cloud computing to make health care service more authoritative.

Health Care Service Providers Responsibilities

Healthcare providers are expected to provide a quality improved patient care with limiting cost. Cloud Computing service model is to enable convenient, on -demand software as a service (SaaS) and Storage as a Service to Hospitals and patients through Health Care Cloud Service Providers. Cloud Computing Service model rapidly reduce Management efforts and services. Cloud Computing services initiated in Health care have the following features:

THE IIOAB JOURNAL ISSN: 0976-3104

- Elastically increase the infrastructure such as servers and storage
- Moving huge data such as radiology images, clinical data from Health care IT department to Cloud reduces troublesome task
- Software and Patient record are on demand and 24 x 7
- Maintenance risks are substantially reduced
- Single point of failure is totally avoided
- Web based easy access to Patients and Doctors
- Improves the skill to analyze and track patient information
- Provides facility to sharing lifesaving information quickly among doctors in various geographical areas during emergency and reduces the need for duplicate testing

Hence health care cloud service providers provide preventive, curative health care services through storage, service request, report generation and Analysis during emergency and for out-patients (OP). However, during the integration of Cloud service to Health care service providers, two properties need to be considered: Expectations of users from Service providers and Trust value of Cloud Service Providers [4]. Hence, Qualifying criteria need to be chosen carefully to choose appropriate cloud service providers.

Qualifying Health Care Service Providers

Hospitals and patients wanted to use service providers have all authority to choose the best providers. Hence, CSPs are responsible to get trust certificates from some centralized authority called Trust Managers. The role of Trust Manager is to match the expectations of users and trust value of CSPs.

Expectation of Users from Cloud Service Providers

Users or Doctors use cloud services 24 x 7 to access patient health records, medicine details and Doctors can access other doctor's assistance through these services. Hence users expect from cloud service providers that the service should be available 24 x 7, the records need to be available where ever they go, their records should be maintained securely and hence privacy of patients health records is playing significant role. It is the responsibility of cloud service users to choose cloud service provider with high reputation value. It provides the facility to the patients to utilize their database when they move from one Health Service Organization to another.

Trust Value of Cloud Service Providers

Trust value of cloud service provider is calculated based on the parameters considering Security, Capability and Behaviour of Cloud Service Providers. The criteria Security, Capability and Behaviour are chosen from the user's feedback, which we have taken manually by giving the forms to various hospitals and patients who are all interested to go for external service and for keeping their records in cloud. The parameters are assigned with weightage value to differentiate the significance of each parameter.

Security: The service provider has to maintain the privacy of patient's record, be accessed only by the authorized patient and doctors and data should be recovered immediately in case of failure. Hence, the attribute security is highly deciding parameter for health service cases. Therefore, 50% of weightage is assigned to security parameter alone.

Capability: Health care service is a basic responsibility of any Government to treat the world population without considering their financial background. Web services and cloud services makes the total system simple, economic and easy accessible. To provide cloud service to every patient, who need service from anywhere, the capacity is the next important factor of CSP. To make the system highly elasticity, the high capacity server, storage and RAM are vital factors. Server and Storage makes the system to store more patient data. The RAM capacity makes the system quick access.

THE IIOAB JOURNAL ISSN: 0976-3104

Behaviour: Service level agreement advises the possibilities of services the provider can offer to users. Hence, the cloud service users need to decide the level of CSP based on their SLAs. Also the user has to obtain the feedback from other cloud users to choose the best service provider. Hence, every cloud users have to register their feedback in particular time interval about CSPs. The Trust value calculation is based on Fuzzy logic for cloud service providers. The proposed system uses Fuzzy based model which takes input from the user's feedback and SLAs of individual CSPs and the model gives the Trust rating as output. The fuzzy model chosen is Mamdani FIS, Mamdani FIS and Defuzzifier generates the output as a crisp value. The implementation consists of two stages: Fuzzy based calculation for Trust expectations of users and Fuzzy based weightage calculation of CSPs based on three criteria Security, Capability and Behaviour.

The rest of the paper is organized as follows. Section 2 presents the related works to cloud based health care service in the literature. Section 3 describes Fuzzy based trust evaluation model (F-TEM) used in this paper; In Section 4 the results achieved in this work are discussed. Finally Section 5 is conclusion and future enhancements to be carried out. The list of references to build the work is furnished at the end.

RELATED WORK

There is no comprehensive trust mechanism existing among leading cloud service providers like Amazon, Google and Microsoft. So cloud service users are really struggling to identify suitable service providers. [5] proposed Feedback rating based reputation to evaluate trust. Feedback collector module is developed for collecting feedback from users about cloud service provider. But, more detailed classification of context is missing in this paper. [6] proposed credibility based trust model which evaluated the credibility of trust feedbacks given by cloud service users also this model identifies malicious trust feedback from hackers. [7] proposed SLA based trust model to help cloud service users to identify trustworthy cloud service providers based on their SLA agreement. However some important parameters outside SLA agreement cannot be monitored and evaluated which reduces the overall effectiveness. [8,9] proposed user behavior based trust evaluation. The author divided the user behavior into four categories and under each category users behavior is monitored and evidences are maintained. Based on some formula the level of deviation with normal behavior is computed and trust decision is made. [10] proposed game theory based trust model which helps cloud user applications to map appropriate service providers. However this model failed to answer uncertain nature of applications and resources. The proposed game theoretic based approach for trust evaluation model can be applied only for the first time user and provider; it could not be used for evaluation of HSP with continuous evaluation model. [11] proposed trust based mechanisms for the integration of cloud and sensor network, but the technique considers few attacks on trust such as good mouthing, bad muting, collusion and white washing. Fuzzy inference system is used for trust evaluation in [12-16], but dynamically changing behavior is not considered by those authors.

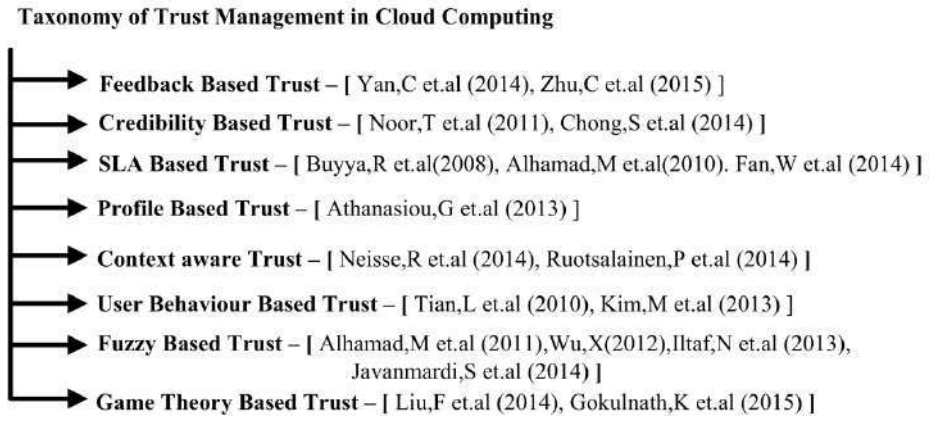


Fig: 1. Taxonomy of Trust Management in Cloud Computing

[17] developed a hierarchical fuzzy inference system for service selection process. In hierarchical inference system output of low level are given to input of high level models. It helps the system to compute partial solutions. The proposed system uses min-max structure of Mamdani model to forward output of one level to other level in same category. [18] used fuzzy comprehensive evaluation method to validate trust evaluation

THE IIOAB JOURNAL ISSN: 0976-3104

value is close to trust level or not. The model works well only the granularity of trusted value is great. The fuzzy comprehensive evaluation model is a decision making method. And the developed model is not applied on any real time experiments. But Fuzzy inference system is proposed method is the process of formulating the mapping process from input values to the output using fuzzy rules.

PROPOSED FUZZY BASED TRUST EVALUATION MODEL

Figure- 2 illustrates the general overview of our proposed system. There are three major level of input steps involved in the usage of F-TEM Engine: The first step, Users requirements submission to F-TEM Engine, Second, Service Catalogues and SLA Submission by Service providers and finally dynamic feedback collection from Users.

F-TEM overview

A software framework of F-TEM consists of three functional units: F-TEM Engine, Cloud based HSP Layer and HSU layer. The CSU and CSPs are mentioned as HSU and HSP layers. The F-TEM makes final decisions based on the inputs such as Cloud user's feedback, SLA and Service catalogue. It provides its resultant decisions to HSU layer as Service advertisements. The F-TEM is responsible to check the identity of Cloud users and Cloud service provider users from Identity Provider (IdP). The F-TEM engine contacts Knowledge base to retrieve and store details of HSU layer feedbacks, Service catalogue and SLA of individual Cloud service providers. The Knowledge Base is a storage engine to store complex unstructured information used by the F-TEM Engine. It is a centralized repository for information: details of HSU and HSP layer. This knowledge base assists the F-TEM to analyse and take decisions on Trust requirements of Cloud users. The trust decisions about Cloud providers will also be stored by Knowledge base for later retrieval.

The HSU layer is a set of cloud users who need health care services through online during emergency and for report generation. Initially, The HSU layer communicate its requirements to F-TEM Engine. The HSU layer expectation is to identify a quality service provider based on its requirements. The service advertisements from F-TEM help out users to identify a competent provider who matches the requirement of cloud users. After identifying the Cloud service provider, HSU layer directly interact with the service provider for all its Health care issues and decisions.

The Cloud Based HSP layer is a set of cloud providers who have different SLAs, models, performance rates in terms of security, capability and behaviour. It provides its service catalogues and capability list through SLA. The Cloud providers are assigned with trust value by the F-TEM. The trust values are assisting the users to choose its provider. The proposed work focuses only the F-TEM Engine Role.

F-TEM Architecture

The components of F-TEM as shown in **Figure- 3** are Input Layer, Fuzzifier, F-TEM Inference Engine, Knowledge Base, COG defuzzifier and Trust Result. The Input layer considers three parameters: Security, Capability and Behaviour. For example, the input values of attributes namely authentication, access control, Data security and Data recovery of Security are given as in terms of Low, Medium and High to Fuzzifier.

The steps of Fuzzy Inference Engine are : Compare the input variables with the membership function to obtain the membership values. This is done by Fuzzifier. Fuzzy Inference Engine generate the qualified resultant value either crisp value or fuzzy based. Aggregate the resultant value to produce a crisp output. This step is done by defuzzifier. A rule base containing number of IF-THEN rules and Database jointly called as knowledge base.

THE IIOAB JOURNAL
ISSN: 0976-3104

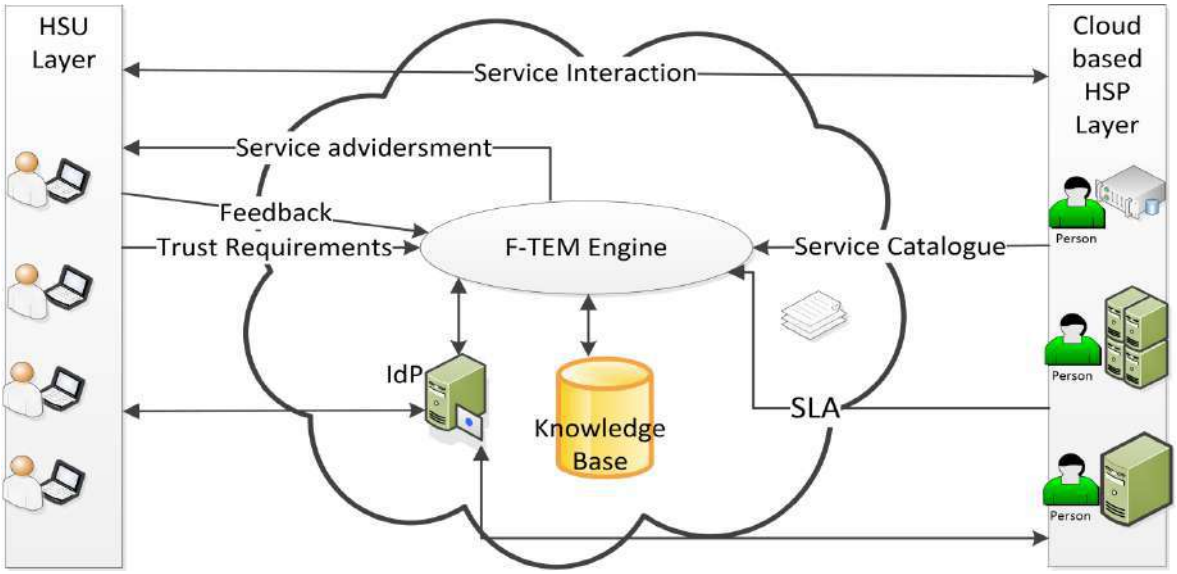


Fig: 2. F-TEM Overview in Cloud Environment

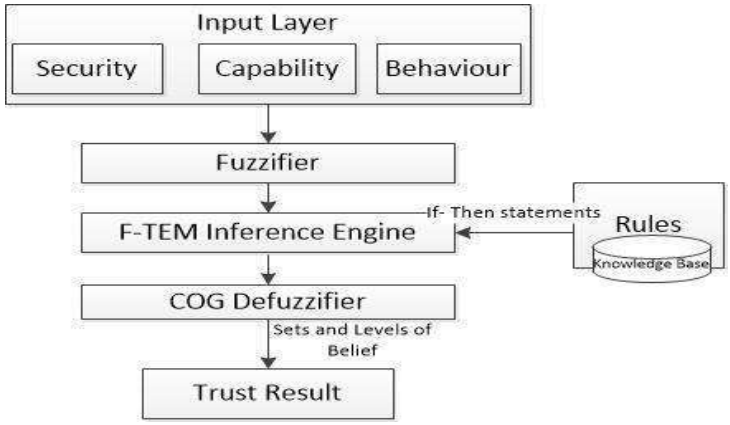


Fig: 3. F-TEM Architecture

Fuzzy based Trust Evaluation Model

The evaluation of the trust value for CSP and user expectation for cloud computing users, comprises of two stages: trust value for CSP and Trust Expectation from Users. Fuzzy inference system is classified into direct Fuzzy inference system and Indirect fuzzy inference system. Mamdani and Sugeno-type are direct method and direct methods are easy to implement, but indirect methods are complex in nature. In this paper, the model based on Mamdani is used to calculate trust value dynamically for cloud service provider. Mamdani method and Sugeno methods are widely accepted fuzzy models for capturing expert knowledge. The most fundamental difference between Mamdani-type FIS and Sugeno-type FIS is the way the crisp output is generated from the fuzzy inputs. The Mamdani-type FIS uses the technique of defuzzification of a fuzzy output. Due to the interpretable and intuitive nature of the rule base, Mamdani-type FIS is widely used in particular for decision support application [19]. Also the Mamdani type is most suitable for human input. The purpose to choose mamdani model is it simple min-max structure. The parameters of each category rules are combined with min structure and the categories rules are connected by Max structure. These simple min-max structures of mamdani are easily incorporated into this trust evaluation model to combine the two levels of structure (Parameters and Categories). We classified the health care service providers trust evaluation model

THE IIOAB JOURNAL ISSN: 0976-3104

into two levels, three categories of HSP and each category is further classified into set of parameters. Hence the Mamdani inference system with min-max structure is well suited for Fuzzy based trust evaluation model. The user expectation model has three or four input parameters in each category mapped to one output. The input and parameter has three member functions: low, medium, high and output parameter has four membership functions: very low, low, medium and high. [20] claimed that "Fuzzy logic is a logic of approximate reasoning whose distinguishing features are (i) fuzzy truth-values expressed by linguistic terms, (ii) imprecise truth-tables..." Didier Dubois, Toulouse (France) and with Stefan Lehmke, Dortmund stated as

Fuzzy logic = many value logic+ partial belief.

I.e. extend the weighted formula in possibilitic logic and many valued logic conjointly [paper] into a genuine fuzzy logic. This paper uses the weighted parameters as multi value logic with rules of possibilitic logic made the final model as fuzzy based trust evaluation model.

Trust Expectations of Users

Trust based Health Care system is developed using Mamdani-Type Fuzzy Inference Systems model. CSP trust value has three categories: Security, Behaviour and Capability. The Category 1 consists of four attributes as inputs from user feedback on the subject of authentication, authorization, data security level and data recovery level. The system has one output for the category 1 that decides Trust value for the CSP on the subject of Security. The same is repeated for Capability and Behaviors as shown in **Table- 1**. We have collected user feedback from Health centre to fix the weightage of Categories and we have given options to users to choose parameters under categories as given in section 1.3.2. The collected feedback has been given as data to train the system. The authentication, authorization, data security and data recovery are taken to be in range of low, medium, high based on SLAs issued by CSPs and feedback from users. The fuzzification (fuzzifier) process converts crisp inputs to non-crisp (fuzzy) outputs. The Trust value range for Authentication is fixed based on user requirement as shown in **Table- 2**. For example, if user requirement for authentication is Login & Password based, then trust value range is low, medium for One Time Password (OTP) and high for biometric authentication. The Attributes of three categories and its Scenario and Range are stated in **Table- 2**.

Table: 1. Parameter's Weightage and Attributes

Parameters & Weightage	Attributes
Security - 50%	Authentication
	Authorization
	Data Security Level
	Data Recovery Level
Capability - 20 %	Server Capacity
	Storage Capacity
	RAM Capacity
Behaviors - 30 %	User Feedback
	SLA accomplishment
	Availability

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 2. Range of Attributes

Attributes	Range	Scenario	Attributes	Range	Scenario
Authentication	Low (1)	Password based	Storage Capacity	Low (1)	Limited with <= 1 GB
	Medium (2)	OTP based		Medium (2)	Limited up to 1 TB
	High (3)	Biometric based		High (3)	Unlimited
Authorization	Low (1)	Role based	RAM Capacity	Low (1)	<=2GB
	Medium (2)	Attribute based		Medium (2)	2 GB to 8 GB
	High (3)	Semantic based		High (3)	More than 8 GB
Data Security Level	Low (1)	http	User Feed back	Low (1)	Less aggregate Feedback value
	Medium (2)	https		Medium (2)	Avg. aggregate feedback
	High (3)	https with high key value		High (3)	Good aggregate Feedback
Data Recovery Level	Low (1)	No recovery	SLA Accomplishment	Low (1)	Not at all Accomplished
	Medium (2)	Partial recovery		Medium (2)	Partially Accomplished
	High (3)	Complete recovery		High (3)	Completely Accomplished
Server Capacity	Low (1)	Non-replicated	Availability	Low (1)	50% to 69%
	Medium (2)	Replicated-min		Medium (2)	70 % to 89%
	High (3)	Replicated-max		High (3)	90% and above

The following is a simple calculation which takes input as attributes of security as mentioned in **Table- 2**, and the rule is framed to find the output for Security. The output ranges are very low, low, medium and high. The sample rules for the security are described in **Table- 3**.

Rule to calculate output of Security

We have collected expert's feedback from cloud users, software industry and Health care professionals for setting the rules to calculate the output for Security, Capability and Behaviors. For example security can be fixed as high only if total is higher than 9, i.e. if one parameter may be low but other parameters should be high to pass security category as output the result – high. The same rules we fixed for behavior and capability.

Total =Authentication + Authorization + Data Security + Data Recovery

if total>=4 and total <=5

security = very low

else if (total >= 6 and total<=7)

security = low

else if(total>=8 and total <=9)

security = medium

else

security = high

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 3. Sample Rule Base of Mamdani FIS for the Parameter - Security

Authentication	Authorization	Data Security	Data Recovery	Security
Low(1)	Low(1)	Low(1)	Low(1)	Very Low
Low(1)	Low(1)	Low(1)	Medium(2)	Very Low
Low(1)	Low(1)	Low(1)	High(3)	Low
Low(1)	Medium(2)	High(3)	High(3)	Medium
Low(1)	High(3)	High(3)	Low(1)	Medium
Low(1)	High(3)	High(3)	Medium(2)	Medium
Low(1)	High(3)	High(3)	High(3)	High
Medium(2)	Low(1)	Low(1)	Low(1)	Very Low
Medium(2)	Low(1)	Low(1)	Medium(2)	Low
Medium(2)	High(3)	Medium(2)	High(3)	High
Medium(2)	High(3)	High(3)	Low(1)	Medium
Medium(2)	High(3)	High(3)	Medium(2)	High
Medium(2)	High(3)	High(3)	High(3)	High
High(3)	Low(1)	Low(1)	Low(1)	Low
High(3)	Low(1)	Low(1)	Medium(2)	Low
High(3)	Low(1)	Low(1)	High(3)	Medium
High(3)	High(3)	Medium(2)	Low(1)	Medium
High(3)	High(3)	High(3)	Medium(2)	High
High(3)	High(3)	High(3)	High(3)	High

The following is a simple calculation which takes input as attributes of Behavior and Capability and the rule is framed to find the output for Behavior and Capability.

Rules to calculate output of Behavior

total = user feedback +SLA accomplishment + availability

if total<=4

behavior = very low

else if (total == 5)

behavior = low

else if(total>=6 and total <=7)

behavior = medium

else

behavior = high

Rules to calculate output of capability

total =server capacity + storage capacity + RAM capacity

if total<=4

capability = very low

else if (total == 5)

capability = low

else if(total>=6 and total <=7)

capability = medium

else

capability = high

Triangular Membership Functions

The rules included for the Trust based Health Care system for Category Security are described in **Table- 3** and other rules are not explicitly mentioned in the paper. For the calculation of Trust value of CSP using Mamdani model, the inputs are taken from security as shown in **Figure- 4**, Capability as shown in **Figure- 5** and Behavior as shown in **Figure- 6** and produces output as shown in **Figure- 7** that decides the trust value of each CSP. The output membership functions are very low, low, medium and high.

THE IIOAB JOURNAL
ISSN: 0976-3104

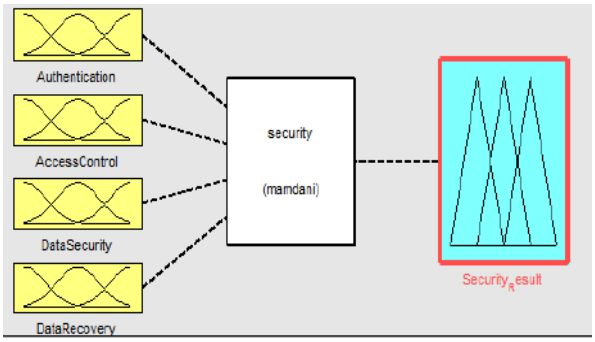


Fig: 4. Security Membership Function

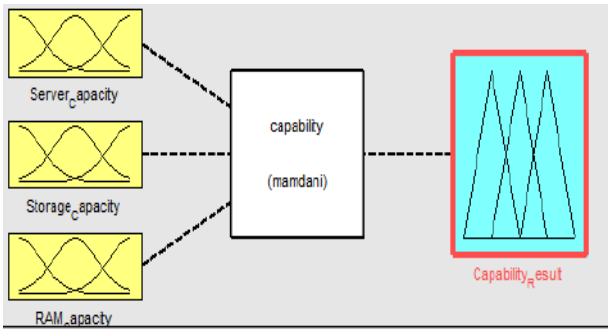


Fig: 5. Capability Membership Function

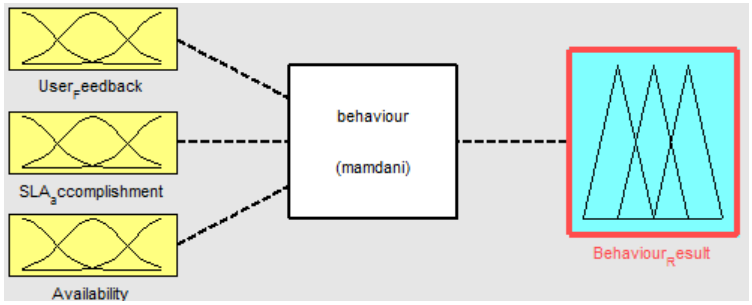


Fig: 6. Behavior Membership Function

The outputs of Security, Behavior and Capability membership functions are given as input to the Membership ship for trust expectation of users. The crisp value for output of parameters is assigned as shown in **Table- 4**.

Table: 4. Membership Function of Parameters - Security, Behavior and Capability

Parameter Output	Constant Value
Very low	0.25
Low	0.50
Medium	0.75
High	1.00

The inputs for the calculation of resultant membership function as constant value x weightage/100. For example, if the output of security parameter is high, 1 x 50/100 = 0.5 is given as crisp input for the resultant membership function. The sample outputs of Triangular membership function for Trust value of CSP are as shown in **Table- 5**. The crisp value generated from F-TEM is given as Trust value of CSP in percentage to the CSUs. The CSUs can choose their service provider according to their expectations.

Table: 5. Sample Trust Values of CSP

Security	Behavior	Capability	F-TEM Output	Constant Value in %
Very low	Very low	Very low	Very low	25
High	Medium	Medium	Medium	75

THE IIOAB JOURNAL
ISSN: 0976-3104

Medium	low	low	low	50
Very low	Very low	low	Very low	25

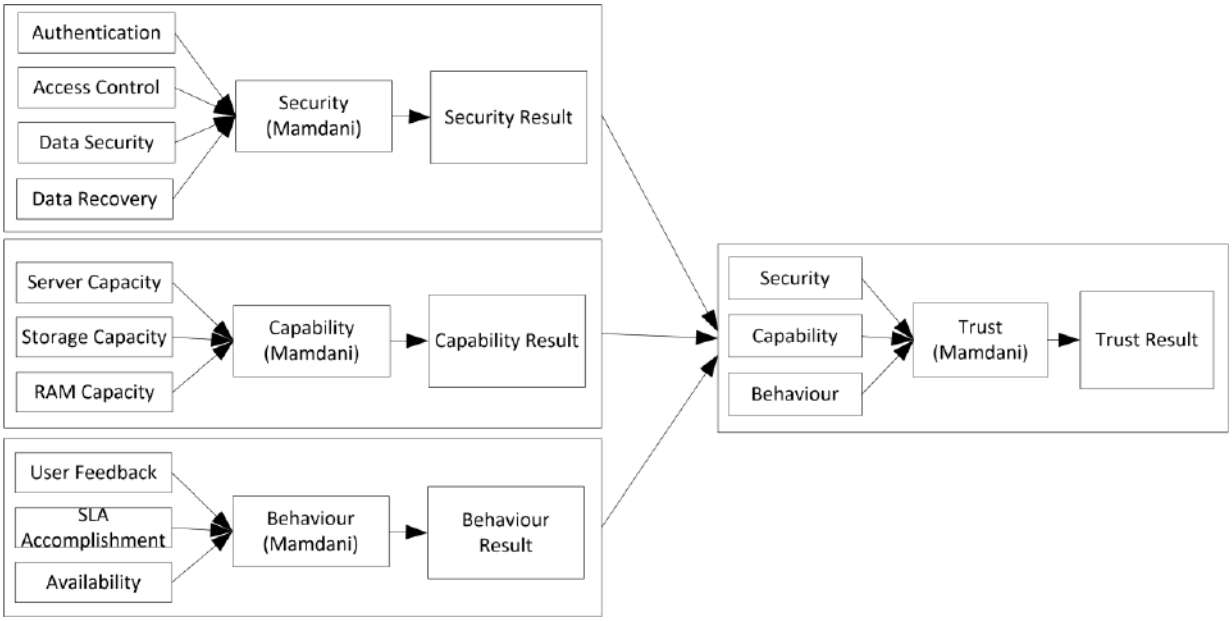


Fig: 7. Membership Function for Trust value of CSP

RESULTS AND DISCUSSION

This section shows our proposed method results to verify the cloud based trust calculation. The proposed model is provided with the results of the experiment using Matlab. The Matlab toolbox for fuzzy logic is used to implement our methodology. This toolbox has ready functions and calculation for fuzzy inference systems Mamdani and Sugeno fuzzy inference system. We have used Mamdani model inference system for calculating Trust value for parameters Security, Behaviors and Capability and final Trust value of CSP. We have collected real time datasets from maximum of 100 different users from Health Centers and trained the system. The service catalogue details from 10 to 20 CSPs and the feedback from users on the subject of CSPs are fixed randomly. The output of membership function is calculated as shown in **Figure- 7.** Number of CSUs and CSPs in total varies from 13 to 110 and time taken by the F-TEM is calculated. The following results show that time taken by the F-TEM to calculate trust value will not be much varied by number of CSUs and CSPs. Hence it is proved that F-TEM is not creating any time complexity overhead in choosing the right CSP by using fuzzy logic inference system.

Table :6 .Time Taken by the F-TEM

Sl.No	No. of CSU and CSP	Time in ms
1	13	5.0
2	23	5.2
3	36	6.3
4	46	6.6
5	58	7.1
6	68	7.2
7	110	8.0

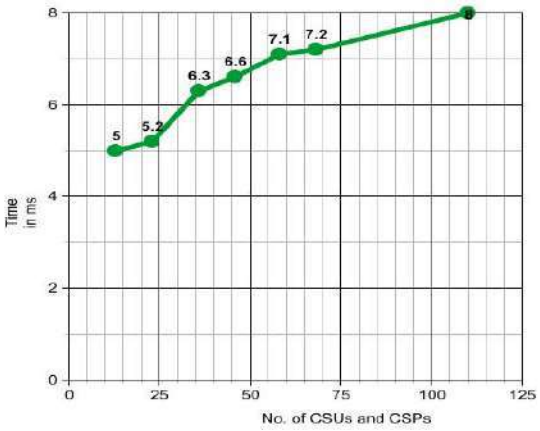


Fig: 8. Time Taken by the F-TEM

We have implemented the proposed method on fuzzy logic in three methods: user feedback based (without Fuzzy), Fuzzy logic, Fuzzy with dynamic feedback and the results are compared with traditional method in **Figure- 9.** In traditional method (feedback), we have taken average of the values of parameters. The second results are feedback with Fuzzy logic and third results are one the subject of fuzzy logic with dynamic feedbacks from CSUs who uses CSPs. The dynamic feedback on the subject of CSPs added greater strength to choose right CSP. The category Behavior is fixed based on user's feedback on SLA accomplishments, feedback percentage and availability during access. The trust value results are taken from F-TEM with and without dynamic user feedback. The results shows that the fuzzy logic and fuzzy logic with user's feedback creates greater impact than traditional method.

Table: 7. Comparison of F-TEM with Traditional Model

Techniques	CSU and CSP - % of match
Feedback based	50%
Fuzzy logic Methods(without dynamic Feedback)	80%
F-TEM model with dynamic user feedback	87.5%

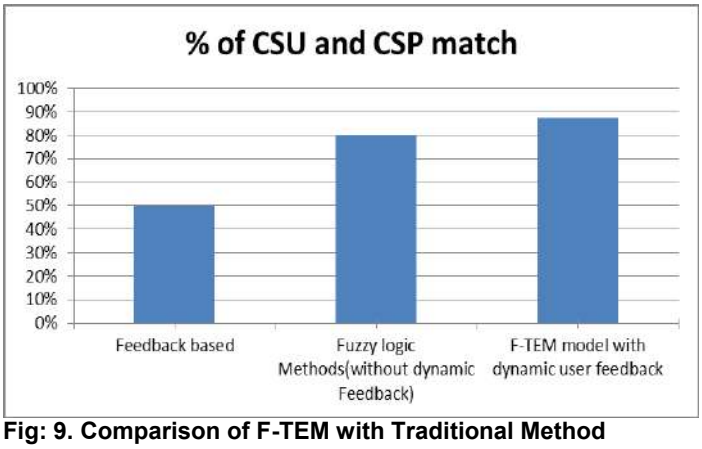


Fig: 9. Comparison of F-TEM with Traditional Method

THE IIOAB JOURNAL ISSN: 0976-3104

CONCLUSION

In this paper, a trust evaluation scheme based on fuzzy logic inference system for Health care service is described. The proposed system evaluates the trustworthiness of cloud service providers. We have collected user feedback from Health centre to fix weightage of Categories and we have given options to users to choose parameters under categories. The collected feedback has been given as data to train the system. The results show that time complexity is not the major concern when we calculate trust value with F-TEM. Hence it is proved that F-TEM is not creating any time complexity overhead in choosing the right CSP by using fuzzy logic inference system. This F-TEM method experiments on Matlab show that the dynamic feedback from the users, assigning proper weightage(Security 50%, Behavior 30% and Capability 20%) and Fuzzy logic inference system assist cloud service users to choose right CSPs. The work can be extended as F-TEM product or web service, which is used to choose any service providers such as Internet service providers, cloud service providers based on dynamic user's feedback.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

ACKNOWLEDGEMENT

None

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

REFERENCES

[1] Bamiah MA, Brohi SN, Chuprat S, Manan JA,[2014] TRUSTED CLOUD COMPUTING. *J Comput Sci,*10(2):240–250.

[2] Buyya R, Yeo CS, Venugopal S.,[2008] Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities. 2008 10th IEEE IntConf High Perform ComputCommun .*IEEE*,5–13.

[3] Wul R, Ahnl G.[2012] Secure Sharing of Electronic Health Records in Clouds. 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing, Collaboratecom 2012 Pittsburgh, PA, United States, 711–718.

[4] Buyya R, Ranjan R, Calheiros RN. [2009]. Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities. *Proceedings of the 7th High Performance Computing and Simulation Conference*, Germany. Leipzig, Germany,1-11

[5] Yan C, Qi L, Ni J.[2014] Building Feedback Rating-based Reputation System for Trusted Delivery of Cloud Services. Proc 2014 IntConf Mechatronics, Electron Ind Control Eng. Paris, France: *Atlantis Press*;(Meic):684–687.

[6] Noor TH, Sheng QZ.,[2011] Credibility-Based Trust Management for Services in Cloud Environments. ICSOC 2011, LNCS 7084, *Springer-Verlag Berlin Heidelb,*328–343.

[7] Alhamad M, Dillon T, Chang E.,[2010] SLA-Based Trust Model for Cloud Computing. 2010 13th IntConf Network-Based InfSyst .*Ieee*, 321–324.

[8] Tian L, Lin C, Ni Y, [2010] Evaluation of user behaviour trust in cloud computing. *ComputApplSyst.* 2010;(Iccasm 2010):567–572.

[9] Kim M, Park SO.[2013] Trust management on user behavioral patterns for a mobile cloud computing. *Cluster Comput*, 16(4):725–731.

[10] Gokulnath K, Uthariaraj R.,[2015] Game Theory Based Trust Model for Cloud Environment. *Sci World J*.2015(2015):1–10.

[11] Zhu C, Member S, Nicanfar H, Leung VCM, Yang LT.[2015] An Authenticated Trust and Reputation Calculation and Management System for Cloud and. *IEEE Trans Inf FORENSICS Secur*. 2015,10(1):118–131.

[12] Xu Wu., [2012] A Fuzzy Reputation-based Trust Management Scheme for Cloud Computing. *Int J Digit Content Technol its Appl* .2012 Sep 30, 6(17):437–445.

[13] Alhamad M, Dillon T, Chang E. [2011] A Trust-Evaluation Metric for Cloud applications. *Int J Mach Learn Comput* .2011; 1(4):416–421.

[14] Iltaf N, Ghafoor a.,[2013] A fuzzy based credibility evaluation of recommended trust in pervasive computing environment. 2013 *IEEE 10th ConsumCommunNetwConf.Ieee*, 617–620.

[15] Javanmardi S. FR trust, [2014] a fuzzy reputation–based model for trust management in semantic P2P grids. *J Grid*,6(1):57-66

[16] Lin G, Lin C, Chou C, Lee Y. [2014] Fuzzy Modeling for Information Security Management Issues in Cloud Computing. *Int J Fuzzy* 16(4):529–540.

[17] Qu C, Buyya R.,[2014] A Cloud Trust Evaluation System Using Hierarchical Fuzzy Inference System for Service Selection. 2014 IEEE 28th IntConfAdvInfNetwAppl.,850–7.

[18] Zhou Z, Luo Y, Guo L, Sun L.[2013] Assessment of P2P Trust Model Based on Fuzzy Comprehensive Evaluation. *J Softw* 2013 Nov,8(11):2711–2714.

[19] Arshdeep Kaur, Amrit Kaur,[2012] Comparison Of Mamdani-type And Sugeno-type Fuzzy Inference Systems For Air Conditioning System. *International Journal Of Soft Computing And Engineering.,* 2(2):323-325.

[20] LA Zadeh,[1975] Fuzzy logic and approximate reasoning. *J Synthese* 3(30):407–428.

THE IIOAB JOURNAL
ISSN: 0976-3104

ABOUT AUTHORS

Mr. K. Mohan *is a research scholar in computer science and engineering department at Sathyabama University. He is also working as an Assistant Professor in School of Computer Science and Engineering at VIT University. His research interest includes cloud security, healthcare system, Internet of Things. He can be reached at meetmohan.k@gmail.com*

Dr. M. Aramudhan *is an Associate Professor in IT department of PKIET, karaikal. His research interest includes cloud security, semantic web, Internet of Things, web mining. He can be reached at aranagai@yahoo.co.in*

Dr. Sasikala Ramasamy *is an Associate Professor in School of Computer Science and Engineering, VIT University. Her research interest includes cloud security, semantic web, Internet of Things, Big data Analytics and social networking. She can be reached at sasikala.ra@vit.ac.in*

Dr. Swarnalatha P *is an Associate Professor in School of Computer Science and Engineering, VIT University. Her research interest includes Image Processing, Artificial Intelligence, Remote Sensing, Software Engineering BigData and Internet of Thing. She can be reached at pswarnalatha@vit.ac.in*

ISSN: 0976-3104

SPECIAL ISSUE (SCMDSA)
Gupta et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

SPEECH GUIDED FEATURE EXTRACTION AND BRAIN STORM OPTIMIZATION TO CLUSTER OBJECTS USING FUZZY LOGIC

Utkarsh Gupta*, Swarnalatha P, Prateek Chharia
**Department of SCSE, VIT University, INDIA*

ABSTRACT

In this paper a novel object recognition technique is proposed which is based on fuzzy clustering and Brain Storm Optimization Algorithm. The aim is to create classifiers based on clustered data. Object features are extracted from real time video frames guided by speech recognition. The proposed feature extraction works in two phases, first phase deals with extracting average pixel intensities of Red, Green and Blue channels respectively from the sample object image along with illuminance reading of Lux Meter and name of the object recognized by speech engine. These features are then stored as primary feature vector set. Second phase deals with extraction of keypoints using robust local feature detector algorithm called as SURF (Speeded-Up Robust Features) which will be stored as secondary feature vector set. FREAK (Fast Retina Key-point) descriptor has been combined with SURF detector algorithm for comparison with SIFT (Scale Invariant Feature Transform). Brain Storm Optimization helps in optimization and minimization of cluster distances. In our proposed technique we perform clustering using fuzzy C-Means and BSO only on primary feature vector set. The aim is to reduce keypoints matching time complexity. Computing distances like mahalanobis distance between primary feature vector and test object features will reduce the candidate rows of feature set. Applying keypoints mapping on fewer records will reduce the complexity of recognition algorithm. 65.8% reduction in time has been observed using this strategy over the conventional method of mapping keypoints of complete dataset with test object. Clustering algorithm has 86.9 per cent accuracy for the primary feature vector set consisting of 56 real time object data points.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 29h – March-2016
Published on: 10th–June-2016

KEY WORDS

Object Recognition, Image Processing, Speech Recognition, Brain Storming Optimization

***Corresponding author:** **Email:** utkarsh.satishg2014@vit.ac.in **Tel:** +91-9944702033

INTRODUCTION

Multimodal recognition is one of the most important fields of robotics. Multimodal feature extraction will increase the accuracy of recognition. Suppose an image of pen is captured, then image processing is done to extract features. In this case system may get confused with other cylindrical objects. Similarly if a user says “It is pen”, and speech engine detects words, there is a possibility that acoustic model of system recognizes the word “pan” and not actually “pen”. Therefore multiple modalities are required to increase the accuracy of the system. Reconsidering the previous example using multimodal inputs where system captures image and user’s speech input, “It is pen”. In this case, system can parallely know that the object is cylindrical in shape and its name/category is pen, so it can filter the data set, and map cylindrical shape with keyword “pen”. Thus the accuracy will be really improved if we fuse the speech with real time video frame to classify object [1]. The proposed technique splits feature set into primary vector of features and secondary vector of features. Primary vector contains generic and less complex numerical data. Whereas secondary vector contains more complex data. Example: primary vector set contains average pixel intensities of various channels of image and secondary vector set contains more complex keypoints extracted from image (using SURF with FREAK). Separating feature vectors helps in reducing the overall matching time. A comparative study between SURF with FREAK and SIFT has also been conducted and results have been included in the experiments and discussions section.

