SPECIAL ISSUE

THE IIOAB JOURNAL

VOLUME 11 : NO 2 : JULY 2020 : ISSN 0976-3104

Institute of Integrative Omics and
Applied Biotechnology Journal

Dear Esteemed Readers, Authors, and Colleagues,

I hope this letter finds you in good health and high spirits. It is my distinct pleasure to address you as the Editor-in-Chief of Integrative Omics and Applied Biotechnology (IIOAB) Journal, a multidisciplinary scientific journal that has always placed a profound emphasis on nurturing the involvement of young scientists and championing the significance of an interdisciplinary approach.

At Integrative Omics and Applied Biotechnology (IIOAB) Journal, we firmly believe in the transformative power of science and innovation, and we recognize that it is the vigor and enthusiasm of young minds that often drive the most groundbreaking discoveries. We actively encourage students, early-career researchers, and scientists to submit their work and engage in meaningful discourse within the pages of our journal. We take pride in providing a platform for these emerging researchers to share their novel ideas and findings with the broader scientific community.

In today's rapidly evolving scientific landscape, it is increasingly evident that the challenges we face require a collaborative and interdisciplinary approach. The most complex problems demand a diverse set of perspectives and expertise. Integrative Omics and Applied Biotechnology (IIOAB) Journal has consistently promoted and celebrated this multidisciplinary ethos. We believe that by crossing traditional disciplinary boundaries, we can unlock new avenues for discovery, innovation, and progress. This philosophy has been at the heart of our journal's mission, and we remain dedicated to publishing research that exemplifies the power of interdisciplinary collaboration.

Our journal continues to serve as a hub for knowledge exchange, providing a platform for researchers from various fields to come together and share their insights, experiences, and research outcomes. The collaborative spirit within our community is truly inspiring, and I am immensely proud of the role that IIOAB journal plays in fostering such partnerships.

As we move forward, I encourage each and every one of you to continue supporting our mission. Whether you are a seasoned researcher, a young scientist embarking on your career, or a reader with a thirst for knowledge, your involvement in our journal is invaluable. By working together and embracing interdisciplinary perspectives, we can address the most pressing challenges facing humanity, from climate change and public health to technological advancements and social issues.

I would like to extend my gratitude to our authors, reviewers, editorial board members, and readers for their unwavering support. Your dedication is what makes IIOAB Journal the thriving scientific community it is today. Together, we will continue to explore the frontiers of knowledge and pioneer new approaches to solving the world's most complex problems.

Thank you for being a part of our journey, and for your commitment to advancing science through the pages of IIOAB Journal.

Yours sincerely,

Vasco Azevedo

Vasco Azevedo, Editor-in-Chief
Integrative Omics and Applied Biotechnology
(IIOAB) Journal

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

MULTI-CRITERIA SELECTION OF TESTING METHODS FOR SEPARATE SOFTWARE MODULES

Vladimir V. Lomakin*, Natalya P. Putivtseva, Tatyana V. Zaitseva, Olga P. Pusnaya

Institute of Engineering Technology and Natural Science, Belgorod State National Research University, Belgorod, 85 Pobedy st., 308015, Belgorod, RUSSIA

ABSTRACT

The paper presents the results of the testing methods selection for individual software modules using the ELECTRE method and the simplified analytic hierarchy process. The selection criteria were formulated, and the main approaches to testing separate software modules were considered as alternatives. It is planned to use the results obtained within the framework of the project to create high-tech production "Development of a methodology and tools for creating applications, supporting the life cycle of information technology and decision-making software for the effective implementation of administrative and management processes within the established authorities."

KEY WORDS
Software testing, multi-criteria selection, ELECTRE method, simplified analytic hierarchy process

Received: 3 Oct 2019
Accepted: 22 Nov 2019
Published: 4 Jan 2020

***Corresponding Author**
Email:
lomakin@bsu.edu.ru

INTRODUCTION

Software testing is a checking the correspondence between the actual and expected behaviour of the program, carried out using a finite set of tests selected in a certain way [1, 2].

You should test some small modules that form program before to start testing this program as a whole. The reasons for this approach are: 1) it becomes possible to control the combinatorics for testing; 2) the process of detecting the place of error and correcting the program text is facilitated; 3) concurrency is allowed, which makes testing several modules at the same time possible.

The testing of modules is mainly focused on the “white box” principle which is explained by the fact that the subsequent stages of testing are focused on detecting errors of various types not necessarily related to the program logic, but arising, for example, due to a program mismatching with the user's requirements. When testing separate modules of a software product, the best option for a number of criteria is to select a modified sandwich method.

There are six main approaches that can be used to merge modules into larger units [3].

1) Ascending testing. The program is assembled and tested from the bottom to up. Only the modules of the lowest level are tested in isolation, i.e. autonomously. Then, those modules are tested that directly call the already tested ones, but no longer autonomously, and together with the already tested lower-level modules. The process is repeated until a top point would be reached. This completes both unit testing and program pairing testing.

2) Top-down testing. The program is assembled and tested from top to bottom. Only the head module is tested in isolation. Then, the modules directly called by the head module are connected to it, one by one, and the resulting combination is tested. The process is repeated until all modules are assembled and tested. In the case when the module under test calls a module from a lower level, which does not exist at the moment, “faceplace” modules are programmed to simulate the functions of the missing modules.

3) Modified top-down testing. Using top-down testing, it is often impossible to test errors or exceptions, as well as defensive checks. The modified top-down testing method requires that each module undergo stand-alone testing before connecting to the program.

4) The "big jump" method. In accordance with this method, each module is tested autonomously. At the end of testing the modules, they are integrated into the system all at once. When using this approach, it is necessary to take into account that the modules are not integrated until the very last moment, serious errors in the interfaces may remain undetected for a long time, and “faceplates” and drivers are necessary for each module.

5) The sandwich method [4]. In this method, they simultaneously start the ascending and top-down testing, assembling the program both from below and from above and meeting somewhere in the middle. The method saves the beginning of system integration at an early stage. Since the top-point of the program comes into operation early, early on we get a working program framework. Since the lower levels of the program are created by the ascending method, the problems of impossibility are removed and some conditions are tested in the depths of the program.

IIOAB JOURNAL
ISSN: 0976-3104

6) The modified sandwich method. When testing by the sandwich method, a problem arises that it is impossible to thoroughly test individual modules. In the modified sandwich method, the lower levels are tested strictly "from the bottom up" and the modules of the upper levels are first tested in isolation, and then assembled in a downward manner.

MATERIALS AND METHODS

We will solve the problem of choosing one of the software testing approaches using two methods: the ELECTRE method [5] and the simplified analytic hierarchy process [6].

Consider the solution of the problem by the ELECTRE method, which consists in the following. An integer w, characterizing the importance of the criterion, is assigned to each of the N criteria. It is assumed that the criteria weights are given by the decision maker. The index of agreement with the hypothesis of superiority of alternative x over alternative y is determined as follows:

$$C(x, y) = \sum_{i \in I^+ \cup I^-} \omega_i / \sum_{i=1}^{N} \omega_i \ ,$$

Where ω i is the weight of the i-th alternative.