MULTITHREADED SYSTEM ARCHITECTURE

Multithreaded architecture ensures parallel feature extraction and thus reduces overall running time. One thread uses speech engine for object’s name/category/type recognition and other thread does the image processing over real time video frames. This will decrease the computation time of the system. **Figure- 1** shows the flow of information representing multithreaded architecture. Speech recognition engine and video capture engine are working parallel. Speech engine converts speech to text using voce library. In case of video processing engine, camera continuously captures video frames, from which features are extracted. Features are extracted using the openCV libraries.

THE IIOAB JOURNAL
ISSN: 0976-3104

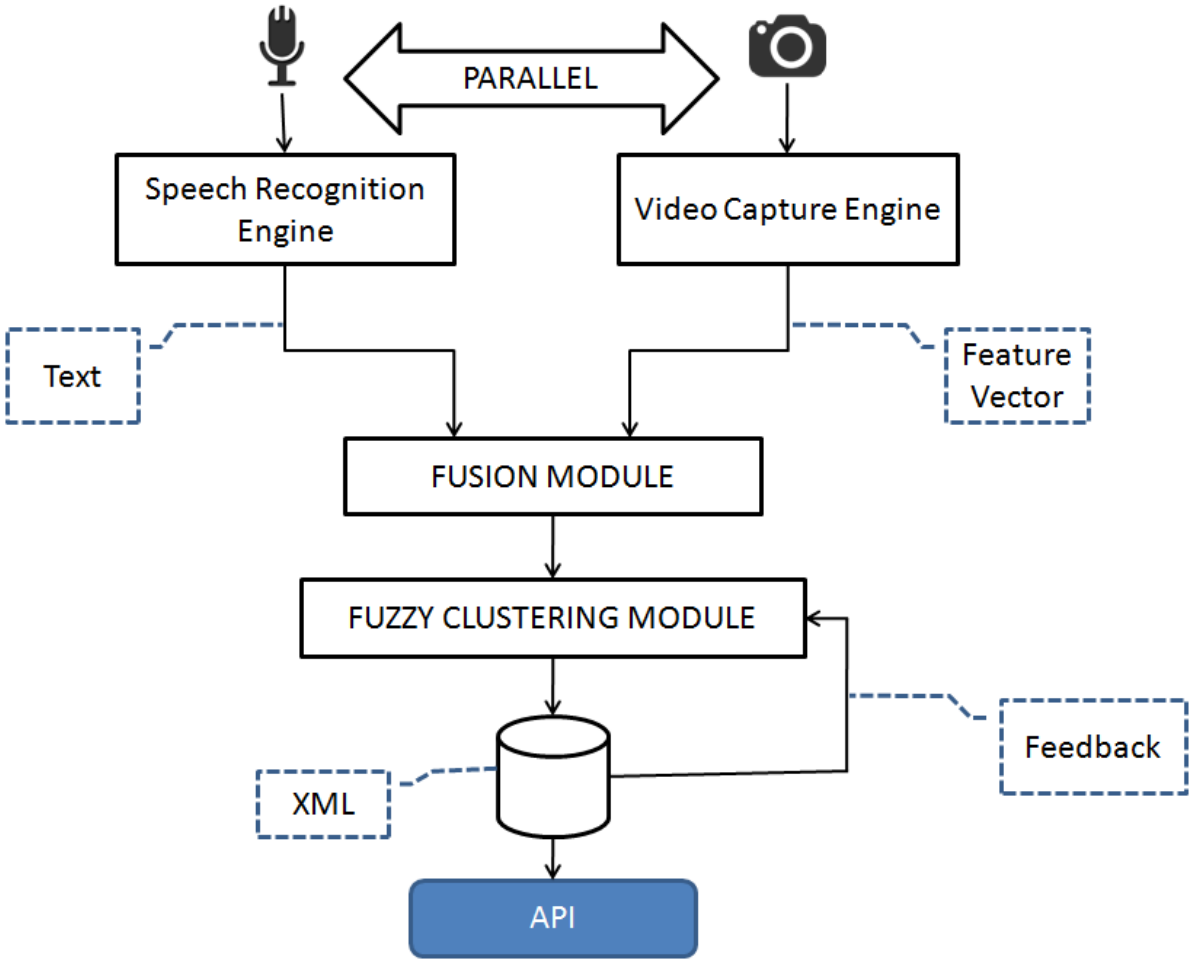


Fig: 1. Architecture showing multithreaded processes and modules

BRAIN STROM OPTIMIZATION

In order to solve a very difficult problem, different people from different backgrounds get together to brain storm. Such methodology helps in generating a large number of ideas because of collaborative thinking. Great ideas originate because of interactive sharing of information. Brain storming focuses initially on bulk ideas generation then eliminating ideas of less importance. Here in our application brain storming is applied to generate new feature vectors and adding them as new objects as per their fitness (calculated by computing distance values).

The BSO algorithm [2, 3] first finds out random cluster centers then applies FCM (Fuzzy C-Means), then it finds best data points (best ideas) in each cluster formed. Next the algorithm generates new data points (new ideas) on the basis of experimentally derived probability attribute. Then for the complete set of data points, the algorithm either selects single cluster center or two cluster centers (again on the basis of experimentally derived probability values). Then for each selected cluster, it finds new data points on the basis of activation function like sigmoid function. Finally the newly generated data points are checked for the fitness among several clusters. If their fitness is as per threshold then the new data points are added or replaced in the data set. Brain Storming Process is illustrated in the following steps in **[Table-1]**.

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 1. BRAIN STORMING PROCESS STEPS

STEP I	Assemble a group of people from different backgrounds and disciplines.
STEP II	Produce several ideas as per rules in **[Table-2]**.
STEP III	Select certain number of owners of the problems to generate better ideas to solve the problem.
STEP IV	Follow the ideas generated in Step III with greater probabilities to engender new ideas as per the rules in **[Table-2]**.
STEP V	Again owners must select certain nearer ideas as done in Step III.
STEP VI	In this step random selection of objects is made. The looks and functionalities of the objects can be used to generate further new ideas as per rules in **[Table-2]**.
STEP VII	Inform the owners to again select better ideas.
STEP VIII	Final step deals with merging or replacement of newly generated ideas with old ideas.
STEP I	Assemble a group of people from different backgrounds and disciplines.

Table: 2. IDEA GENERATION RULES GIVEN BY OSBORN

Rule or Pattern 1	Suspend Judgement.
Rule or Pattern 2	Anything Works.
Rule or Pattern 3	Cross-Fertilize (Piggyback)
Rule or Pattern 4	Achieve Quantity

In a process of brain storming, generally there is enabler, a group of problem members (people) the brain storming of ideas, and various owners of the problems. The function of enabler is to enable the generation of idea by imposing the group to adapt the Osborn's 4 rules of generation of ideas in a process of brainstorming [4]. The Osborn's 4 rules are presented in **Table- 2** below. The enabler is not be implied in generation of ideas, but alleviating the process of brain storming only. The road map for choosing enabler is to have a good enabler who has prior experience but has less expertise on the background knowledge related to the problem to be solved and who can help in alleviation. The aim of this is that generated ideas should have less, if not zero, biases from the enabler.

EXPERIMENTS AND DISCUSSIONS

We carried out experiments using standard IRIS dataset. We used single iteration of Fuzzy C-Means Algorithm [11, 12] to create initial clusters. Then applied the BSO on created clusters to generate new data points. IRIS dataset consists of 3 classes called as Setosa, Verginica and Versicolor. The blue color (cluster_0) in below given figures represents Setosa, red color (cluster_1) indicates Verginica and yellow color (cluster_2) indicates Versicolor. Figure 2 represents clusters created after applying BSO. **Figure- 2** is plotting of membership values on the graph of PW (Petal Width) vs PL (Petal Length). The DUNN's index was observed to be 2.908. **Figure- 3** is plotted on PW vs SW (Sepal Width). **Figure- 4** is plotted on PL vs SW.

Yellow color cluster (cluster_1) is Setosa; Blue color cluster (cluster_0) is Verginica; and Red color cluster (cluster_2) is Versicolor. Points with black color boundary represent cluster centers. Fuzzy index used in the algorithm was considered to be 2. New data point generation is computed using Equation 1.

$$X_i^d = X_{selected} + \xi * \mathcal{N}(\mu,\sigma) \quad (1)$$

$X_i{}^d$ is the d^{th} dimension of the individual data point chosen to generate new individual data point, $X_{selected}$ is is the dth dimension of the newly generated data point. $\mathcal{N}(\mu,\sigma)$ represents the Gaussian random function with μ as mean and σ as variance.

THE IIOAB JOURNAL
ISSN: 0976-3104

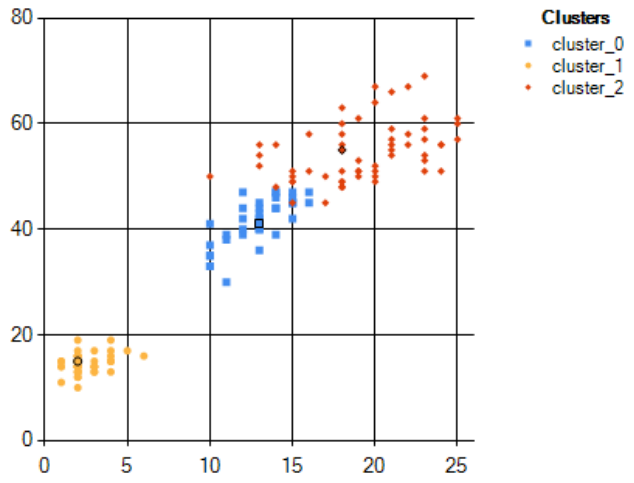


Fig: 2. Plot of membership values on graph of PW vs PL

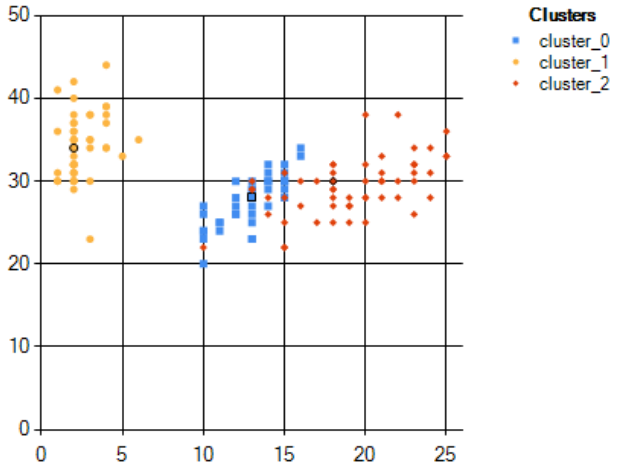


Fig: 3. Plot of membership values on graph of PW vs SW

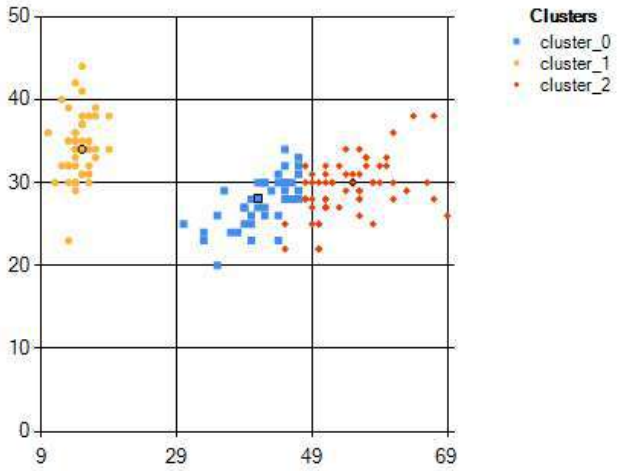


Fig: 4. Plot of membership values on graph of PL vs SW

THE IIOAB JOURNAL ISSN: 0976-3104

$$\xi = \text{logsigmoid}\,(\,0.5 * \text{max\_i} - \text{curr\_i} / k) * \text{rand}(0,1) \quad (2)$$

k represents change factor for logsigmoid() function's slope. In our experimentation we considered k as 20. max_i is the maximum number of iterations and curr_i represents current iteration number.

Our experimental setup consists of a quad core processor Intel(R) Core(TM) i7-4700MQ having capacity to reach 2.34GHz (each core), an integrated web camera and integrated microphone. We created a multithreaded application to train objects' images through camera and speech through microphone simultaneously. We created a dataset of objects containing 56 tuples, each tuple represents a particular object given as input to system via camera and microphone. First 20 tuples of the dataset consist of "faces", next 20 tuples consist of "hands" and remaining 16 were of "watches". We extracted following features as primary feature vector set: (1) Name of the object obtained from speech recognition engine (2) Average pixel intensity of red plane (3) Average pixel intensity of green plane (4) Average pixel intensity of blue plane (5) Average pixel intensity of canny edge plane (6) Number of keypoints extracted using SURF (Speeded Up Robust Feature) extraction algorithm and FREAK descriptor of OpenCV (7) Capacity of SURF keypoints vector (8) Illuminance reading abtained from LUX Meter. Apart from this primary feature vector set, we stored keypoints extracted from video frames of object using SURF as secondary feature vector set. **Figure- 5** shows the dataset in chronological ordering i.e. the numbering is done from left to right in each line of the figure. We applied Brain Storm Optimization algorithm over all attributes except the first attribute i.e. name of the object. **Figure- 6** represents the graph of membership values plotted on "average pixel intensity of red plane" vs "average pixel intensity of green plane". **Figure- 7** represents the graph of membership values plotted on "average pixel intensity of red plane" vs "average pixel intensity of blue plane". **Figure- 8** represents the graph of membership values plotted on "average pixel intensity of green plane" vs "average pixel intensity of canny edge plane". **Figure- 9** represents the final clusters of the objects (data points). Blue color in the figures represent "Faces", yellow color represent "Hands" and red color represent "Watches". There is 1 outlier in cluster 1 i.e. "Faces", there are 4 outliers in cluster 2 i.e. "Hands" and 4 outliers in cluster 3 i.e. "Watches". Remaining points lie in correct clusters. The aim of separation of primary feature vector set from secondary feature vector set is to minimize the computational complexity required to match a large number of keypoints of every data point with every other data points.

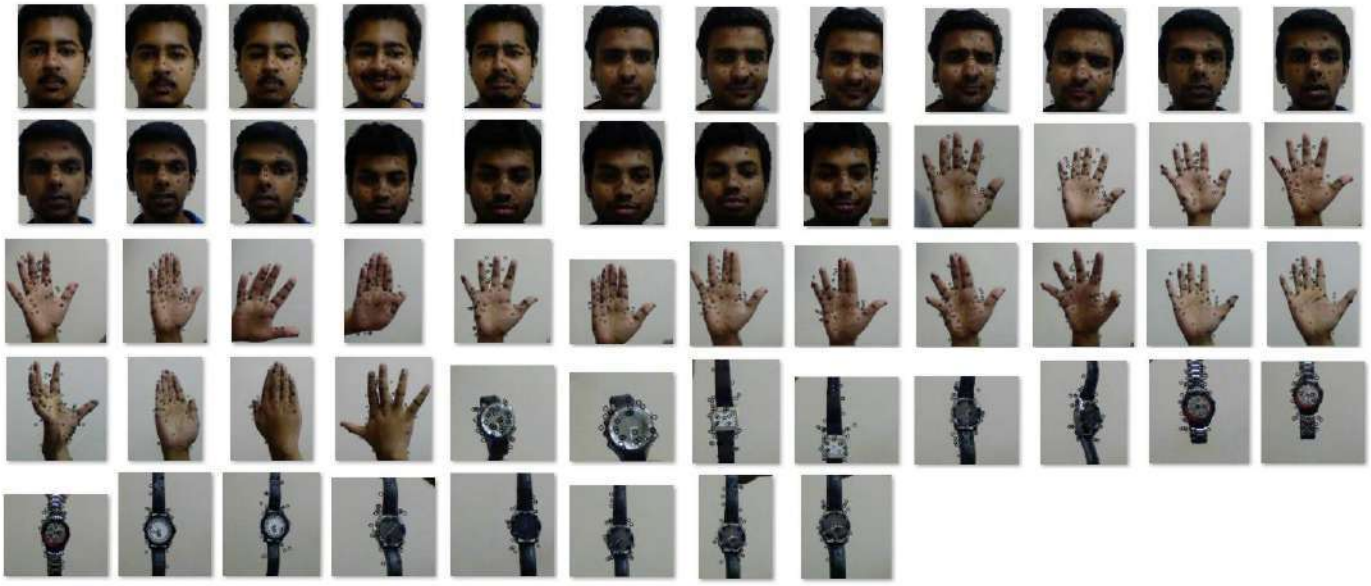


Fig: 5. Chronological Ordering of Data Set representing keypoints extracted from SURF algorithm

THE IIOAB JOURNAL
ISSN: 0976-3104

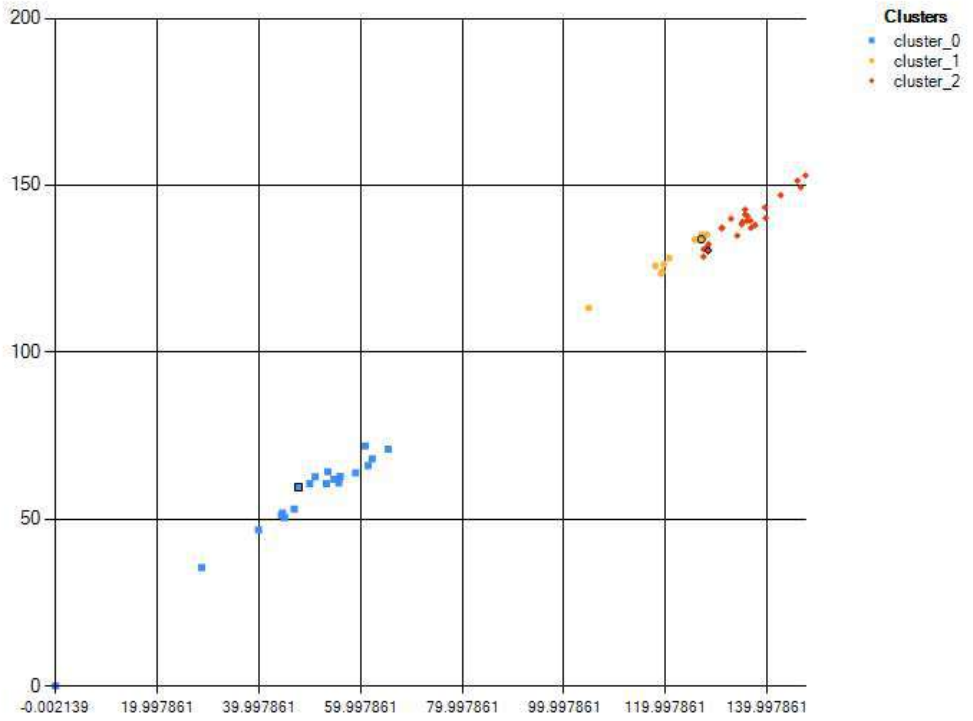


Fig: 6. Plot of membership values on graph of average pixel intensity of red plane vs average pixel intensity of green plane

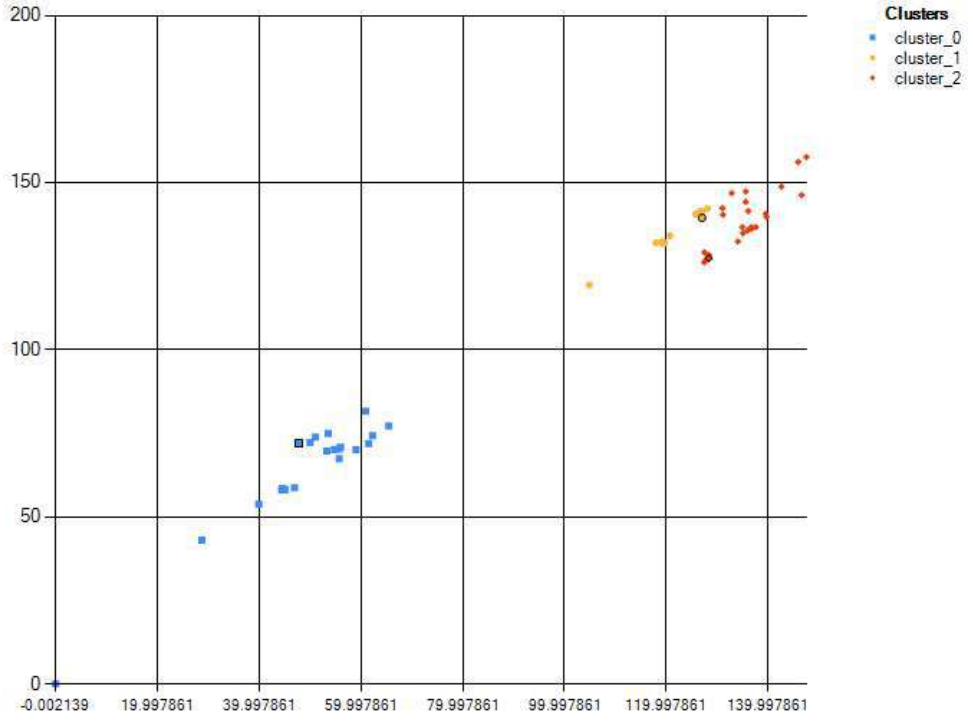


Fig: 7. Plot of membership values on graph of average pixel intensity of red plane vs average pixel intensity of blue plane

THE IIOAB JOURNAL
ISSN: 0976-3104

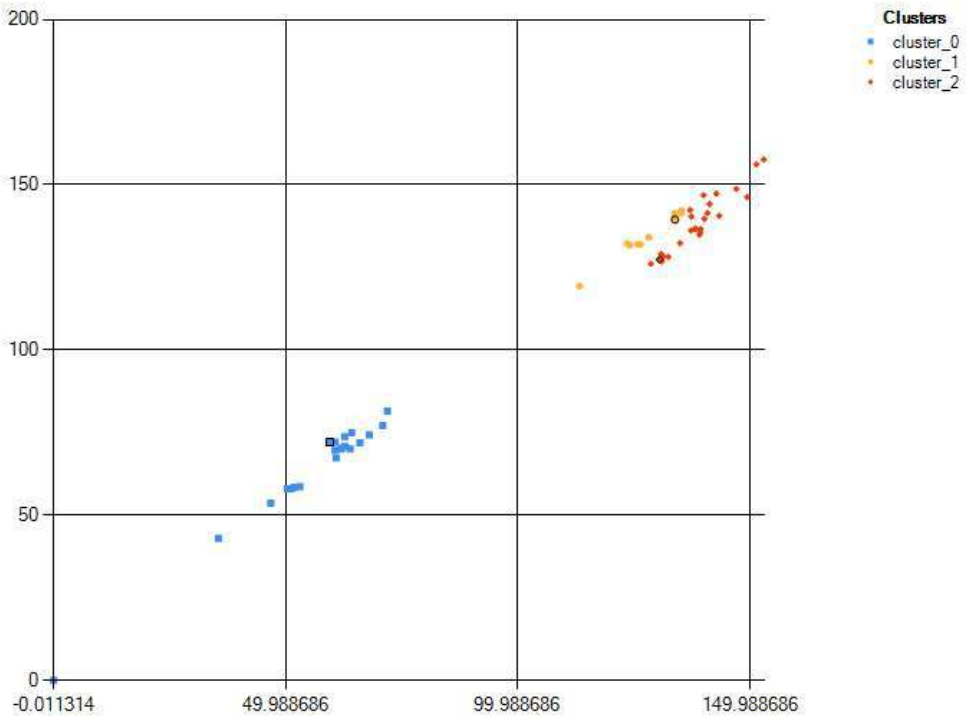


Fig: 8. Plot of membership values on graph of average pixel intensity of blue plane vs average pixel intensity of canny edge plane

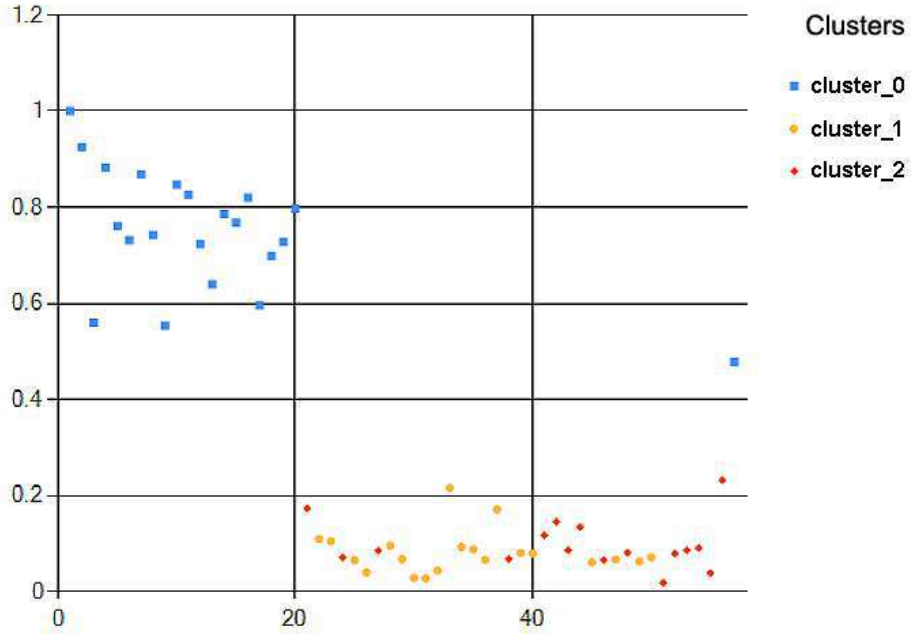


Fig: 9. Final Clusters - X Axis represents data point number, Y Axis represents membership value of data point for first cluster.

Since there are total 9 outliers among the dataset of 56 objects, the accuracy of the clustering system is computed to be 83.92%. Following table contains the data of 30 objects out of 56 objects, 10 from each category. In the following **table- 3**, Type 0 indicates "face", Type 1 indicates "hand" and Type 2 indicates "watch".

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 3. Features extracted from 3 types of objects

Type	Avg red intensity	Avg green intensity	Avg blue intensity	Avg edge intensity	Num keypoints	capacity	lux
0	47.86284	59.492586	72.006016	11.103232	46	63	43
0	51.042227	62.684196	73.82555	10.616559	45	63	43
0	60.908019	71.887839	81.503608	9.881328	66	94	43
0	50.014296	60.59837	72.156533	11.147912	54	63	43
0	53.570218	64.173977	74.944223	11.456442	63	63	43
0	54.747319	61.883129	70.098111	9.620038	38	42	43
0	55.986561	62.738898	70.834437	9.007336	44	63	43
0	53.298458	60.609684	69.636722	9.697591	40	42	43
0	65.459998	70.869158	77.104781	9.7257	67	94	43
0	55.705207	62.016665	70.288868	10.167026	53	63	43
1	104.872065	113.241252	119.362969	10.864501	60	63	43
1	147.529959	152.940015	157.625289	7.561342	72	94	43
1	145.91001	151.363548	156.174345	7.327393	74	94	43
1	126.614514	133.733009	141.312034	8.491996	63	63	43
1	132.854051	139.959366	146.840013	7.80492	67	94	43
1	128.094414	135.174938	142.229197	7.048164	69	94	43
1	119.110055	123.521824	132.29523	8.331642	53	63	43
1	119.293145	124.118101	131.703569	6.810675	87	94	43
1	135.600186	142.708154	147.335056	7.691119	76	94	43
1	127.035292	135.138251	141.312565	7.727407	79	94	43
2	146.597138	149.358333	146.258618	7.885305	44	63	43
2	135.942973	139.288143	135.654857	13.260759	33	42	43
2	136.67944	139.34492	136.652582	6.887209	46	63	43
2	139.538909	143.370271	140.620213	5.783088	40	42	43
2	137.617936	138.075293	136.660811	7.148363	71	94	43
2	136.799375	137.246696	136.17469	6.865798	62	63	43
2	139.680335	140.177728	139.716264	8.818973	73	94	43
2	134.933946	138.309244	136.6316	7.511284	47	63	43
2	127.459702	130.817212	129.102416	9.615451	78	94	43
2	128.349193	130.488091	127.352004	6.542601	69	94	43

Feature vector matching in FREAK descriptor [13] involves following steps:

1. The descriptor uses varied scales of Difference of Gaussians (DoG) which extract the object information. It contains binary symbols set.
2. FREAK descriptor simulates the topology of retina [13, 14].
3. Gaussian is used to smooth sampling points which are distributed on concentric circles where Gaussian kernel size is proportional to the radii of current sampling point's concentric circle.
4. Hamming distance is used a measure of similarity between sampling points.

Our algorithm using SURF with FREAK descriptor gives 22.3% more accuracy than using SIFT.

THE IIOAB JOURNAL ISSN: 0976-3104

CONCLUSION

Proposed algorithm of division of features into primary and secondary feature sets help in reduction of complexity of mapping large number of keypoints of all objects. Our experiments show that BSO based clustering over the dataset yields in 83.92% accuracy, which indicates that the system can strongly eliminate the problem of large number of keypoints mapping. As per the experiments conducted over a quad core processor with total 8 logical processors having capability of reaching 2.34GHz each, the total running time for matching SURF keypoints of a test object with that of complete dataset (56 objects) takes 4.2 minutes. Whereas our algorithm takes 1.43 minutes which corresponds to 65.8% reduction in time.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

We would like to thank our dean SCSE Department, Vice Chancellor and our parents because of whom, we had the opportunity to perform research in this field.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Kate Saenko, Trevor Darrell,"Object Category Recognition Using Probabilistic Fusion of Speech and Image Classifiers" Springer Berlin Heidelberg vol. 4892 ISSN 0302-9743 pp 36-47 2008

[2] Yuhui Shi, Brain Storm Optimization Algorithm, ICSI 2011, Part I, LNCS 6728, pp. 303–309, 2011.

[3] Zhi-hui Zhan; Jun Zhang; Yu-hui Shi; Hai-lin Liu.[2012] A modified brain storm optimization, Evolutionary Computation (CEC), 2012 *IEEE Congress* on , 1(8): 10-15

[4] Smith R. The 7 Levels of Change, 2nd edn. Tapeslry Press (2002)

[5] Tyler Streeter. Open Source Speech Interaction with the Voce Library".

[6] Tsontzos G, Orglmeister R. CMU Sphinx4 speech recognizer in a Service-oriented Computing style," Service-Oriented Computing and Applications (SOCA), 2011 *IEEE International Conference on* , pp.1,4, 12-14 Dec. 2011.

[7] Hong Kook Kim Cox, RV,Rose RC. [2002] Performance improvement of a bitstream-based front-end for wireless speech recognition in adverse environments," Speech and Audio Processing, *IEEE Transactions* on , .10(8):591,604, Nov.

[8] Sheikhzadeh H, Li Deng, [1994]Waveform-based speech recognition using hidden filter models: parameter selection and sensitivity to power normalization," Speech and Audio Processing, *IEEE Transactions on* , 2(1):80-89, Jan. 1994.

[9] Anderson S, Kewley-Port D. [1995] Evaluation of speech recognizers for speech training applications, Speech and Audio Processing, *IEEE Transactions on* , 3(.4):229-241, Jul 1995.

[10] Yao X, Liu Y, Lin G. [1999] Evolutionary Programming Made Faster. *IEEE Transactions on Evolutionary Computation* 3: 82-102

[11] Swarnalatha Purushotham, BK Tripathy. [2015]A Comparative Analysis of Depth Computation of Leukaemia Images using a Refined Bit Plane and Uncertainty Based Clustering Techniques", *Cybernetics and Information Technologies*, ISSN: 1314-4081 15(1):.126-146.

[12] Tripathy BK, P Swarnalatha, et.al.[2013] Rough Intuitionistic Fuzzy C-Means Algorithm and a Comparative Analysis, Proceedings of the 6th ACM India Computing Convention, COMPUTE '13, Aug 22-24, 2013 ACM 978-1-4503-2545-5/13/08.

[13] Wu Yanhai, Zhang Cheng, Wang Jing, Wu Nan.[2015] Image registration method based on SURF and FREAK, in Signal Processing, Communications and Computing (ICSPCC), 2015 *IEEE International Conference* on , 1-4, 19-22 Sept.

[14] Križaj J,Štruc V, Dobrišek S, Marčetić D, Ribarić S. SIFT vs. FREAK.[2014] Assessing the usefulness of two keypoint descriptors for 3D face verification, in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on ,.1336-1341, 26-30 May 2014

ABOUT AUTHORS

Utkarsh Gupta *is a Student of Master of Technology at VIT University in Vellore, Tamil Nadu, India. He has worked in the field of image processing and has published research work in the subject of Face Recognition and Multimodal Biometric System.*

THE IIOAB JOURNAL
ISSN: 0976-3104

Prof.Swarnalatha Purushotham *is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D. degree in Image Processing and Intelligent Systems. She has published more than 50 papers in International Journals/International Conference Proceedings/National Conferences. She is having 14+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.*

Prateek Chharia *is a Student of Master of Technology at VIT University in Vellore, India.*

ISSN: 0976-3104

SPECIAL ISSUE (SCMDSA)
Sree et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

ANALYSIS OF THE APPLICATION OF DATA MINING TECHNIQUES IN THE FIELD OF EDUCATION

Kaitha Sai Sree*, BMurali Manohar, Swarnalatha P

VIT University, Vellore, Tamil Nadu, INDIA

ABSTRACT

In olden days students exposed to teaching and learning was confined only to classroom. Students in the 21st century are connected well with lot of literature and learning from various open source platforms. Hence the students find it difficult to identify the right content for their learning. In this context data mining tools which are evolved in the past decade are of great use in filtering the large volumes of data and to pick the appropriate content for the student. In the previous years a considerable amount of time has been spent on student profiles and not on the factors responsible for the varied performances of the students. Data mining plays an important role in this field and can be used to mine relevant data for further optimization. This paper explains about how various techniques of educational data mining can be used to identify the student profiles and their behavior on the social media by taking many factors into consideration, to forecast the students' performance and also identify the best suited curriculum structure for them, to understand the pitfalls in teaching-learning environment etc. All these predictions can be used to help the 'at risk' student's category. The aim of the paper is to contribute to the literature on the application of data mining in the field of education by providing an overview of the various mining techniques that can be used, challenges faced in implementing the techniques and a comparative study of all the techniques.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26th-March-2016
Published on: 10th–June-2016

KEY WORDS

Knowledge Discovery in Databases (KDD); K-Means Clustering; Apriori Algorithm; Educational Data Mining (EDM); Feature Selection

***Corresponding author:** **Email:** kaitha.saisree2013@vit.ac.in; **Tel:** +91-8608556133

INTRODUCTION

Large volumes of data are added in every field of life on the internet by people across the globe through various platforms. It becomes a challenge for any person to identify the relevant information. The selection of useful information from the huge amount of data is known as Data Mining. Data Mining also refers to the application of algorithms for extracting patterns from data without the additional steps of Knowledge Discovery in Database (KDD) process. The results of this can be used to predict the most contributing attributes towards academia by making the good use of the enlarging amount of data in the field of education. Educational Data Mining (EDM) is concerned with developing new methods and techniques to understand the students ability of learning.[1]Due to the use of more computer based learning , the amount of data being generated has increased to a great extent , an interest to develop techniques for analyzing this data has also increased. These techniques are very useful in raising the educational standards and managements. It helps us to understand how people can become skilled in educational sector. Data Mining has many techniques like classification, Clustering, Association, Decision Tree, Neural Networks etc and the results obtained will help the educators to redesign their curriculum and to improve their learning methods, used to discover patterns which characterize learners across groups based on their choice of study and goals, used to attract the desired set of students, used to maintain student base and prevent dropouts by identifying students at risk and taking necessary actions etc. This paper deliberates about the various methods and algorithms that can be used to mine the data for educational purposes.

MATERIALS AND METHODS

Education is the process of learning, development of knowledge to enlighten oneself. The methods of education are teaching, discussing, training etc. Educational technology is the use of technological methods to improve the learning experience for the learners and the educators. These technologies have been changing and improving every day from many years. Initially a wooden board with lessons was used as the teaching medium, later after many decades the black board and chalk was introduced. This has been a means to teach the students for many hundreds of years. Distance learning was most preferred by

THE IIOAB JOURNAL ISSN: 0976-3104

the people before the PCs were introduced into the education system. In the early 20th century TV (television) were used as means of teaching , later in the mid-20th century the PC(Personal Computer) was introduced and the same is used as a means of teaching. Later many gadgets were developed for the ease of the learners like the calculator etc. By the end of the 20th century WWW(world wide web) came into existence.

Distance-Learning and E-Learning

Distance education or the distance learning is a means of learning where the students are not physically present in the classroom. The teaching can be done using the internet or it can partly have classes on site which is known as courses conducted through partly distance learning. Students were provided with the material by the institution which offers the distance learning programs. For clarification the student can reach through the student counselor in person at a specified place. With the advent of technology i.e. Internet, E-learning etc. the learning experience is brought into the door steps of aspirants to gain knowledge and enhance skills by various open source platforms like coursera, function space etc. These platforms provide a experiment , to interact with course provide by raising questions through discussion forum. Data mining plays a critical role for customising the courses according to the students learning pace and convenience. This results in customisation of learning modules rather than standardised modules.

Web Mining

The experience of distance learning is improved by applying the mining techniques to the Wold Wide Web(WWW) which is known as web mining.With the amount of data increasing on the internet , mining of this data can help us discover interesting patterns in distance learning and can help in customizing the courses for the students.

In the past many decades, the educational practices have hardly been changed. With the implementation of the distance learning there are some major changes that will take place like the relationship between the student and the teacher, the teacher will be a coach now and will lead the group of students i.e. the coach teaches the students and keeps track of their performance and gives feedback. The students will be more independent and tend to learn them and discuss with the peers and helps them develop individually. They make the actual use of the technology like the computer, internet etc.

These distance education platforms like udacity, coursera etc., gather large volumes of information which is generated automatically by the web servers and are collected and stored in the servers. Some other sources of the information can be taken based on the references the students browse. Such data can be helpful for the institution to improve the structure of the learning process by understanding the learning patterns of the students. [2]

The mining techniques like clustering can be applied to group students based on the skills they possess or other criteria. Other techniques can be used to predict the performance of the students. For example, if a student has attempted many practice tests these results can be used to predict the performance of the students in the final tests and estimate their performance to check for the areas which needs improvement and work on those areas and generate rules. By applying these rules, if we can find simple association rules it can helps us in improving the structure.