The disagreement index d(x,y) with the hypothesis that alternative x is superior to alternative y is defined as follows:

$$d(x, y) = \max_{i \in I^-} \frac{F_i(x) - F_i(y)}{L_i}$$

Where Fi(x), Fi(y) are the estimates of the alternatives x and y according to the i-th criterion,
Li is the length of the i-th criterion scale.

The ELECTRE method sets a binary relationship of superiority between levels of agreement and disagreement. If C(x, y) ≥α & d(x, y) ≤ β, where α and β are given levels of agreement and disagreement, then alternative x is declared to be superior to alternative y. A kernel of non-dominated elements is distinguished for a set of alternatives at the given agreement and disagreement levels. With decreasing α and increasing β, a smaller kernel (contained in the previous one) is identified in the given kernel, etc. An analyst offers to a decision maker a series of possible solutions to the problem in the form of various kernels. In the end, we can get one better alternative.

When solving the problem, the developed ELECTRE program was used. When the program starts, a window appears in which the user enters the number of alternatives and criteria, as well as their values. After entering all the parameters, the user clicks the "Remember" button and goes to the next tab, on which he/she needs to enter estimates of alternatives according to the criteria. After entering, the user must click the "Execute" button, and the program will calculate the values of the parameters L and ω. After that, the user switches to another tab. He/she calculates the agreement / disagreement indices there. Indexes are calculated by the program automatically if to click the "Calculate" button. After that, the tab "Matrix of values of a kernel" is switched on. Here on the left there are two fields where the user enters the coefficients α and β for calculating the kernel of solutions. Below the input fields there are two tables "Alpha" and "Beta", where the user is shown sorted lists of calculated agreement / disagreement indexes. After entering the coefficients, the user must click the "Find a kernel" button and the program will calculate the kernel of solutions using the binary superiority formula between the levels of contest (agreement) and disagreement. The calculations will be displayed on the left in the table, and the kernel of solutions, in which the best solutions will be displayed at the bottom for the given coefficients α and β. The user can change the values of the coefficients in order to achieve one best solution.

IMPLEMENTATION

After analyzing the software testing methods, we can identify the main criteria that are considered when choosing a method. When selecting the criteria, the features of the IC ISKU (developed as part of the joint activities of an enterprise and the educational institution of the Platform) were taken into account [7, 8]. The following criteria were identified: Cr 1 - Assembly; Cr 2 - Time until the appearance of a working version of the program; Cr 3 - Need for drivers; Cr 4 - Need for "faceplates"; Cr 5 - Concurrency at the beginning of work; Cr 6 - Ability to test individual paths; Cr 7 - Ability to plan and control the sequence. For convenience, their comparative analysis is presented in [Table 1].

Table 1: Comparative analysis of software testing methods for selected criteria

Criterion	Ascending Method (A1)	Top-Down Method (A2)	Modified Top-Down Method (A3)	Big Jump Method (A4)	Sandwich Method (A5)	Modified Sandwich Method (A6)
Cr 1	Early	Early	Early	Late	Early	Early
Cr 2	Late	Early	Early	Late	Early	Early
Cr 3	Yes	No	Yes	Yes	Partially	Yes
Cr 4	No	Yes	Yes	Yes	Partially	Partially
Cr 5	Medium level	Weak	Medium level	High	Medium level	High
Cr 6	Easily	Difficult	Easily	Difficult	Medium	Easily
Cr 7	Easily	Difficult	Difficult	Easily	Difficult	Difficult

When solving the problem by the ELECTRE method, initial values were entered (7 criteria and 6 alternatives), the input results are presented in [Fig. 2].

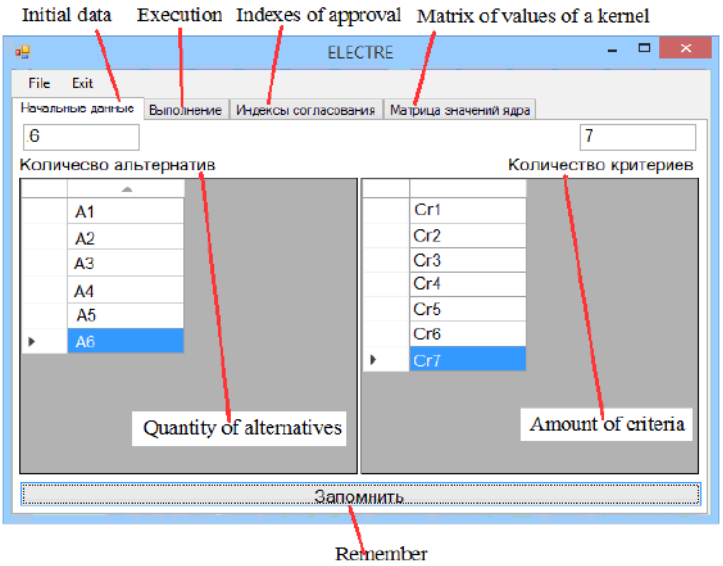

Fig. 2: Filling in the initial values.

Next, estimates of alternatives by criteria are entered [Fig. 3].

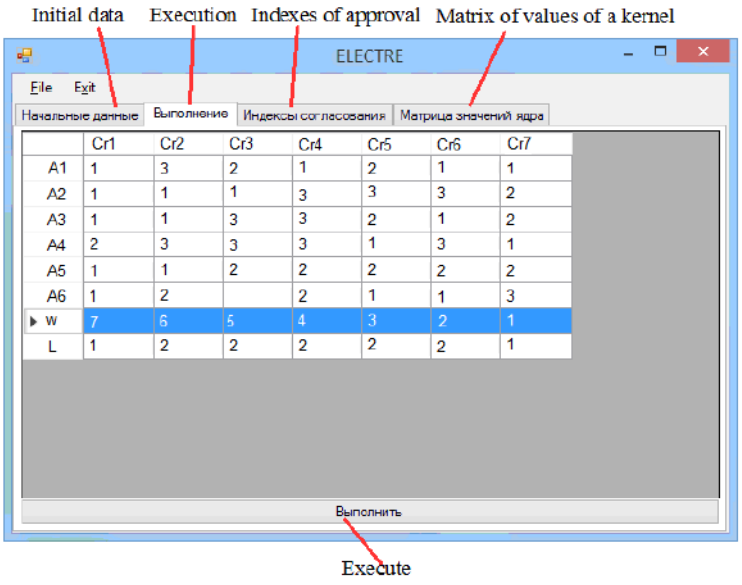

Fig. 3: Filling in the values for comparison of alternatives by criteria.

The following figure shows the calculated indices of contest and disagreement [Fig. 4].