The data collected from the students can be improved further by asking them about their prior experiences and collect more extensive data and apply the mining techniques on that data to discover interesting patterns about the student behavior and improve the courses offered accordingly and customise them for the students. [3]

This web based learning methods collect a lot of information on user patterns etc. Data mining techniques can be applied on this data to predict the final grade of the students, and faculties can use this information to concentrate more on students who need help and advise them accordingly. This will be of more use in classes that have many people.

The data generated in the field of education has drastically increased after the E-Learning has been introduced. The data generated through this is very huge and plays a very important role if converted into knowledge in improving the learning experience for the learners and the educators. The sources of data include, data generated from the educational settings like, from the universities, intelligent tutoring systems etc. At a higher level, the data in this field needs to be explored in a proper way in order to discover new perceptions on how the data generated can be used.

E-Learning (Electronic Learning) refers to learning conducted via electronic media, generally the internet; it also refers to intentional use of networked information, communication technology information and communication technology in teaching and learning. This can be one form of distance learning, and it involves the use of internet to download materials, interact with the instructor etc. [4]

Common Attributes of E-Learning System

The common attributes for a E-Learning System can be identified in many levels and sections.[5]
Some common attributes of E-Learning System are mentioned in the following table **[Table-1]**.

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 1. Common Attributes of E-Learning System

S.No	Attributes	Description
1	Visited	If the unit, document or web page has been visited
2	Total_time	Time taken by the student to complete the unit
3	Score	Average final score for the unit
4	Knowledge_level	Student's initial and final level in the unit
5	Difficulty_level	Difficulty level of the unit
6	Attempts	No. of attempts before passing the unit
7	Chat_messages	No. of messages sent/read in the chat room
8	Forum_messages	Number of messages sent/read in the forum

Procedure for selecting E-Learning Resources

Step by step procedure to select E-Learning Resources

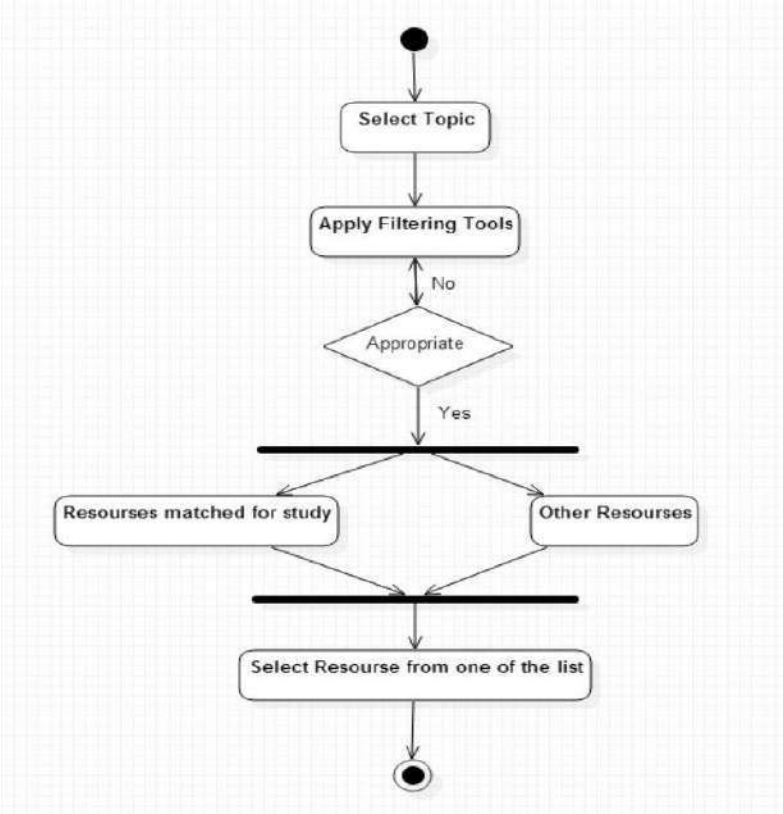
1.**Filtering:** The students are first asked to select a topic of choice, then some filtration rules are applied on the topic. Later depending on the results the topic is declared to be appropriate or not appropriate for the student.

2.**Prediction**: According to the topic chosen other related results are also predicted and shown.

3.**Decision**: Two rank lists are made one based on the topics that are suggested for the student to study and the other based on the knowledge level of the students and difficulty level of the topic.

4. **Adaptation**: After the lists are made, the students select a topic from one of the lists. [6]

The steps involved in selecting E-Learning resources is shown in the following flow chart in **Figure-1**



\Fig: 1. Step by step procedure to select E-Learning Resources

THE IIOAB JOURNAL ISSN: 0976-3104

E-learning can be of great advantage to students from rural areas to take up many courses which aren't available due to lack of resources like such as lack of qualified faculty, infrastructure in their academic institutions. Many working executives won't find time to go to college to take up additional skill development, certification courses so they make use these online portals to complete their courses. Many courses are offered by global institutions across the world where anyone interested can take the course.

With the scope of E-Learning being increased, many e-learning platforms have been developed causing the amount of data generated to increase. This generated data plays a crucial role in optimizing the e-learning methodologies. The data produced can become the knowledge and help in improving the e-learning process in all the possible ways.

Feature Extraction Models

Feature extraction models are used to create new feature from the existing original features whereas feature selection models are used to select the best subset of features from the given feature set.

Feature selection models are generally used in the field of data mining to explain in detail about the tools and techniques that help to reduce large amount of data into manageable size for further processing and analysis. These models are developed with an aim to discover the combinations of attributes that have high classification accuracy. Complex data have a large number of variables. When large number of variables are involved the memory then the computation power required will be high and may cause classification algorithms to over fit. Hence this model is used to reduce the unnecessary data into reduced set of features.[7]

Among the popular feature extraction models RELIEF and FOCUS algorithms were found to give better combinations of feature subsets than other wrapper approaches. RELIEF algorithm assigns a relevant weight to each attribute of feature vector by computing difference between selected test instance nearest hit and nearest miss training instances. [8]

$$w_i=w_i\text{-}diff(x_i,near\text{-}hit_i)^2+diff(x_i,near\text{-}miss_i)^2 \quad (2.2.1)$$

The wrapper model finds out the subset using n-fold cross validation method in which the data set is portioned into equal sized partitions, where the unique partition is the data set and the rest n-1 partitions is the training data set.

With the help of the induction methods John et al. has extracted feature subsets using machine learning. They have researched about improving the accuracy of the prediction, it majorly used two versions namely the forward and backward version. The forward version started with all the features where as in the backward version the features weren't known at all. [9]

Martin Sewell did a survey on feature extraction which explained that subsets containing minimum dimensions contribute most to the accuracy. [10]

A model has been proposed by Micheal Fire to predict the success of a student in a course by taking the previous data related to the course and applying network analysis methods. Parameters like personal information, friends grades etc.

Azwa Abdul presented a survey to predict student performance focusing on three elements i.e. parameters, method and tools. The extraction of patterns has been done using the Naive Bayes Classifier. [11]

Feature selection is a commonly used preprocessing step in machine learning. It is categorized into filter and wrapper methods. In the filter method, from a set of all the features a best subset of features is selected, then the algorithm is learned and that results in performance. In the wrapper method from the set of all the features selecting a subset involves both generating a subset and learning algorithm and this result in performance. In wrapper methods the computation time is high when the number of variable is more.

The student data sets generally have his/her personal data, details of parent's education, occupation etc. and academic attributes (grade in high school, higher secondary etc.) tot behavioral attributes (gender, cast etc.).They also assess based on the evaluation parameters like total number of answers which were correct and the ones that were correct in the first attempt etc. The parameters considered in the case of a drop-out student were difficulties in grasping during the class, improper selection of academic preferences, financial factors, feminine problems etc. These prediction results are used in increasing the capability of the students during the period of their studies.
A method has been proposed by the author to predict student performance using historic patterns and help them by contributing to their efforts. This method has the following phases.

Phases in Predicting the Students Performance

Phase 1: Current Scenario of Attributes

Attributes influencing were gathered from various sources. These Attributes were of two types
1. External Attributes: These are changeable with academic effort **[Table-2]** 2. Internal/Inherent Attributes: These are not changeable with academic effort **[Table-3]**.

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 2. External Attributes

S.No	Variable	Explanation	Values
1	Attendance	Attendance of the student during the semester	Percentage
2	Assignment and Project marks	Marks scored out of 20	Marks on a scale of 20
3	Internal Score	Marks scored in 2 internals out of 30	Marks on a scale of 30
4	No of subjects	No of internal papers which the student has appeared for out of 7 registered courses	Number
5	Computer	Knowledge of using a computer	Low/Medium/High

Table: 3. Internal Attributes

S.No	Variable	Explanation	Values
1	Gender	Sex of the student	Male/Female/Other
2	Cast	Cast f the student	Gen/ST/SC/OBC
3	Medium	The medium of instruction	Hindi/English/Other Language
4	Food Habit	Food Behavior	Vegetarian/Non Vegetarian
5	Settlement area	Residence of the student	Village/Town/City
6	Residence	Accommodation in hostel or home	Yes/No
7	Class10 grade	Grades of 10^{th} class	92-100 - A1 83-91 - A2 75-82 - B1 67-74 - B2 59-66 - C1 51-58 - C2 43-50 - D1 35-42 - D2 <35 - E
8	Class 12 grade	Grades of 12^{th} class	91-100 - A1 81-90 - A2 71-80 - B1 61-70 - B2 51-60 - C1 41-50 - C2 33-40 - D 21-32 - E1 00-20 -E2
9	Fathers Qualification	Education of father	12^{th} /UG/PG
10	Mothers Qualification	Education of mother	12^{th} /UG/ PG
11	Family status	Type of family	Joint/Small
12	Fathers Occupation	Job of the father	Govt/Private
13	Mothers Occupation	Job of the mother	Govt/Private
14	Family Income	Yearly income of the family	<50000 50000-100000 100000-200000 200000-300000 300000-400000 400000-500000 >500000
15	Percentage of dropouts	No of dropouts in previous year	%

Phase 2: Preprocessing Of Data Sets

In this phase of the feature extraction model deal mostly with preprocessing of the data sets. Some of the inherent attributes like gender, location, medium of education also seem to have an effect on the results. So these are taken into consideration during prediction of "at risk" students.

Phase 3: Working Model

This phase involves the classification and the fitness evaluation tests with the criterion that the "at risk" category students will obtain less than 40% of the total score. By applying Naive Bayesian Classification the prediction is done.

THE IIOAB JOURNAL ISSN: 0976-3104

Phase 4: Experimental Observations

In this phase we need to check how the internal and external factors affect the prediction. From the experiment it has been observed that external attribute i.e. students attendance was found to contribute the most for academic assessment which is then followed by internal assessment scores and assignment credit etc.

Results:

The objective of the experiment was to find out how the external attributes affect the academic performance of the students and the results of this prediction can help the management of the institution to take actions accordingly and improve their grades in upcoming examinations [8].

Clustering and Prediction

Prediction models are used to predict the event that is to occur in the future. It guesses the probability of an outcome based on the given input. Clustering is used to group the students according to similarities and this helps to improve the educational process and helps us in getting more specific models for the student's response. Artificial Neural Networks (ANN) is one of the best way to predict the performance of the student with a very simple and easy to use interface.

Student Performance Evaluation System (SPES)

Anju and Robin have done a research work on Decision tree classification algorithms like ID3 (Iterative Dichotomiser), C4.5 and CART were used to generate a decision tree from a data set, which is used to predict the performance of the student.ID3 accepts only categorical data whereas C4.5 which is the improved version of the ID3 accepts both categorical and continuous attribute and is highly efficient and accurate [12]..Ajai and Saurabh have proposed a method to extract knowledge using ID3, C4.5.Accurate result is not given by ID3 when there is noise and pruning is also not supported by ID3 whereas C4.5 is successful in identifying the students who are most likely to fail by using gain ratio [13].Kalpesh et al. discussed about the performance of students using ID3 and C4.5 i.e. the algorithms to generate the decision tree from data sets. They proposed and developed a system which can show the achievements of the students based on the previous performance[14].Brijesh Kumar et al has discussed about the decision tree methods for the classification which helps in identifying the dropout students and also the students who need special care and counseling[15].Narayana Swamy and Hanumanthappa have discussed about decision tree classification technique and predicted the academic success of the student enrollment which was very accurate[16].

The author suggested the various techniques which have been developed to see the data in different perspectives. The curriculum for the course is one of the main aspect which is required to improve the teaching of a particular course and make it more successful. The different perspectives suggested by the author can be used to develop a proper curriculum and there by improve the course. The author concludes saying that of the both algorithms ID3 and C4.5 which are used to identify the various categories of students , C4.5 algorithm is of higher superiority.

Apriori Algorithm and K-Means Clustering Algorithm

Apriori algorithm

Apriori algorithm is used to find out the frequent item sets in a transactional database. The principle of the algorithm says that the subset of a frequent item set must also be frequent. The most important use of this algorithm is in Market Basket Analysis.

Apriori algorithm majorly has 2 steps they are,

STEP 1: All the frequent item sets are found in this mode using the recursive mode.
STEP 2: The item sets which satisfy the confidence condition are acquired here.

Apriori algorithm is generally used on databases which contain transactions. Bottom up approach; breadth first search and Hash tree structure are used to count the candidate item sets appropriately. This is used when there is large item set property. It is easy to implement and can be paralleled easily. The main disadvantage of this algorithm is that it has to do many database scans and assumes that the database is memory resident. It is used in situations where we have the transactions of which students are taking particular subjects and the probability that they will take another subject as well in common can be found out using this algorithm.

K-Means Clustering Algorithm

Clustering is the process of partitioning the group of data point into small number of clusters. Lloyd's Algorithm also known as K-Means algorithm aims in partitioning 'N' observations or data objects into 'K' clusters in which each observation will belong to the cluster with nearest mean.
The center of the clusters is initialized first then the attribute that is closest to cluster of each data point. The position of the cluster is set to the mean of the data points which belong to the cluster, these steps are repeated until convergence.

This is used for classification of lot of information based on its own data. It is based on the comparison of numerical values that result the data based on the distance, hence it is considered as an unsupervised data mining algorithm. This will classify data into clusters and indicate which patterns belong to the class. This is the best method to be used, if the variables are huge and is

THE IIOAB JOURNAL ISSN: 0976-3104

comparatively faster than other hierarchical methods. The main disadvantage of this algorithm is the difficulty in K value prediction. This is used in situations where we have to classify the students or related things into clusters accordingly.

These techniques are used not only to analyse the students but also be used by the faculties to see the students previous records and modify the course content accordingly for easy understanding of the students and this can be used by the school authorities to prepare a curriculum that is suitable for all the students [17].

Crisp DM Methodology (Cross Industry Standard Process for Data Mining)

Crisp DM Methodology is a well-made methodology that provides a structured approach to plan a data mining project. In this process many automated analysis techniques are used to extract the knowledge. This model is independent of the industry in which it is applied. This describes the different approaches the experts commonly use to deal with problems in data mining. The Methodology is a wider concept compared to the KDD (Knowledge Discovery in Database)
This methodology is mainly divided into 6 stages. The order in which the stages have to be performed is not particular.

1. Business Understanding

This phase focuses on understanding the objectives and the requirements of the project from the business point of view, and then converting the problem into data mining problem definition. In the step the objectives of the KDD should be set depending on the field of business for which the analysis is being conducted. E.g. the financial sector uses this to detect fraud, the retails industry for ensuring customer satisfaction etc. This can also be used in the education system in predicting the success of the student, optimization of the courses, attracting new students etc. To be up to date in the market one should understand the latest changes being made in the field and should know the changes in the customer profiles. Social Media should be used in the marketing strategy of the companies. From the results of the analysis on the data obtained from the social media, the desired group of customers should be targeted and the offers should be changed according to the customer preferences.

2. Data Understanding

This phase begins with collection of initial data in a suitable format followed by activities to get a clear understanding about the data so that the problems can be identified and to detect the hidden information.
The average grade of the student is affected by their behavior on the social media.
-The activities the students generally do when they are on the social media sites.
-Most of the students who use social media for educational purposes are successful.

- The survey done by the author broadly divided the questions into 3 categories.
- The first set was related to the general information from which the demographic information can be obtained.
- The second part is related to how familiar the respondents are to the social media and how frequent they visit the social media and the purpose of networking and sharing.
- Third they examine the use of social media for the selection of college and education needs.

Additional questions related to their activities on blogs, forums etc. and the reasons for not being active on the social media. The data collected is entered into a sheet for the easy understanding and manipulation. According to the analysis done on the data collected, it was observed that the respondents would prefer websites that were more flexible in communication and that offer a more entertaining way to communicate with people. The common reasons for using social media were found to be to connect with friends, watch and listen to videos, communicate with fellow students regarding materials etc. , to follow school activities , share videos , songs etc. and other conclusions included expressing views through social media.

3. Data Preparation

In this phase all the activities related to constructing the final data set from the initial data is covered. The tasks in this phase are performed multiple times without any particular order to be followed. The tasks include data selection, data cleaning, data construction, data integration and formatting data.
Many attributes are removed for some of the following reasons,

- Part of the data collected is used for statistical processing about the future students general information.
- Most of the data for a particular attribute has the same value, since this does not bring any knowledge it can be removed from the further analysis.
- For the missing data substitute the default or global value, or substitute with the mean etc. to complete the incomplete data. The data subsets are selected based on the data mining goal.

4. Modeling

In this phase different modeling techniques are selected and used for building the data mining model. The method selected depends on the problem that has to be solved. There are many solutions available to solve a particular problem. The techniques used depends on the staff available, quality of the data, it also depends on the time and other factors. For the purpose of clustering K-Means clustering can be used.
Grouping students based on activities they perform has resulted in four major groups,
Group 1: This group has the students who regularly check the comments in the forums and comment on posts, videos but are not the active creators of the content on internet.
Group 2: These groups of people are active creators of content on the group.

THE IIOAB JOURNAL ISSN: 0976-3104

Group 3: These groups of students majorly use social media occasionally for the purpose of entertainment. They hardly leave any comments on the posts, videos etc. These students listen to songs, watch videos once in week approximately.
Group 4: Students in this often visit the blogs, forums but do not have the freedom to express their opinion on social media.

5. Evaluation

By this stage the model has been built and should be reviewed thoroughly before deploying the model. The main objective is to find if there are any issues that have to address.

6. Deployment

The project does not end with simple creation of the model. Depending on the requirements of the user, the model has to be deployed as needed by the user, it can be in the form of a simple report or can be as complicated as data mining process [18].

The results on using this methodology are, the essential step in this is to identify the various profiles of students as various users produce variety of groups and profiles. By understanding the behavior of the students on social media and internet there can be many advantages to this field as the data generated can be used for many purposes especially in publicizing about the institution. The results show that there are many distinct patterns among the students. There are students who use social media only for the purpose of entertainment and do not influence others whereas a portion of the students do get influenced by word of mouth which is also considered as publicity.

Data mining in higher education

Higher education is a way of learning provided by the universities through teaching, research work and other practical experiments. The quality of education provided by the universities depends on the faculties or the teaching staff that they recruit. The HR team of the university plays a very important role in selecting the members. The data mining techniques can be applied in this field to recruit the best staff. The mining techniques like clustering , classification , prediction play a very important roles in improving the performance of the organisation.In large organisations , performance evaluation is done on a yearly basis where the performance of each individual is done based on the qualities , abilities they possess. Based on the results of this test , the areas which need improvement can be figured , and the workforce is divided according to the knowledge they possess and assigns appropriate people to a particular job. This improves the overall performance of the education sector by achieving the organisational goals.

The performance evaluation of the employee is done taking all the factors that contribute to the performance into consideration. According to the general approach used the score of the employee is a simple number on given scale. But this score does not depict the various inter dependencies among the attributes used which could be more helpful in the performance determination. Few factors like responsibility, contribution in achieving the organisation goal etc. that contribute in measuring the performance are not directly measurable. Thus the performance evaluation is a tedious task if many attributes are considered to the evaluation. The author mainly focuses on the data mining techniques available, the model that supports both classification and prediction and its performance analysis.

According to the authors S.Anupama Kumar and M.N Vijayalakshmi , various mining techniques like classification and prediction can be applied on the student data to predict the performance of the students. This paper concludes that different methods or techniques have an advantage in different areas. [19]. M.Sukanya et al.the performance of the student can be improved in many ways by applying the mining techniques clustering, classification and prediction. The factors like psychological, social and personal play an important role in the performance of the student. By mining the information gathered can be useful to manage the next batch of students in a more efficient way and helps in improving the performance of the students [20].Data mining is a tool which helps on allocating the resources and the staff appropriately and helps in managing the resources. In higher education, the employees play a very important role and their management is very important for an effective performance. The performance of the employees is evaluated based on various contributing factors. The supervised and unsupervised learning techniques are used to build the models for performance prediction[21].Data Mining plays a very important role in Human Resources(HR) department as they can analyse the skills of the applicant and choose employees according to the need of the organisation[22].

With the changing methods of teaching and learning in higher education there are many issues that the field is facing. The employee should bear the role of being a mentor , educator , researcher etc. By considering only the teaching in classroom or the feedback given by the students which is biased , the performance of an employee cannot be predicted instead the performance should be predicted taking all the possible factors into consideration. About more than 50 factors like Attitude, Teaching skills, Communication skills etc. are considered in evaluating an employee. By evaluating all these attributes the performance of the employee is understood and are rated on a scale.[23]

THE IIOAB JOURNAL ISSN: 0976-3104

Comparison of the Techniques/Algorithms/Methodologies

Table: 4. Comparison of various technologies

Technique/Algorithm/Methodology Used	When is this technique used?	Principle	Advantages	Disadvantages
Apriori Algorithm	This is used when we have to find the frequent item set from the given transactions	The subset of a frequent item set must also be frequent.	▪ Easy to implement. ▪ Easily paralleled. ▪ Uses large item set property.	▪ Assumes the transaction database to be memory resident. ▪ Requires many database scans.
K-Means Clustering Algorithm	This is used when we have to classify a lot of information based on its own data.	Partitions 'n' observations into 'k' clusters in which each observation belongs to the nearest mean.	▪ Computes faster than hierarchical clustering if k value is small. ▪ Produce tighter clusters than hierarchical clustering	▪ Difficult to predict the K-Value. ▪ Different initial partitions can result in different final clusters.
Feature Extraction Model	This is used when we have to reduce the amount of resources required to describe the large set of data.	When the input data is too large and is suspected to be unnecessary then it is transformed into reduced set of features.	▪ The memory and computation power required can be reduced. ▪ Has a very important application in image processing.	▪ Sometimes can be computationally expensive. ▪ They may fail to remove the redundant features.
Crisp DM Methodology	This model describes the commonly used approaches that experts use to tackle problems.	This methodology breaks the process of mining into six major phases namely Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment.	▪ Makes large data mining projects faster, cheaper, more reliable and more manageable. ▪ Even small scale data mining can be benefited. ▪ Provides unique framework for guidelines and experience documentation.	▪ There are many risks at every stage that have to be taken care off.
ID3 Algorithm	This is used when we have to construct a decision tree.	Entropy and Information Gain is used to select the most useful attribute for classification. Entropy is used to calculate the homogeneity of the sample. Information gain is based on the decrease in entropy after the data set is split on an attribute.	▪ Understandable prediction rules are created form the training data. ▪ Builds the fastest tree. ▪ Builds a short tree. ▪ Whole data set is searched to create a tree. ▪ Finding leaf nodes enables test data to be pruned, reducing number of tests.	▪ Data may be over-fitted or over classified if used on a small sample. ▪ Only one attribute at a time is tested to make a decision. ▪ Classifying continuous data may be expensive as many trees have to generate.
C4.5 Algorithm	This is used when we have to generate a decision tree.	The concept of information entropy is used in building the decision tree, i.e. It is based on the expected value of the information.	▪ Handle noisy data, missing data better. ▪ More memory efficient. ▪ Deals with continuous and discrete attributes. ▪ Easy to implement.	▪ Does not work well on small training set. ▪ Small variation in data can lead to different decision trees.

All the methods, algorithms, methodologies mentioned above are compared and presented in **[Table-4]**.This gives a clear understanding of the different methodologies available in the literature till date.

Visual Data Mining Techniques

The presentation of the data in a way that can be understood by the users is called as visualization. In the traditional method of evaluation response is obtained by checking out the previous works and collecting related materials. [24]This response plays a crucial role in the evaluation and shouldn't be ill-judged. The main aim of this visualization is to improve the feedback between the

THE IIOAB JOURNAL
ISSN: 0976-3104

user and graphs and to make it more interactive so that the system of evaluation can be more adjustable, dependable and user friendly.

A model has been proposed by the author, which follows a multi-tier architecture including the application layer, middle layer and the data layer. The main functions of this model are as follows.

Step1: The data is obtained by interactive means, and then it is sorted interactively and expressed in view.
Step2: The algorithms like decision trees, clustering, and association rules can be used according to the requirement. New algorithms can be added, if they support the format. Generally C++ is used to compile these algorithms as it is more stable.
Step3: The large amount of data is expressed so that the users can understand the data and can find out the hidden information. The visualization technologies used are
1.Column Graphs - This graph tells about the situation of an element at different points of time. The horizontal axis has the time and the vertical axis has the changing values. E.g. the performance of the students in different years can be depicted by this graph.
2.Scatter Graph - These graphs show how much one variable is affected by the other variable. The value of one variable shows the position on the horizontal axis, and the other variable gives the vertical axis position. Taking these two coordinates the point is marked on the graph. E.g. The relation between the capacity of the lung of a person and how long the person can hold his breadth.
3.Parallel coordinates - In this method of representation the n dimension space can be shown on a two dimension plane with n-parallel coordinate at equal distances.

By representing the data using these methods, the role of the users can be utilized to a maximum extent and unwise decisions can be reduced making good use of the knowledge possessed by the users [25].

CONCLUSION AND SCOPE FOR FUTURE STUDY

This paper has examined the concepts of data mining in the field of education and simultaneously gave the overview of the application of various mining techniques on how the data can be used in this field. Though a sizable amount of research is being carried out in this field , there is wide area which still has to be uncovered. Like the experimental data collected in many of the researches is just a small amount and the results obtained are only for that. To get a clear view these techniques have to be applied on huge amount of data collected intensively by interviews, questionnaires etc. so that the results can be more apt.

Researchers can focus on the application of data mining techniques in the field of education, so that it can help the educators to develop better curriculum's for the students. Apart from the internal and externals attributes mentioned many other parameters can be taken into consideration and used for the prediction of more detailed information about the student. This prediction can be used to advise the student on the choice of his/her major based on the parameters. They can also focus on the weighing the data mining techniques so that it can be useful to both the educators and learners. In the future the researches can collect extensive data by making questionnaire and then apply the techniques mentioned above to get the valuable results. The current research will require more in depth study to discover more valuable rules to get a clear picture of the applications of data mining in the field of education for its optimization .Furthermore this can be used to make groups among the students in the easier way, identify hidden patterns, find undesirable behavior of students etc. The results of all these can help the faculty to improve relationship with the students and help them reach out to students who are in need of their help. Data mining can be a part of the technologically advanced techniques. The authors would like to propose a hybrid algorithm in the future which will combine the existing clustering algorithms to reduce the gaps identified in the paper.

FINANCIAL DISCLOSURE
This research is self assisted financially.

ACKNOWLEDGEMENT
The authors thank VIT University, Tamil Nadu , India for providing all the facilities required.

CONFLICT OF INTERESTS
Authors declare no conflict of interest.

THE IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] RamandeepKaurBhinder el al. [2015] A Review on Using Cryptography Techniques for Securing User Data in Cloud Computing Environment. *International Journal of Computer Science & Communication (*IJCSC),.6:83–86.

[2] NiteenSurv et al. Framework for Client Side AES Encryption Techniques in Cloud Computing. *International Advance Computing Conference (IACC)*, 525– 528.

[3] Periyanatchi S, Chitra.K. [2015] Analysis on Data Security in Cloud Computing-A Survey. *International Conference on Computing and Intelligence Systems* 04:1281 – 1284.

[4] LovepreetKaur et al. [2015] A Survey on the Encryption Algorithms in the Cloud Security Applications. International journal of Science Technology & Management (IJSTM), pp.1– 9.

[5] Neha A Puri et al. [2014] Deployment of Application on Cloud and Enhanced Data Security in Cloud Computing using ECC Algorithm. 1667– 1671.

[6] Thiyagarajan B, Kamalakannan R. [2014] Data Integrity and Security in Cloud Environment Using AES Algorithm. Information communication and Embedded Systems. 1– 5.

[7] CharanjeetKaur et al. [2015] Data Security Algorithms In Cloud Computing: A Review. *International Journal For Technological Research In Engineering* 2:372– 375.

[8] Sana Belguith et al. [2015] Enhancing Data Security in Cloud Computing Using a Lightweight Cryptographic Algorithm. *The Eleventh International Conference On Autonomic and Systems*. 98– 103.

[9] Tembhurne S et al. [2015] An Improvement In Cloud Data Security That Uses Data Mining. *International Journal of Advanced Research in Computer Engineering & Technology* 4: 2044– 2049.

[10] Nikhitha K, Navin K S. [2015] A Survey On Various Encryption Techniques For Enhancing Data Security In Cloud. *International Journal of Advanced Research Trends in Engineering and Technology* 194– 197.

[11] Rashmi S, et al. [2015] Architecture for Data Security In Multi-cloud Using AES-256 Encryption Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication* 157-161.

[12] Masthanamma V, et al. [2015] An Efficient Data Security in Cloud Computing Using the RSA Encryption Process Algorithm. *International Journal of Innovation Research in Science, Engineering and Technology* 4: 1441– 1445.

[13] SaiSindhuTheja R et al. [2015] Data Security in Cloud for Medical Sciences using AES 512-bit Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication* 1746– 1749.

[14] Nasrin K, ZurinaMohd. [2014] A Framework Based on RSA and AES Encryption Algorithms for Cloud Computing Services. *IEEE Conference on Systems, process and Control* pp. 58-62.

[15] Sugumaran M et al. [2014] An Architecture for Data Security in Cloud Computing. 2014 World Congress on Computing and Communication Technologies. pp. 252– 255.

[16] Arockiam L, Monikandan. [2014] Efficient Cloud Storage Confidentiality to Ensure Data Security. *International Conference on Computing Communication and Information*. 5:1– 5.

[17] Anuj Kumar et al. [2014] Cloud Data Security using Authentication and Encryption Technique. *International Journal of Innovative Research In Technology* 1: 388– 391.

[18] Vanya divan et al [2014] Cloud security solution: comparison among various cryptographic Algorithms. *International journal of advanced research in computer science and software Engineering* 4:1146– 1148.

[19] pradeep Kumar et al [2014] An authentication approach for data sharing in cloud environment for dynamic group. International conferences on issues and challenges in intelligent computing techniques(ICICT)9:262–267.

[20] Meenakshi et al. [2014] Data security analysis in cloud environment. *International journal of innovations & advancement in computer science* 2:14– 19.

[21] Aized Amin Soofi et al [2014] Encryption Techniques for cloud data confidentiality. International Journal of Grid Distribution computing. 7:11–20. .

[22] Vishwanth, S.Mahalle et al [2014] Enhanced the data security in cloud by implementing hybrid(RSA&AES)encryption algorithms. *International conference on automation and communication*. pp.146–149.

[23] Prashantrewagad et al. [2013] Use of digital signature with Diffie Hellman key Exchange and AES encryption algorithm to enhanced data security in cloud computing. *International conference on communication systems and network technologies*. 3:437-439.

[24] Ching-Nung yang, jia-bin lai. [2013] Protecting data privacy and security for cloud computing based on secret sharing. *International symposium on biometrics and security technologies* 7:259–266.

[25] ParsiKaplana ,sudha. [2012] Data security in cloud computing using RSA algorithm. *International journal of research in computer and communication technology,* vol.1.

[26] Rohit ,sunil [2012] A proposed secure framework for safe data transmission in private cloud. *International journal of recent technology and engineering*, vol.1

[27] Aleksandar et al. [2012] Data confidentiality using fragmentation in cloud computing. *International journal of networks and distributed system*, 1:85–90.

[28] Mohand M et al [2012] Enhanced data security model for cloud computing. International conference on Informatics and systems. vol.36, pp.cc-12.

[29] PachipalaYellamma et al. [2013] Data Security In Cloud Using RSA. 4'th International Conference on Computing Communication and Networking Technologies, pp. 1-6.

[30] Sonia sindhu. [2015] A survey of security algorithms in cloud computing. International journal of Advanced Research in Computer Engineering & Technology,.4(5):2368–2371.

[31] Ramesh K, Ramesh S. [2014] Implementing One time password based security mechanism for securing personal health records in cloud. International Conference on control, Instrumentation, Communication and computation technologies. pp. 968–972.

[32] Subbhiah S, Selva S [2015] Distributed data security for data prevention in cloud computing using One time password for user authentication. Journal of Environmental Science, *Computer Science and Engineering & Technology* 4:752–758.

[33] Priyanka Nema [2014] An Innovative Approach for dynamic Authentication in Public cloud: Using RSA, Improved OTP and MD5. *International Journal of Innovative Research in Computer and Communication Engineering*. 1(11):6697–6702.

[34] RandeepKaur, SupriyaKinger. [2014] Analysis of Security Algorithms in Cloud Computing. *International Journal of Application or Innovation in Engineering & Management* 3: 171–176

[35] SunithaSharma et al. [2013] Enhancing Data Security In Cloud Storage. *International Journal of Advanced Research in Computer and Communication Engineering* , 2: 2132–2134

[36] Vijendra et al. [2014] Data Storage Security in Cloud Environment with Encryption and Cryptographic Techniques. *International Journal of Application or Innovation in Engineering & Management,* 3: 209–213

[37] Jay Singh et al. [2012] Improving Stored Data Security In Cloud Using RC5 Algorithm. Nirma University International Conference on Engineering. pp. 1–5.

[38] DeepikaVerma, Karan Mahajan. [2014] To Enhance Data Security in Cloud Computing Using Combination of Encryption Algorithms, 2: 41–44.

[39] Honey Patel, JasminJha. [2012] Securing Data in Cloud Using Homomorphic Encryption. *International Journal of Science and Research*. 4, :1892–1895.

[40] Jayanthi M et al. [2014] Analysis on Secure Data Sharing using ELGamal's Cryptosystem in Cloud. *International Journal of Computer Science and Electronics Engineering*, 4:50–55.

[41] Raghul et al. [2015] Data Security in Federated Cloud Environment using Homomorphic Encryption Technique. *International Journal of Emerging Technology and Advanced Engineering*, 5:137–141.

[42] Vishal Paranjape, VimmiPandey [2013] An Approach towards Security in Private Cloud Using OTP. *International Journal of Emerging Technology and Advanced Engineering* 3:.683–687.

[43] Abhishektripathy, TarunGoyal. [2014] Cloud Data Security Using Encrypted Digital Signature & 3D Framework. *International Academic of Science, Engineering and Technology* 3:114–121.

[44] ShikhaChoksi. [2014] Comparative Study on Authentication Schemes for Cloud Computing. *International Journal of Engineering Development and Research*, 2: 2785–2788.

[45] HanumanthaRao et al. [2013] Data Security in Cloud using Hybrid Encryption and Decryption. *International Journal of Advanced Research in Computer Science and Software Engineering*. 3: 494–497.

[46] Roshani et al. [2015] Data Security in Cloud through Confidentiality and Authentication. *International Journal for Scientific Research & Development* 3: 1735–1738.

[47] Dimpi Rani, Rajiv. [2014] Enhanced Data Security of Private Cloud Using Encryption Scheme with RBAC. *International Journal of Advanced Research in Computer and Communication Engineering* 3: 7330–7337.

[48] Pradeep et al. [2012] Enhancing Data Security in Cloud Computing Using 3D Framework & Digital Signature with Encryption. International Journal of Engineering Research & Technology. 1:1–8.

[49] Ranu S, Hasna. [2015] Biometric Based Approach for Data Sharing in Public Cloud. *International Journal of Advanced Research in Computer and Communication Engineering*. 4:95-97.

[50] SuchitaKolhe et al. [2015] Five-Level Authentication Security in Cloud Computing. *International Journal for Research in Emerging Science and Technology,* 2:116–118.

[51] Sasi E, Saranyapriyadharshini.[2015] Secured Biometric Authentication In Cloud Sharing System. *International Journal of Computer Science and Mobile Computing,* 4: 572–577.

[52] Sudhansu & Biswaranjan [2014] Enhanced data security in cloud computing using RSA encryption and MD5 Algorithm. International *Journal of Computer Science Trends and Technology* 2(2):60–64.

ABOUT AUTHORS

Ms. Kaitha Sai Sree *is a 4th year , M.Tech Software Engineering student , in the School Of Information Technology , VIT University , Vellore.She interned at two major stat ups Axinovate Technologies and Think Tankers Innovative Solutions at Hyderabad where she worked on Back-end technologies and Mobile App development.Her current research interest include Data Mining , Data Analytics and Big Data*

Dr. B. Murali Manohar, Senior Professor - VIT Business School*, V I T University.He was a Visiting Professor for six months at CMIS, University of Cologne, Germany sponsored by DAAD. He completed the UNDP assignment for a period of 2 years- employed by Ministry of Education, Govt. of Ethiopia.at Debub University.he is actively involved in the activities of NBA accreditation, ISO –9002 from DNV, Netherlands and Deemed University status from the ministry of HRD, Govt. of India at VIT. He pursued his Ph. D in* ***E-Commerce.****He is having 2 years of Industrial experience and more than 20 years of teaching experience both at UG and PG level Management Programs.Received Fellowship from Rotary Int, U S A to visit East Yorkshire, U K.He published more than 65 papers in International Journals/International Conference Proceedings/National Conferences. He has reviewed 3 Management books and published the same.Two research scholars have completed their Ph. D under his guidance and five more Ph.D scholars are pursuing their research.Currently a reviewer/member for many of the reputed International/National Journals.*

THE IIOAB JOURNAL
ISSN: 0976-3104

Swarnalatha Purushotham *is an Associate Professor, in the School of Computing Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Systems. She has published more than 57 papers in International Journals/International Conference Proceedings/National Conferences. She is having 15+ years of teaching experiences. She is a senior member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE.She is an Editorial board member/reviewer of reputed International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.*

ISSN: 0976-3104

SPECIAL ISSUE (SCMDSA)
Reddy and Pricilda

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A REVIEW ON THE CONCEPT OF SENTIMENT ANALYSIS AND ITS ROLE IN MARKETING STRATEGIES FOR E-COMMERCE

Bhanu Sree Reddy* and Uma Pricilda Jaidev

VITBS, VIT University, Tamil Nadu, INDIA

ABSTRACT

Sentiment Analysis is considered as a critical research technique for collecting and analyzing textual data that is available on social websites. SA is also used as research method that systematically evaluates the consumer opinions in real sense in an efficient and effective manner. Marketing experts can collect rich data on attitudes and opinions about the products and services which help them in taking a relook at features which are positive and negative and use this as feedback system to rectify anything. By identifying the polarity of opinion either positive or negative will help businesses to identify right strategies to make their brands appreciated better by consumers. This review paper focuses on evolution of sentiment analysis, synonyms and why it is used and how research has happened related to how SA could be used in e- commerce.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26– March-2016
Published on: 10th –June-2016

KEY WORDS

Sentiment analysis, Opinion Mining, Market research, polarity of opinions

***Corresponding author: Email:** dbhanusreereddy@vit.ac.in **Tel:** +91 8903311849

INTRODUCTION

Even though Sentiment Analysis is looked at as active tool in data mining, web mining and text mining in Computer Science, its role is business especially in e-commerce is very critical. Basically SA is a type of natural language processing (NLP) for understanding the attitude or opinion of the people about a particular product or service or even a topic [1]. Due to constant development in online media, users are seen to be more dynamic and consistent in giving out and sharing their opinions on products and services with several platforms such as blogs, product reviews, wikis and Twitter. This specific information could be defined as User Generated Content (UGC). UGC "refers to media content produced by users to share information and/or opinions with other users". In E-UGC both positive and negative information is separated and made available for information searchers to read. Furthermore, this E-UGC is more keenly available through internet for product review [2]. According to Floyd et al., (2014) E-UGC enables customers to obtain more elaborate information on a product. Many Businesses to Customer (B2C) websites offer video and photo upload features in their review sections to feel the pulse of the customers.

The growth of Social media such as Blogs, forum discussions, Twitter and other social networks has made Sentiment Analysis very important to the business decisions. Managing the huge volume of opinionated data recorded in digital form has become very critical to business houses in terms of product launch to feedback on the product preferences by customers. In marketing the Sentiment Analysis helps in judging the success of a product or service by reviewing the opinions posted by users in social media. This also helps in identifying the demographics which like or dislike particular features of a product.

Sentiment analysis is a part of Natural Language Process (NLP) that draws a tactics from information recovery and computational syntax to identify opinions expressed in text. It is considered as a specific type of text mining [3], and it has been called opinion mining. Text mining is majorly used to detect the sentiment (reaction) of customers towards the products and services or in the related domain of sentimental computing. While the terms evaluation abstraction or review mining have also been applied, they are not always completely correct [4]. The

THE IIOAB JOURNAL ISSN: 0976-3104

main goal of sentiment analysis is to identify positive or negative overall attitudes or opinions toward a brand, product or service based on text comments [3].

ORIGINS OF SENTIMENT ANALYSIS

EUCLIDEAN DISTANCE

The rapid growth of Web information led to an increasing amount of user-generated content, such as customer reviews of products, forum posts and blogs. Sentiment analysis—automatically identifying the emotions conveyed by a text, and in particular distinguishing positive from negative valence—has become one of the most popular research areas in computational semantics both because of the interest of the field in the interplay among emotion and cognitive abilities, and because of its understandable applications (e.g., companies could analyze social networks to determine customer response to their products). Such research however requires collecting judgments about the valence of sentences and possibly lexical items, and simply asking subjects often results in low inter-annotator agreement levels. Sentiment analysis seems to originate from opinion extraction called synonymously as opinion mining, sentiment mining, and subjectivity analysis. Opinion mining also called as SA is a process of finding users opinion about particular topic or a product or a problem. A topic could be a news, event, product, movie a hotel in a specific location. User's opinion is expression of their emotion, attitude on anything in social media which gives opportunity to people to voice their feelings. A thought, view, attitude or comment without a reason but with an emotion is sentiment [5].

According to Khoo et al., (2012) data mining and machine learning algorithm are used to detect sentiment strength and sentiment analysis. These algorithms are rationally effective for sentimental analysis [6]. Even simple algorithms are shown to work well with large data sets, as in the case of the naive Bayesian approach. Most difficult algorithms are also needed for specific circumstances.

Taking opinion on a movie, a product, public sentiment or politics and prediction of election outcomes etc help in analyzing the positivity or negativity of people attitude. Emotions include feelings like anger, sadness, joy, fear, shame or proudness in human being. Moods are diffused non-caused low intensity long duration change in subjective feelings which are cheerfulness, gloomy, irritation, depression or buoyant. Attitudes are enduring affectively colored beliefs, dispositions towards objects or persons include feelings like liking, loving, hating, valuing and desiring. Personality traits are stable personality dispositions and typical behavioral tendencies influence how people attitudes are formed. They are nervousness, anxiety, and recklessness, morose, hostile, and jealous. SA helps in detecting the attitudes towards Holder (source) of attitude, Target (aspect) of attitude and type of attitude in a polarity between positive, negative and neutral.

Table: 1. Title: Hierarchy of Data Mining

THE IIOAB JOURNAL
ISSN: 0976-3104

LITERATURE REVIEW ON SENTIMENT ANALYSIS

A study was made on Sentiment analysis by Jeevanandam Jotheeswaran, Dr. S. Koteeswaran (2015) highlighted the classification of opinion mining as corpus based approach: determining words emotional affinity to learn their probabilistic affective scores from a large corpus. Dictionary Based Approach is used in lexical resources like Word Net to automatically obtain emotion-related words for emotion classification. Feature based opinion mining: Using Opinion Mining, a review is evaluated at 3 levels- document, sentence and feature levels. Evaluating a review at document level, the entire review is classified positive or negative based on opinions expressed in that review.Haseena Rehamath (2014) [7] has identified product purchase, quality improvement, market research, recommendation to other customers, opinion spam detection, policy making, decision making as application areas for which Sentiment analysis can be used effectively.

Darena and Burda, (2012) derives, coarse-grained and fine-grained are the two ways to conduct sentimental analysis in the field. Aim of Coarse-grained sentiment analysis is to predict the overall sentiment opposition of a document by using statistical method, unsupervised, semi-supervised or supervised machine learning methods (Moghaddam and Ester, 2010; Turney, 2002; Chen et al., 2012; Kim and Lee, 2014; Sayeedunnissa et al. , 2013). Kanayama and Nasukawa, (2012) describes, Fine-Grained sentimental analysis is used to spot the sentiment strength and its opposition towards particular product features in word level. Machine Learning and semantic analysis are the two common methods used for sentimental analysis (Pang et al., 2002; Zhang et al., 2011; Yu and Hatzivassiloglou, 2003; Ding et al., 2008). From the former unsupervised method attains the classification accuracy of 74% on average for 410 reviver's view, where the accurateness of movie, bank, and automobile reaches 66% and 80% respectively.

There will be a change in the performance depends upon the domains (Turney, 2002). According to Thelwall et al., (2010) the senti-strength can detect the positive feeling at 61.03% and negative feelings at 73.56% accurately in Myspace short text. Hanshi Wnag et, all (2014) have proposed a feature based Vector model for Chinese reviews to understand the sentence structure and content to solve ambiguity of emotions which helps in identifying polarity of the opinion. They have successfully used this model to study opinions on E-journal, electronic devices and hotel service. [8]

Ravendra Ratan Singh Jandail (2014) [9] applied different levels of SA. He used document level, sentence level, entity and aspect level to study positive and negative, interrogative, sarcastic, good and bad functionality, sentiment without sentiment, conditional sentence and author and reader understanding points. Yajun Deng et all [10]vhave presented a dimension based Sentiment Analysis modal for E-Commerce reviews where dimensions mapping and sentiment word disambiguation are main challenges. They have used a real database of 28 million e-commerce product reviews. Svetlana et all (2014) have used SA to detect the sentiment in short informal textual messages as tweets and short message service (SMS) and sentiment in a word or phrase. [11]

WHY SENTIMENT ANALYSIS

Chuang. Any one wants to buy a product or searching for a hotel to stay while touring, they usually check the blogs or reviews made by the consumers who have used the product or availed a particular service. There are three levels of opinion mining at document, sentence and phrase levels [12]. The negative, positive, sarcastic, interrogative, good and bad functionalities, conditional sentences and author and reader understanding viewpoints are studied thoroughly in these levels. The opinions posted by people are retrieved from review websites and are processed through phases like opinion retrieval, opinion classification and opinion summarization about a product or topic. The techniques used in opinion mining are supervised machine learning, unsupervised learning, and case based reasoning.

Online product reviews are done for electronic goods, real estate, movie reviews, automobiles and also Durable consumer goods. For example sentiment for a movie review could be taken from tracking sentiment flow from one sentence to the next [13]. Hanshi wang et al have done a study on evaluation of effectiveness of sentiment analysis in Chinese language related to customer reviews in three domains i.e., electric devices, E-journal and hotel. They have extracted 1375 documents which consist of both positive and negative sentiments.

THE IIOAB JOURNAL ISSN: 0976-3104

Table: 2. THE SIZE OF THREE SENTIMENT CORPORA

Date sets	Sentiment	Documents
Electronic devices	positive	356
	Negative	306
Hotel	Positive	180
	Negative	218
E-Journal	Positive	139
	Negative	158

E-commerce websites are been widely used by the customers to know about the product reviews and there service feedback, which is used to guide the customers to buy the product and services. Majorly 75% of the customers are visiting the e-commerce websites for review guidance (Gretzel and Yoo, 2008). E-commerce leads to the essence of eWOM: where the buying behavior of one customer is affected due to informal communication via internet reviews (Litvin et al., 2008) and eWOM users can freely share their views on various brands, products and services through online [14].Due to this eWOM consumer – generated content, there will be a direct influence on product sale (Zhu and Zhang, 2006). Thus, online reviews are more important to the consumers and sellers to know about the feedback on product and services. Various studies have been undergone on product recommendation by one customer to other via online reviews; such reviews will act as a most influence part in customers buying decision. According to Ghose et al., (2009) ratings or numberings are more important to review the feedback of the product and services. Ghose and Ipeirotis (2011) describe that user perseverance and influence of sale using online comment is found through readability, subjectivity and linguistic correctness of text comments which make a difference on it.

Text comments will be priorities and can be quickly added on the website, such that the user cannot handle the full range of comments, for example: recent comments of the product and services will lead the review display. It is impractical for the prospective consumers to go through all the large number of reviews about a product quality and services through online reviews. This problem can be solved, where the large number of reviews and comments are synchronized to a clear picture with numerical ratings will make the consumer much easier to analyses about the product quality and its services. These are the major disadvantage of text comments on online reviews. Where, sentiment analysis seems well suited to this task, but it has yet to be validated for converting text opinions into numbers.

THE ROLE OF SENTIMENT ANALYSIS IN SOCIAL NETWORKS

The role of sentiment analysis is analyzed in social networks like Facebook, Twitter, LinkedIn, you tube, Myspace, Blogs by in their study where data collection process, text pre-processing, classification algorithms and performance evaluation results are used to achieve accuracy levels in polarity of opinion [15]. In their study on Sentiment analysis in social streams, Hassan Saif, F et al have explained how SA tasks are more difficult in social streams than other classic text based opinion mining techniques. The social streams use vocabularies and expressions that are not in the formal language expressions and may lead to complex dynamics in the individual and community Opinions and sentiments [4]. It is an established fact that social communities have profound influence on consumer behavior which is accepted by market researchers and academicians. They can collect large amount of data and unobstructed with less cost and errors while reliability and validity are enhanced [13], [7]. Few examples of how sentiment analysis helps marketers to understand the polarity of opinions.

THE IIOAB JOURNAL
ISSN: 0976-3104

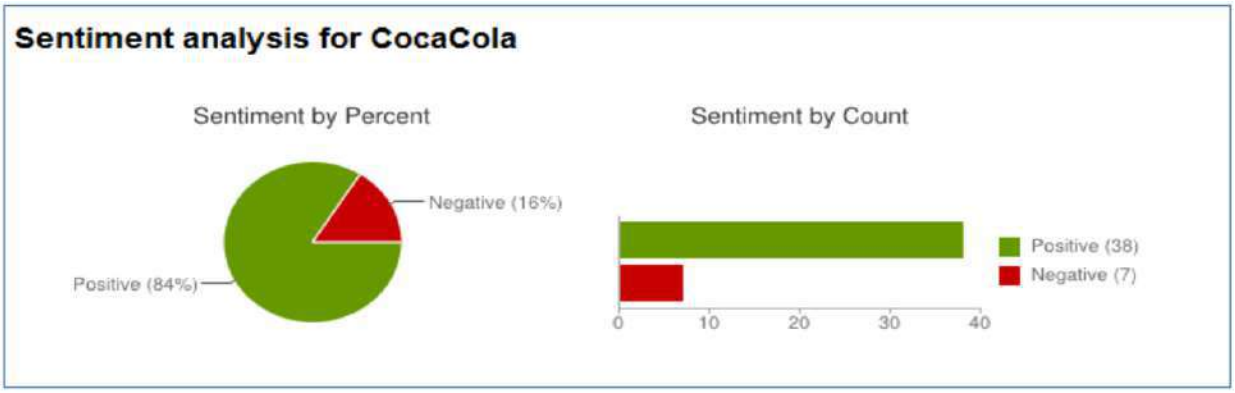


Figure 3: Summary of sentiments on a single product

Source: http://www.sentiment140.com/

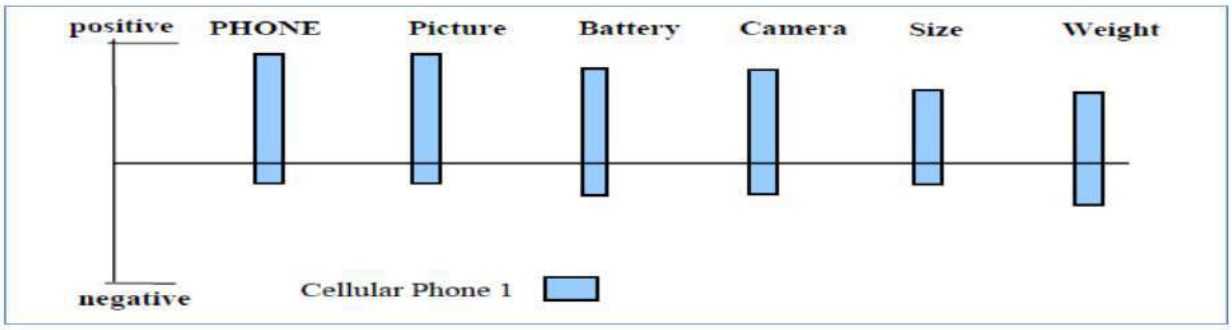


Figure 4: Summary of sentiments on features of a single product

(Source: Lu, 2010 p.8)

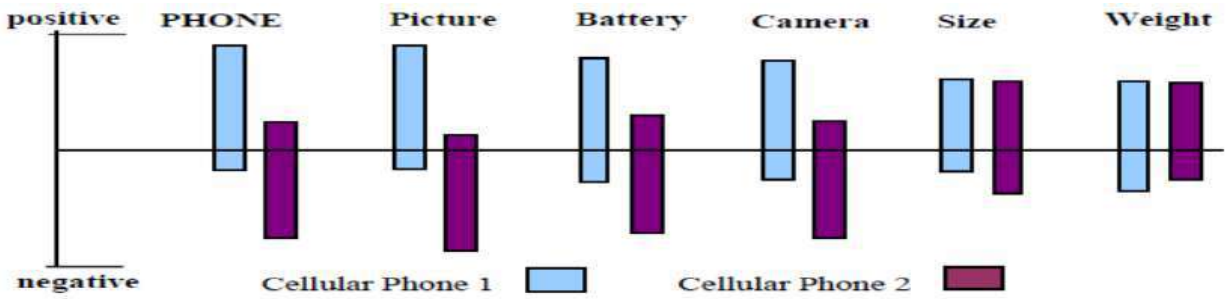


Figure 5: Summary of sentiments comparing two products

(Source: Lu, 2010 p.8)

SENTIMENT ANALYSIS IN MARKETING STRATEGIES IN E-COMMERCE

E commerce plays a very vital role in marketing now. The consumer behaviour in online buying decision is influenced by both internal and external factors. External influences include demographics, socio economics, technology and public policy, culture-sub culture, reference groups and marketing. Internal factors include variety of psychological processes like attitude, learning, perceptions, motivation, self-image and semiotics. The internal influences play a key role in sentiment analysis [16]. The traditional shopping depends on factors like one stop

THE IIOAB JOURNAL
ISSN: 0976-3104

shopping place to save time, free parking place and offer of wide variety of products at a comfortable price. But online shopping gives freedom to consumer in terms of convenience and interactive services.

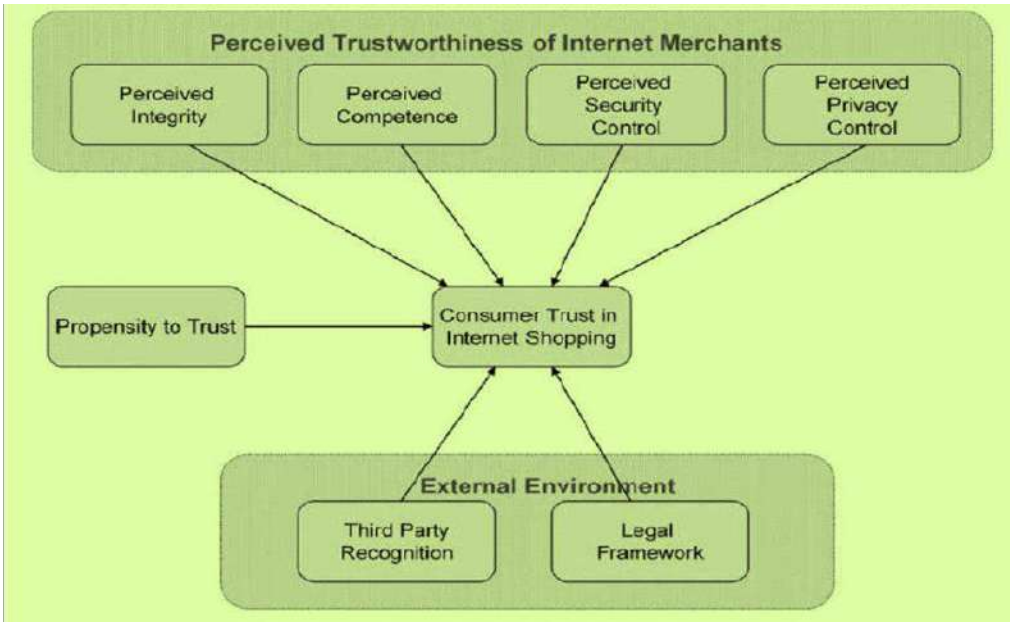


Fig: 6. An Integrative model on Consumer trust in internet shopping. Source: Arjun Mittal, Global Journal of Business And Management Studies P- 135

……

Many companies are now more interested to use sentiment analysis to develop marketing strategies by assessing and predicting public attitudes towards their brands. The company websites are gearing up to provide tools that track public view points by offering graphical summarization f trends and opinions in blogosphere [5]. The consumers usually prefer to compare specific features of different products as the Apple i-phone versus Samsung galaxy or to that extent compare sub features rather than giving opinions on whole item. The researchers should be able to construct comprehensive common knowledge bases to spot features and common sense bases to find polarity. The concept of "small room" is a negative for a hotel review but small que is a positive for Bank. Same way the sentiments should be read based on the context as Sentiment Analysis provides more insights to strategists to identify the gaps in their product or service.

E-Commerce is not only buying and selling of goods and services but also transmitting funds or data over an electronic network. E-Commerce in India has grown in sales touching nearly 20 billion dollars in 2015 with a growth rate of 700% in Asia Pacific region. Snap deal and Flipkart are two major players in India's e-commerce and now into mobile commerce. The Tweets were studied to identify the polarity of opinion on these two highly popular e-commerce companies which helped them to identify that the tweets made positive had a positive impact on the business they do[17]

The study made by [18] identified that more than 50% of the younger generation feel that social media plays a positive role in betterment and creating awareness about the latest trends in living styles. Social media sites are utilized by youngsters in terms of finding jobs, products for their needs, connect people from faraway places and also help in enhancing their skills. Hence Sentiment Analysis in social media sites is best bet for companies to connect with youngsters who form major chunk of their business.

The present generation customers spend more time on online to see the price variations of products and how face the negative messages have more impact on purchase intention rather than the positive messages [5]. The consumer intention to purchase is influenced by the blogs made many companies to move to create their own social media sites to host blogs. A customer experience is generated from a set of interactions between a customer and product, a company or part of its organization which provokes a reaction [8]. When the Word of Mouth

THE IIOAB JOURNAL ISSN: 0976-3104

Marketing Association (2014) asked consumers what sources "influence your decision to use or not use a particular company, brand, or product," 72% claimed reviews from family members or friends exert a "great deal" or "fair amount" of influence [13]

SCOPE FOR FUTURE RESEARCH

The rapid growth of online communication has attracted the attention of Market researches to the tremendous value that opinion mining provides to vendors in analysing their products feedback provided by the users. In many Western countries the companies use Data Mining and sentiment analysis vividly in their market research. They use Sentiment Analysis to create and maintain review and opinion aggregation websites where the system continuously gathers a variety of information from the opinions expressed on the net such as product reviews, perceptions on brands and image issues.

The reviews written by the customers on blogs or social media sites have economic value as their opinions of positive or negative will influence the buying decision of the prospective consumers. But many studies have limited their scope to the influence factor of those reviews but not focussed on the predictive value of those expressions [19].

Once, the concept of “Word of mouth” by the customers has been carefully researched by marketing analysts to track public sentiments on a product or a service. Now opinion mining provides them a reliable, rich and quick data regarding the same. More than Market researches this opinion mining was critically researched and utilized by the Information Technology departments by using various techniques like Domain Driven Data Mining (D3m), Review mining, lexical affinity etc.

Companies will depend more and more on developing opinion tracking system as it helps them in analysing the information from blogs and websites about the trends in customers’ opinions about products and services. Erik Cambria et al, have mentioned about the Multimodal Sentiment analysis which consists of linguistic, acoustic and video information. There is a great scope for further research on designing multimodal Sentiment analysis by corpora to find out broader sent of emotions and polarity of opinion [20].

CONCLUSION

Around 35 articles have been reviewed to understand the role of Sentiment analysis in designing marketing strategies for e- commerce. It is evident that in the era of social media networks the e-commerce is thriving because of netizens who spend more time on net to browse, find, review a product or service they want to purchase. In this process they go through the earlier experiences of customers who have already used the product or availed a service and read their reviews in various social sights.

The polarity of opinion of the customers expressed as messages is taken into consideration by companies to identify the opinions of customers which in turn is collected as data and analysed to understand what a customer is signalling the company indirectly. Hence sentiment analysis has become a very important research tool for businesses for designing their marketing strategies and it is done in a wise manner by lots of them on websites.

CONFLICT OF INTERESTS
Authors declare no conflict of interest.

ACKNOWLEDGEMENT
None.

FINANCIAL DISCLOSURE
None.

THE IIOAB JOURNAL ISSN: 0976-3104

REFERENCES

[1] Bhagyashri Ramesh Jadhav. Manjushri Mahajan. [2013] Review of Dual Sentiment Analysis, *International Journal of Science and Research* (IJSR) ISSN (Online): 2319-7064, pp-2323-2326

[2] Creswell John. [2007] Qualitative Inquiry and Research Design. Choosing Among Five Approaches. 2nd ed. Sage Publications Inc: California.

[3] Hanshi Wang, Lizhen Liu*, Wei Song, Jingli Lu. [2014] Feature-based Sentiment Analysis Approach for Product Reviews", *JOURNAL OF SOFTWARE,*. 9(2) 274 t- 279.

[4] M.Nick Haiji .[2011] A study of the Impact of Social Media on Consumers, *International Journal of Market Research*, 56(3)

[5] Ari, Aslihan Nasir, [2014] A critical review on online purchase intentions, Ezgi Advances in Business Related Scientific Research Conference 2014, 26-28th March, Venice, Italy.

[6] Haseena R Rehamath, oponint mining and sentiment analysis(2013), - Challenges and application", *International Journal of Application on Innovation in Engineering and Management*, 3(5):.401-403

[7] Davis, A. & Khazanchi, D. [2008] "An Empirical Study of Online WOM as a Predictor for Multi Product Category e-Commerce sales", Electronic Markets 18(2): 130-141.

[8] Devang Jhaveri, Aunsh Choudary, Lakshmi Kurup. [2015] Twitter Sentiment Analysis on E-commerce Websites in India,(2015) *International Journal of Computer Applications* (0975 – 8887) 127(.18):14-18.

[9] J Ashok Kumar, S Abirami, S Murugappan. performance analysis of the recent role of omsa approaches in online social networks, pp-21-32

[10] Neha A Kandalkar, Prof. Avinash Wadhe. [2008] Review Paper - Expressive Sentiment Analysis of Online Reviews , *International Journal of Engineering Research and General Science* 3(2) March-April, 2015ISSN 2091-2730, 695 www.ijergs.org

[11] Liu, Wei Song, Jingli Lu.[2014] Feature based sentiment analysis for product reviews, Hanshi Wang, Lizhen, *JOURNAL OF SOFTWARE*, 9(2): 274-279

[12] G Vinodhini et al, [2012] Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), 2(6).

[13] Arjun Mittal,[2013] E-Commerce and its impact on consumer Behaviour, Global Journal of Management and Business Studies. ISSN 2248-9878 3(2): pp. 131-138 © Research India Publications http://www.ripublication.com/gjmbs.html.

[14] Seyed Rajab Nikhashemi, Laily Paim, Saeideh Sharifi, [2013]The Effectiveness of E-Advertisement towards Customer Purchase Intention : Malaysian perspective, IOSR Journal Of Business Management e-ISSN: 2278-487X, Vol-10, Issue 3 (May-Jun2013 PP 93-104

[15] G Angulakshmi, R ManickaChezian. [2014] An Analysis on Opinion Mining: Techniques and Tools, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July, ISSN (Online) : 2278-1021 ISSN (Print) : 2319-5940 pp-7483-7487

[16] Melville, Wojciech Gryc,[2009] Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD"09, Paris, France. Copyright 2009 ACM 978-1-60558-495-9/09/06.

[17] D. Gruhl, R. Guha, R. Kumar, J Novak, and A Tomkins. [2005] —The Predictive Power of Online Chatter,‖ Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), pp. 78-87,.

[18] Erik Cambria, Bjourn Schuller et a.l [2013] New Avenues in Opinion Mining and Sentiment Analysis, 1541-1672/13© 2013 IEEE 15 Published by the IEEE Computer Society. Pp-15 -21

[19] Hongwei Wang Wei Wang. [2014] Product Weakness Finder- and Opinion – Aware System through Sentiment Analysis, *Industrial Management and Data Systems*, 114(8):1301-1320.

[20] Antony M. Hooper Maria Uriyo, [2015]"Using Sentiment Analysis to Review Patient Satisfaction Data Located on the Internet, *Journal of Health Organization and Management*, 29(2):221-233.

[21] Alen Yee Long chong Boyin et al,[2016] Predicting Online Product Sales Via Online Reviews, Sentiments And Promotion Strategies", *International Journal Of Operation And Production* Management, 36(4):358-383.

[22] Erik Cambria, Björn Schuller, Catherine Havasi, [2013] New avenues in Sentiment analysis and opinion mining" , Knowledge based approaches to Concept level sentiment Analysis, *IEEE Computer Society* pp- 15-21

[23] Ghulam Shabir , Yousef Mahmood Yousef Hameed , Ghulam Safdar , Syed Muhammad Farouq Shah Gilani,(2014), The Impact of Social Media on Youth: A Case Study of Bahawalpur City, , *Asian Journal of Sciences & Humanities,* Vol. 3(4).

[24] Han J, Kamber N, Pei, J. [2011] Data Mining : Concepts and Techniques, MORGAN KAUFMANN, WALHAM, MA.

[25] B Liu, M Hu, and J Cheng. [2005] Opinion observer: Analyzing and comparing opinions on theWeb,‖ in Proc. 14th Int. Conf. World Wide Web, 342–351

[26] Meena Rambocas , João Gama.[2013] Marketing Research: The Role of Sentiment Analysis, FEP working papers, n. 489 April 2013 ISSN: 0870-8541 pp- 1-21

[27] Pang B , Lee L. "Optioning Mining and Sentiment Analysis", *Foundation and Trends in Information Retrieval*. 2(1-2): 1-135.

[28] Ravendra Ratan Singh Jandail. [2014] A proposed Novel approach for Sentiemnt Analysis and opinion Mining, *International Journal of UbiComp* (IJU), 5, (.1/2)April 2014 Pp:1-10

[29] Rutilio Rotolfo Lopez Birbosa et al, [2015] Evaluating Hotels Rating Prediction Based on Sentiment Analysis Services", *ASLIB Journal of Information Management*, 67(4):392-407.

[30] Sentiment Analysis in Social Streams, Hassan Saif, F. Javier Ortega, Miriam Fern´andez, Iv´an Cantador,

[31] Thomas. [2014]The Relationship Between Fashion Blogs and Intention to Purchase and Word of Mouth Behavior, Thesis by Cassidy L. Vineyard University of Nebraska-Lincoln, , Spring 4-18-2014 pp-1-137

[32] Yi Mao and Guy Lebanon. [2006]Sequential models for sentiment prediction, In ICML Workshop on Learning in Structured Output Spaces.

[33] Yajun Deng, Lizhou Zheng, Peiquan Jin. [2014] Dimension based Sentiment polarity detection for E-Commerce Reviews, Advanced Science and Technology Letters Vol.45 (CCA 2014), pp.55-59, http://dx.doi.org/10.14257/astl.2014.45.11

THE IIOAB JOURNAL
ISSN: 0976-3104

ABOUT AUTHORS

Dr, D. Bhanu Sree Reddy *is in VIT Business School as Senior Professor in General Management area. Her teaching area includes Business Consulting, Business Plan Development, International Business and Logistics. She guided 4 PhDs successfully and another 3 scholars are completing the thesis writing. She published around 24 research papers in reputed journals including SCOPUS and presented papers in national and international conferences. She is mentor for International Programs of University of Michigan, USA and International Projects at the Business School.*

Dr. Uma Pricilda Jaidev *is currently an Associate Professor of Marketing at VIT Business School. Her teaching and research interests include Advertising and Promotions Management, Brand Management and Consumer Behavior. She has been teaching for more than 15 years now and has also occupied Administrative roles. She has also convened two International Conferences and has published in peer reviewed journals and edited books.*

ISSN: 0976-3104
SPECIAL ISSUE (SCMSA)
Vijayakumar et al

ARTICLE OPEN ACCESS

A NEURAL NETWORK CLASSIFICATION APPROACH FOR OIL SPILL DETECTION ON SAR IMAGES

Vijayakumar S[1]*, Swarnalatha P[1], and Rukmini S[2]
[1]*School of Computing Science and Engineering, VIT-University, INDIA*
[2]*Electronics and Communication Engineering (VLSI), AITS, JNTU-A, INDIA*

ABSTRACT

Synthetic Aperture Radar (SAR) is one of the coherent technique, that has been shown to have a great potential for marine surveillance applications such as oil spill and ship tracking detection. In this paper, we proposed an approach called a grey level co-occurrence matrix (GLCM) based texture feature, are applied on the SAR images to spill the oil from selected region. This uses artificial neural networks which classifies each pixel of selected region of interest (ROI) on SAR image('s). The performance development is studied by utilizing a time series of SAR images in a single ocean surface region acquired from SAR scene time series. Thereby, the remote sensing setup is trained on preliminary image of the time series and then pragmatic to consequent images from the radar satellites. After applying the classification method, the result of accuracy is at least 75% depending on the choice of our oil spill regime type. It is possible only with the radar satellite incidence projection views from the training dataset, is similar to that the classified image('s). Our proposed approach computational cost is moderately considering for classification operational procedure, near-real time service of SAR image processing.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26– March-2016
Published on: 10th –June-2016

KEY WORDS

Remote Sensing, SAR Imagery Analysis, Oil Spill, Neural Network, Classification, Statistical Decisions

***Corresponding author:** Email: vkchinna5a8@gmail.com

INTRODUCTION

In generally, Synthetic Aperture Radar (SAR) imagery technique is considering for all intentions and purposes used for controlled surveillance of oil spill monitoring on coastal surface areas. The SAR imagery technique is using two types of sensors called as airborne and spaceborne sensors. However, most probably space-borne SAR surveillance is comparably independent of day-night and cloud coverage conditions. While, utilizing the SAR airborne sensors cannot be adverse under coastal weather conditions, are may simply unavailable over remote Arctic regions. But, the spaceborne SAR system can be acquire images over named regions on a regular and reliable basis. In particularly, the radar satellites are using SAR technique (such as RADARSAT, ERS, ENVISAT, etc...) have been employed to investigating and controlling for the navigational purposes. This can be monitoring for environmental conditions and environmental changes, for example, ice drift, iceberg detection, ice concentration, weather changes, oil spill detection, ship wake detection, shoreline detection, marine surrounding object identification, water quality analysis, and soon [1] – [4].

In marine environment surroundings, the oil spills spread over the sea surface regions for varying the gradation of lager oil tanker accidents, especially this type of variations are occurred the coastal beaches. As of now the oil pollution problem is a major aspect in observing the sea surface regions and marine surrounding environments. In addition, the small-scale release of oil into the sea surface regions, is recognized as a slicks and while large-scale ones are called spills [5]. However, the accurate detection of oil spill in a timely manner, would be beneficial to resource management for monitoring of oceanography. As of now, the remote sensing method is one of the most effective operation that can be performed on sea monitoring and marine environment for object identifying.

Most of the radar satellites are monitoring for coastal regions using SAR remote sensing techniques for detecting object's (i.e., oil spills). Basically, the satellite remote sensing system are two types, namely passive and active remote sensing techniques. The passive remote sensing techniques are used optical sensors, infrared/ultraviolet

THE IIOAB JOURNAL ISSN: 0976-3104

systems, and microwave radiometer. The active remote sensing techniques are used radar satellite systems and laser fluoro-sensors [6], [7]. In the middle of these systems SAR can provide the valuable details about the region of specific location and size of the oil spill. It is due to the wide area coverage and day/night and all-weather conditional capabilities [8]. At this situation, the object identification is not accurately shown, due to the large area covering.

In generally, most of spaceborne SAR imagery systems are observing for detecting the oil spills, which is follows three steps: 1) dark spot identifying; 2) extraction of physical and geometrical features about the dark spot characteristics and properties; and 3) classify and discrimination about the dark spot as an oil spill or look-alike (i.e., internal waves, natural organics, jelly-fish areas, algae, threshold wind speed less than 3 m/s, rain cells, and kelp beds) [9], [10], [11]. Usually, these measurable features can be classifying as the performance of those operations, it can be operating based on manual or automatic methods [12] – [16]. In many existing systems are proposed the different approaches for detecting dark regions. But, most of the approaches are used manual selection method by cropping a broader area containing the dark formation [16], [17], regions such as adaptive threshold algorithms [12], [15], [18] – [20], marked point and statistical rule based thresholding methods [21], [22], wavelet based methods [23] – [25], fractal dimension estimation [26], [27], support vector machine [28], and neural networks [29] – [32].

In this paper, a fast, robust and effective automated approaches have been developed for oil spill monitoring. This approach is proposed a neural network classification algorithm that is used for achieving this goal. Before classifying the neural network segmentation capabilities have been already proofed in marine environment [33], that can be based on knowledge base system and it couldn't be tested for dark spot detection.

This paper is organized in four sections. Section II contains a description of the proposed method with their principle descriptions. In section III, a detailed description of each step in the dark-spot detection and the experimental analysis are obtained using ERS-2 SAR images. The conclusion and future scope is discussed in section IV.

METHODOLOGY

In our proposed methodology, the designing and development phase having two main difficulties, which can be occur for detection of dark spot is as: 1) the SAR imagery removes speckle noise due to its constructive and destructive inferences of reflections from sea surfaces regions of oil spill object detection and 2) the SAR image contrast between the dark region and its background can vary from specified area of image dark spot. And also, SAR image dark region vary from the local area of sea surface, and the spatial resolution of incidence angles [34], [35].

Pre-Processing

In this proposed methodology, the SAR imagery of full resolution is 1200 X 1800 pixels and each pixel is prohibitive computations of extracting feature vectors using radar technique for remote locations. And also, it has to take the textural information that is relevant feature responsible of spill type, which is contained on a wide coverage of SAR image dark region. It containing the low wind speed of radar resolution encompasses for spill type classification implementation. Designed for this analysis, we like better to choose using low resolution of the complete SAR image, that destruction subsequently applied using low-pass filtering. The consequence result of the image size in our algorithm is about 2200 X 3000 pixels [38], [44]. In some of artefacts are commonly find on SAR images, that is overlapping the regions of separate the image beams for several impedes and also any kind of texture analysis. Thus, the classification will occur on the separating of beams that encompass one full SAR image. The calibration as outlined of X-band SAR products in the manual operation accomplished to arrive at ρ_0 values according to the following formula:

$$\rho_0 = CalFactor * P^2 \sin\theta \text{.....(1)}$$

Where θ is the resident frequency approach of pixel of image, P denotes the pixel arithmetical value backscatter about the intensity value, and *CalFactor* is denotes the calibration factor as per the distribution.

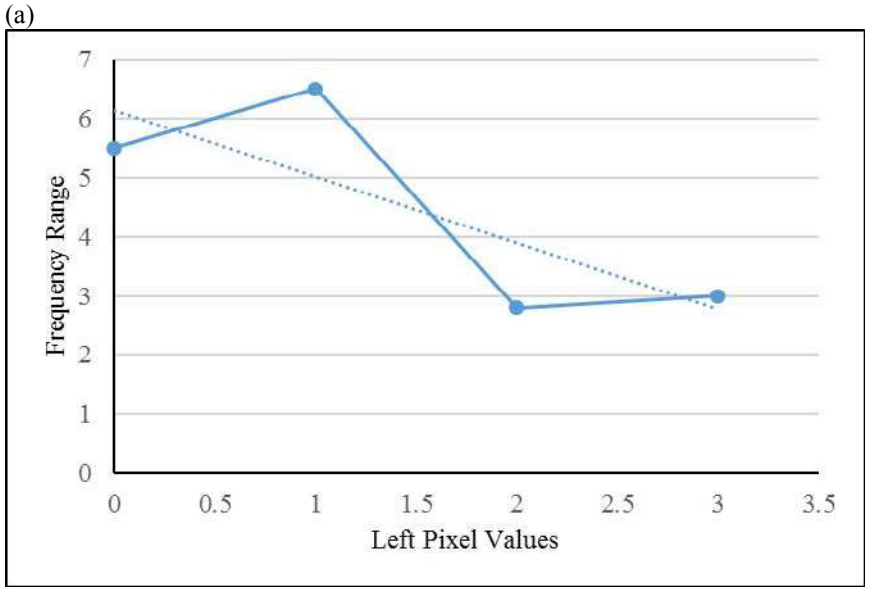
THE IIOAB JOURNAL ISSN: 0976-3104

Texture Extraction

In this analysis, the set of features are an instant feature of SAR imagery region as well as grey-level co-occurrence matrix (GLCM) features. Usually, these features are investigating for texture information and it has been explored in abundant periodicals, especially in SAR exploration [35] – [37]. Naturally, these features are containing only for first-order logic, that can be obtain the information about the image object of each grey values on each sliding window. But, the higher statistical orders can be described as usually for computing the histogram pixel values, pixel triples, etc. After this process the determination of histogram results are worked with 64 dim levels. In such a way that, we would process the enough region of interest on images for histogram results. In our numerical analysis, lower number of dark levels (e.g., 32 and 16) prompted to enhance the image that shows with less visualized pixels of interest. Such arrangements are along with these lines regarded less valuables of dark region. On account of GLCM, the first picks a parameter is based on the sliding window size.