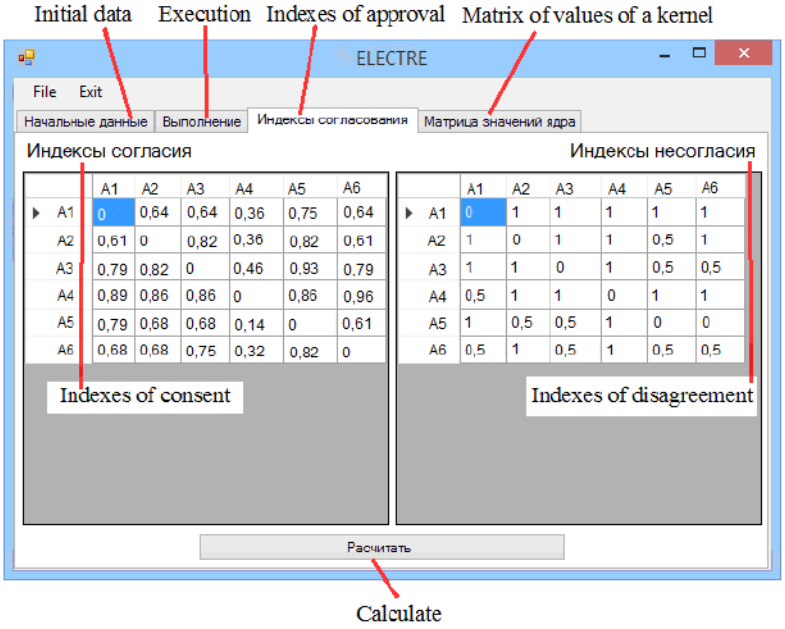

Fig. 4: Calculation of agreement and disagreement indexes.

Next, the values of the coefficients α and β are set [Fig. 5] and the alternatives included in the kernel of the solution are determined [Fig. 6].

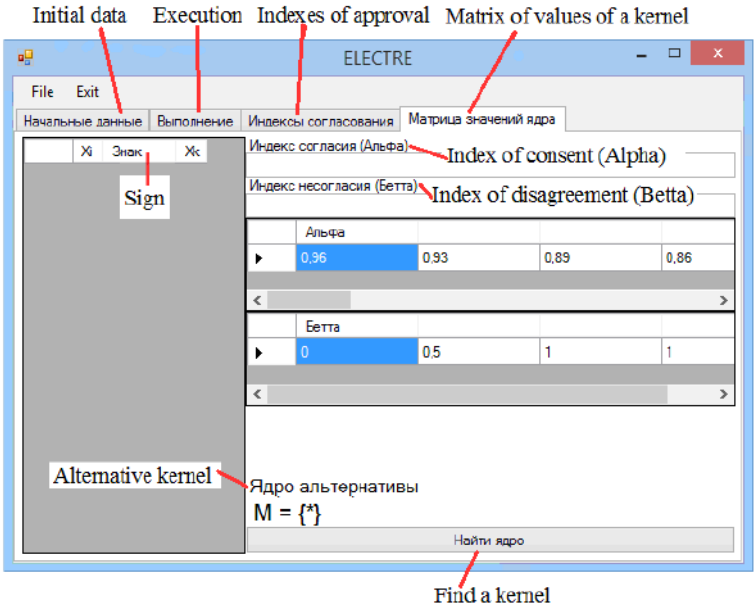

Fig. 5: Selection of coefficient values for finding the solution kernel.

Since, as a result of the experiment, two variants of the sandwich method (classical and modified) were proposed as the best alternatives, which only partially confirms the hypothesis, a second experiment was conducted using the simplified analytic hierarchy process to determine the best alternative according to the selected criteria.

The use of a simplified analytic hierarchy process was considered in [6]. When solving the problem with the simplified analytic hierarchy process, the “Simplified method” module of the program “System for calculating automation of the hierarchy analysis method” was used [9].

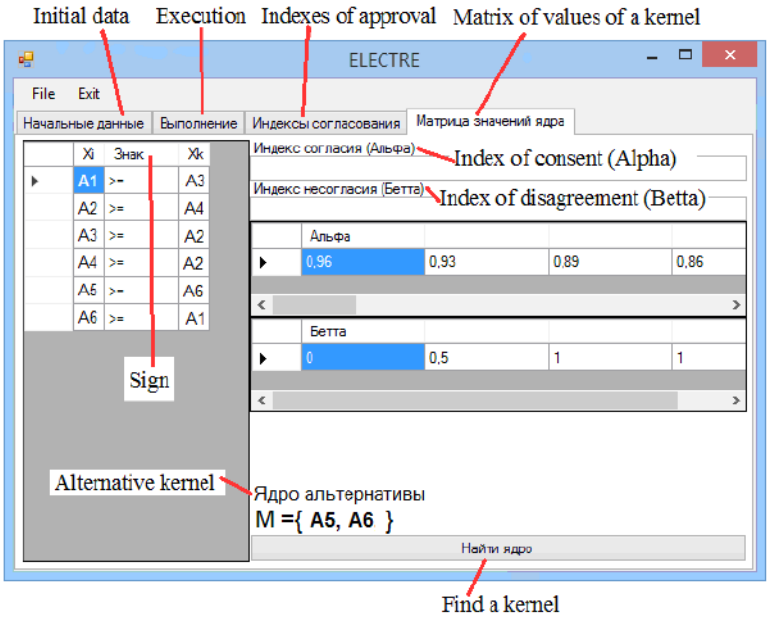

Fig. 6: Definition of alternatives included in the kernel solution.

In the first step, fill in the first row of the pairwise criteria comparison matrix, the remaining rows are filled automatically. Next, the pairwise alternative comparison matrices for each criterion are filled. Fig. 7 shows the result of filling in the pairwise alternative comparison matrix according to the first criterion, while the user filled in only the first row of the matrix, and the remaining values in the rows were calculated.

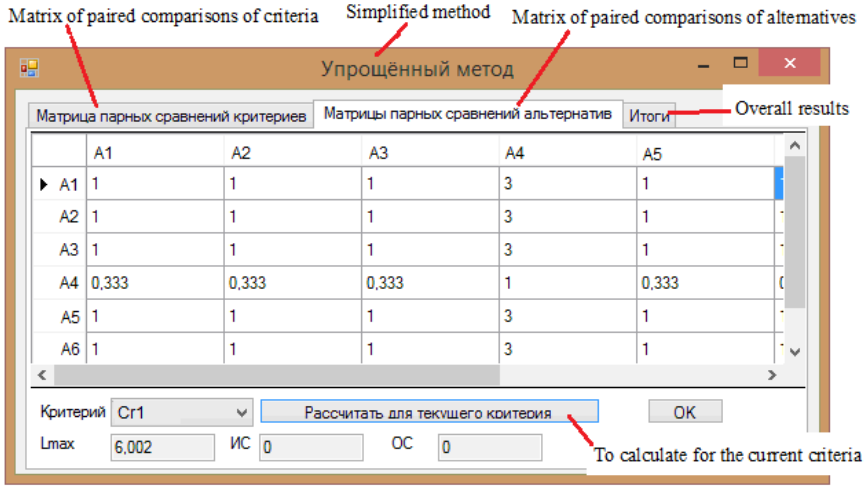

Fig. 7: Filling in the pairwise alternative comparison matrix by the criterion "ease of calculation".

After filling in all the matrices, the total weights of the alternatives were calculated, which is shown in [Fig. 8], and all summary data are presented in [Table 2].

Table 2: Summary table

Alternatives	Criteria							Weights of alternatives
	Cr1	Cr2	Cr3	Cr4	Cr5	Cr6	Cr7	
A1	0.19	0.07	0.1	0.36	0.14	0.25	0.3	0.1878
A2	0.19	0.21	0.4	0.09	0.05	0.06	0.1	0.1527
A3	0.19	0.21	0.1	0.09	0.14	0.25	0.1	0,1605
A4	0.06	0.07	0.1	0.09	0.27	0.06	0.3	0.1374
A5	0.19	0.21	0.2	0.18	0.14	0.12	0.1	0.1554
A6	0.19	0.21	0.1	0.18	0.27	0.25	0.1	0.192

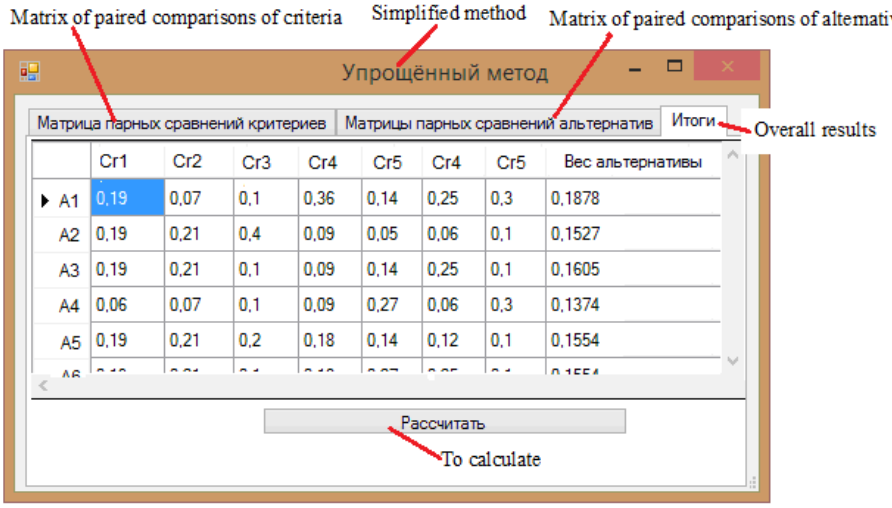

Fig. 8: Calculation of the total weights of the alternatives.

CONCLUSION

According to the ELECTRE method, the following results were obtained: the alternatives are ranked in descending order of importance: A5, A6, A1, A3, A2, A4; The kernel of solutions includes two of the most preferred alternatives: A5 - Sandwich Method and A6 - Modified Sandwich Method; since the ELECTRE method did not give an unambiguous result, and it is also impossible to establish the degree of priority 5 of the alternative over 6, the hypothesis is considered confirmed, since both sandwich methods are recognized as the best alternatives.

According to the simplified analytic hierarchy process, the following results were obtained: the alternatives are ranked in descending order of importance: A6, A1, A3, A5, A2 and A4; the best alternatives are A6 - Modified Sandwich Method and A1 - ascending testing; since the alternative A6 is the most preferable, the hypothesis is considered confirmed.

Thus, comparing the results of two experiments, we can conclude that when testing separate modules of a software product, the best option is to choose a modified sandwich method.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
Completed as part of implementing a comprehensive project on creation of high-tech production “Development of a methodology and tools for creating applications, supporting the life cycle of information technology and decision-making software for the effective implementation of administrative and management processes within the established authorities”, 2017-218-09-187; Decree of the Government of the Russian Federation dated April 9, 2010 No. 218.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] IEEE Guide to Software Engineering Body of Knowledge, SWEBOK, [2004].

[2] Software testing - basic concepts and definitions [Electronic resource]. - Access mode: http://www.protesting.ru/testing/ (accessed September 13, 2019)

[3] Soloviev SV, Tsoy RI, Grinkrug LS. [2011] Application software development technology, Academy of Natural Sciences. RU

[4] Sandwich Testing [Electronic resource]. [2019] Access mode: https://www.geeksforgeeks.org/sandwich-testing-software-testing/

[5] Larichev OI. [2000] Theory and decision-making methods, as well as the Chronicle of events in the Magic Countries: A Textbook, 229.

[6] Nogin VD. [2004] A simplified version of the simplified analytic hierarchy process on the basis of nonlinear convolution of criteria, Journal of Computational Mathematics and Mathematical Physics, 44(7):1261–1270, Comput Math Math Phys, 44(7):1194–1202.

[7] Putivtseva NP. [2017] Solving the problem of choosing Russian corporate information systems using the analytic hierarchy process. Bulletin of the Voronezh State University, Series: System Analysis and Information Technologies, 4:85 - 91.

[8] Zaitseva TV. [2018] To the question on the composition of an integrated high-level development tools complex and the corporate-level information system functioning environment. Collection of articles of the XV International Scientific and Practical Conference, Moscow: “Scientific and Publishing Centre, 81-82.

[9] Calculation automation system for the hierarchy analysis method: Certificate on registration of databases and computer programs No. 2018615850 dated 05/17/2018.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

MULTICRITERIA ANALYSIS OF MATHEMATICAL MODELS FOR SOFTWARE RELIABILITY EVALUATION

Vladimir V. Lomakin*, Natalia P. Putivtseva, Tatyana V. Zaitseva, Olga P. Pusnaya
Institute of Engineering Technology and Natural Science, Belgorod State National Research University, Belgorod, 85 Pobedy st., 308015, Belgorod, RUSSIA,

ABSTRACT

The paper presents the results of an experiment to select the most preferred model for software reliability evaluation. Since the mathematical models used to assess the reliability of software have a number of characteristics that are not the same in importance, it is advisable to apply multi criteria assessment methods to select the most suitable approach to reliability evaluation. A hierarchy analysis was chosen in the capacity of such a method, which allows us to obtain the results of comparing objects in the form of numerical weights. The paper presents the solution to the problem of choosing a dynamic model for evaluating software reliability using classical and simplified hierarchy analysis methods.

KEY WORDS
software reliability, mathematical model, multivariate analysis, analytic hierarchy process (ahp).

Received: 12 Oct 2019
Accepted: 24 Nov 2019
Published: 4 Jan 2020

***Corresponding Author**
Email:
lomakin@bsu.edu.ru

INTRODUCTION

We understand the reliability model of a Platform software as a mathematical model that reflects the dependence of a given software tool reliability on a number of parameters which values are either previously known or can be measured in the process of observing the system operation or in an experimental study of the Platform functioning [1-3].

Depending on the mathematical tools, analytical and empirical reliability models are distinguished. Two groups of the models are allocated depending on the need to consider the time factor in evaluating the reliability: dynamic and static [1-6]. In dynamic models, the behaviour of the software under test over time is considered. In static models, the appearance of failures takes into account only the dependence of the number of errors on the number of test runs or the dependence of the number of errors on the characteristics of the input data.