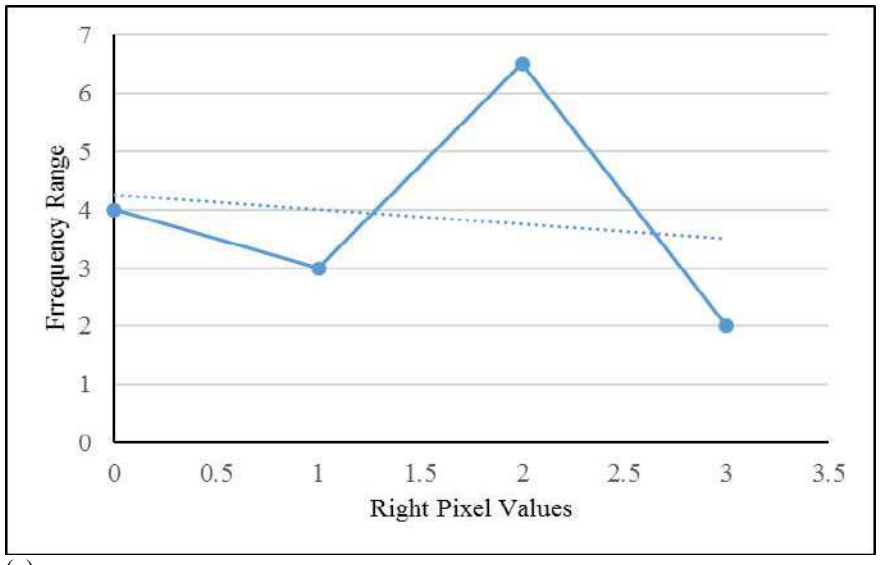
1	2	0	0	0
1	2	0	0	0
3	1	2	0	0
3	3	1	2	1
3	1	2	1	2

(a)



(b)

THE IIOAB JOURNAL
ISSN: 0976-3104



(c)

Fig. 1: Grey-level co-occurrence of matrix (GLCM) illustration for 2-bit grey-level SAR image – (a) grey values in a 5X5 sliding window. (b) Resulting GLCM left pair direction, pair distance 1. (c) Resulting GLCM right pair direction, pair distance 1.

However, this grey-level co-occurrence matrix window (in the wake of re-binning dim qualities to, e.g., 64 dim levels) represents one of the histogram pixel set processes on SAR image region. This operation can be selected based on the image region of two neighbouring pixels, with a settled inter-pixel separation of the two pixels and altered introduction of the pair pixels [38]. The subsequent histogram operation could be done on geometric request of image pixel values, is the assumed GLCM. As half the outline for the GLCM of one 5 X 5 sliding window is shown in following **Figure- 1**.

Thus, the altered inter-pixel separation is specified image region on one of pixel, is processed from left and right pixel values. Which can be shown as the above illustration of GLCM pair axis with introduction about the image region of left and right pixel values. The other conceivable introductions for the inter-pixel separation contain no further measurable data and are subsequently not generally processed (cf., [35]). For this processing performs the operation to designing the frameworks for four bearings are then included and the subsequent grid is signified by (C (i, j)). On this joined grid (C (i, j)), we process the five normal GLCM highlights:

1) Entropy 2) dissimilarity 3) contrast 4) homogeneity 5) energy

Entropy: $$E = -\sum_{i,j} C(i,j)\log(C(i,j)) \quad \ldots\ldots(2)$$

Contrast: $$C = \sum_{i,j} |i-j|^2 \, C(i,j) \quad \ldots\ldots(3)$$

Dissimilarity: $$D = \sum |i-j| \, C(i,j) \quad \ldots\ldots(4)$$

Homogeneity: $$H = \frac{1}{1+|i-j|^2} C(i,j) \quad \ldots\ldots(5)$$

THE IIOAB JOURNAL ISSN: 0976-3104

Energy: $$E = \sum_{i,j} C^2(i,j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

The texture extracting features algorithm was computed using the IDL programming language (IDL 5.0) and run on a HP Core i5 machine utilizing one CPU core. In this computation operation is performed using the ERS-2 SAR image for texture information extraction. For this analysis, each ERS-2 SAR image is computing time is up to 10 min for our texture features extraction [43]. And, its processing of all bands was near real-time results and that data processing is done through the satellite station using remote sensing techniques.

NEURAL NETWORK CLASSIFICATION

In this paper, the classification is used a neural network classifier and it has been shown without prior knowledge of data. But, the classification inconsistency is superior than the statistical methods. Especilly, it is used for dynamic neural network adoptation for classification of image region. In particularly, this phenamenon is used for image region of the object detection and classification. The ocean region monitoring is a major aspect for object identifying, which is based on the part of very low gravity and gravity–capillary waves (from one centimeter to decimeteres). The locality of an oil slick on the ocean surface moistures is categories of influences in light of the echo sort conduct of the thick versatility surface movies portrayed by the Marangoni damping hypothesis [25], [29]. So, this is the vicinity of an oil slick on the ocean surface radically decreases the deliberate backscattering vitality of ocean region, bringing about darker regions in SAR symbolism. In any case, the monitoring image investigation is required in light of the fact that dim territories may likewise be created by locally low winds or by normal ocean slicks, as appeared in **Figure – 2,** where diverse concurrent occasions, which may bring about low backscattering, are noticeable.

Basically, the most of SAR image classifiers having three steps: first pre-processing the SAR image, second to extracting the features from selected SAR image region, and finally, classify the SAR image region as follows:

Step 1: In this step, we processing to remove the background of an SAR image and performs normalization opearation.

a. Initially, it eliminates the background noise from the SAR image using region based segmentation method.
b. And, then the SAR image size is converted for normalizing the neural network classier.

Step 2: In this step, we performing the feature extration as an oil spill or look-alike and their appropriate extraction of texture feature values.

a. At starting of this process is proforming single-level transformation, that is narmolized image acquazation of shape and texture information from the SAR image.
b. Then, it is based on the RGB color component model to using the image pixel intensity values.
c. Finally, we choose the two opposite corner in SAR image region to joining, that is given for the better classification results from different texture features on image and also, this step extracting a physical feature intensity values with two appropriate points.

Step 3: finally, this step was performing the classification method based on the neural network classifier operation. For this using 12 values of physical stuctural features from the existing knowledge base system.

a. This module performs the neural network classifier operation, that is trained from input SAR image data using the knowledge data base.

Algorithm for oil spill detection

The neural network analysis was utilized as a part of this study on looks like from the one officially depicted in [36]. A self-loader apparatus permits the extraction of some remarkable components which portray the chosen dull spot in the picture and that will be incorporated into the neural netwrok system as an information vector. These information vector elements are differenciating into three unique sorts. Some of them contains data on the backscattering force

THE IIOAB JOURNAL
ISSN: 0976-3104

(ascertained in dB) inclination along the outskirt of the dissected dim spot: Max Gradient (Gmax), Mean Gradient (Gme), Gradient Standard Deviation (GSD); others concentrate on the backscattering oblivious spot and/or out of sight: Object Standard Deviation (OSD), Foundation Standard Deviation (BSD), Max Contrast (ConMax), Mean Contrast (ConMe); a third classification considers the geometry and the state of the dim spot: Area (A), Perimeter (P), shape Complexity (C), Spreading (S) regarding a longitudinal hub. More exact definitions can be found in [38].

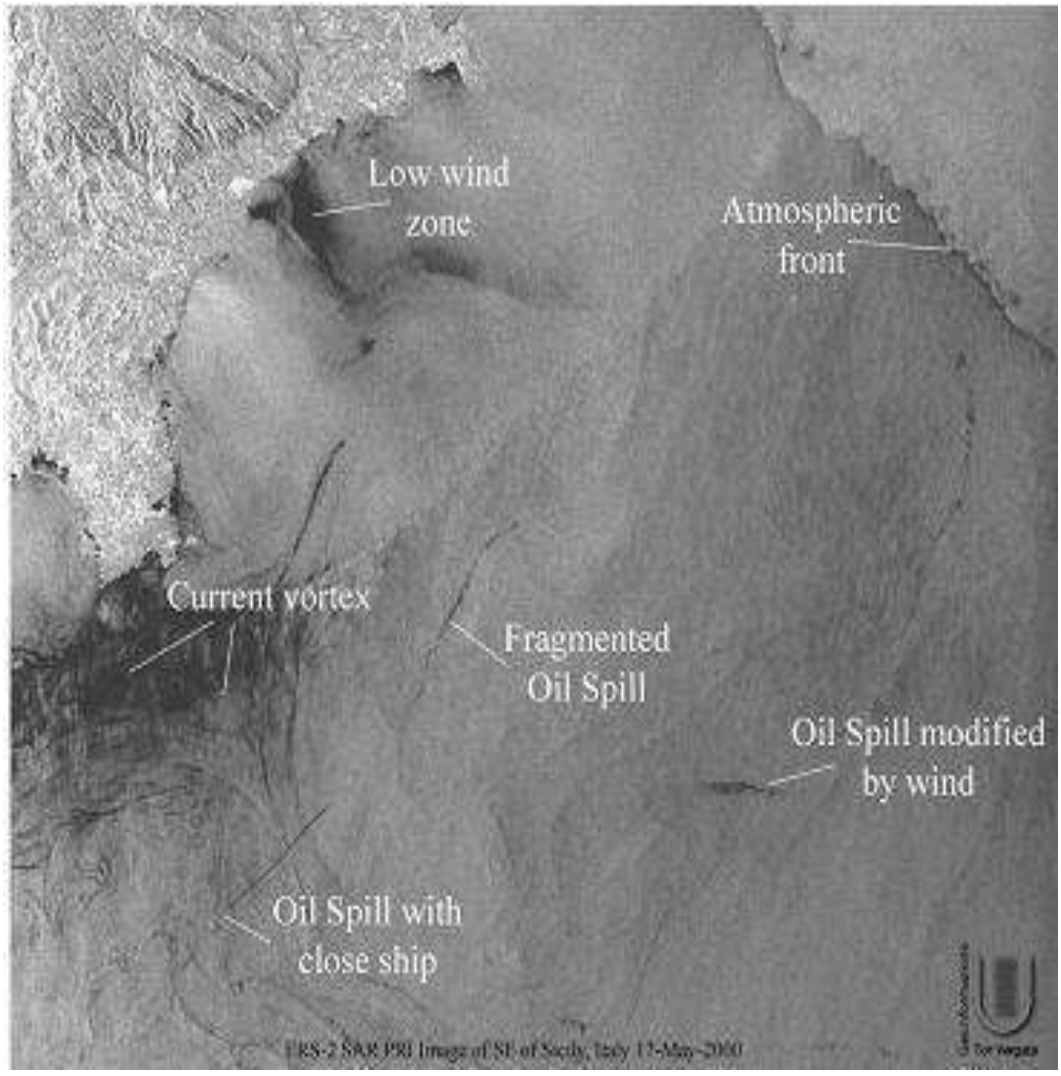


Fig: 2. Different analysis events were caused a reduction of the energy backscattering © ESA

Mainly, the ERS-2 SAR image archives over the environment of dark spot, is classified as oil spill or look-alike. For example, number of ground-truth values are available, else the discrimination was based on the independent decision making analysis of ERS-2 SAR images. The main statistical decision parameters are described from the extracted features of ERS-2 SAR image data values are show in **Table - I.**

In this proposed method, additionally considering the image object feature information is derived from the local wind speed vectors. This can be performed using backscattering gravity capillary waves and then the strong influence can appear as an oil slick on SAR image. It can be considering the wind speed conditions (i.e., less than 2-3 m/s) and it can be visualize based on the wind nature. Naturally, the wind speed is greater than 7-8 m/s and less than 15 m/s to identifying the slicks are appeared as an oil spill or look-alike. But, the wind speed computations are done by the inversion of CMOD4 model, it can be modelled by ESA to transforming the wind vectors from the ERS-2 SAR C-band measurements.

THE IIOAB JOURNAL ISSN: 0976-3104

Table: I – The Main Decision Parameters in ERS-2 SAR Data about oil spill or look-alike classification

Feature	Oil Spill					Look-alike			
	Max	Min	Mean	SD		Max	Min	Mean	SD
A (Km^2)	32.94	0.30	4.40	5.89		146.31	0.43	14.91	23.80
P (Km)	142.89	2.31	23.26	25.72		304.03	4.23	47.22	54.67
C	7.03	0.93	2.93	1.43		7.54	1.27	3.54	1.45
S	36.15	0.00	6.62	7.87		39.05	0.00	11.86	10.62
OSD (dB)	4.58	0.72	1.99	0.74		4.87	0.61	2.43	0.94
BSD (dB)	1.97	0.62	0.92	0.22		4.12	0.88	1.60	0.52
ConMax (dB)	17.85	4.21	9.51	3.09		18.56	4.77	12.45	2.87
ConMe (dB)	12.71	1.62	4.76	2.01		11.92	1.55	6.52	2.26
GMax (dB)	15.19	3.21	7.43	2.57		15.41	3.92	9.01	2.81
Gme (dB)	7.42	1.42	2.92	1.12		6.32	1.24	3.24	1.12
GSD (dB)	3.07	0.63	1.49	0.56		3.01	0.60	1.77	0.65
WS (m/s)	7.16	2.09	3.52	0.72		10.23	1.61	3.01	1.33

For this analysis many challenges have been initializing to accurately well excellent the number of units to be considered in the hidden layer of neural network. This topology 12-4-4-1 is shown in below **Figure - 3**. And finally this performance analysis should be in terms of both classification accuracy and training data set time.

The traditional neural network training method is as target vectors and then deciding class membership functions based on the maximum features are extracted from the selection of image region. It can be processed in the sense of the Bayesian optimal neural network classifier. Even though, the neural network assumes based on the network global weight space and its global minimum solution, that can be modelled from the specified network topology. And also, the neural network classification pattern can be considering as a non-parametric method for estimation of probabilities. This estimation process is considering the heuristic view of essential characteristics and its interpretation of level-of-confidence. From these concepts the classification confidence error estimating probability depends on the maximum occurrence of object features. These estimation processes are clearly performed on the specific network structure, hidden layer weights, as well as the dynamic learning strategies and its network topology consistency of image.

This provide a huge quantitative about the measuring directions in differentiating the input image distributions. This choice is defined as Kullback-Leibler distance strategy in [42]. And also, measuring the neural network approximation of given Bayesian classifier from specified region of image. Note, this function is differentiating the actual distance of the image region as well as image posterior knowledge base.

THE IIOAB JOURNAL
ISSN: 0976-3104

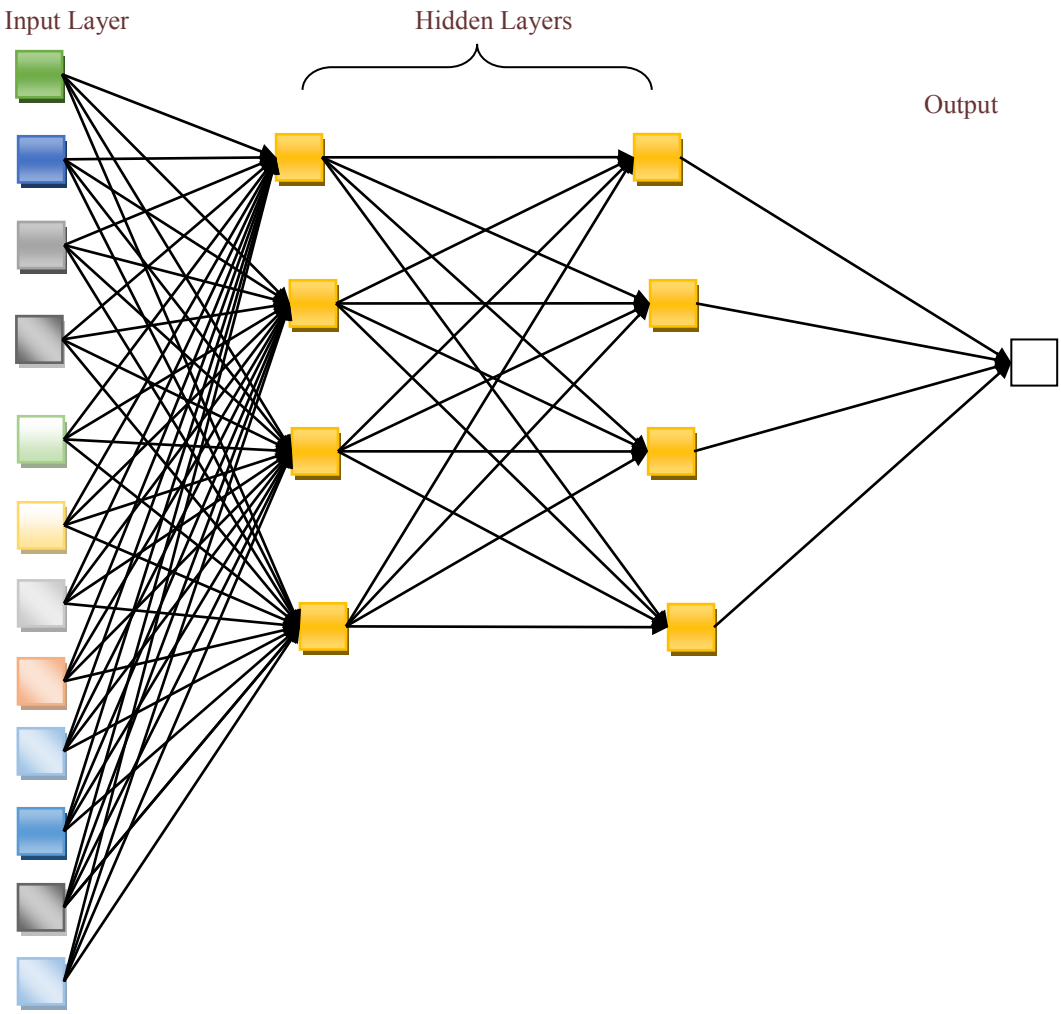


Fig: 3. Illustration of neural network classification topology

However, the neural network has been trained to return 1 i.e., specifies the dark spot as an oil spill, otherwise it is a look-alike. For responsible of training dataset procedure should be stopped according to the "stopping algorithm" [38]. As indicated by this calculation, the execution of the net amid the preparation (knowledge) stage be present assessed either on the preparation set or on an alternate free approval set. In the preparation set the general error in the recovery of the right yield continues diminishing with the preparation, drawing nearer an estimation of joining. On the other hand, the error on the approval set achieves a base quality after which it will begin expanding on the off chance that we proceed with the preparation. As of now the present learning stage must be interrupted, while validation of the errors. The dependence of the error on the quantity of periods (preparing cycles) for the preparing and the approval set is appeared in **Figure- 4**.

After the preparation taking into account the learning set and applying the "early ceasing" calculation we discovered a root mean square blunder (rmse) of 0.227 on the test set (3 mis-classified samples out of 60).

THE IIOAB JOURNAL
ISSN: 0976-3104

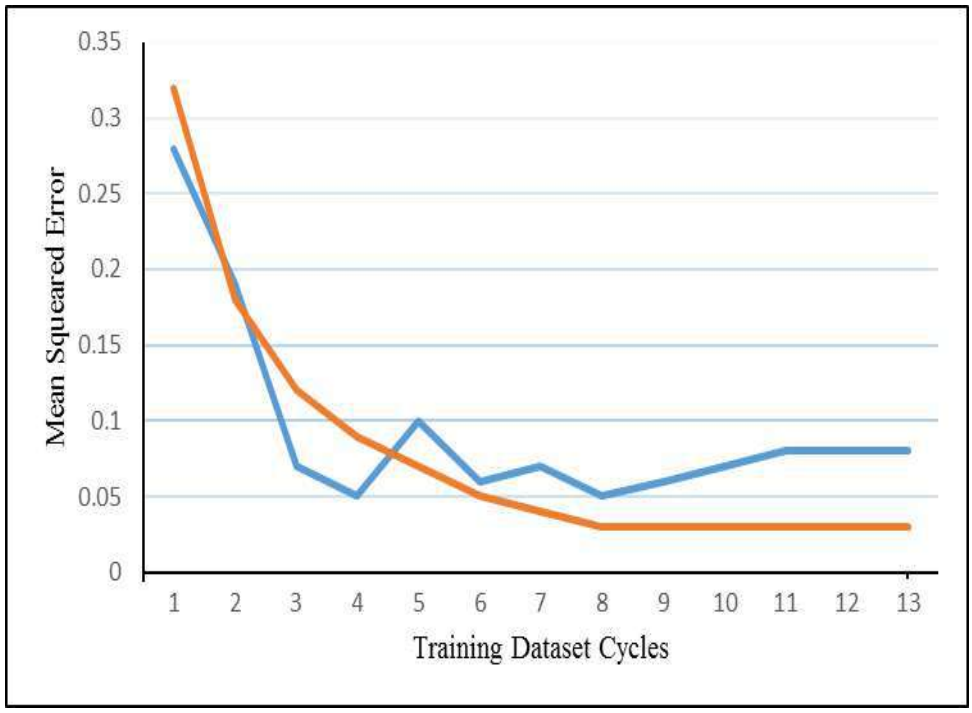


Fig: 4. The total number of training dataset cycles to be dependence of the errors for training and validation test

Sensitivity Analysis

As talked about in [39], the unwavering quality of the qualities for the components figured by the self-loader apparatus relies upon the capacity of the client dealing with installed edge discovery utilities. In this point of view, the sensitivity analysis completely programmed based on instruments and then the sensible analysis is assuming that the S/N proportion describing the estimations of elements might be smaller. For this situation, the determination of inputs from the neural network system, on the base of adequacy of their image data content in assessing the yield, may be prescribed to arrange of misdirecting inputs. These system operations with less inputs and has less versatile parameters to be resolved, which require a littler preparing set to be valid constraints. This system prompts are enhanced with speculation properties that giving smoother mappings. Likewise, a system with less weights may be speedier to prepare.

In the endeavour to look at which includes, in the picked setting, contain less data for the arrangement assignment we considered two routines. In a first investigation we assessed the system execution, both regarding the RMSE and the misclassification rate, for 12 distinct situations where, on turn, one of the parts of the at first considered information arrangement was absent. In a brief moment examination, we delayed the pruning system (portrayed before) to the information layer [40], [41], [42], until 11 of the 12 segments of the data vector were evaluated (we remind that an info or concealed unit is uprooted when it has lost every one of its associations).

CONCLUSION AND FUTURE WORK

This study takes after the work portrayed where the possibilities. The paper concludes with efficient results which recognises the oil spills in satellite SAR symbolism using gray level co-occurrence matrix (GLCM) based texture feature of neural networks. Thereby, the computational procedure focuses on local wind speed direction on the SAR images and provides an additional information about the ERS-2 SAR image as an input for neural networks known to evolution of the oil spills.

The future work deals with local low wind speed for knowledge based system which is significant for short gravity and the gravity-capillary waves. Finally, the low wind speed can strongly influence the appearance of the oil spill in a SAR image effectively.

THE IIOAB JOURNAL
ISSN: 0976-3104

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

This work carried out in ESA under ERS-2 SAR has been done as for oil spill detection and this work is to supported by the experienced person of Dr. Juerg Lichtnegger

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] D. Hamidi, S. Lehner, T. König, and A. Pleskachevsky, [2011] "On the sea ice motion estimation with synthetic aperture radar," in Proc. 4th TerraSAR-X Meeting, DLR, Oberpfaffenhofen, Germany, CAL0166, pp. 1–10.

[2] J. Karvonen, [2012] "Operational SAR-based sea ice drift monitoring over the Baltic Sea," Ocean Sci., vol. 8, pp. 473–483.

[3] L. Kaleschke et al., [2013] "IRO-2 Eisvorhersage und Eisroutenoptimierung," 25. Internationale Polartagung 2013-03-17. Hamburg, Germany: Deutsche Gesellschaft für Polarforschung (DGP), 22.

[4] R. Kwok, G. Spreen, and S. Pang, [2013] "Arctic sea ice circulation and drift speed: Decadal trends and ocean currents," J. Geophys. Res. Oceans, vol. 118, no. 5, pp. 2408–2425.

[5] R. H. Goodman, [1989] "Application of the Technology in the Remote Sensing of Oil Slicks", A. E. Lodge, Ed. Hoboken, NJ, USA: Wiley, 1989.

[6] M. Fingas, [2005] "The Basics of Oil Spill Cleanup. Boca Raton, FL, USA: Lewis Publ., 2001.

[7] C. Brekke and A. H. S. Solberg, "Oil spill detection by satellite remote sensing," Remote Sens. Environ., vol. 95, no. 1, pp. 1–13, Mar. 15.

[8] G. Ferraro, S. Meyer-Roux, O. Muellenhoff, M. Pavliha, J. Svetak, D. Tarchi, and K. Topouzelis, [2009] "Long term monitoring of oil spills in European seas," Int. J. Remote Sens., vol. 30, no. 3, pp. 627–645.

[9] C. Brekke and A. H. S. Solberg, [2005] "Oil spill detection by satellite remote sensing," Remote Sens. Environ., vol. 95, no. 1, pp. 1–13.

[10] A. T. Jones, M. Thankappan, G. A. Logan, J. M. Kennard, C. J. Smith, A. K. Williams, and G. M. Lawrence, [2006] "Coral spawn and bathymetric slicks in synthetic aperture radar (SAR) data from the Timor Sea, NorthWest Australia," Int. J. Remote Sens., vol. 27, no. 10, pp. 2063–2069.

[11] M. Thankappan, N. Rollet, C. J. H. Smith, A. Jones, G. Logan, and J. Kennard, [2007] "Assessment of SAR ocean features using optical and marine survey data," presented at the ENVISAT Symp., Montreux, Switzerland, Apr. 23–27.

[12] The assessment of SAR image features related Available: https://earth.esa.int/envisatsymposium/proceedings/contents.html

[13] F. Nirchio, M. Sorgente, A. Giancaspro, W. Biamino, E. Parisato, R. Ravera, and P. Trivero, [2005] "Automatic detection of oil spills from SAR images," Int. J. Remote Sens., vol. 26, no. 6, pp. 1157–1174.

[14] V. Karathanassi, K. Topouzelis, P. Pavlakis, and D. Rokos, [2006] "An objectoriented methodology to detect oil spills," Int. J. Remote Sens., vol. 27, no. 23, pp. 5235–5251.

[15] I. Keramitsoglou, C. Cartalis, and C. T. Kiranoudis, [2006] "Automatic identification of oil spills on satellite images," Environ. Modell. Softw., vol. 21, no. 5, pp. 640–652.

[16] A. H. S. Solberg, C. Brekke, and P. O. Husoy, [2007] "Oil spill detection in RADARSAT and Envisat SAR images," IEEE Trans. Geosci. Remote Sens., vol. 45, no. 3, pp. 746–755.

[17] F. Del Frate, A. Petrocchi, J. Lichtenegger, and G. Calabresi, [2000] "Neural networks for oil spill detection using ERS-SAR data," IEEE Trans. Geosci. Remote Sens., vol. 38, no. 5, pp. 2282–2287.

[18] J. Lichtenegger, G. Calabresi, and A. Petrocchi, [2000] "Detection and discrimination between oil spills and look-alike," in XIX ISPRS Congress, XXXIII, pp. 193–200.

[19] R. Solberg and N. A. Theophilopoulos, "ENVISYS—A solution for automatic oil spill detection in the Mediterranean," in Proc. 4th Int. Conf. Remote Sens. Marine Coastal Environ., 1997, pp. 3–12.

[20] A. Solberg, G. Storvik, R. Solberg, and E. Volden, [1999] "Automatic detection of oil spills in ERS SAR images," IEEE Trans. Geoci. Remote Sens., vol. 37, no. 4, pp. 1916–1924.

[21] L. Chang, Z. S. Tang, S. H. Chang, and Y. L. Chang, [2008] "A region-based GLRT detection of oil spills in SAR images," Pattern Recognit. Lett., vol. 29, no. 14, pp. 1915–1923.

[22] Y. M. Shu, J. Li, H. Yousif, and G. Gomes, [2010] "Dark-spot detection from SAR intensity imagery with spatial density thresholding for oil-spill monitoring," Remote Sens. Environ., vol. 114, no. 9, pp. 2026–2035.

[23] Y. Li and J. Li, [2010] "Oil spill detection from SAR intensity imagery using a marked point process," Remote Sens. Environ., vol. 114, no. 7, pp. 1590–1601.

[24] S. Derrode and G. Mercier, [2007] "Unsupervised multiscale oil slick segmentation from SAR images using a vector HMC model," Pattern Recognit., vol. 40, no. 3, pp. 1135–1147.

[25] S. Y. Wu and A. K. Liu, [2003] "Towards an automated ocean feature detection, extraction and classification scheme for SAR imagery," Int. J. Remote Sens., vol. 24, no. 5, pp. 935–951.

[26] A. K. Liu, C. Y. Peng, and S. Y. Chang, [1997] "Wavelet analysis of satellite images for coastal watch," IEEE J. Ocean Eng., vol. 22, no. 1, pp. 9–17.

[27] G. Benelli and A. Garzelli, [1999] "Oil-spills detection in SAR images by fractal dimension estimation," in Proc. IGARSS, Hamburg, Germany, Jun. 28–Jul. 2, pp. 218–220.

[28] M. Marghany, M. Hashim, and A. P. Cracknell, [2007] "Fractal dimension algorithm for detecting oil spills using RADARSAT-1 SAR," in Proc. LNCS, pp. 1054–1062.

[29] G Mercier and F. Girard-Ardhuin, [2006] "Partially supervised oil-slick detection by SAR imagery using kernel expansion," IEEE Trans. Geosci. Remote Sens., vol. 44, no. 10, pp. 2839–2846.

[30] K. Topouzelis, V. Karathanassi, P. Pavlakis, and D. Rokos, [2007] "Detection and discrimination between oil spills and look-alike phenomena through neural networks," ISPRS J. Photogramm. Remote Sens., vol. 62, no. 4, pp. 264–270.
[31] K. Topouzelis, V. Karathanassi, P. Pavlakis, and D. Rokos, [2008] "Dark formation detection using neural networks," Int. J. Remote Sens., vol. 29, no. 16, pp. 4705–4720.
[32] K. Topouzelis, D. Stathakis, and V. Karathanassi, [2009] "Investigation of genetic algorithms contribution to feature selection for oil spill detection," Int. J. Remote Sens., vol. 30, no. 3, pp. 611–625.
[33] S. Vijayakumar, V. Santhi, [2015] "Different Approaches for Oil Spill Detection in SAR Images – A Review", International Journal of Oceans and Oceanography, vol. 9, no. 2, pp. 221-228.
[34] K. N. Topouzelis, [2008] "Oil spill detection by SAR images: Dark formation detection, feature extraction and classification algorithms," Sensors, vol. 8, no. 10, pp. 6642–6659.
[35] N. Bartsch, K. Grüner, W. Keydel, and F. Witte, [1987] "Contribution to oil spill detection and analysis with radar and microwave radiometer: Results of the Archimedes II campaign," IEEE Trans. Geosci. Remote Sens., vol. GE-25, no. 6, pp. 677–690.
[36] Bogdanov, A.V.; Sandven, S.; Johannessen, O.M.; Alexandrov, V.Yu.; Bobylev, L.P., [2005] "Multisensor approach to automated classification of sea ice image data," in Geoscience and Remote Sensing, IEEE Transactions on, vol.43, no.7, pp.1648-1664.
[37] M.-A. N. Moen et al., [2013] "Comparison of automatic segmentation of full polarimetric SAR sea ice images with manually drawn ice charts," Cryosphere Discuss., vol. 7, pp. 2595–2634.
[38] C. M. Bishop, [1995] "Neural Networks for Pattern Recognition," Oxford Univ. Press.
[39] Del Frate, F.; Salvatori, L., [2004] "Oil spill detection by means of neural networks algorithms: a sensitivity analysis," in Geoscience and Remote Sensing Symposium, IGARSS '04. Proceedings. IEEE International, vol.2, no., pp.1370-1373 vol.2, pp. 20-24.
[40] Ressel, R.; Frost, A.; Lehner, S., [2015] "A Neural Network-Based Classification for Sea Ice Types on X-Band SAR Images," in Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, vol.8, no.7, pp.3672-3680.
[41] Taravat, A.; Latini, D.; Del Frate, F., [2014] "Fully Automatic Dark-Spot Detection From SAR Imagery With the Combination of Nonadaptive Weibull Multiplicative Model and Pulse-Coupled Neural Networks," in Geoscience and Remote Sensing, IEEE Transactions on, vol.52, no.5, pp.2427-2435.
[42] Wan, E.A., [1990] "Neural network classification: a Bayesian interpretation," in Neural Networks, IEEE Transactions on, vol.1, no.4, pp.303-305.
[43] Prabu Sevugan, RamaKrishnan, R., Manthira Moorthi S., and Swarnalatha P., [2015] "Application of average brightness threshold algorithm for the analysis of relative radiometric correction for Ahmedabad region," International Journal of Earth Sciences and Engineering, Vol.8, Issue 1, pp.nos.341-348.
[44] Swarnalatha P, BK Tripathy,S Prabu, R Ramakrishanan and S Manthira Moorthi. [2014] "Depth Reconstruction using Geometric Correction with Anaglyph Approach for Satellite Imagery," Proceedings of International Conference on Advances in Communication, Network, and Computing, CNC, Elsevier, pp.506-515.

ABOUT AUTHORS

Singanamalla Vijayakumar *received his B.Tech in Computer Science and Engineering from Vaagdevi Institute of Technology and Science ,Jawaharlal Nehru Technological University, Anantapur and M.Tech in Computer Science and Engineering from Madanapalle Institute of Technology and Science,Jawaharlal Nehru Technological University, Anantapur. Presently, he is pursuing his PhD and also working as Teaching cum Research Associate in School of Computer Science and Engineering from Vellore Institute of Technology University, Vellore, Tamil Nadu, India. He has one year teaching experience at PG level. He is doing his research work in data mining analysis and wireless network systems.*

Dr. Swarnalatha P urushotham *is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Sstems. She has published more than 50 papers in International Journals/International Conference Proceedings/National Conferences. She is having 14+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.*

Rukmini Singanamalla *received her B.Tech in Electronics and Communication Engineering from Vaagdevi Institute of Technology and Science, Jawaharlal Nehru Technological University, Anantapur and M.Tech in VLSI Engineering from Annamacharya Institute of Technology and Sciences, Jawaharlal Nehru Technological University, Anantapur. Presently, she is working as VLSI Design Engineer in Gray Matter Electronics Pvt. Ltd., Hyderabad, Telangana. She is doing her research work in Image Processing, Networking and VLSI system Design.*

ISSN: 0976-3104
SPECIAL ISSUE(SCMSA)
Jenisha et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A SURVEY OF NEURAL NETWORK ALGORITHMS USED FOR IMAGE ANNOTATION

Jenisha T. and Swarnalatha Purushotham
School of Computing Science and Engineering, VIT University, Vellore-632014, INDIA

ABSTRACT

In recent years, scalable image annotation has gained more popularity because of its usefulness in various applications like image recommendation system, context aware chat bots , image based query retrieval, object detection, object segmentation, pose estimation, matching patterns, generating synthetic images and many more. The usage of convolutional neural network for object detection and localization became popular in 2012 when then the neural network designed by Alex Krizhevsky et al. had record breaking performance with error rate of 16.4% in top 5 hit% in ImageNet 2012 challenge . In the year 2014 Goog LeNet neural network model from Google outperformed with error rate of 6.66% in ImageNet 2014 test dataset. This statics shows the accuracy is pretty high when deep convolutional network is used for object detection. In medical image the accuracy of detected object is crucial. Object detection using Convolution Neural Networks has given good accuracy of about 96.7% in ImageNet dataset. This research paper does literature review of different architecture styles researchers used to construct Convolutional Neural Network model for object detection, localization, and annotation. Visualizing neurons which was a complete black box till 2012 is described. Convolutional Neural Networks (CNN) is networks that share parameters across space. It also describes the entire process flow of automatic image captioning task. The performance improvements gained by different optimization techniques are described. It starts from data preparation to CNN and ends with automatic sentence generation using CNN.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26– March-2016
Published on: 10th –Aug-2016

KEY WORDS

CNN,annotation,deeplearning,classification,pooling, convolution

***Corresponding author: Email:** jenisha.t2011@vit.ac.in

INTRODUCTION

Current approaches to image annotation make essential use of deep neural networks because of available big data and GPU provided by Nvidia and other GPU infrastructure providers like amazon cloud GPU instance. The current computing performance of GPU K80 is amazingly good which takes only few hours to process 1 million images in contrast to CPU which takes several weeks to compute. To improve performance of annotation, large datasets are trained. Taking large datasets and training them prevents over fitting of data. ImageNet [1] one of the biggest object detection dataset available consists of 500K images and 251 basic categories like dog, tree, person etc. Microsoft Common Objects in Context [2] consists of 300K images and 80 object categories. To learn about millions of objects we need a learning model with good learning capability. Despite of having object detection dataset which consists of millions of images, to automatically annotate all objects present in real world is challenging. This means the training model should have prior knowledge to compensate for the objects not present in training dataset. Convolutional Neural network a supervised learning model comes with such learning capability.

Supervised learning is composed of two important subfields Classification and Regression. Classification is predicting discrete classes. Regression is finding the continuous outcomes (range) of target variables. Example for regression is predicting the house price in particular region. In image classification task for input image Ix predict which class is most suitable. In **Figure- 1** shows linearity layer of CNN. For Image Ix linear model as shown in Eqn. 1 is applied. W and b are computed using Stochastic Gradient Descent or other numerical optimization techniques which are detailed in coming sections.

THE IIOAB JOURNAL ISSN: 0976-3104

$$Y=WI_x+b \quad (1)$$

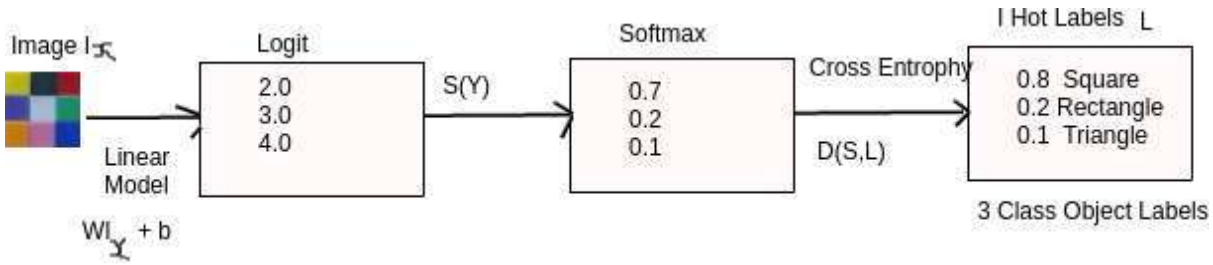


Fig: 1.Linearity Layer

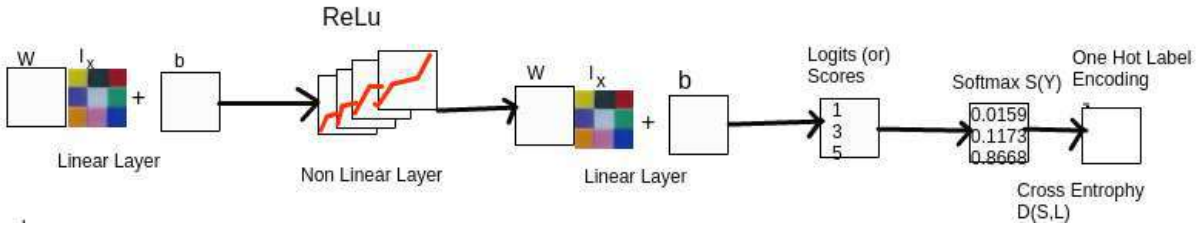


Fig: 2. Combining Linearity and Non Linear Layer to build CNN

The output is scores in terms of logistic regression. These are also called Logits. In CNN two linear models are connected by nonlinear layer. In **Figure-2** shows how two linear layers are connected using ReLu to form non-linear network. The first layer effectively consists of the set of weights and biases applied to X and passed through ReLUs. The output of this layer is fed to the next one, but is not observable outside the network; hence it is known as a hidden layer. The second layer consists of the weights and biases applied to these intermediate outputs, followed by the softmax function to generate probabilities.