Hypothesis: the Mus's model is the most preferred of the dynamic models for evaluating software reliability.

Within the framework of this study, we restrict ourselves to considering only analytical dynamic models due to the fact that often there is a need to obtain data on the occurrence of failures in time, both continuously and discretely.

Consider the following dynamic reliability models [6]: Shooman's model; La Padula's model; Jelinski-Moranda's model; Chic-Walverton's model; Mus's model; Model of transition probabilities.

Shooman's model. The initial data for the Shooman's model is collected during the software testing process. At each time interval, a program is run on a full range of developed test data and a certain number of errors are recorded. Statistics on detected errors are collected. After the end of the stage, the errors found in the previous stage are corrected, test sets are adjusted, if necessary, and a new testing stage is conducted. The reliability function is calculated as the probability of no failure on a time interval from 0 to t.

La Padula's Model. According to this model, a sequence of tests is performed in t stages. Each stage ends with the introduction of changes (corrections) to the software. The reliability of the software during the i-th stage is calculated as the difference between the marginal reliability of the software at this stage and the coefficient of the growth parameter. Being based on the data obtained during testing, the model makes it possible to predict the likelihood of a program running smoothly at subsequent stages of its execution.

Initial data for the Jelinski – Moranda's model is collected in the process of testing software. At the same time, the time until the next failure is recorded. Each detected error is resolved. The software reliability is calculated as a function of the i-th error detection time distribution density, counted from the moment when the (i-1) th error was detected.

The Chic-Walverton's model considers that errors are corrected only after the expiration of the time interval at which they arose. The error rate is proportional to the number of errors remaining in the program after the (i-1) th interval and the total time already spent on testing.

Mus's model. During testing, the program execution time (test run) until the next failure is recorded. It is allowed to detect more than one error during program execution until the next failure occurs. The total

number of failures occurred throughout the entire software life cycle is related with the initial number of errors depending on the number of errors eliminated per failure.

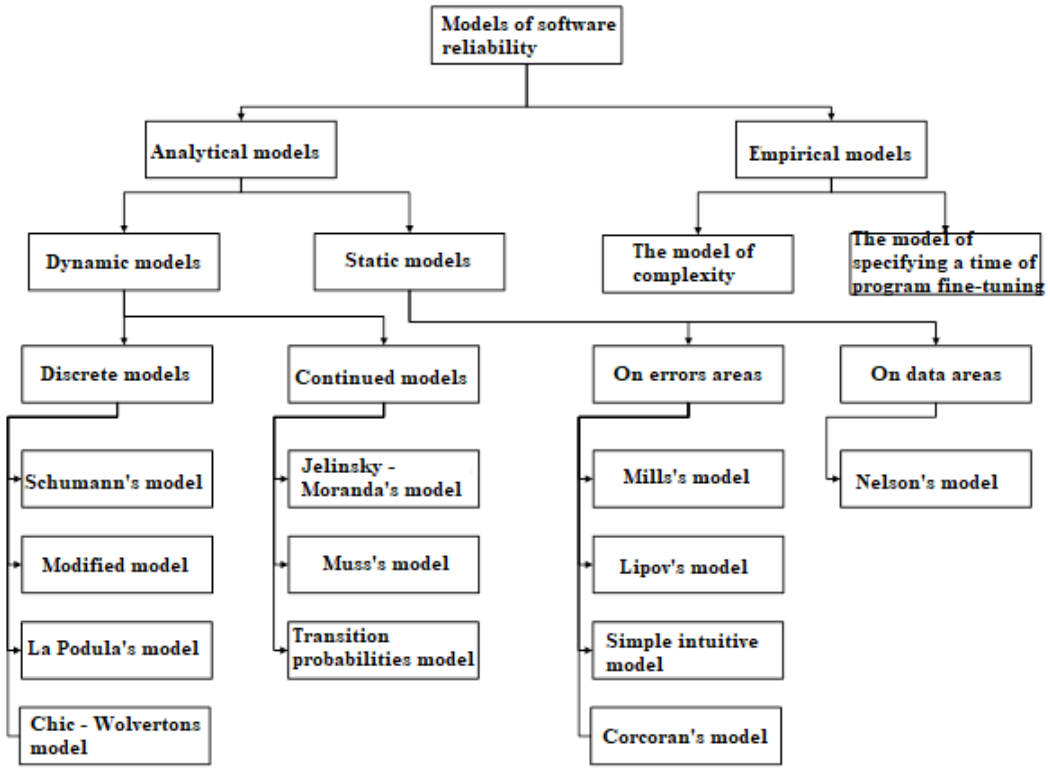

Fig. 1: Classification of reliability models.

The transition probability model is based on a Markov process that takes place in a discrete system with continuous time. It is suggested that during the testing process one error is detected. At any moment in time, the system can be in two possible states - operational or in a moment of another error correction. System readiness is defined as the sum of the probabilities of finding it in an operational state.

METHODS

When solving problems, the classical analytic hierarchy process (Saati method) and the simplified method were used.

Consider the steps of applying the classical method [7]:

1. Hierarchical decomposition of the problem “top-down”.
2. A comparative assessment of the importance of the hierarchical structure elements in relation to the overlying level based on a unified scale.
A set of pairwise comparison matrices of the elements H i and H j of any hierarchical level Ak=||aijk||hxh, aijk=si/sj, where h is the number of compared basic elements, and where the preference of elements for decision-making is defined as H i > H j if a ij k > 1; H i ≈ H j if a ij k = 1; H i < H j if a ij k <1.
3. Calculation of the priority of options by aggregating particular estimates of the hierarchical structure elements in the direction “bottom-up”.

In that case, when the simplified method is used, a simplified procedure is carried out for calculating the relative priorities of the options and criteria when constructing the pairwise comparison matrix A k = || a ij k || hxh of elements Hi and Hj of any hierarchical level [8]. The decision maker selects an element as a reference and assigns it the first number. The remaining elements are given arbitrary numbers. A pairwise comparison of each of the elements Hi, i = 2,3,. .., h with the reference H1 is performed. The element Hi is associated with a certain positive number b1i showing the subjective significance of this element for the decision-maker with respect to the element H1. As a result, all elements H i will be associated with positive numbers b11, b12,. .., b1h, where b11 = 1. The sequence of numbers b11, b12,. .., b1 h is taken as the first row of the square pairwise comparison matrix B = ||bij||hxh of rank h. The matrix B is constructed from the very beginning so that its elements bij satisfy the conditions of inverse symmetry bij x bji = 1 and compatibility bij x bjk = bik to satisfy the equalities bij = b i1 x b1j = b1j / b1i, for all numbers i, j = 2,. .., h, with the help of which all elements of the pairwise comparison matrix B are obtained in a unique way.

A comparative importance indicator for the hierarchy element H i is introduced in the form of si = bi1 / bh1 = b1h / b1i. Then the local priority of the element H i will be determined by the expression vik = vk (Hi) = si / . For an arbitrary element bij of the matrix B, the relation bij = si / sj is satisfied. The column vector s = (s1, s2, .., sh)T composed of the values of comparative importance indicators is an eigenvector of the matrix B, and its eigenvalue is equal to h. Thus, the matrix B is consistent.

IMPLEMENTATION

The following criteria were chosen for solving the problem of choosing an analytical dynamic model: simplicity of calculations; forecasting ability; iteration ability, correction of errors at the time of finding; registration of failure; the average program execution time in a period; ratio of the number of errors eliminated to failures; accumulation of error data; the need for software support. The following dynamic models were considered as alternatives: Shooman's; La Padula's; Jelinski – Moranda's; Chic – Walverton's; Mus's; and transition probabilities model.

When conducting the experiment, the program "System for automated calculating according to the analytic hierarchy process" was used [9]. When working with the tool, at the first step it was necessary to enter the parameters which include the selection criteria and alternative solutions [Fig. 2].

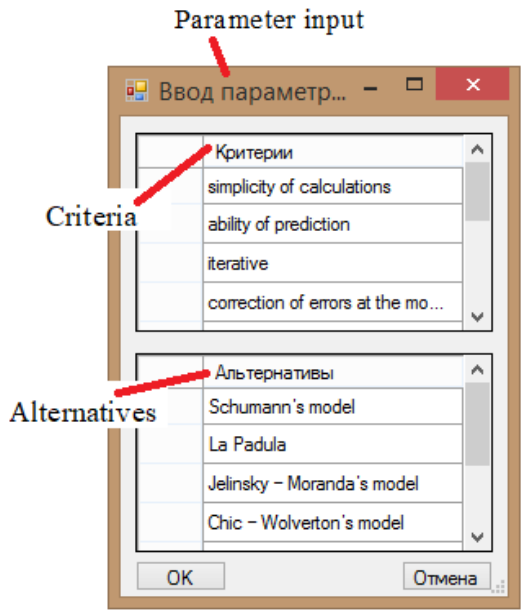

Fig. 2: Entering the decision criteria and their alternatives.

Next, we need to select the method on the basis of which the selection will take place; for the first experiment, the classic Saati's method was chosen [Fig. 3].

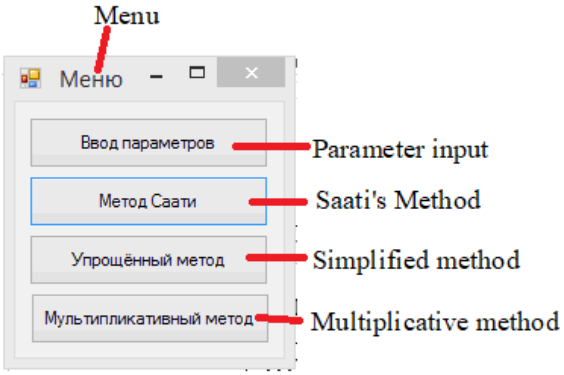

Fig. 3: Choice of a method (Saati method).

In the next step, we fill in the matrix of paired criteria comparisons [Fig. 4].

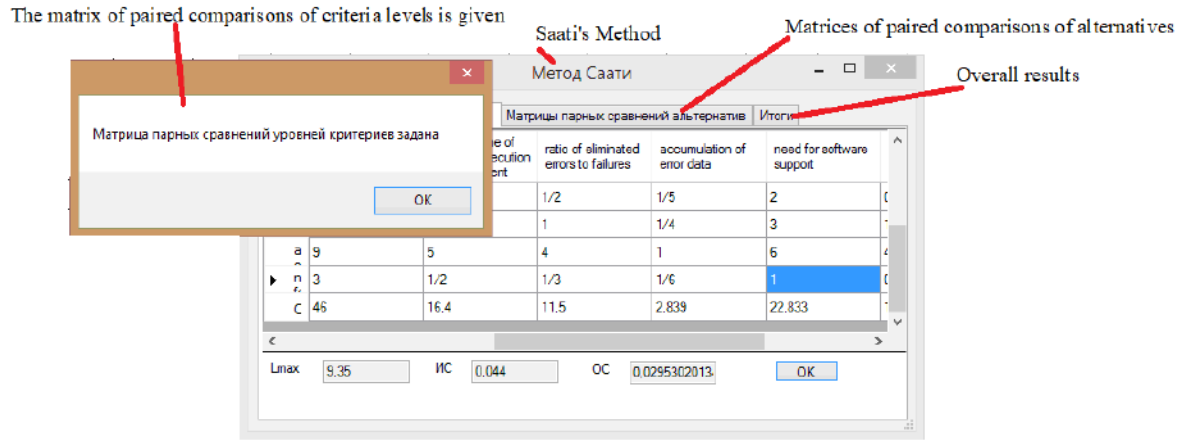

Fig. 4: Filling in the paired criteria comparisons matrix.

Next, the matrices of pairwise alternative comparisons for each criterion are filled. [Fig. 5] shows the result of filling in a pairwise alternative comparison matrix by the criterion of "simplicity of computation". The remaining matrices are filled in the same way.

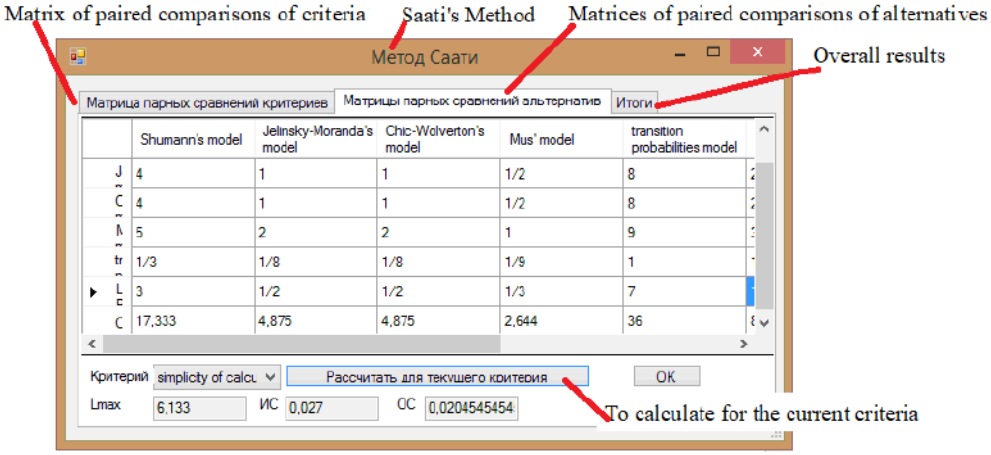

Fig. 5: Filling in the pairwise alternative comparison matrix by the criterion of "ease of calculation".

After filling in all the matrices, the total weights of the alternatives were calculated; the alternatives are shown in [Fig. 6], and all summary data are presented in [Table 1].

When applying the simplified method, we must select the appropriate method in the menu following the stage of filling in the parameters.