Terminologies used in CNN are explained below.

Softmax

Softmax function converts scores into probability. Eqn. 2 gives proper probability for score *i*. Proper probability is probability when summed, all the probabilities of given *i* the value will be 1. When the size of output *Y* is high the confidence will be more for prediction. When the size of *Y* is small the confidence which class given object belongs to decreases.

$$S(y_i)=\frac{e_i^y}{\sum e_j^y} \quad (2)$$

One Hot Encoding

In One Hot Encoding we give one for correct class and zero to other class labels. This method has drawback when we have billions of classes which makes one hot encoding inefficient. A better approach is to use embeddings. In embeddings we convert word to vector and the words that have same context has closer value. For more details about embeddings see section on Automatic Image Captioning Using CNN and RNN.

Cross Entropy

The way to measure distance between two probabilities is called cross entropy. Eqn. 3 is used to compute the distance between two probabilities.

THE IIOAB JOURNAL ISSN: 0976-3104

$$D(S,L)=-\sum_i L_i log(S_i) \quad (3)$$

Cross Entrophy is not symmetric. $D(S,L)\neq D(L,S)$
Composition of computations abstracted as layers in CNN is organised Directed Acyclic Graph(DAG). The different layers in CNN are Convolution layer, Non-linearity layer and Pooling layer. A featurer map is obtained after passing through these layers. These are commonly used layers found in literature. One or more layer like inception layer as discussed in subheading inception and a backpropagation layer in subheading backpropagation,and residual block discussed in resnet are introduced by researchers for performance tuning. **Figure-3** shows different layers of CNN.

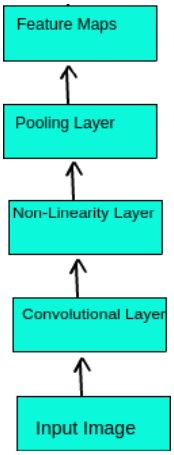


Fig:3. Different Layers of CNN

.....

Computing Weight and Bias

When computing weight W and bias b, the value should be more for correct class and less for incorrect class. The distance should be low for correct class and high for incorrect class. We need to minimize distance between similar labels and maximize distance between dissimilar labels. Eqn. 4 gives training loss over the entire training set. Training loss is computed by measuring the distance averaged over the entire training sets for all inputs and all labels. If we have 10,000 training images and 100 labels then distance between 10000 images and each label is calculated and average distance is taken as training loss L.

$$\mathcal{L}=\frac{1}{N}\sum_i D(S(WX_i+b),L_i) \quad (4)$$

Loss is function of weight and biases. We want to minimize loss. A simple numerical optimization approach to minimize loss is to use Gradient Descent method. To obtain numerical stability we subtract and divide each pixel by 128 as shown below.

RedPixel R= (*R*-128)/128,

GreenPixel G= (*G*-128)/128,

BluePixel B= (*B*-128)/128)

While computing Gradient Descent weights are initialized by randomly drawing weights from Gaussian distribution with mean μ zero and equal variance. The smaller values for standard deviation leads to uncertainty and larger values for standard deviation leads to overconfidence. In training its better to choose very small value as variance σ. Eqn. 5 and Eqn. 6 shows computing weight W and bias b used in 8

$$W=w-\alpha\Delta_w L \quad (5)$$

$$b=b-\alpha\Delta L' \quad (6)$$

THE IIOAB JOURNAL ISSN: 0976-3104

Alternate numerical optimization methods to find optimum weight and bias are using Stochastic Gradient Descent (SGD), Ada Grad, Adam, Nesterov's Accelerated Gradient, RMSprop.Ada Grad is modification of SGD which implicitly does momentum and learning rate decay. Using AdaGrad makes less sensitive to hyper parameters.

This literature survey is organized as follows. Section 1 gives overview of CNN, section 2 describes evolution of CNN, section 3 describes data preparation for CNN, followed by each layers in CNN, then section on optimization next error computation methods are described. Next popular CNN architecture are described in section 12. Different CNN network models found in literature, then performance evaluation used to determine accuracy of object detection finally conclusion based on comparative study of different approaches.

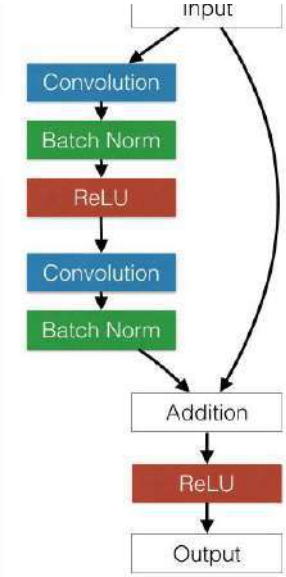
EVOLUTION OF CONVOLUTIONAL NEURAL NETWORKS

The research in neural network started early 1980's but lack of computing power the model didn't gain popularity. In the year 1998 [4] LeNet -5 model was designed to classify MNIST dataset . This model is used to read the digits in cheque and successfully used in banking industry. LeNet-5 model has 60,000 training and 10,000 test samples hand written digits are taken.Test error rate is 0.8 %. In 2012 AlexNet [5] was modeled with 7 hidden weight layers, 650K neurons and 60M parameters, and 630M connections to classify 1.2M images in 1000 categories. Imagenet 2012 dataset is used to train AlexNet. 4 layer CNN with activation unit ReLU is used which converged 6 times faster than equivalent network using tanh as activation unit. Alexnet decreased state of art error rate 47.1 % to 37.5 % for top 1 result and the error rate was further reduced from 28.2 % to 17 % for top 5 results. In the year 2014 GoogLeNet [7] introduced deep CNN architecture named as 'inception'. In 2015 Microsoft deep residual network model is designed based on the principle of residual learning. This model had 152 layers which is 8x deeper than state-of-the-art CNN layers. **Figure- 4** shows evolution of Convolutional Neural Networks. **Table- .1** shows summary of error rate decresed in Imagenet dataset over a period of time without using any external dataset. The error rate decresed as the cnn layer goes deeper and deeper.

Table: .1: Top 5 Error Rate and CNN Models

CNN Model	Year	Error Rate Top 5 in %
AlexNet	2012	16.4
Clarrifai	2013	11.7
GoogLeNet	2014	6.66
ResNet	2015	3.57

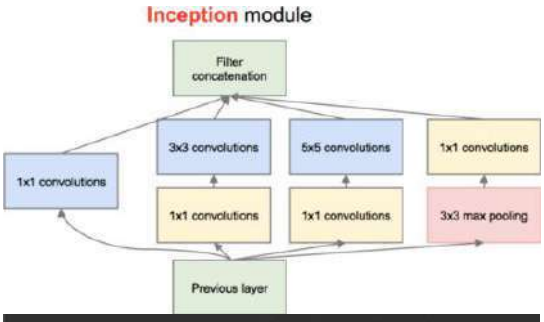
2015 ResidualNet



152 layers

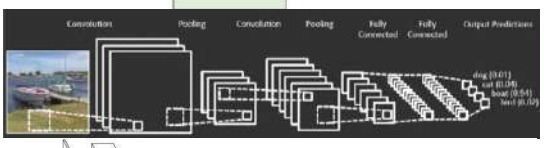
THE IIOAB JOURNAL
ISSN: 0976-3104

2014 GoogLeNet



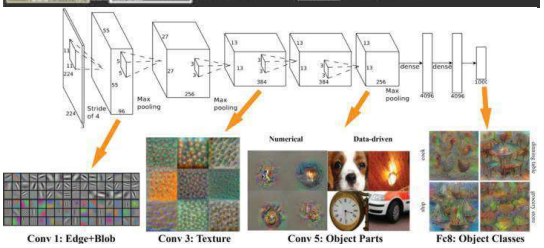
22 layers

2013 Clarifai Net



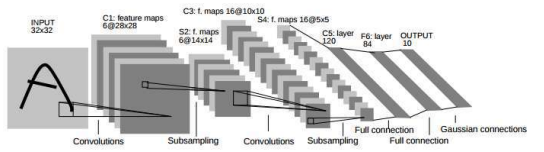
6 layers

2012 AlexNet



12 layers

1998 Lenet-5



6 layers

Fig: 4. Evolution of CNN

.....

Data Preparation for Convolutional Neural Network

Color images have height, width and depth. In RGB color channel an image is represented as Red, Blue and Green intensity. The intensity value of R, G, B are represented as three matrices of same height and width. **Figure- 5** is RGB image of size 38x38x3.When this image is represented in matrix form there are 3 matrix one for each R,G,B and matrix size is 38 x 38.

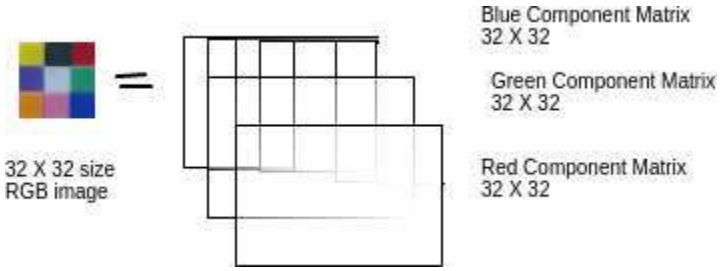


Fig: 5. RGB Components of an Image

.....

If color doesn't matter, it might help to reduce the complexity of the problem by combining color channels into a single monochromatic channel. Taking the average (R+G+B)/3 is one way of doing it; however there are other transformations that might be more effective/closer to, how human's perceive color (e.g. converting to YUV and using the Y channel). In luminosity method of converting RGB to monochrome green is given more weightage as green is more sensitive for human eye as shown in Eqn. 7.

THE IIOAB JOURNAL ISSN: 0976-3104

$$I_{(gray)} = 0.21R + 0.72G + 0.07B \quad (7)$$

The gray image is represented as single column vector. For CNN both color image and monochromatic image can be used. The images taken should be translation invariance meaning even if the image is rotated the features extracted from the image should be able to detect object correctly. When the extracted features are translation invariant, the position of the object in the image doesn't matter. The object located anywhere top, left, bottom, right can be detected with translation invariant features.

Dataset Used to Benchmark CNN Models

There are few standard dataset that can be used to benchmark our algorithms with state-of-art algorithms. Commonly used dataset for natural image annotation are Microsoft COCO dataset, ImageClef dataset, CIFAR dataset.

1. Microsoft COCO Dataset with Image Captions is used by researchers for image recognition, Image Segmentation and Image Captioning.
2. Imagenet dataset can be used for object detection,object localization,Object detection from video and Scene Classification for Video.

Data Layer for CNN

While training the CNN model common practice is 70 % of image is used during training, 20 % during validation and 10 % during testing. HDF5 file format is popularly used while handling large image data.

Data_layer= AsyncHDF5DataLayer(name=``train_data``,source=``data/train.txt``,batch_size=64,shuffle=true")

Above code is to read HDF5 format image data in TensorFlow Framework. It reads 64 images in single batch from data/train.txt. Shuffle the image is set to true to get faster convergence.

Convolution Layer

Natural image has stationary properties. That means the statistical properties of images remains the same in different parts of the images. This suggests that the features we have learned in one part of the image can be used to learn in other parts of the image. This is done using convolution. Suppose we have learnt features from 8 X 8 patch sampled from 1024 X 1024 pixel image. Train a sparse encoder on 8 X 8 or a X b patch small patch sampled from bigger image X. From the sampled patch x_{small} the number of features learned is k

$$f=\sigma(W^{(1)}x_{small}+b^{(1)}) \quad (8)$$

σ - sigmoid function given by weight $W^{(1)}$ and $b^{(1)}$ from visible unit to hidden units.
For every x_{small} in image I_x compute f_s as shown in Eqn. 8. The size of convoluted image $f_{convoluted}$ array is computed according to Eqn. 9

$$f_{convoluted}=kX(r-a+1)X(c-b+1) \quad (9)$$

where, r - number of pixel array rows in image I_x

c - number of pixel array columns in image I_x

a - number of pixel array rows in sampled patch of the image x_s

b - number of pixel array columns in sampled patch of the image x_s

k – number of image channel used.
For patch size 7 X 7 pixel from 38 X 38 image, k=3 ,then array size of convoluted image will be 3 X 32 X 32 when using valid convolution. In valid convolution those part of convolution computed without using zero padded edges is returned. Its observed that CNN first learns lines, then edges, shapes and finally objects.

THE IIOAB JOURNAL ISSN: 0976-3104

Non-Linearity Layer

Two linear netwoks are connected using nonlinear function such as sigma, tanh or ReLu as shown in Eqn. 10 & 11.An activation function layer or non-linearity layer applies an entry wise non-linearity map. More useful insights are obtained by increasing depth of the layers than increasing width of layers.

$$Sigmoid\sigma(x)=\frac{1}{1+e^{x}} \quad (10)$$

$$ReLu\sigma(x)=max\{x,0\} \quad (11)$$

Pooling Layer for CNN

Pooling is done for dimensionality reduction and make image translation invariant. A convoluted image is divided into *m*m* sub regions. Usually the value of *m* is less than 5 for large image. From each sub regions take max value or maxpooling or average value for meanpooling. Pooling also avoids overfitting. The outcome of pooling is pooled convoluted features. **Figure- 6** shows sample of pooling image.

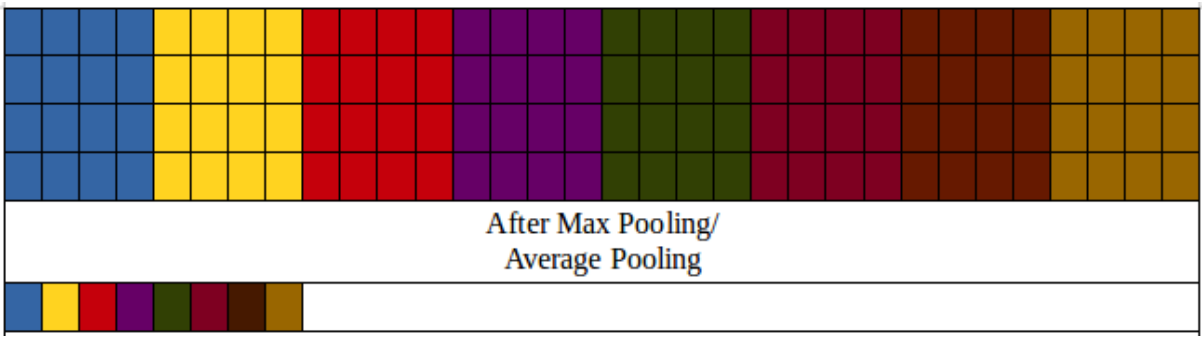


Fig: 6. Pooling Convoluted Image of size 32 X 32

.....

Poolinglayer oolingLayer(name=``pool``,kernel=(2,2),stride(2,2),bottoms=[:conv],tops=[:pool])

Network Layer for CNN

Fully connected layer is also known as dense layer or inner product layer or linear layer is computed as following

$$y_i=\sum w_{ij}x_j \quad (12)$$

Backpropagation Layer

Backpropagation layer takes advantage of chain rule.

$$[g(f(x))]'=g'(f(x))\times f'(x) \quad (13)$$

It decompose convolution, nonlinear and pooling operations into $f^{conv}, f^{nonlinear}, f^{pool}$ elements whose derivatives with respect to inputs are known by symbolic computations. It backpropagate error signals corresponding to a differentiable cost function by numeric computation. The propagated error signal is used to compute local minima.

Error Computation for CNN

The common criteria for error computation are recall with respect to precision and carried out for different image transformations. In **Figure-7** shows precision and recall computation is shown.

THE IIOAB JOURNAL
ISSN: 0976-3104

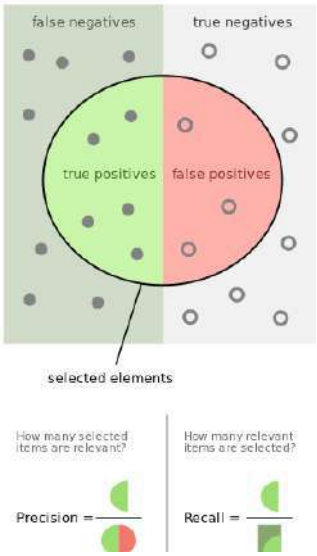


Fig:7.Precision and Recall.Image Credit : Wikipedia, Precision and recall,2016

Eqn.14 shows recall calculation. Recall is used to measure sensitivity or true positive rate. Its equivalent to computing hits or how many images classified correctly.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (14)$$

Eqn. 15 shows precision calculation. Precision is used to measure Positive Predicted Value (PPV).Its equivalent to computing correct rejection.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (15)$$

The other approaches are computing true negative rate and accuracy. Eqn. 16 is used to compute true negative rate.

$$TrueNagativeRate(TNR) = \frac{TrueNegative}{TrueNegative + FalsePositive} \quad (16)$$

Eqn. 17 is used to compute Accuracy.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative} \quad (17)$$

A combination of prediction and recall is called F-score. Eqn. 18 shows computing F-score.

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (18)$$

Optimization Techniques in CNN

CNN model success is heavily on designing correct training loss function which can be optimized in less time. Challenge is the optimization models are very difficult to scale. Few optimization methods used in CNN are Gradient Descent, Stochastic Gradient Descent(SGD), Ada Grad and L-BFGS, Nesterov's Accelerated Gradient and Adam.

Stochastic Gradient Descent Optimizing Weights W and Bias B

If computing Training Loss ∇L takes 1× memory computing Gradient Descent takes 3× memory. One way is to choose random sample. Then compute Training Loss ∇L for that sample. Next take derivative of that sample. Repeat this step for different random samples multiple times till numerical stability is obtained. This is called Stochastic

THE IIOAB JOURNAL
ISSN: 0976-3104

Gradient Descent which performs better than Gradient Descent and scales well. But in practice it has it's own limitation. **Figure- 8** shows SGD with Momentum which can easily navigate to local minimum even if the area of surface curves more steeply in one direction.

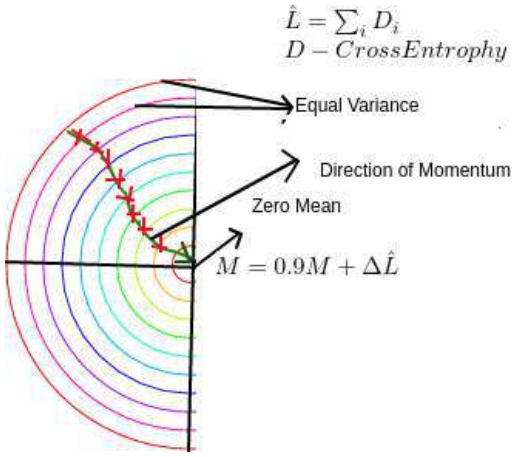


Fig: 8. SGD with Momentum

REGULARISATION

Deep learning technique is successful only if we have right amount of data to train the model. Network with right size of data is very hard to optimize. In practice we train network that is very big for model then find to try the best. In **Figure- 9** shown when to stop training process a time when the accuracy rate doesn't change considerable even after repeated training.

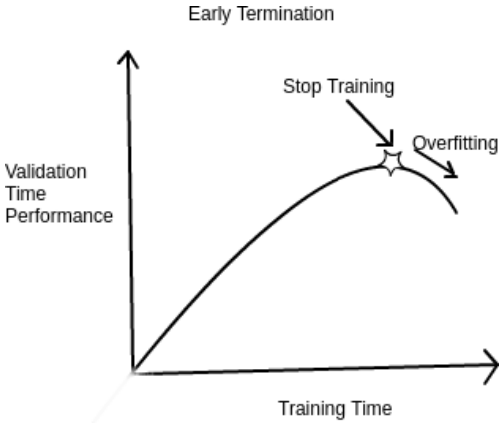


Fig: 9. Early Termination of Training Data in CNN

Regularization is adding artificial constraints to network which implicitly reduces network parameter. In L2-Regularization L2 Norm is added to training loss $\hat{L}$ as shown in Eqn. 19

$$\hat{L} = L + \beta \frac{1}{2}\|w\|_2^2 \qquad (19)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

L2 Norm is sum of the squares of individual weights in the vector.

$$L2Norm = \frac{1}{2}(w_1^2 + w_2^2 + w_3^2 + w_4^2 + ... + w_n^2) \quad (20)$$

Regularizing through early stopping , results in fast training and good generalization performance. A small-enough step-size w.r.t learning rate is often sufficient for state-of-the-art performance. One-vs-rest strategy is a flexible option for large-scale image classification.

Dropout

The parameter that passes from one network to another network is called activators. Take half activations flowing through network randomly and destroy them randomly again. If one activation get smashed there is always one or more activations which do the same job that are not destroyed. Dropouts prevent overfitting and make it more robust . If dropout doesn't work we should increase network size. Randomly destroy these activators in each layer. In **Figure- 10** the value 0.2 and 0.9 are eliminated during training. Scale all activation parameters by 2 and subtract them at output to determine accuracy improved using dropout.

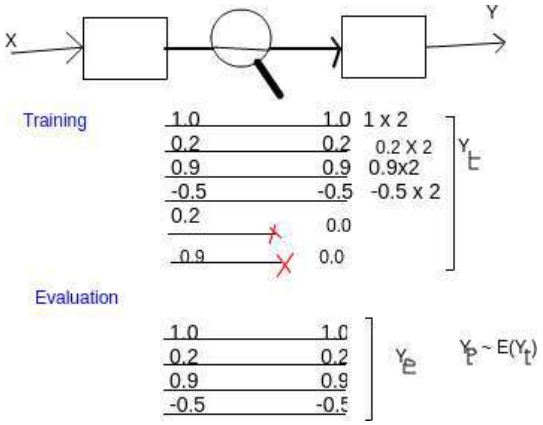


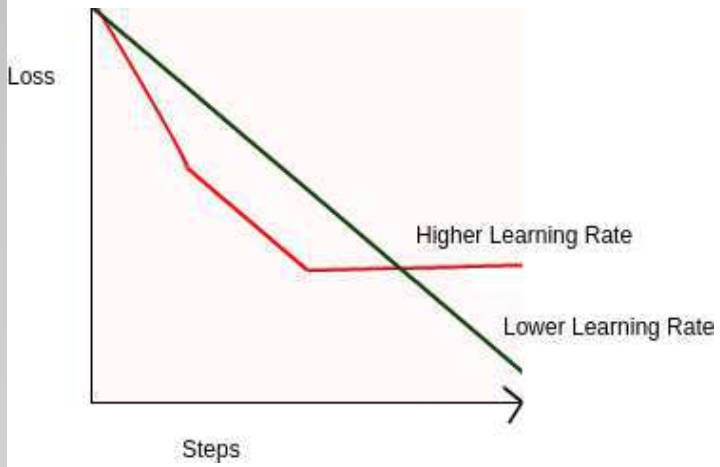
Fig: 10. Droupout: Redundant Activation Smashed to Aoid Overfitting

.....

Dropout's doesn't work well with RNN and LSTM. **Momentum** Momentum techniques leads to better convergence rate. Take running average of gradient to find momentum.

Learning Rate Tuning

In practice the training is done with lower learning rate 0.2. With higher learning rate the image classification model tends to learn quickly but doesn't converge. Training loss is average cross entropy. We do to minimize distance between similar labels and maximize distance between dissimilar labels. Training loss is computed as shown in Eqn. 4. As shown in **Figure- :11** lower learning rates stabilize over a period of time.



THE IIOAB JOURNAL ISSN: 0976-3104

Fig:11. Learning Rate Turning

CNN MODELS IN LITERATURE

Convolutional Neural Networks are designed to take raw pixel as input and visualize the pattern present in the image with minimal preprocessing. The CNN learns at each layer starts from learning lines then edges then detect shapes and finally objects. Ross Girshick, Donahue et al. [3] proposed model to combine region information to localize and segment images. This approach is called R-CNN. A detailed description of famous CNN models described below.

LeNet

Lecun,Bottou et al. in 1998 [4] proposed LeNet5. The Lenet-5 model is capable of recognising hand written digits with extreme variability. Graphical depiction of LeNet-5 model is shown in **Figure- :12**. Convolution Layer and max pooling layer forms the core of LeNet5 model. 32 X 32 handwritten image is convoluted and 6 features map is formed. Then maxpooling is done to reduce size of feature map to 14 X 14. A non linearity layer is inserted between layer 1 and layer2. Tanh function is used as nonlinearity function. After nonlinearity function feature map is convoluted and 10 X 10 size feature map is formed. The last layer is fully connected layer. It creates a fully connected bipartite graph between input layers and output layers which are class labels in case of supervised classification of images.

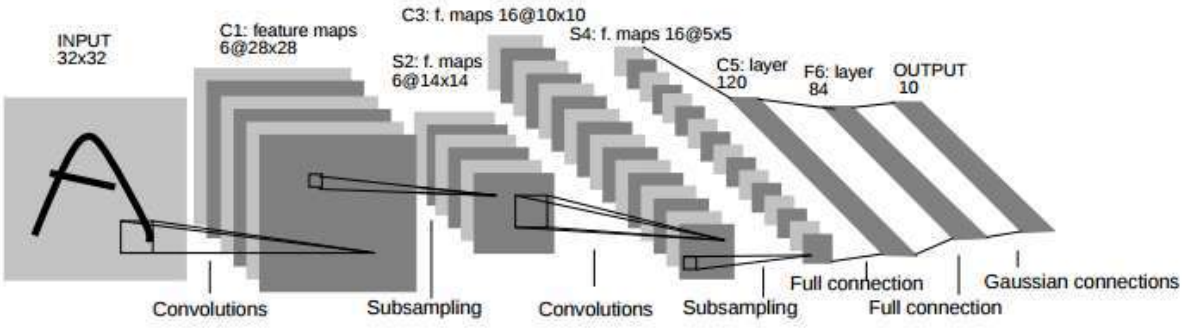


Fig:12. LetNet5 Architecture Image Credit: [4]

AlexNet

Krizhevsky, Sutskever et al. in 2012 [5] proposed Alexnet which won ImageNet competition.It uses the same model as LeNet-5 but a bigger model, mored data, and GPU implementation. It was able to classify 1000 objects. The accuracy of this model is high when compared to previous years hand engineered features which existed for decades. It had 7 hidden layers and 60,000,000 parameters trained on GPU for 2 weeks. **Figure- 13** shows graphical depiction of Alexnet architecture with visualization of features learned below. Neurons learned to detect simple edge and blob in layer1, texture pattern in layer 3 and object classes in layer 5.

THE IIOAB JOURNAL
ISSN: 0976-3104

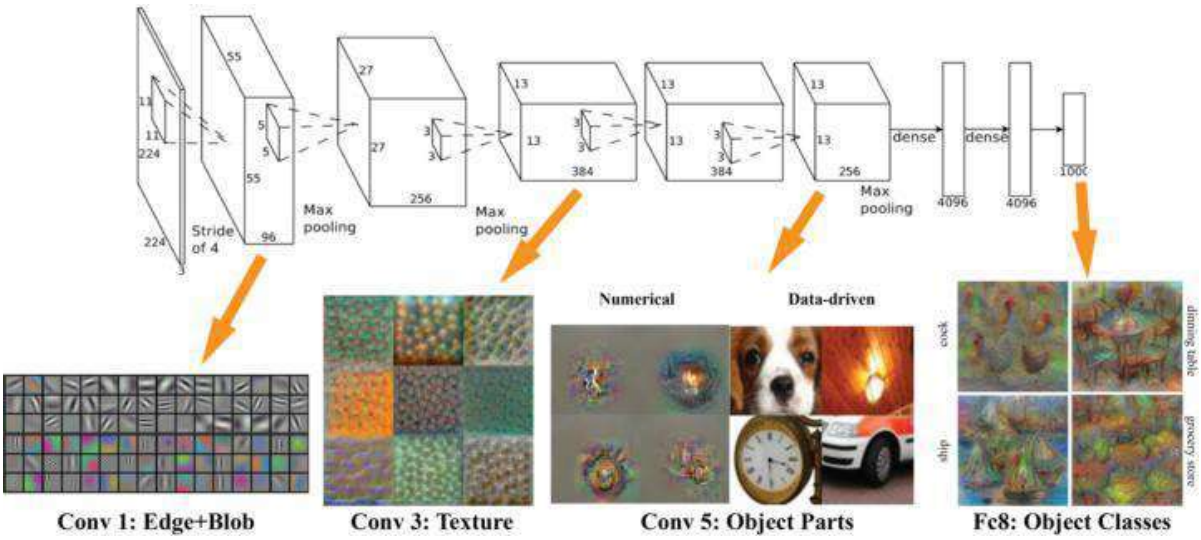


Fig:13. AlexNet Architecture Src: [6]

.....

GoogLeNets

Inception Module Szegedy,Liu et al. [7] which are parallel convolution and pooling layers merged with previous convolution stage. **Figure- 1** is graphical depiction of GoogLeNet Architecture is shown. Idea is at each layer of convnet you can make a choice have a pooling operation, have a convolution, then decide is it 1 X 1 convolution or 3 X 3 convolution or 5 X 5 convolution. All of these are actually beneficial to the modelling power of the network.Average pooling and 1 X 1 convolutions are better than convnets that simply use a pyramid of convolutions.

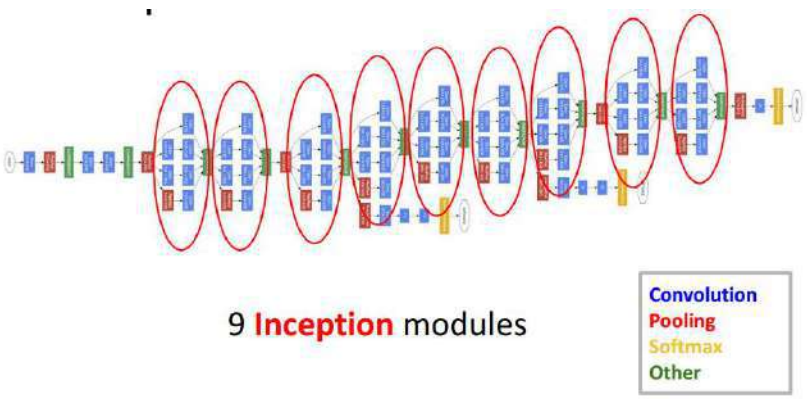


Fig: 14: GoogLeNet Architecture Src: [7]

.....

ResNets

In 2015 Kaiming He, Xiangyu Zhang et al. [8] proposed Residual Neural Networks wich solves the probem of optimizing Feed Forward Neural Network beyond certain depth.They add residual block to overcome this challenge. A residual block is constructed by by passing few convolutional layers at a time. **Figure- :15** is graphical depiction of ResNet architecture at the left and a modified approach at middle and without Relu at right.

THE IIOAB JOURNAL
ISSN: 0976-3104

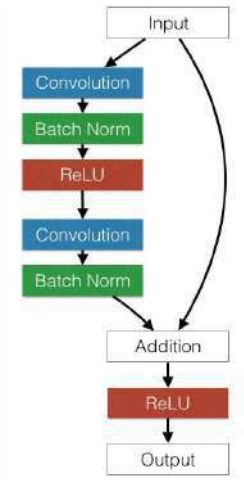


Fig:15.ResNets Architecture Image Credit: Tourch, Training and investigating Residual Nets [8]

Decision Forest Deep Neural Networks DNDF and other CNN models

Kontschieder,Fiterau et al. [9] model with 12 layers beats GoogLeNet's 6.67% top-5 error on ImageNet to 6.38% accuracy.In [10] Oquab, Bottou et al. mentions weakly supervised model is used to recognize objects from cluttered scenes. Saliency inspired neural network model by Erhan, Szegedy et al [11] is used to localize images. Wang,Huang et al [12] used weekly supervised model using latent category learning. Ontology based hierarchial image annotation was proposed by Zarka,Ammar et al. [13]. Part based CNN method was used by zhang2014,Donahue et al [14] for category detection.

AUTOMATIC IMAGE CAPTIONING USING CNN AND RNN

In 2015, Klein,Lev et al [15] used Fisher Vectors derived from HGLMMs (Hybrid Gaussian-Laplacian Mixture Model) to represent sentence. Donahue,Anne et al [16] used Learning Long term dependency called LSTM to avoid the problem of vanishing gradients.A problem which occurs if we need to generate long sentence. This problem was solved by using memory unit which stores past state. Text embedding is done to convert words into vectors that can be used for auto image captioning.

Text Embedding

There are two challenges in text embedding. First is all words used in real world may not be present in training sets. For example the word 'tiffin'meaning light meal appears chiefly in Indian documents. While training a restaurant recommendation system the training data contains the word 'lunch' and 'breakfast' but not the word 'tiffin'. To do this mapping perfectly we need to know word context. The second challenge is sharing the weights between semantically related words. For example the words 'daughter' and 'princess' are related to each other. In **Figure-16** the semantic similarity similar words for the word 'daughter' is depicted. The word 'princess', 'sister' shares weight 0.7697988749 and sister shares weight 0.7702152133 which is very close to the weight of the original word 'daughter' which has weight of 0.7847946882 as shown in **Table- 2**

Table: 2.Word Similarity Based on Semantic Meaning for Word 'Daughter'

Word	Vector
princess	0.7697988749

THE IIOAB JOURNAL
ISSN: 0976-3104

sister	0.7702152133
cousin	0.7743015289
grandmother	0.775909245
daughters	0.7847946882
niece	0.7886790037
mother	0.7911210656
widow	0.8072112799
son	0.8378984928
wife	0.8501636982

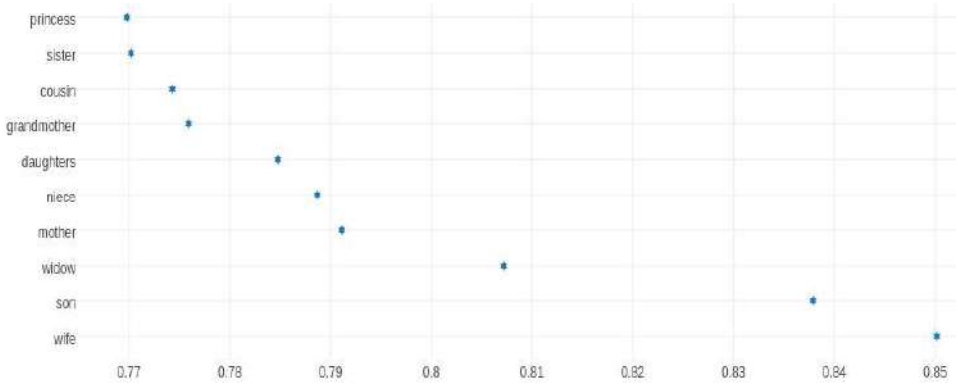


Fig: 16. Word2Vec similarity for word 'daughter'

In. Eqn.21 computing distance between two words are shown.

$$CosineSimilarity = \frac{V_{w1} V_{w2}}{||V_{w1}||||V_{w2}||} \qquad (21)$$

In sampled softmax take random samples of the target and predict its nearest neighbors. This increases the performance of the model. To generate sentence Recurrent Neural Networks are used. CNN share weights across space. RNN shares weight across time. Gated Recurrent Network (GRU) model is recently becoming popular to generate sentence. In **Figure- 17** the process flow of word2vec model is depicted.

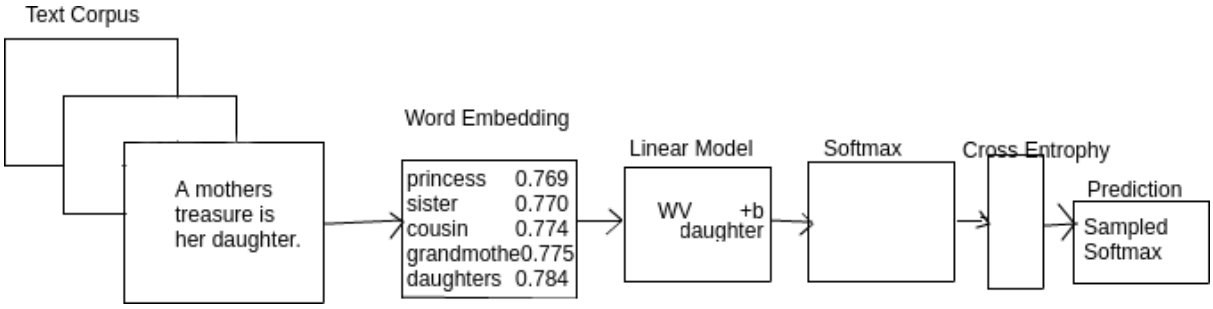


Fig: 17. Word2Vec Process Flow for word 'daughter'

Sutskever,Vinyals et al. [17] sequence to sequence mapping is done using LSTM. The sentence generated can also be used for translation. Luong,Sutskever et al [18] out-of-vocabulary (OOV) word emitted during sentence generation is fed to french dictionary for translation. The advantage is the model can be used to translate in different languages at no additional cost. Vinyals, Toshev et al. [19] has solved the challenge of connecting computer vision to natural language processing. They have proposed model named NIC that can automatically view an image and generate description in English by connecting CNN with RNN as shown in **Figure- 18**

THE IIOAB JOURNAL
ISSN: 0976-3104

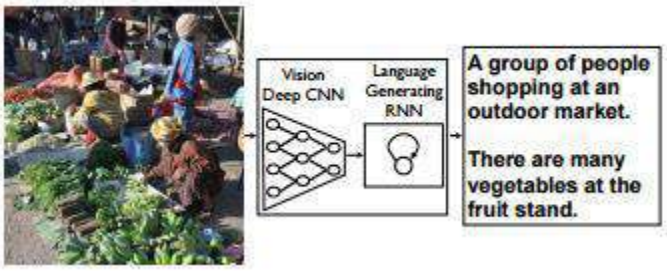


Fig: 18: CNN followed by Language Generating RNN. Image Credit : Vinyals, Toshev et al. 2014 [19]

Karpathy,Li et al. [20] proposed alignment model a combination of CNN image regions and bidirectional RNN over sentences for automatic image description. This is used for region level annotation.Lebret,Pinheiro et al. [21] has combined CNN with phrases to generate sentence. Using sematic information for video annotation is shown in [22].

VISUALIZING CNN

In [23] has shown a method to deconvolve neural network and plot the output. As the images passes through these layers depending on the features seen in input image the corresponding neuron gets activated as shown in **Figure- 6** At each layer the transform function from one neuron to other neuron is highlighted. **Figure- 19** shows learned features at Layer 4 and Layer 5. It has learned lines, edges, shapes and finally objects. Visualizing CNN has improved performance by fine tuning parameters at each layers. Clarifai model which first introduced Visualizing CNN is constructed using six layers. It used backpropagation technique to deconvolve and plot intermediate layers.