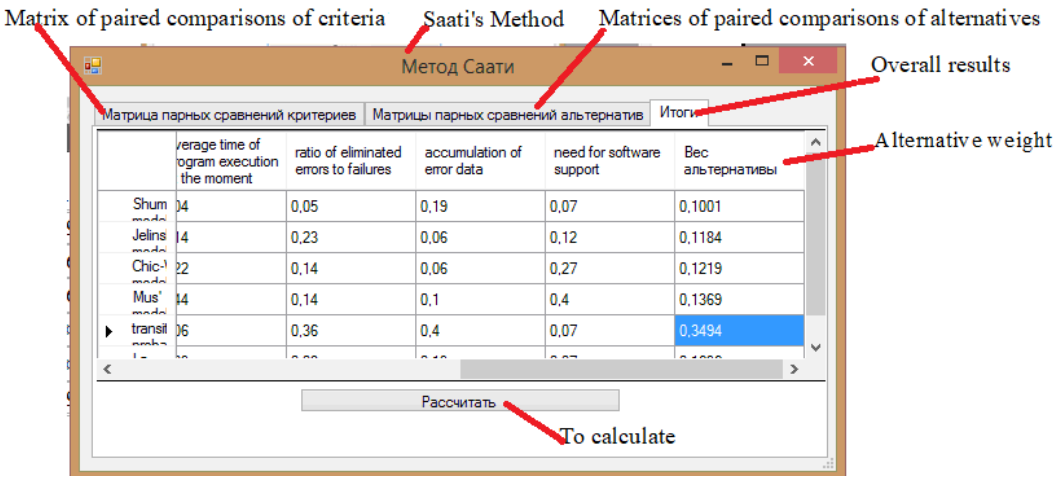

Fig. 6: Calculation of the total weights of the alternatives.

Table 1: Summary (Saati method)

Alternatives	Criteria									Alternative weight
	simplicity of calculations	forecasting ability	Iteration ability	correction of errors at the time of finding	failure record	average program execution time in a period	ratio of the numbers of errors eliminated to failures	error data accumulation	need for software support	
Shooman's Model	0.06	0.05	0.29	0.06	0.04	0.04	0.05	0.19	0,07	0,1001
La Padula's Model	0.22	0.05	0.06	0.22	0.16	0.14	0.23	0.06	0.12	0.1184
Jelinsky – Moranda's Model	0.22	0.05	0.06	0.22	0.25	0.22	0.14	0.06	0.27	0.1219
Chick – Walverton's Model	0.35	0.09	0.11	0,03	0.1	0.44	0.14	0.1	0.4	0.1369
Mus's Model	0,03	0.48	0.18	0.35	0.38	0.06	0.36	0.4	0,07	0.3494
Transition Probability Model	0.13	0.26	0.29	0.13	0.06	0.09	0.08	0.19	0,07	0.1696

In the next step, fill in the first row of the paired criteria comparisons matrix; the remaining rows are filled in automatically [Fig. 8].

Next, the matrices of pairwise alternative comparisons for each criterion are filled. The user fills only the first row of the matrix, and the rest of the values in the rows are calculated by the program. After filling in all the matrices, the total weights of the alternatives which are shown in [Fig. 8] were calculated; and all summary data are presented in [Table 2].

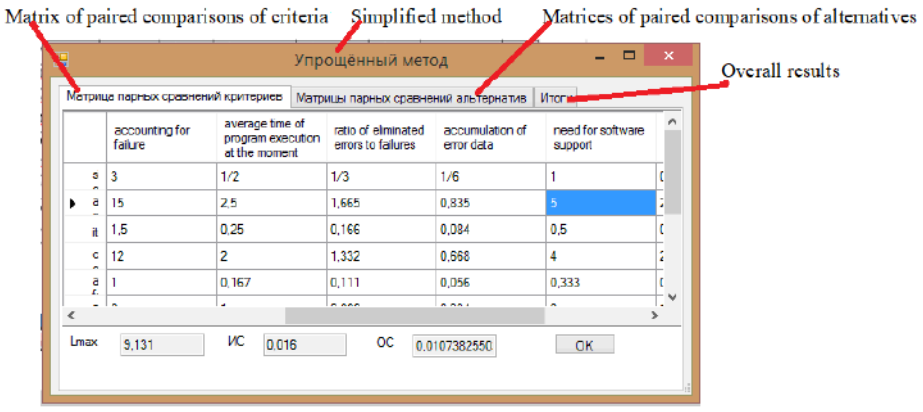

Fig. 7: Filling in the matrix of paired criteria comparisons.

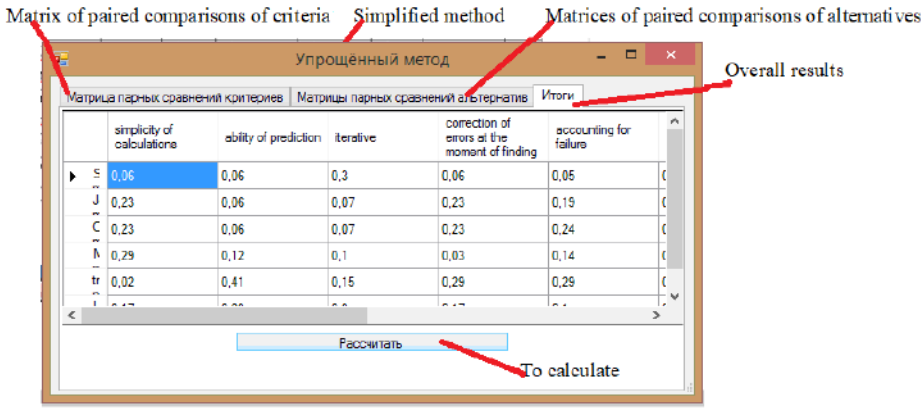

Fig. 8: Calculation of the total weights of the alternatives.

Table 2: Summary (simplified method)

Alternatives	Criteria									
	simplicity of calculations	forecasting ability	iteration ability	correction of errors at the time of finding	failure record	average program execution time in a period	ratio of the number of errors eliminated to failures	error data accumulation	need for software support	Alternative weight
Shooman's Model	0.06	0.06	0.3	0.06	0.05	0.05	0.05	0.19	0,07	0,096
La Padula's Model	0.23	0.06	0,07	0.23	0.19	0.18	0.22	0.06	0.14	0.133
Jelinsky – Moranda's Model	0.23	0.06	0,07	0.23	0.24	0.23	0.17	0.06	0.29	0.138
Chick – Walverton's Model	0.29	0.12	0.1	0,03	0.14	0.32	0.17	0.1	0.36	0.138
Mus's Model	0.02	0.41	0.15	0.29	0.29	0.09	0.28	0.39	0,07	0.298
Transition Probability Model	0.17	0.29	0.3	0.17	0.1	0.14	0.11	0.19	0,07	0.187

RESULTS

An analysis of the results obtained using the classical and simplified methods shows that:

- The most important criterion is "the ratio of the number of errors eliminated to failures", as well as "correction of errors at the time of finding" and "average program execution time in a period".
- Shooman's and transition probabilities models are the best alternatives according to the criterion of "iteration ability".
- The Chick – Walverton's model is the best alternative according to the criteria: "simplicity of calculations", "average time of program execution in a period", and "need for software support".
- The Mus's model is the best alternative according to the criteria: "forecasting ability", "error correction at the time of finding", "accounting for the presence of a failure", "ratio of the number of errors eliminated to failures" and "accumulation of error data".
- The best alternative (with a wide margin from the next - 0.3494 versus 0.1696 for the classical method and 0.2978 versus 0.1873) is the Mus's Model.