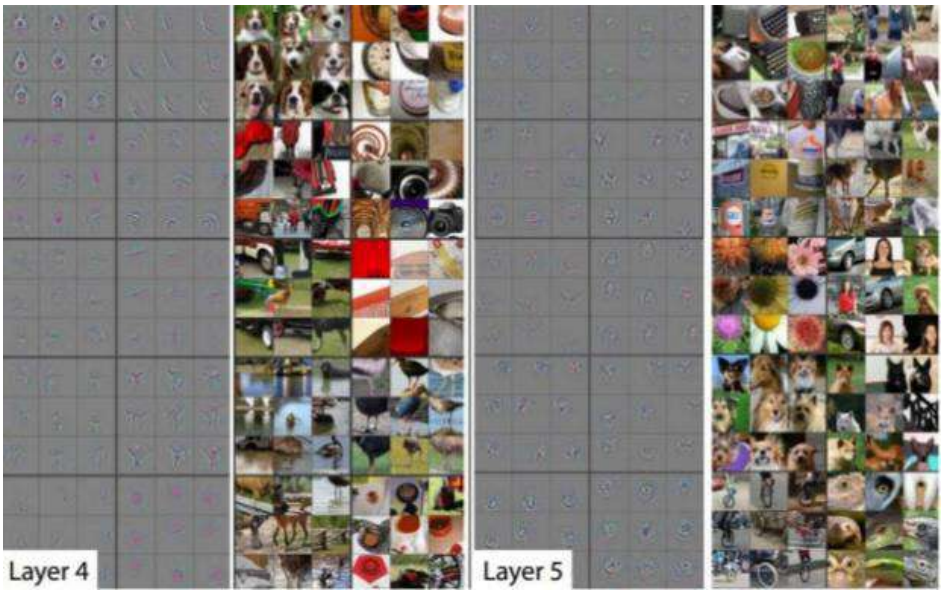


Fig: 19. Visualizing and Understanding CNN. Image Credit : zeiler & Fergus, ECCV2014 [23]

THE IIOAB JOURNAL ISSN: 0976-3104

PERFORMANCE EVALUATION FOR CNN

Entire dataset is divided into training, testing and validation. Testing set is subset of training set which is not used during training. Change the parameters and measure the performance with validation set. Most researches use 70% of data for training set, 10% or more than 30,000 images whichever is higher as validation set and 20% as testing set. A change of 30 image in validation set to improve the accuracy by 1% or more is statistically significant size of validation set. To change the accuracy from 92% to 95% minimum 90 image parameters in validation set has to be changed. The math is as follows $\frac{3\times3000}{100}$=90 when we achieve 3% improvement in accuracy with 90 image parameters fine turned that indicates that accuracy is increasing.

APPLICATIONS OF CNN

CNN is used for gaming, image completion, object detection, object segmentation, to do pose estimation, for pattern matching and artificial image/music synthesis. Mastering the game of Go using neural networks and Monte-Carlo tree search [24] which achieved 99.8% winning rate. It used Monte-Carlo simulation to play sub games. Research on CNN for Image Completion is useful for morphing images and reconstructs images. CNN is used in identifying drugs.

CONCLUSION

The abundance of research on Convolutional Neural Network suggests that without much preprocessing robust neural network can be modeled for object detection and localization. After 2012 more papers has been published using Convolutional Neural Network model. Instead of detecting general category say dog, the species can also be detected like 'German Shepard'. These annotated images are used for automatic scene description, image retrieval based on query and many more. Localize the image and then annotate improves performance of image annotation. The error rate of Image annotation has significantly reduced to 6.7% in 2014. More number of research papers contributed for object detection and localization using convolutional neural networks in last two years when compared to other approaches like graph based object detection or object detection through statistical models. Also the number of participants in ImageNet challenge is also showing upward trend. Results get better with more data, bigger models, more computation, better algorithms, new insights and improved techniques. Automatic image annotation achieves higher accuracy rate by combining CNN output with RNN and Natural Language Processing techniques gives state-of-the art results. The challenge is correctly understanding all the regions in image and understanding the context to generate sentence are still active research areas. The open problems are unsupervised learning in CNN, Reinforcement learning in CNN, Highly multi-task and transfer learning, Automatic learning of model structures.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] UNC Vision Lab. Large scale visual recognition challenge 2015 (ilsvrc2015). http://www. image-net.org/challenges/LSVRC/2015/results, 2015. [Online; accessed 18-Dec-2015].

[2] Tsung-Yi Lin, Michael Maire, Serge J Belongie, et al[2014]. Microsoft COCO: common objects in context, CoRR, abs/1405.0312

[3] Ross B Girshick, Jeff Donahue et al. [2013]. Rich feature hierarchies for accurate object detection and semantic segmentation, 10.1109/CVPR.2014.81

[4] Y Lecun, L Bottou, Y Bengio et al .[1998] Gradient-based learning applied to document recognition. Proceedings of the *IEEE*, 86(11):2278–2324.

[5] Alex Krizhevsky, Ilya Sutskever et al.[2012] Imagenet classification with deep convo- lutional neural networks. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., p. 1106–1114

[6] Donglai Wei et a, mNeuron: A Matlab Plugin to Visualize Neurons from Deep Models, http://vision03.csail.mit.edu/cnn_art/index.html#v_single, . [Online; accessed 19- Jun- 2016].
[7] Christian Szegedy, Wei Liu et al. [2014], Going deeper with convolutions. CoRR, abs/1409.4842,p.1-9.
[8] Kaiming He, Xiangyu Zhang, et al. [2015], Deep Residual Learning for Image Recognition, CoRR, abs/1512.03385
[9] P. Kontschieder, M. Fiterau, A. Criminisi and S. R. Bulò, [2015], "Deep Neural Decision Forests," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, p. 1467-1475.doi: 10.1109/ICCV.2015.172
[10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, [2015], Is object localization for free? - weakly-supervised learning with convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference p. 685–694
[11] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. [2014], Scalable object detection using deep neural networks. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference , p 2155–2162
[12] Chong Wang, Kaiqi Huang, Weiqiang Ren, Junge Zhang, and S. Maybank[2015], Large-scale weakly supervised object localization via latent category learning. Image Processing, IEEE Transactions on, 24(4):1371–1385.
[13] Mohamed Zarka, Anis Ben Ammar, and Adel M. Alimi. Regimvid [2015],Imageclef scalable concept image annotation task: Ontology based hierarchical image annotation. In Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, p.8-11.
[14] Ning c, Jeff Donahue, Ross B Girshick. Trevor Darrell.[2014] Part-based R-CNNs for Fine-grained Category Detection, CoRR, abs/1407.3867
[15] Klein, B. and Lev, G. and Sadeh, G. and Wolf, L.[2015], Associating neural word embeddings with deep image representations using Fisher Vectors,Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, p.4437-4446
[16] Jeff Donahue, Lisa Anne Hendricks et al. [2015] ,Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389.
[17] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le.[2014] Sequence to sequence learning with neural networks. CoRR, abs/1409.3215
[18] Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. [2014] Addressing the rare word problem in neural machine translation. CoRR, abs/1410.8206
[19] Oriol Vinyals , Alexander Toshev, Samy Bengio, Dumitru Erhan [2014] Show and Tell: {A} Neural Image Caption Generator, CoRR, abs/1411.4555
[20] Andrej Karpathy and Fei-Fei Li.[2014] Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306
[21] R´emi Lebret, Pedro O. Pinheiro, and Ronan Collobert [2015] Phrase-based image captioning. CoRR, abs/1502.03671
[22] Virginia Fernandez Arguedas, Qianni Zhang, Krishna Chandramouli, Ebroul Izquierdo [2013] Vision Based Semantic Analysis of Surveillance Videos. Semantic Hyper/Multimedia Adaptation, p.83-125
[23] Matthew D. Zeiler and Rob Fergus [2013] Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013.
[24] David Silver, Aja Huang, Chris J Maddison, et al.[2016] , Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):p.484–489,
[25] Vincent Vanhoucke, Udacity, Deep Learning, ud730, https://classroom.udacity.com/courses/ud730/lessons/6370362152/concepts/63798118150923, 2015, [Online; accessed 29-Jun-2016].

ABOUT AUTHORS

Jenisha T *is a research scholar at VIT University, Vellore, India. Her area of interest includes Machine Learning,Computer Vision, Deep Learning, Cognitive Computing and Big Data Analytics. She did her B.Tech from St. Xaviers Catholic College of Engineering, Nagercoil,India. She completed her M.Tech in Information Technology in the year 2010.*

Dr. Swarnalatha Purushotham *works as Associate Professor in the School of Computer Science and Engineering,VIT University,Vellore - 632 014,Tamil Nadu, India. Her area of interest includes Digital Image Processing and Artificial Intelligence. She has published more than 40 papers in National and International Journals.*

ISSN: 0976-3104
SPECIAL ISSUE (SCMSA)
Agrawal and Swarnalatha

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

COMBINING SEMANTIC DATA STORE AND BIG DATA FOR PHARMA USE CASE

Pragya Agrawal* and Swarnalatha P.

School of Computing Sciences and Engineering, VIT University, Vellore, Tamil Nadu, INDIA

ABSTRACT

In the pharmaceutical R&D procedure the increased generation of data has failed to generate the estimated returns in terms of better efficiency and pipelines. The failure of existing integration methodology to systematize and apply the available knowledge to the range of real scientific and business issues which influence on not only efficiency but also transparency of information in crucial safety and regulatory applications. The new range of semantic technologies based on ontologies enables the proper integration of knowledge in a way that is reusable by several applications across businesses, from discovery to corporate affairs. This paper supports the use of Semantic Web technologies across health care, life sciences, clinical research and translational medicine which help to increase the accuracy of information mining, retrieve complex entities, combine structured and unstructured analytical queries and create comprehensive queries.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26– March-2016
Published on: 10th –Aug-2016

KEY WORDS

Semantic web technologies;Real World Evidence; clinical pharmacy database; big data; Allegrograph

***Corresponding author: Email:** ag.pragya796@gmail.com, **Tel:** +91-9944912522

INTRODUCTION

Data analysis acts as significant role in recognizing productive use cases in any field. Analytics in Pharmaceutical industry orbits around determining better drugs, management of supply chain, and other competitive reward. Drug detection has so far been based on the historical records of the company, and approximately market research with help from researchers. Though, there is more information in the real world than what is really available with the company alone, for example, EMRs, Insurance claims, Prescriptions, etc. When these real world data place to utilize for analytics would carry a stronger evidence subsequent in better and more dedicated inventions. In simple words, RWE involves assembling and scrutinizing data in what way a drug is actually utilized in the real world, as contrasting to what occurs in a structured clinical research situation with protocols and highly motivated physicians.

(RWE) Real World Evidence [1] can be stated as "insights from anonymous patient-level data using sound commercial and scientific analytics". (RCTs) Randomized controlled trials is the golden way for demonstrating security and effectiveness before launch, but stakeholders are observing for more. They entail outcomes and information about the holistic patient journey. Growing healthcare data, produced through hospital reports and payer claims EMRs could offer that significant addition to RCTs. Applicable sets of that information, combined with clinical, commercial and scientific expertise could allow organizations to support and prove importance during the product lifecycle.

Factors driving RWE –

- Patient Centricity - A drug must be approved to a patient also based on his medical background instead of the disease for which he is being treated presently.
- Peer Trends - Every corporation wants to outperform their peer, get improved medicine in terms of both efficacy and cost leading to an improved market. Competitors' information requirement to be associated with real world data to fulfill the visions.
- Clinical Trials - are RCTs which doesn't provide the same outcomes of analysis exterior to the tested investigational conditions, for e.g., Medications that were tested on one background may not work on the other.

IIOAB JOURNAL
ISSN: 0976-3104

- Observational Data - Each observation of a case (patient) when documented offers more understanding than what is obtainable from the conventional information sources.
- Regulatory Requirements - Pharma corporations compete to enter their drugs into Tier-1 formularies, subsequently drugs in this tier get instant endorsement. To attain this, corporation's intention is on cost and drug effectiveness. Insurance corporations also favor paying for medications in the Tier-1 list which has indication of execution well.

Life sciences corporations are using RWE [1] to provision advances in data technology, external and internal healthcare decision making, engagement and analytics prototypes could further its scientific and commercial influence all functions must study how to join the power of its visions. Fig 1. Represent the multiple data sources that can produce an evidence-based patient journey.

Who Is Using RWE Today?
• Epidemiology, drug safety researchers and HEOR, to increase faster visions from richer patient datasets.
• Pricing and market access colleagues, to inform payers and HTAs with evidence of their products' performance
• Brand and franchise teams, to understand their markets, differentiate their products, and improve stakeholder engagement.
• Clinical development teams, to design trials based on actual treatment practices versus dated and inconsistent guidelines

Fig: 1. Patient Journey: Multiple data sources can create an evidence-based patient journey.

RWE [2] helps us to define subsets of patients who are most benefitted from a medicine, based on their background of genetics, social aspects and variations in disease. RWE [3] may help us decide where the most demanding medical requirements are. It also helps drug companies to terminate the drug candidates that are not going to effectively follow with existing treatments. It also makes us to choose "safety signals" much faster, thereby notifying companies and the public to the hazardous side effects of some medicines.

Real World Data means variety offered in a huge scale streamed at high rates. Data from diverse sources as well as those from patient observations lead to the existence of world data in different forms. Most of these are scarce. The data volume dumped steadily from various sources such as social networks, clinical trials, etc. are so enormous to handle. Semantically identical texts may be characterized differently. So, Natural Language Processing (NLP) needs to implement which is highly complex. NLP will help in bringing diverse data sources together. Semantic web technologies and big data thus comes into the picture as a solution to these.

Semantic Web technologies [4] – is a group of very particular technology standards from the (W3C) World Wide Web Consortium that are reflected to define and relate information inside enterprises and on the web. These standards include:

- a flexible information prototypical (RDF[5-7]),
- ontology languages and schema for defining concepts and relationships (OWL[8] and RDFS),

THE IIOAB JOURNAL ISSN: 0976-3104

- a query language (SPARQL[9-11]),
- a rules language (RIF),
- A language for marking up data inside Web pages (RDFa) and more.

The term "ontology" can be stated as “an explicit specification of conceptualization”. Ontologies establishes the building of the domain, i.e. conceptualization. This contains the model of the domain with probable limitations. The conceptualization defines knowledge about the domain and not approximately the certain state of relationships in the domain. In other words, the conceptualization is not varying, or is varying very seldom. Ontology is then specification of this conceptualization: the conceptualization is specified by using particular modelling language and particular terms. Formal specification is required in order to be able to process ontologies and operate on ontologies automatically.
Ontology [8][12] defines a domain, while a knowledge base defines certain state of relationships. Each knowledge based system or agent has its own knowledge base, and only what can be communicated using an ontology can be kept and used in the knowledge base. When an agent wants to communicate to another agent, he uses the constructs from some ontology. In order to understand in communication, ontologies must be shared between agents.

Medicine informatics is defined as the “field of information science concerned with the analysis, use and dissemination of medical data and information through the application of computers to various aspect of health care and medicine” [13].

In 2007, the (ASHP) American Society of Health System Pharmacists released a position paper that defined this subspecialty area of pharmacy practice as the use , designates the pharmacist’s role in informatics, integration of knowledge, data, technology, information, and automation in the medication-use process for the determination of refining health consequences [14]. The term “big data” has been coined and is defined “as the emerging use of rapidly collected, complex data “[15]. Big data is a term well-defined in three V’s: volume, velocity, and variety. Other dimensions may also include complexity and variability. The Centre for US Health System Reform released a paper that defines the revolution of big data in health care and cites four major sources of big data that include:

- Pharmaceutical research and development from pharmaceutical companies and academia, clinical trials, and high-throughput screening libraries.
- Clinical data provided by the electronic medical record (EMR) that contain patient-specific data on treatment outcomes.
- Claims and cost data from payers and providers that contain utilization of care and cost estimates
- Patient behaviour and sentiment data that come from consumers and stakeholders outside of health care(for instance, from retail exercise apparel and exercise monitoring equipment) [16]

In this paper the use of Semantic Web technologies and Real World Evidence across health care, life sciences, clinical research and translational medicine will help to increase the precision of information mining, retrieve complex entities, combine structured and unstructured analytical queries and create comprehensive queries.

MATERIALS AND METHODS

Prediction based data aggregation- a survey

The energy management is one of the major issues in wireless sensor networks. A sensor utilizes high energy for communication rather than sensing and processing. The redundant communication in noisy channels causes the depletion of network energy. The prediction based data aggregation approach reduced unnecessary data transmission and so energy expenditure in communication subsystem was minimized. Hyuntea Kim et al., [4] exploited linear data prediction method to improve communication efficiency and to minimize energy consumption with data correlation. As the model is designed considering some factors such as the selective transmission, it reduced data accuracy and adjustments in aggregation period caused the network to meet the additional delay. Guiyi Wei et al., [7] proposed a method that saves network energy and eliminates redundant communication by exploiting prediction based data aggregation protocol. However, in this method synchronization time increased due to synchronization has to be done prior to each transmission. Guorui Li et al., [9] proposed an Auto Regressive Integrated Moving Average Model (ARIMA) that predicts the next time value based on the previous observed values. When the prediction error is less than the preconfigured threshold value the aggregator would not transmit the data sensed by the source node. Otherwise, it transmits the data to sink node. Therefore ARIMA model reduced the amount of data transmitted between the ordinary sensor node and aggregator node. Since this method performed aggregation on the ordinary sensor node and aggregator node it increased the computational complexity and reduced accuracy. Rajesh G et al., [5] proposed the data fusion method using Simpson’s 3/8 rule to forecast next time data based on the early sensed information. When prediction error is

ISSN: 0976-3104

greater than the prediction threshold the cluster head transmits the actual sensed value to the base station. Otherwise, it would not transmit data to the base station. This method reduced unnecessary transmission between cluster head and base station. However, this method provides less prediction accuracy since the deviation error is increased between subsequent values. There are several data fusion techniques in Wireless sensor networks. The main features of the proposed work are that it, Improves the performance of the forecast and Performs less computation to obtain the forecasted data.

Chaos theory based data aggregation (CTAg) technique

The typical features of chaos include: 1) Nonlinearity. If it is linear, it cannot be chaotic. 2) Determinism. It has deterministic underlying rules every future state of the system must follow. 3) Sensitivity to initial conditions. Small changes in its initial state can lead to radically different behavior in its final state. Long-term prediction is mostly impossible due to sensitivity to initial conditions. A dynamic system is a simplified model for the time-varying behavior of an actual system [17]. These systems are described using differential equations specifying the rates of change for each variable. A dynamical system of dimension N system first–order differential equations for N variables $x_1(t), x_1(t) \ldots x_N(t)$ evolve with time t according to,

$$\hat{x}_1 = f_1(x_1, x_2, \ldots x_N, t) \quad (3)$$

$$\hat{x}_2 = f_2(x_1, x_2, \ldots x_N, t) \quad (4)$$

$$\hat{x}_N = f_N(x_1, x_2, \ldots x_N, t) \quad (5)$$

Where f_1, f_2 are assigned functions and a dot is a derivative with respect to time.

The system following Characteristics of a Chaotic System:

- Sensitivity to initial conditions
- Non-linear
- Dynamic and mixed topology system and Continuous or periodic time.

So that the Chaos is the aperiodic long–term behavior in a deterministic system that exhibits sensitive dependence on the initial condition. These characteristics enables chaos theory based data aggregation (CTAg) prediction method is suitable for eliminating data redundancy in WSNs.

Considered a hierarchical wireless sensor network G (SN, E) where, SN represents the sensor nodes and E represents links connecting the nodes. These sensor nodes collect weather monitoring data (Temperature, Humidity) periodically. Each node transmits data to sink node through the intermediate node or aggregator node (A). The aggregator (A) will perform data fusion by eliminating redundant data using chaos theory before transmitting the gathered data towards the base station. This will minimize the amount of data transmitted between aggregator node and sink node.

Steps for Ontology Learning or Enrichment-

1. Convert xml document into owl.
2. Perform computational mapping of rawterms to ontology through NLP and get intelligent raw data.
3. Through intelligent raw data get unmapped terms and mapped terms through NLP.
4. Manually add the unmapped terms using Ontology editor. Mapped terms will be added computationally to the ontology.

Fig 2, shows the ontology enrichment where new tuples are inserted or updated as and when required to achieve the desired result.

THE IIOAB JOURNAL
ISSN: 0976-3104

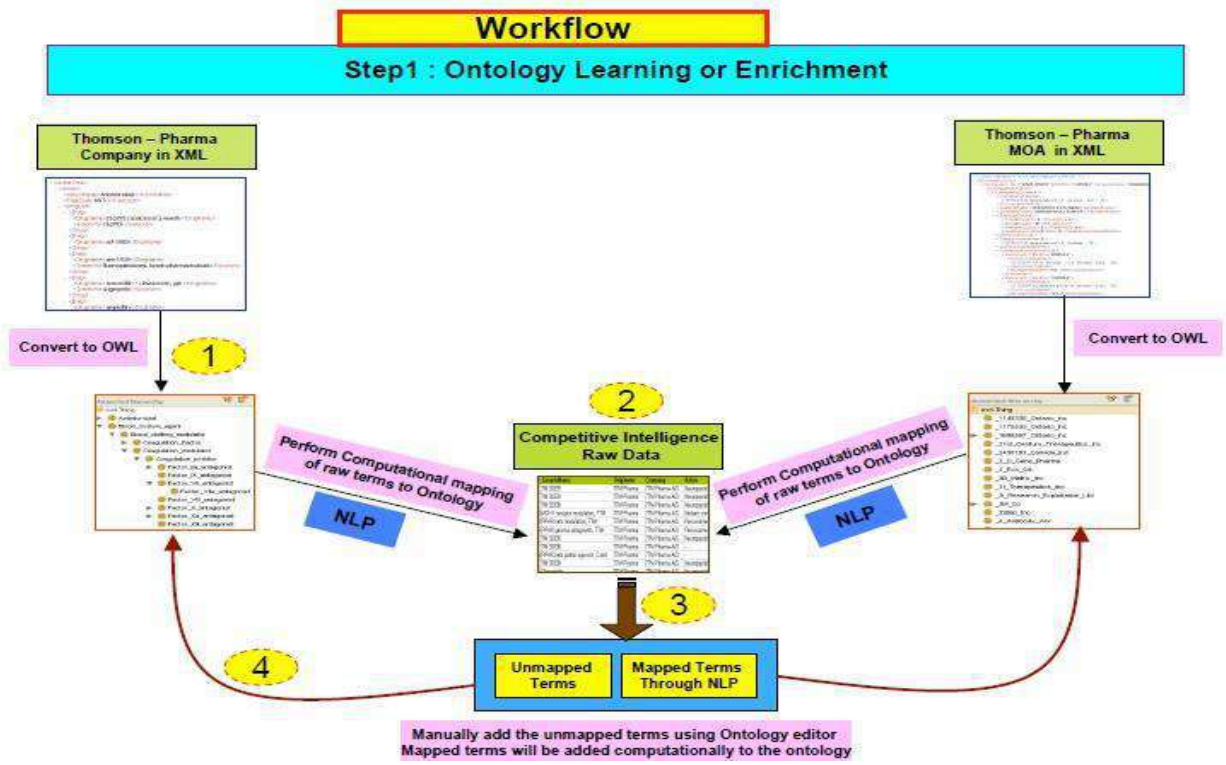


Fig: 2.Ontology Learning or enrichment

Steps for RDF conversion, loading and querying-

1. Ontology is coverted to rdf.
2. RDF is coverted using Jena Programming API.
3. Coverted data is uploaded to Allelograph Native RDF triple store.
4. After uploading, data is queried using SPARQL

Figure 3, shows RDF conversion where xml file is converted to rdf file. After conversion, the rdf file is loaded to Allegrograph and sparql query is performed.

IIOAB JOURNAL
ISSN: 0976-3104

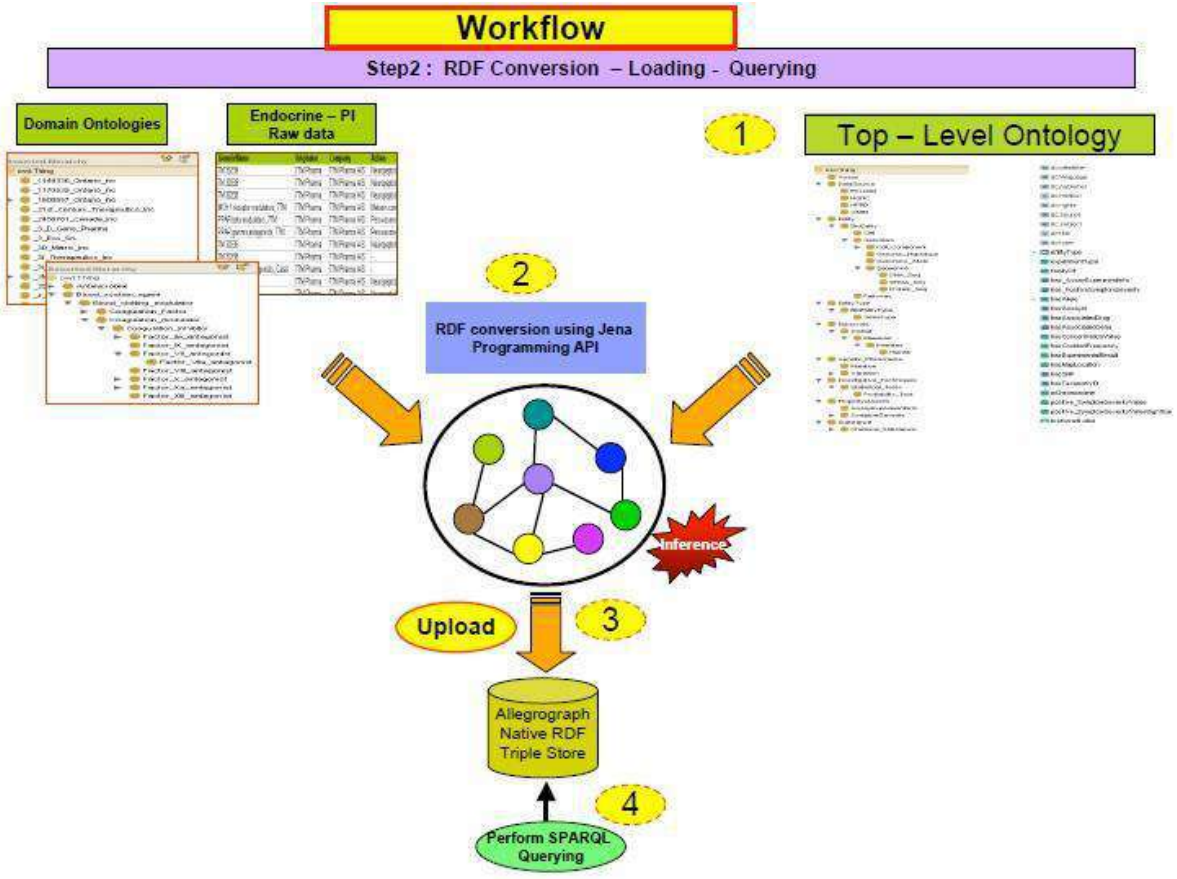


Fig: 3.RDF conversion

……

Finding the market share of drugs.

- Provides details on which brand has a better market share on every state, which could indicate the possibilities of another brand with the same ingredients to improve their market share on that state.
- Data Considered
  - CMS[25] for drug name, total drug cost, provider city, provider state and specialty description.
  - RxNorm[24] for preflabel and tradename.

CMS: As part of the Obama Administration's struggles to create our healthcare system more apparent, reasonable, and responsible. The (CMS) Centers for Medicare & Medicaid Services have organized a public data set, the Part D Prescriber Public Use File ("Part D Prescriber PUF"), with information on prescription drug events (PDEs) incurred by Medicare beneficiaries with a Part D prescription drug plan. The Part D Prescriber PUF is organized by National Provider Identifier (NPI) and drug name and contains information on drug utilization (claim countsand day supply) and total drug costs.

Data Content

- NPI – National Provider Identifier (NPI) for the performing provider on the claim.
- NPPES_ENTITY_CODE – Type of entity reported in NPPES. An entity code of 'I' identifies providers registered as individuals and an entity type code of 'O' identifies providers registered as organizations
- NPPES_PROVIDER_LAST_ORG_NAME – individual (entity type code='I'), this is the provider's last name. Entity type code = 'O', this is the organization name.
- NPPES_PROVIDER_FIRST_NAM
- NPPES_PROVIDER_GENDER
- DESCRIPTION_FLAG – A flag variable that indicates the source of the specialty_description.

THE IIOAB JOURNAL ISSN: 0976-3104

- DRUG_NAME – The name of the drug filled. This includes both brand names and generic names.
- GENERIC_NAME – A term referring to the chemical ingredient of a drug rather than the advertised brand name under which the drug is sold.
- BENE_COUNT – The total number of unique Medicare Part D beneficiaries with at least one claim for the drug. Beneficiary counts fewer than 11 are not displayed.
- TOTAL_CLAIM_COUNT – The number of Medicare Part D claims. This includes original prescriptions and refills. Claims counts fewer than 11 are not displayed.
- TOTAL_DAY_SUPPLY – The aggregate number of days' supply for which this drug was dispensed.
- TOTAL_DRUG_COST – The aggregate total drug cost paid for all associated claims. This amount includes ingredient cost, dispensing fee, sales tax, and any applicable vaccine administration fees.
- BENE_COUNT_GE65 – The total number of unique Medicare Part D beneficiaries with at least one claim for the drug where the beneficiary is 65 or older. Beneficiary counts fewer than 11 are not displayed.
- TOTAL_CLAIM_COUNT_GE65 – The number of Medicare Part D claims where the beneficiary is 65 or older. This includes original prescriptions and refills. Claims counts fewer than 11 are not displayed.
- DAY_SUPPLY_GE65 – The aggregate number of days' supply for which this drug was dispensed, where the beneficiary is 65 or older.
- TOTAL_DRUG_COST_GE65 – The aggregate total drug cost paid for all associated claims where the beneficiary is 65 or older. This amount includes ingredient cost, dispensing fee, sales tax, and any applicable vaccine administration fees.

npi	nppes_provider_last_org_name	nppes_provider_first_name	nppes_provider_city	nppes_provider_state	specialty_description	description_flag
1003889502	AACHI	VENKAT	SAN JOSE	CA	Physical Medicine and R	S
1336172998	AAMODT	DENISE	ALBUQUERQUE	NM	Family Practice	S
1255569273	AANDERUD	PAUL	CLACKAMAS	OR	Dermatology	S

drug_name	generic_name	bene_count	total_claim_count	total_day_supply	total_drug_cost	bene_count_ge65
ACETAMINOPHEN-CODEINE	ACETAMINOPHEN WITH CODEINE		14	366	$277.42	
ACETAMINOPHEN-CODEINE	ACETAMINOPHEN WITH CODEINE		19	458	$271.90	
ACETAMINOPHEN-CODEINE	ACETAMINOPHEN WITH CODEINE	18	19	39	$70.68	

bene_count_ge65_redact_flag	total_claim_count_ge65	ge65_redact_flag	total_day_supply_ge65	total_drug_cost_ge65
*		*		
*		*		
#		#		
#		#		
#		#		
*	19		307	$358.59

Fig 4: CMS database

……

RxNorm[17] Ontology:

- Created by NIH's National Library of Medicine.
- Combines several different drug vocabularies.
- A standardized nomenclature for drug names.
- Unifies vocabularies around RXCUI, a concept unique identifier

THE IIOAB JOURNAL ISSN: 0976-3104

Data Content-

Properties

- Notation
- PrefLabel
- RXAUI
- RXCUI
- Cui
- Tui
- hassty
- constitutes
- RXN_STRENGTH
- RXN_IN_EXPRESSED_FLAG
- RXN_AVAILABLE_STRENGTH
- altLabel
- RXTERM_FORM
- RXN_HUMAN_DRUG

Properties

- RXN_HUMAN_DRUG
- RXTERM_FORM
- NDC
- RXN_QUANNTITY
- ORIG_SOURCE
- RXN_ACTIVATED
- RXN_OBSOLETE
- contains
- ORIG_CODE
- RXN_BN_CARDINALITY
- ORIG_CODE
- ORIG_TTY
- ORIG_VSAB

Relationship

- Has_ingredient
- Inverse_isa
- isa
- Has_dose_form
- form_of
- precise_ingredient
- has_tradename
- consists_of
- Has_part
- Ingredient_of
- Part_of

Drug-Drug interactions

The aim is to find the drug to drug interactions that lead to adverse reactions among patients and report such cases as a notification to doctors, pharma companies, etc. for alerting them of such interactions which would lead to better prescription knowledge amongthe doctors. To make inflight adjustments, recommend Concomitant drugs, and drug label enhancements by the pharmaceutical companies is achieved. New cases of drug interactions would also be used for enrichment of the Ontologies dealing with drug-drug interactions.
Data Sources required:-
(I) The FAERS data provided by the FDA has information on the adverse events and outcomes, a patient has undergone and the drugs consumed during each of these events by the patient.
There are also information on:-
1. Patient details like age, sex, demography of event, etc.
2. Drug details like name, brand, active ingredient, dose form, etc.
3. Adverse reactions like headache, chest pain, etc.
4. Indications for which the drug is to be taken like headache, nausea, etc.
5. Outcomes like hospitalization, death, etc.
6. Source of Information - Whether it is from the doctor, consumer, distributor, study material, etc.
7. Dates on which the therapy has taken place.
(II) The DrugBank data/ontology which has the information on drug-drug interactions. DrugBank is available in xml. It can be converted to RDF and uploaded into any triple store like Allegrograph.

RESULTS

Market Share of drugs

This evaluation has been done in Allegrograph where a federated session was created between RxNorm and CMS.A drug from CMS was referred in RxNorm to get the GENERIC Name of the drug; other brand drugs that were having the same generic name were then listed; this list was returned to CMS for finding the total drug cost of each of those drugs in the list generated from RxNorm.
Actamin has the Generic Name Acetaminophen. Drugs like Tylenol, Dolphin, Hydrocodone, etc. might behaving the same base component generic name Acetaminophen. Then all these drugs are searched for their total market share from CMS to return the brand that has the best share in the market and could also be filtered for an area like New York "NY".There were problems of exact string matches between the multiple databases, so NLP was proposed to be used to solve the problem

THE IIOAB JOURNAL ISSN: 0976-3104

100 Results in 22.594 s | Warnings

prefLabel	tradename_of	gen_rx_name	has_tradename	tradename	drug_name
"Actamin"	161	"Acetaminophen"	1052413	"Pamprin Max Formula"	"ABILIFY"
"Actamin Oral Product"	1152842	"Acetaminophen Oral Product"	1437472	"Dolofin Infantil Oral Product"	"ABILIFY"
"Actamin Pill"	1152843	"Acetaminophen Pill"	1187315	"Tylenol Pill"	"ABILIFY"
"Acetaminophen 325 MG Oral Tablet [Actamin]"	313782	"Acetaminophen 325 MG Oral Tablet"	209384	"Acetaminophen 325 MG Oral Tablet [Tycolene]"	"ABILIFY"
"Acetaminophen 325 MG [Actamin]"	315263	"Acetaminophen 325 MG"	570402	"Acetaminophen 325 MG / butalbital 50 MG [Marten-Tab]"	"ABILIFY"
"Acetaminophen 500 MG [Actamin]"	315266	"Acetaminophen 500 MG"	1101750	"Acetaminophen 500 MG / pamabrom 25 MG [Midol Teen]"	"ABILIFY"

100 Results in 22.594 s | Warnings

drug_name	gen_cms_name	provider_state	provider_city	specialty_description	total_drug_cost
"ABILIFY"	"ARIPIPRAZOLE"	"NY"	"GENESEO"	"Internal Medicine"	"6.48193E3"
"ABILIFY"	"ARIPIPRAZOLE"	"NY"	"GENESEO"	"Internal Medicine"	"6.48193E3"
"ABILIFY"	"ARIPIPRAZOLE"	"NY"	"GENESEO"	"Internal Medicine"	"6.48193E3"
"ABILIFY"	"ARIPIPRAZOLE"	"NY"	"GENESEO"	"Internal Medicine"	"6.48193E3"
"ABILIFY"	"ARIPIPRAZOLE"	"NY"	"GENESEO"	"Internal Medicine"	"6.48193E3"
"ABILIFY"	"ARIPIPRAZOLE"	"NY"	"GENESEO"	"Internal Medicine"	"6.48193E3"

Fig: 5. Shows result drug name and total drug cost of Geneseo city

………..

Drug-Drug interaction

It provides how one drug reacts in the presence of other drug.

Figure 6. Shows how drug Lipitor react in the presence of drug aliskiren (Atorvastatin may increase the serum concentration of Aliskiren.)

ISSN: 0976-3104

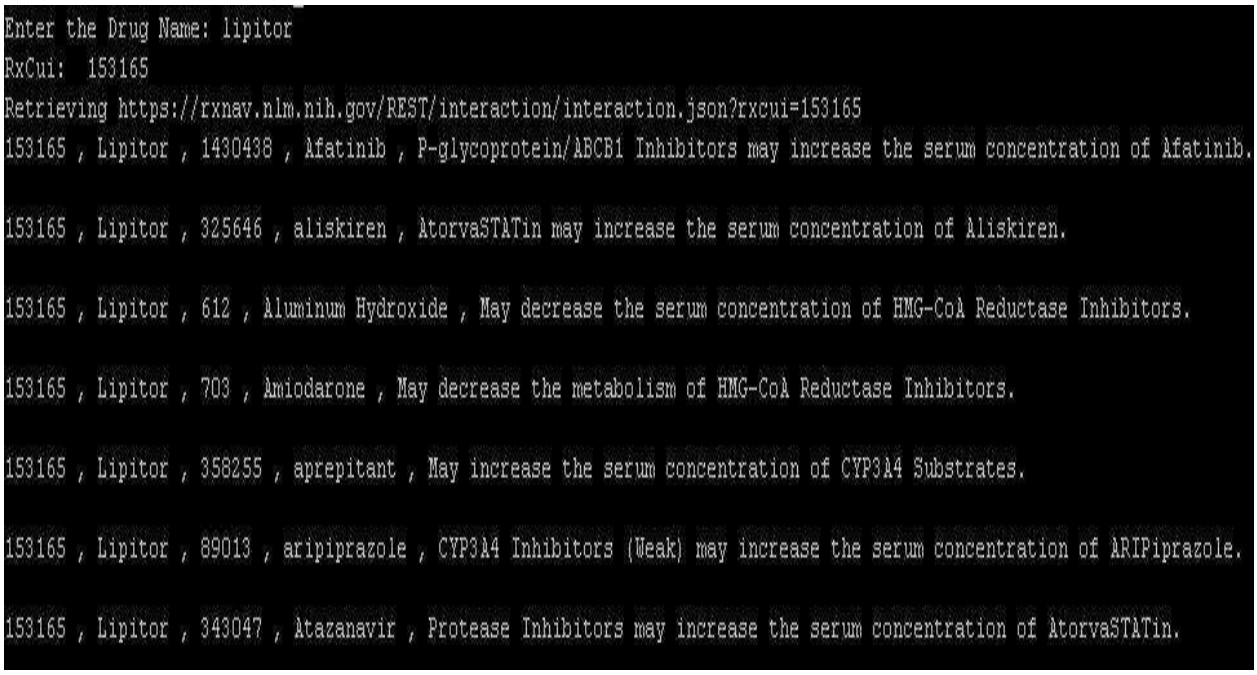


Fig: 6. Shows how Lipitor interact with other drug.

Here one drug interact with other drug and causes some adverse reaction to the patient.

Figure- 7, shows when one patient take drug named "6-(3'-5' Dimethylbenyl)-1-ethoxmethyl-5-isopropyluracil" with other drug named "Dabrafenib" may decrease the excretion of amephetamines.

CONCLUSION

The Real World Evidence in the pharmaceutical domain is an ultimate goal towards achieving better healthcare as per the Obama-Care and a platform that would showcase the implementation of this would be the first step in at least visualizing the RWE. RDF data model combined with Semantic Integration (instance mapping using NLP) was effective in answering questioning Competitive Intelligence. Ontologies provide a powerful framework in providing dictionaries and taxonomical relations that help to reason and inference the data for knowledge discovery. Manual curation is a tedious, error prone and labor intensive task. A semi-automated intelligent computer based solution that utilizes Ontologies, Semantic Integration and NLP could drastically reduce manual curation process and maintain high quality information.

ISSN: 0976-3104

Fig: 7.Drug-Drug Interaction

……..

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] Andreas M. Pleil "Using Real World Data in Pharmacoeconomic Evaluations: Challenges, Opportunities, and Approaches" Worldwide Medical and Outcomes Research Pfizer, Inc., La Jolla, CA

[2] Mack C and Lang K. Using Real-World Data for Outcomes Research and Comparative Effectiveness Studies. Drug Discovery & Development. Nov. 4, 2014. Accessed at: http://www.dddmag.com/articles/2014/11/using-real-world-data-outcomes-research-and-comparative-effectiveness-studies.

[3] Christel M. "Evidence" Trail Elusive. Applied Clinical Trials. June 30, 2014. Accessed at: http://www.appliedclinicaltrialsonline.com/evidence-trail- elusive-0.

[4] John Davis, Rudi Studer, Paul Warren "Semantic Web Technologies: trends and research in ontology based system". USA, Willey;2006 ISBN: 978-0470025963

[5] Hyunsik Choi , Jihoon Son , YongHyun Cho , Min Kyoung Sung , Yon Dohn Chung, SPIDER: a system for scalable, parallel / distributed evaluation of large-scale RDF data, Proceedings of the 18th ACM conference on Information and knowledge management, November 02-06, 2009, Hong Kong, China [doi>10.1145/1645953.1646315].

[6] Resource Description Framework(RDF): Concepts and Abstract Syntax, http://www.w3.org/TR/rdf-concepts/

[7] Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C, G. Klyne and J. J. Carroll. Resource Description 2004.

[8] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, and L.A. Stein. OWL Web Ontology Language Reference. Technical report, W3C, 2004.

[9] SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql- query/.

[10] Bernstein, M. Stocker, and C. Kiefer. SPARQL Query Optimization Using Selectivity Estimation. In Poster Proceedings of the 6th International Semantic Web Conference (ISWC), Lecture Notes in Computer Science. Springer, 2007.
[11] Olaf Hartig , Ralf Heese, The SPARQL Query Graph Model for Query Optimization, Proceedings of the 4th European conference on The Semantic Web: Research and Applications, June 03-07, 2007, Innsbruck, Austria [doi>10.1007/978-3- 540-72667- 8_40]
[12] Making a Semantic Web Business Case at Pfizer
[13] National Library of Medicine. Collection development manual: medical informatics. Available from: http://www.nlm.nih.gov/tsd/acquisitions/cdm/CDMBook .pdf. Accessed on June 5, 2015.
[14] Proceedings of the Pharmacy Practice Model Summit: An invitational consensus conference conducted by ASHP and the ASHP Research and Education Foundation, November 7–9, 2010, Dallas, Texas. Available from http://www.ashp.org/DocLibrary/PPMI/PPMISummit/Proceedings.aspx
[15] Erickson AK, editor, Pharmacy Today. Pharmacy:Harnessing the power of big data. American Pharmacists Association. http://www.pharmacist.com/pharmacy. Accessed March 20, 2015.
[16] Groves P, Kayyali B, Knott D, Van Kuiken S. The “big data” revolution in healthcare: accelerating value and innovation. New York, NY:Center for US Health System Reform, Business Technology Office, McKinsey and Company; 2013.
[17] RxNORM, http://bioportal.bioontology.org/ontologies/

ABOUT AUTHORS

Pragya Agrawal pursued her Master degree from VIT University, Vellore in Computer Science and Engineering.She has worked as a Project Trainee In Wipro Technologies for duration of 10 months. She has keen interest on Big Data Analytics and had attended short term training program on Big Data Analytics organized on 26th, 28th and 30th September 2014 conducted by School of Computing Science and Engineering and Centre for Ambient Intelligence and Advanced Networking Research at VIT University, Vellore. She got All India Rank-97 in Nationwide Interactive Science Olympiad 2007 in class 12th. She also has keen interest on Semantic Web technologies, Data Mining, Artificial Intelligence and Image Processing.

Swarnalatha Purushotham is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Systems. She has published more than 50 papers in International Journals/International Conference Proceedings/National Conferences. She is having 14+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.

ISSN: 0976-3104

SPECIAL ISSUE (SCMSA)
Purushotham and Tripathy

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

CLUSTER ANALYSIS USING HYBRID SOFT COMPUTING TECHNIQUES

Swarnalatha Purushotham*, BK. Tripathy
VIT University, Vellore-Tamil Nadu, INDIA

ABSTRACT

In a given set of data values with several attributes, similar data points can be clubbed together using a clustering architecture that uses global prototypes. These subset prototypes are exchanged so that a communication link is established between different clustering units. In this paper, a detailed clustering methodology is developed by combining both rough and fuzzy set techniques. This methodology shall be used to formulate a grouping of randomly generated unsupervised data that considers the integration of collaborative clustering in fuzzy data sets.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26th-March-2016
Published on: 10th–Aug -2016

KEY WORDS

Cluster Centre Matrix, Fuzzy Membership, Lower and Upper Approximation, Objective Function, Satellite Images

***Corresponding author: Email:** pswarnalatha@vit.ac.in **Tel.** +40-9001010010; **Fax:** +40-9001010012

INTRODUCTION

In Cluster is a collection of data objects which are similar to one another within the same cluster but dissimilar to the objects in other clusters. The problem is to group N patterns into c possible clusters with high intra-class similarity and low interclass similarity by optimizing an objective function. In objective function-based clustering algorithms, the goal is to find a partition for a given value of c. The c-means algorithm represents each cluster by its center of gravity [1].

The aim of collaborative clustering is to make different clustering methods collaborate, in order to reach at an agreement on the partitioning of a common dataset. As different clustering methods can produce different partitioning of the same dataset, finding a consensual clustering from these results is often a hard task. The collaboration aims to make the methods agree on the partitioning through a refinement of their results. This process tends to make the results more similar. In this paper, after the introduction of the collaboration process, we present different ways to integrate collaboration into already existing methodologies.

The implementation of fuzzy clustering has to be dealt with imprecise data that takes into consideration soft computing algorithms like c-means clustering. The fuzzy data is specifically used to deal with overlapping of data points. Whereas, the rough c-means incorporates the idea of vagueness and it is used cluster imprecise data.

Rough sets are purposed at defining clusters in terms of upper and lower approximations, which are identified by a pair of parameters while computing cluster prototypes. It is to be noted that RCM assigns objects into two distinct regions, viz., lower and upper approximations, such that objects in lower approximation ensures that the object is absolutely in the cluster while those in the upper approximation indicate possible inclusion in it. Since there is no concept of membership involved, therefore any measure of closeness of patterns to the clusters cannot be determined.

The paper [2] deals with a comparative study using RIFCM [3] with other related algorithms from their suitability in analysis of satellite images with other supporting techniques which deals with proving the superiority of RIFCM with RBP in clustering with other clustering methods and other supporting metrics with and without refined which integrates judiciously RIFCM with RBP. Finally, the superiority of the RIFCM using RBP is demonstrated, along

with a comparison with other related algorithms, on satellite images with NASA.org images(Hills, Drought) and national geographic photographic images(Freshwater, Freshwater valley). Several papers have used image segmentation through clustering with various applications in view [4], [5], [6], [7], [28]. A family of clustering algorithms has been established with the use of the kernel function instead of the Euclidean distance [8], [9], [10], [11], [12], [13]. Algorithms have been devised to use mode as the measure of central tendency instead of mean some more clustering algorithms have been devised [14]. Using the possibilistic approach to clustering some algorithms have been proposed [15], [16], [17], [18]. Using covering based rough sets instead of basic rough sets some algorithms have been devised [19]. Some efforts have been done to improve the speed of existing algorithms like in [20]. Clustering of time series data is done in [21]. The initial assignment of input is done arbitrarily in almost all the above algorithms. But using genetic algorithms like the firefly algorithm an algorithm is proposed in [22].

In this paper, we present a novel collaborative clustering through the use of rough–fuzzy sets that is further expanded by means of incorporation of fuzziness powered grouping [23]. The use of rough sets is designed at restricting the effect of uncertainty among patterns that belongs to the upper and lower approximations, during collaboration between the modules. Incorporation of membership, in the RCM framework, is seen to enhance the robustness of clustering as well as collaboration. FRFCM framework is designed such that it is structured at finding data set collaboration.

CLUSTERING ALGORITHMS

Here we are going to describe different clustering algorithms, like c-means, fuzzy c-means, rough c-means and ant colony clustering. We are going to compare and contrast between them.

A. Hard C-means Clustering: Literature Overview

In this algorithm we partition N objects into c clusters. During each iteration centroids of each cluster is calculated. The algorithm goes as follows:

STEP 1: Fix c ($2 \le c < n$) and initialize the $U^{(0)}$ matrix

STEP 2: For r = 0, 1, 2 . . . do

Calculate the c center vectors $v_i^{(r)}$, i =1, 2…with $U^{(r)}$

STEP 3: Calculate the updated characteristic functions (for all i, k) using the formula

$$\chi_{ik}^{(r+1)} = \begin{cases} 1, \ if \ d_{ik}^{(r)} = \min\{d_{jk}^{(r)}\}, for \ all \ j = 1,2,...c \\ 0, otherwise \end{cases} \quad (1)$$

STEP 4: If $\left\| U^{(r+1)} - U^{(r)} \right\| < \varepsilon$ (the pre-assigned value then STOPS

Else r = r+1 and go to STEP 2 (The $\| \ \|$ norm here is the Euclidean norm)

The paper must have proposed system, results, discussion to infer the quality of the research paper. All the figures, equations and etc. must be in high resolution and in good quality.

B. Fuzzy C-Means (FCM) [5]

FCM is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by [4] and improved by [5]) is frequently used in pattern recognition. It is based on minimization of the following objective function.

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} \mu_{ij}{}^{m} \left\| x_i - c_j \right\|^2, 1 \le m < \infty \quad (2)$$

where μ_{ij} is the degree of membership of the object x_i in the jth cluster, c_i is centre of the i[th] cluster, and ||*|| is any norm expressing the similarity between data and center [6].

The Fuzzy c-means algorithm has the following steps:

STEP 1: Fix c ($2 \le c \le n$) and select a value m'

Initialize the partition matrix $U^{(0)}$

$$v_i^{(r)}, i = 1,2,...c$$

For r = 0, 1, 2… Do

STEP 2: Calculate the 'c' centres using the formula

$$v_{ij} = \frac{(\sum_{k=1}^{n} \mu_{ik}^{m'} . x_{kj})}{(3)\sum_{k=1}^{n} \mu_{ik}^{m'}}$$

STEP 3: Update the partition matrix for the r[th] step $U^{(r)}$ to $U^{(r+1)} = (\mu_{ik}^{(r+1)})$, where

Taking $I_k = \{i \mid 2 \le c \le n; d_{ik}^{(r)} = 0\}$

$$\mu_{ik}^{(r+1)} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{2/(m'-1)} \right]^{-1}, if\ I_k = \phi, \quad (4)$$

$$= 0, \text{where} \quad i \in I_k' = \{1, 2, ...c\} - I_k$$

STEP 4: If $\left\| U^{(r+1)} - U^{(r)} \right\| \le \varepsilon_L$ STOP

Else go to STEP 2,

where ε_L is a termination parameter lying in (0, 1) and k represents the iteration step. This procedure terminates after J_m reaches a local minimum.

C. Rough C-Means (RCM)[3]

The rough set model was introduced by Pawlak in 1982 [24] as another model of imprecision and since then has been found to be useful in many practical situations. [25]. The concept depends upon classification of the universe of discourse, which is equivalent to the notion of equivalence relation on it. For mathematical reasons, Pawlak took equivalence relations to define the model. Here, every subset of the universe is associated with two crisp sets called its lower and upper approximation and the region in between is the region of uncertainty being called as the boundary region associated with the set. The set is said to be rough if the lower and upper approximations are not equal and definable otherwise. Suppose X is a subset of a universe U and R is an equivalence relation defined over U. Then the lower and upper approximations of X with respect to R are denoted by $\underline{R}X$ and $\overline{R}X$ being defined as follows:

$$\underline{R}X = \{x \in U \mid [x]_R \subseteq X\} \text{ and } \overline{R}X = \{x \in X \mid [x]_R \cap X \neq \phi\}$$

X is R-rough iff $\underline{R}X \neq \overline{R}X$ and R-definable otherwise.

A schematic diagram for the different notions associated with the definition of rough set is presented in **[Figure- 1]**.

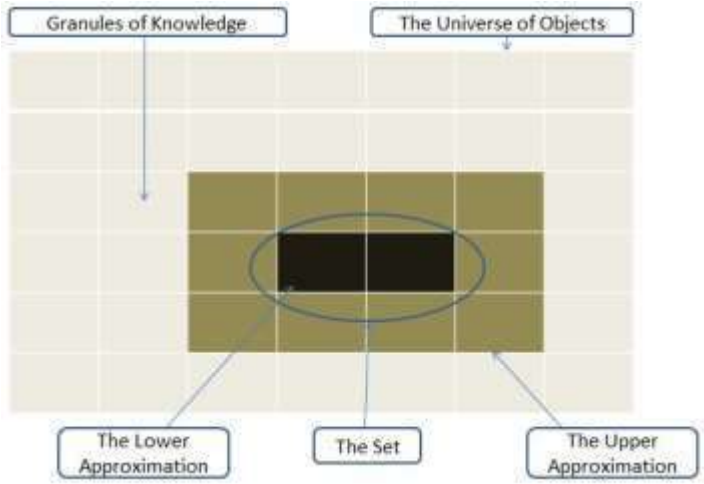


Fig: 1.Rough Set

In the rough *c*-means algorithm, the concept of *c*-means is further devised such that the cluster of data sets can be identified as an interval in rough data sets. A rough set *X* is characterized by its lower and upper approximations B*X* and B*X*, respectively, with the following properties.

1. An object x_k can be a part of at most *one* lower approximation.
2. If $x_k \in \underline{B}X$ of cluster X then simultaneously it also belongs to $\overline{B}X$.
3. If x_k is not a part of any lower approximation then it belongs to two or more upper approximations.

   i. In both the lower and upper approximations;
   ii. Only in lower approximation;
   iii. Only in upper approximation of more than one clusters.

When a cluster contains object in both lower and upper approximations then cluster prototype has to be generated using both the weighing factor. When a cluster contains objects only in its lower or in its upper approximation, the cluster prototype is computed in the classical manner without scaling down by w_{low} and w_{up}. This prohibits drifting of prototypes from their desired location. This explains the formulation of the prototype by RCM in the equation. Note that the computation of the new cluster prototype is weighted by w_{low} and w_{up} only when both its approximations are nonempty. The actual algorithm is outlined as follows:

Assign initial means v_i for the *c* clusters.

STEP 1: Assign each data object (pattern) x_i to the lower approximation $\underline{B}U_i$ or upper approximation $\overline{B}U_i$, $\overline{B}U_j$ of cluster pairs U_i and U_j by computing the difference in its distance $d_{ik} - d_{jk}$ from the cluster centroid pairs v_i and v_j.

STEP 2: Let d_{ik} be minimum and d_{jk} be the next to minimum.

If $d_{ik} - d_{jk}$ is less than some threshold, then $x_k \in \overline{B}U_i, x_k \in \overline{B}U_j$, and x_k cannot be a member of any lower approximation

Else $x_k \in \underline{B}U_i$ such that distance d_{ik} is minimum over the *c* clusters.

STEP 3: Compute new centre for each cluster U_i using (5).

STEP 4: Repeat Steps 2 to 4 until convergence, i.e., there are no more new assignments of objects.

It is observed that the performance of the algorithm is dependent on the choice of w_{low}, w_{up} and the threshold. We use $w_{up} = 1 - w_{low}$, $0.5 < w_{loe} < 1$ and 0 < threshold < 0.5.

HYBRID CLUSTERING

In this section we introduce collaborative rough-fuzzy c-means algorithm, this is done by collaboration between different partitions or subpopulations.

A. Rough Fuzzy C-Means (RFCM)[1][8]

This algorithm allows us to incorporate fuzzy membership value u_{ik} of a sample x_k to a cluster mean v_i relative to all other means v_j for all $j \neq i$, instead of distance d_{ik} from the centroids. Fuzzy membership enables efficient handling of overlapping partitions while rough set deals with uncertainty, vagueness and incompleteness in terms of upper and lower approximation [8].

$$\mathbf{v}_i = \begin{cases} w_{\text{low}} \dfrac{\sum_{\mathbf{x}_k \in \underline{B}U_i} \mathbf{x}_k}{|\underline{B}U_i|} + w_{\text{up}} \dfrac{\sum_{\mathbf{x}_k \in (\overline{B}U_i - \underline{B}U_i)} \mathbf{x}_k}{|\overline{B}U_i - \underline{B}U_i|}, & \text{if } \underline{B}U_i \neq \emptyset \wedge \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ \dfrac{\sum_{\mathbf{x}_k \in (\overline{B}U_i - \underline{B}U_i)} \mathbf{x}_k}{|\overline{B}U_i - \underline{B}U_i|}, & \text{if } \underline{B}U_i = \overline{B}U_i - \underline{B}U_i \neq \emptyset \\ \dfrac{\sum_{\mathbf{x}_k \in \underline{B}U_i} \mathbf{x}_k}{|\underline{B}U_i|}, & \text{otherwise} \end{cases} \quad (5)$$

Incorporation of membership in the RCM framework enhances the robustness of the algorithm. Previously in RCM, one never had the idea of how similar a sample was to the given cluster in the absence of any similarity index. RFCM solves this problem with the help of membership values. Following are the steps of the algorithm.

ISSN: 0976-3104

STEP 1: Assign initial means v_i for the c clusters.

STEP 2: Calculate u_{ik} for c clusters and N data objects.

STEP 3: Assign each data object x_k to the lower approximation $\underline{BU}_i$ or upper approximations $\overline{BU}_i$, $\overline{BU}_j$ of cluster pairs U_i and U_j by computing the difference in its distance $u_{ik} - u_{jk}$ from the cluster centroid pairs v_i and v_j.

STEP 4: Let u_{ik} be maximum and u_{jk} be the next to maximum.

If $u_{ik} - u_{jk}$ is less than some threshold, then $x_k \in \overline{BU}_i$ and $x_k \in \overline{BU}_j$ and x_k cannot be a member of any lower approximation, else $x_k \in \underline{BU}_i$ such that membership u_{ik} is maximum over the *c* clusters.

STEP 5: Compute new centre for each cluster U_i using (6).

STEP 6: Repeat Steps 2 to 5 until convergence, i.e., there are no more new assignments of objects.

As indicated earlier we use $w_{up} = 1 - w_{low}$, $0.5 < w_{low} < 1$, m = 2 and 0< threshold < 0.5.

B. Hybrid FCM and RFCM (FRFCM-Fuzzy Rough Fuzzy C-Means)

Let us consider a dataset divided into P subpopulations or modules. Divide and conquer strategy is used to cluster this dataset. Each module or subpopulation is clustered individually to discover its structure. Collaboration is incorporated by exchanging information between the modules regarding local partitions in terms of collection of prototype computed within the modules. This strategy enables efficient handling of large datasets [9]. Hence this algorithm has strong communication levels resulting in presentation of information in small granules of prototypes.

Number of samples in the boundary region of clusters depend on the threshold value, higher the threshold value greater the number. Hence stronger collaboration between different modules is achieved resulting in the movement of clusters towards each other. This implies that the cluster modules are moving independently towards each other due to overlapping regions of corresponding clusters. Since the modules correspond to partitions from same large dataset it stabilizes the data towards efficient determination of globally existent structure.

There exists two phases in the algorithm.

Generation of FCM or RFCM clusters within the modules, without collaboration. Here we employ 0.5<w_{low}<1, thereby giving importance to samples lying within the lower approximation of clusters while

1. computing their prototype locally.

$$\mathbf{v}_i = \begin{cases} w_{\text{low}} \dfrac{\sum_{x_k \in \underline{BU}_i} u_{ik}^m \mathbf{x}_k}{\sum_{x_k \in \underline{BU}_i} u_{ik}^m} + w_{\text{up}} \dfrac{\sum_{x_k \in (\overline{BU}_i - \underline{BU}_i)} u_{ik}^m \mathbf{x}_k}{\sum_{x_k \in (\overline{BU}_i - \underline{BU}_i)} u_{ik}^m}, & \text{if } \underline{BU}_i \neq \emptyset \wedge \overline{BU}_i - \underline{BU}_i \neq \emptyset \\ \dfrac{\sum_{x_k \in (\overline{BU}_i - \underline{BU}_i)} u_{ik}^m \mathbf{x}_k}{\sum_{x_k \in (\overline{BU}_i - \underline{BU}_i)} u_{ik}^m}, & \text{if } \underline{BU}_i = \emptyset \wedge \overline{BU}_i - \underline{BU}_i \neq \emptyset \\ \dfrac{\sum_{x_k \in \underline{BU}_i} u_{ik}^m \mathbf{x}_k}{\sum_{x_k \in \underline{BU}_i} u_{ik}^m}, & \text{otherwise} \end{cases} \quad (6)$$

2. Collaborative FCM or RFCM between the clusters, computed locally for each module of the large dataset. Now we use 0<w_{low}<0.5 with a lower value providing higher precedence to samples lying in the boundary region of the overlapping clusters.

a) In collaborative FCM, a cluster U_i may be calculated with an overlapping cluster U_j.

b) Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j. This procedure converges to a local minimum or saddle point of J_m.

c) In case of collaborative RFCM U_i can be considered for merging with U_j

if $\sum_{x_k \in \underline{BU}_i} u_{ik} \leq \sum_{x_k \in \left(\overline{BU}_i - \underline{BU}_i\right)} u_{ik}$

and v_i is closest to v_j in the feature space being the maximum among all overlapping clusters.

The entire algorithm is summarized below.

STEP 1: Split the large dataset into P modules.

STEP 2: For each module p=1, , , , , ,P do

STEP 3: For each module p do collaboration.

a) Assign each pattern x_k to lower or upper approximation of the C (= c*P) collaborative FCM or RFCM clusters,

with $0<W_{low}< 0.5$.

b) Merge overlapping clusters pairs while

i. Compute new prototype for merged clusters U_i and U_j as the mean of v_i and v_j.

ii. Reduce number of clusters C by one.

iii. Reassign each pattern x_k to lower or upper approximation of the C collaborative RCM or RFCM clusters.

COMPARATIVE ANALYSIS

The density of the collaboration between clusters or imprecise set of data values can be statistically evaluated in terms of separate indices and PSNR and RMSE values. These indices are Davies-Bouldin Index, Partition Co-efficient Index, Classification Entropy and Silhouette Statistical Index. In this section, we bring upon these measures to position our collaborative clustering framework. It is also needed to be mentioned that the number of clusters in a module needs to remain fixed but is generalized to be unique both before and after collaboration. Memberships of data objects are computed both before and after collaboration, with respect to the cluster prototypes. Since overlapping clusters can very well be merged by means of collaboration, the final cardinality of indices within different modules are often indication enough towards the degree of the clustering. This lead us to use the maximum membership value max $u_{ik(p)}$ of a data point x_k of module p, to one of the clusters U_i, during our computation of separate indices. We consider the four indices *Davies-Bouldin Index, Partition Co-efficient Index, Classification Entropy and Silhouette Statistical Index*

The DB [25] is a function of the ratio of the sum of within cluster distance to between-cluster separation. Another index used for measuring clustering efficiency is the D index [26]. The method with lower value of the index bears the greater potential of clustering. Let $\{x_1,...x_{|ck|}\}$ be a set of patterns lying in a cluster U_k. Then, the Davies Bouldin index is defined as

$$DB = \frac{1}{c}\sum_{j=1}^{c}\max\left\{\frac{d_w(U_i)+d_w(U_j)}{d(U_i,U_j)}\right\} \quad (7$$

for 1< j, i < c and within-cluster distance $d_w(U_i)$ is minimized while the between-cluster separation $d(U_i,U_j)$ gets maximized.

Silhouette Index, *S*, computes for each point a width depending on its membership in any cluster where c_i is the average distance between points *i* and all other points in its own cluster. Here, b_i is the minimum of the average dissimilarities between *i* and points in other clusters. Negative index showcases the stability in the collaboration; lower the magnitude greater is the amount of grouping [27].

$$S_k = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i - a_i}{\max(a_i,b_i)}$$

$$S = \frac{1}{c}\sum_{k=1}^{c}S_k$$

The Partition co-efficient is the measure of overlap between the clusters. This index value directly corresponds to the degree of partition achieved [28]. If u_{ij} is taken as membership of data point j in cluster i and c as the number of clusters, then the index is defined as

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\left(u_{ij}\right)^2$$

A similar index named Classification Entropy is designed such that the stability of clustering methodology is calculated. The imprecise data is handled and its fuzziness is considered [28]. The negative entropy indicates a stable arrangement among data sets. With all probabilistic cluster partitions c obeying the rule 0 <1-PC(c) <CE(c), the classification entropy is defined as

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}\log\left(u_{ij}\right)$$

Cluster validity index values **[Table-1]** are calculated on the data set given by [(1,3), (1.5,3.2), (1.3,2.8), (3,1)]

iv.

TABLE 1: Cluster validity index values

Index	FCM	RCM	RFCM	FRFCM
Davies-Bouldin	1.956	2.608	1.872	1.870
Silhouette	-0.370	-0.790	-0.352	-0.350
Partition Coefficient Index	0.984	4.570	4.486	4.605
Classification Entropy	0.017	-0.693	-0.685	-0.681

In **[Figure- 2]**, the different index values are plotted and the aforementioned indices are shown in black, red, green and blue colour respectively. The x-axis values correspond to the algorithms specified.

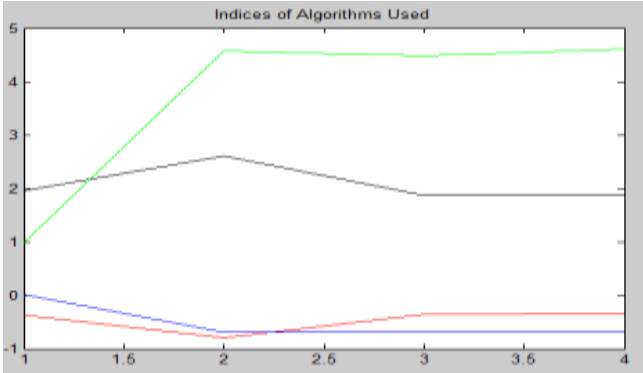


Fig: 2. Comparative Analysis of Clustering Algorithms using different indices

RMSE and PSNR Values

Peak Signal-to-Noise Ratio can be characterized as PSNR, which is the relation with the majority likely power of a signal and the power of corrupting distortions that influence the fidelity of its demonstration.

The PSNR value can be computed through mean squared error (MSE). For an example distortion-free m x n monochrome image 'I' with its noisy approximation 'K'. The RMSE of a model prediction is defined as the square root of the mean squared error: Hence, the PSNR is defined as where MAX_I is the most possible 0's and 1's values of an image. And it will be replaced with 255, as and when the 0's and 1's are given using 8 bits per model. And MSE will become '0'; when the distortion is null indicating that the two input images are same.

Here, MAX_I is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255. In the absence of noise, the two images I and K are identical, and thus the MSE is zero. In this case the PSNR is undefined. And the performance of PSNR and RMSE values as per **Table- 2** & **Figure- 3** of scanned cerebral image [29] [30] resulted efficient result of using hrbrid clustering algorithm given in **Figure-8** compared to other clustering algorithms as per

Original brain image- **Figure-4,** FCM-**Figure- 5** , RCM-**Figure-6,**RFCM-**Figure- 7** and FRFCM-**Figure-8.**

TABLE: 2. PSNR AND RMSE VALUES

Metric/Cluster Techniques	FCM	RCM	RFCM	FRFCM
PSNR	7.8409	9.856	9.9526	8.5964
RMSE	7.0379	6.8821	6.8127	9.421

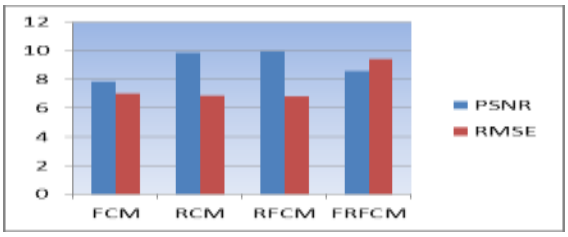


Fig: 3. Performance of PSNR and RMSE Values

RESULTS AND DISCUSSION

The final output of hybrid algorithm is devised to represent membership values in FRFCM framework in scanned cerebral image. The following set of images depicts the level of clustering and finally groups the points with same attributes to bring forward high definition precision with efficient results using metrics of Davies-Bouldin Index, Partition Co-efficient Index, Classification Entropy and Silhouette Statistical Index and PSNR & RMSE values(8.5964 & 9.4210) for FRFCM clustering algorithm in comparison with FCM, RCM, RFCM clustering algorithms.

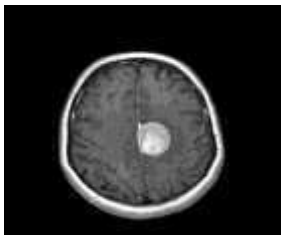


Fig: 4. Original Brain Scan Image

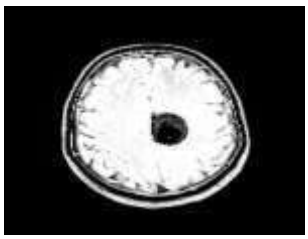


Fig: 5. Clustered Image applying FCM

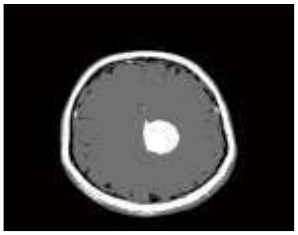


Fig 6: Clustered Image after Applying RCM

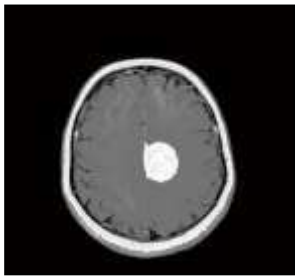


Fig: 7. Clustered Image After applying RFCM

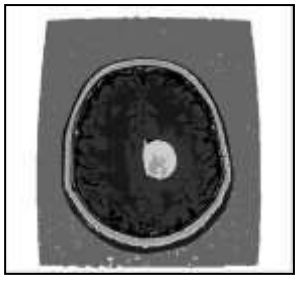


Fig: 8. Clustered Image after applying FRFCM

CONCLUSION

In this paper we have introduced a new clustering algorithm called the Fuzzy Rough Fuzzy C-Means (FRFCM) which is a first of its kind where a hybrid model is formed by combining three models. We know that the hybrid models are more efficient than their individual components. Here we could establish experimentally that as we increase the level of hybridization the efficiency further increases. For this purpose we have taken several measuring indices and also we have taken established images and could show that the segmentation shows improved results than the individual models in the form of individual or another two level hybrid model. This mow opens up a direction of research in which we can increase the hybridization to further higher levels. It would be interesting to find whether this trend will continue. Of course the complexity of these higher levels hybrid algorithms increase.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers and the editor-in-chief of the journal for their valuable guidance which has improved the quality and presentation of the paper.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

REFERENCES

[1] S Mitra, H. Banka and W. Pedrycz, [2006] Rough–Fuzzy Collaborative Clustering, IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics, 36(4):68–73.
[2] BK Tripathy, P Swarnalatha. [2013] A Comparative Study of RIFCM with Other Related Algorithms from Their Suitability in Analysis of Satellite. Images using Other Supporting Techniques, Kybernetes, 43(1).
[3] R Bhargav, B. K. Tripathy, A Tripathy, E. Verma, Raj Kumar, P Swarnalatha: Rough Intuitionistic Fuzzy C-Means Algorithm and a Comparative Analysis, COMPUTE‟13, Aug 22-24, Vellore, Tamil Nadu, India Copyright 2013 ACM 978-1-4503-2545-5/13/08.
[4] P Swarnalatha, B K Tripathy [2016]: ANALYSIS OF DEPTH USING CLUSTERING TECHNIQUES WITH BIT PLANE FILTER FOR SATELLITE IMAGERY, IIOABJ, 7(5): 108-125.
[5] B. K. Tripathy, P Swarnalatha [2014] A Comparative Study of RIFCM with Other Related Algorithms from Their Suitability in Analysis of Satellite Images using Other Supporting Techniques, Kybernetes, 43(1):53-81.
[6] BK Tripathy, P Swarnalatha [2015] A Comparative analysis of Depth computation of Leukemia Images using a refined Bit Plane and Uncertainty based clustering Techniques, Cybernetics and Information Technologies,15(1):126- 146.
[7] B. K. Tripathy, P Swarnalatha. [2013] A Novel Fuzzy C-Means Approach with Bit Plane Algorithm for Classification of Medical Images, IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology [ICECCN 2013], pp.360 – 365
[8] Deepthi PH., B.K.Tripathy [2016]: Kernel Based Spatial Fuzzy C-Means for Image Segmentation, the IIOAB journal, 7(5): 150-156
[9] B.K.Tripathy, A Ghosh, G.K. Panda. [2012] Adaptive K-means Clustering to Handle Heterogeneous Data using Basic Rough Set Theory, In the Proceedings of Second Intl. Conf. [CCIT-2012], Bangalore, Advances in Computer Science and Information Technology Network Communications, LNICST series, Springer, 84:193 – 201.
[10] B.K. Tripathy, A Ghosh, G K. Panda [2012]: Kernel Based K-Means Clustering Using Rough Set, Proceedings of 2012 International Conference on Computer Communication and Informatics [ICCCI -2012], Jan. 10 – 12, Coimbatore, INDIA, pp.1–5
[11] B. K.. Tripathy, R Bhargav [2013]: Kernel Based Rough-Fuzzy C-Means, PReMI, ISI Calcutta, December, LNCS 8251, pp.148–157.
[12] B. K. Tripathy, A Tripathy, K.Govindarajulu and R. Bhargav [2014]: On kernel Based rough Intuitionistic Fuzzy C-means algorithm and a comparative analysis, ICACNI 2014, Advanced computing networking and informatics, vol.1, Smart innovation systems and technologies, 27: 349–359.
[13] B. K. Tripathy, A Tripathy, K.Govindarajulu and R. Bhargav [2014]: On kernel Based rough Intuitionistic Fuzzy C-means algorithm and a comparative analysis, ICACNI 2014, Advanced computing networking and informatics, vol.1, Smart innovation systems and technologies, 27: 349–359.
[14] B. K. Tripathy, A Goyal, PA Sourav [2016] A comparative analysis of rough intuitionistic fuzzy k-mode algorithm for clustering categorical data, Research Journal of Pharmaceutical, Biological and Chemical Sciences Volume 7, Issue 5, pp. 2787-2802
[15] B. K. Tripathy, A Tripathy, K.Govindarajulu [2014] Possibilistic rough fuzzy C-means algorithm in data clustering and image segmentation, proceedings of the IEEE ICCIC2014, pp.981-986.
[16] B. K. Tripathy, A Tripathy, K.Govindarajulu [2014] Possibilistic rough fuzzy C-means algorithm in data clustering and image segmentation, proceedings of the IEEE ICCIC2014, pp.981-986.
[17] B. K. Tripathy, A Tripathy, K.Govindarajulu [2014] Possibilistic rough fuzzy C-means algorithm in data clustering and image segmentation, proceedings of the IEEE ICCIC2014, pp.981-986.

[18] B. K. Tripathy, A Tripathy, K Govindarajulu. [2015]: On PRIFCM Algorithm for Data Clustering, Image Segmentation and Comparative Analysis, Souvenir of the 2015 IEEE International Advance Computing Conference, IACC 2015, Article number-7154725, pp.333–336.
[19] P Lingras, G Peters.[2012] Applying rough sets to clustering. Rough Sets: Selected Methods and Applications in Management and Engineering, Advanced Information and Knowledge Processing, DOI 10.1007/978-1-4471-2760-4_2, © Springer-Verlag London Limited.
[20] B. K. Tripathy, S Sobti, V Shah [2014] A refined rough fuzzy clustering algorithm, proceedings of the IEEE ICCIC2014, pp.776–779
[21] B. K. Tripathy, P Prabhavathy [2015]: An Intelligent Clustering Approach for Sequential Data, presented at ICIMR 2015, Bhubaneswar.
[22] B. K. Tripathy, A Tripathy, K.Govindarajulu [2014] Possibilistic rough fuzzy C-means algorithm in data clustering and image segmentation, proceedings of the IEEE ICCIC2014, pp.981-986.
[23] P Lingras, G Peters.[2012] Applying rough sets to clustering. Rough Sets: Selected Methods and Applications in Management and Engineering, Advanced Information and Knowledge Processing, DOI 10.1007/978-1-4471-2760-4_2, © Springer-Verlag London Limited.
[24] Z Pawlak [1982] Rough Sets, International Journal of Information and Computer Sciences, 11:341–356.
[25] S Petrovic.[2003] A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters. NIS Lab, Department of Computer Science and Media Technology, pp.1-12.
[26] JC Dunn [1973] A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, pp.32-57.
[27] P Maji, SK Pal [2007] Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices. Center for Soft Computing Research Indian Statistical Institute, Fundamenta Informaticae 80:475–496.
[28] W Pedrycz [2002] Collaborative fuzzy clustering, Pattern Recognition. Letters, vol. 23, no. 14,1675–1686, Dec., Article in a conference 7670.
[29] LC Robert, JV Dave, JC Bezdek. [1986] Efficient Implementation of the Fuzzy C-Means Algorithm", Mathematics Department, Utah State University, IEEE Transactions of Pattern Analysis and Machine Intelligence, PAMI-8(2).
[30] Swarnalatha P, B. K. Tripathy BK, Nithin, P L and D Ghosh [2014]:: Cluster Analysis Using Hybrid Soft Computing Techniques, CNC-2014International Conference of Network and Power Engineering ,Proceedings of Fifth CNC-2014,pp. 516–524.

ABOUT AUTHORS

Swarnalatha Purushotham is an Associate Professor, in the School of Computer Science and Engineering, VIT University, at Vellore, India. She pursued her Ph.D degree in Image Processing and Intelligent Systems. She has published more than 60 papers in International Journals/International Conference Proceedings/National Conferences. She is having 15+ years of teaching experiences. She is a member of IACSIT, CSI, ACM, IACSIT, IEEE (WIE), and ACEEE. She is an Editorial board member/reviewer of International/ National Journals and Conferences. Her current research interest includes Image Processing, Remote Sensing, Artificial Intelligence and Software Engineering.

B.K.Tripathy is a senior professor in the school of computer sciences and engineering, VIT University, Vellore, since 2007. He has produced 26 PhDs, 13 M. Phils and 03 M.S students so far. He has published around 400 technical papers in different international journals, conference proceedings. Dr. Tripathy has guest edited some journals of repute. He has edited five research volumes for the IGI publications and Taylor and Francis publications. Dr. Tripathy has written two books on Soft Computing and Computer Graphics. He is in the editorial board or review panel of over 70 international journals including Springer, Science Direct, IEEE and World Scientific publications. He is a life member/ senior member/member of 21 international forums including ACM, IEEE, ACEEE, IRSS, ISRR and CSI. His current interest includes Fuzzy Sets and Systems, Rough sets and Knowledge Engineering, Multiset Theory, List Theory, Data clustering and Database Anonymization, Content Based Learning, Neighbourhood Systems, Soft Set Analysis, Image Processing, Cloud Computing, Social Internet of Things, Big Data Analytics, Multi Criteria Decision Making and Social Network Anonymization.