CONCLUSION

Thus, the proposed hypothesis is confirmed by an experiment for both processes of analytic hierarchy. The Mus's model for the chosen criteria for selection alternatives is the best of the analytical dynamic models.

However, the obtained results indicate the unambiguous priority of the Mus's model, from which we can conclude that it is necessary to conduct a practical experiment in which, using the example of one of the Platform modules, to build and evaluate models that have the highest weight values from alternatives, to consider a larger number of alternatives from analytical models, to increase the number of criteria by which alternatives are considered, and to divide all criteria into groups for a more demonstrable picture concerning the influence of criteria on the choice of alternatives.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

Completed as part of implementing a comprehensive project on creation of high-tech production "Development of a methodology and tools for creating applications, supporting the life cycle of information technology and decision-making software for the effective implementation of administrative and management processes within the established authorities", 2017-218-09-187; Decree of the Government of the Russian Federation dated April 9, 2010 No. 218

FINANCIAL DISCLOSURE

None.

IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] Gozzi K, et al. [2005] Fundamentals of software engineering, Translated from English SPb.: BHV-Petersburg, RU.

[2] Orlov SA. [2003] Software development technology: Textbook. St Petersburg, RU.

[3] Shiryaev MV, Andreeva ON. [2014] The direction of increasing the reliability of special software for embedded systems, Bulletin of MSTU MIREA, 2(3):176-183 https://rtj.mirea.ru/upload/medialibrary/9c3/16-shiryaev_andreeva. pdf

[4] Lyubitsyn VN. [2012] The need to develop reliable software as a challenge of our time. Bulletin of SUSU. Series Computer technology, management, electronics. 16(23):26-29.

[5] Pakulin NV, Lavrishcheva EM, Ryzhov AG, Zelenov SV. [2018] Analysis of methods for assessing the reliability of equipment and systems. The practice of applying methods. Proceedings of the Institute for System Programming of the Russian Academy of Sciences, 30(3):99-120. https://doi.org/10.15514/ISPRAS-2018-30(3)-8

[6] Vasilenko NV, Makarov VA. [2004] Models for assessing the reliability of software, Bulletin of the Novgorod state university, 28:126 - 132.

[7] Saati TL. [2008] Decision making with dependencies and feedbacks: Analytical networks. Translated from English, 360.

[8] Nogin VD. [2004] A simplified version of the hierarchy analysis method based on nonlinear convolution of criteria, Journal of Computational Mathematics and Mathematical Physics, 44(7):1261–1270; Comput Math And Math Phys, 44(7):1194–1202

[9] Automation system for calculating acc. to the hierarchy analysis method: Certificate on registration of databases, computer programs No. 2018615850 dated 17/05/2018.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

SOFTWARE QUALITY ASSESSMENT USING FUZZY PARAMETRIC CHARACTERISTICS

Rustam G. Asadullaev, Vladimir V. Lomakin*, Elena V. Ilinskaya
Institute of Engineering Technology and Natural Science, Belgorod State National Research University, Belgorod, 85 Pobedy st., 308015, Belgorod, RUSSIA

ABSTRACT

The article is devoted to the problem of software quality assessment during the trial operation stage. The developed approach is based on the hierarchy of quality assessment parameters according to GOST 28195-89. The approach is proposed to evaluate the quality of software during the trial operation stage with the breakdown of work into several stages and the subsequent aggregation of the resulting assessment. They formalized the set-theoretical description of the software quality assessment system. To evaluate some aspects of software quality that are evaluated by an expert method, the mathematical apparatus of fuzzy logic is used. The division of this process into separate stages is justified by the nature and specificity of quality indicator collection. At the same time, it is proposed to include usability indicator in the comprehensive assessment of software quality, obtained on the basis of user survey results.

KEY WORDS
software quality, fuzzy logic, GOST 28195-89, quality metric, quality factor.

Received: 17 Oct 2019
Accepted: 29 Nov 2019
Published: 5 Jan 2020

***Corresponding Author**
Email:
lomakin@bsu.edu.ru

INTRODUCTION

In the process of software development life cycle management, individual stages of work are distinguished, distributed over time depending on the chosen life cycle model. Quality assessment can be iterative, for example, in a spiral life-cycle management model, when the quality of each software version (PT) is tested taking into account the revised functionality. Testing staff (PT), in most cases, should have skills related to functional testing, performance testing, reporting on progress, planning, test conducting and automation [1].

Software Quality (SQ) - PT ability to meet specified or anticipated needs under specified conditions [2]. Research in the field of SQ assessment is carried out in the following areas: PT metrics, estimation methods, aggregation methods, model context depending on application type and instrumental support [3, 4]. Regardless of the selected life-cycle management model or other methods for PT development, comprehensive tests of PT are carried out at the final stage of work to assess its quality. To evaluate SQ, a pilot operation program is being developed which allows to evaluate all aspects of the system functioning. Some studies are focused on the development of PT quality prediction models. So in [5] a quality forecasting approach is proposed based on the basic components that are the part of any PT. The SQ assessment process is systematic and regulated by standards.

SQ is formed from a variety of indicators evaluated by various methods [6]: measuring, registration, organoleptic, calculated. Moreover, these methods cannot evaluate the whole range of quality indicators, for example, usability assessment or the availability and completeness of program documentation. An expert method is used for this. So in [7], the approach is proposed for the analysis and measurement of usability indicators at the implementation stage. The authors developed the platform with code annotations that are interpreted by the annotation processor to obtain valuable information and automatically calculate usability indicators during compilation.

Regardless of the approach to SQ estimation, there are PT failures that are difficult to detect by code pre-testing. Such failures can be caused by the factors of configuration parameter changes, depending on suppliers and business goal [8].

Thus, the SQ assessment process is a time-consuming task and requires the development of approaches to improve the efficiency of this process.

METHODS

Application of fuzzy logic to the SQ assessment process is not a new approach. In [9], a quality model is proposed that supports five concepts of quality and uses the theory of fuzzy logic to measure quality. A feature of the work is the use of Choquet Integral with fuzzy measures, which allow to take into account the relationship between the criteria. In [10], the assessment of software module quality is considered based on the frequency of checks and the density of errors using fuzzy logic for quantitative evaluation. Fuzzy logic is used as a quantitative assessment mechanism, allowing an expert to conduct a quality assessment in a natural language. So, in [11], they proposed a comprehensive approach to assess the quality of software based on qualitative factors from ISO/IEC 9126. Some researchers propose at the first stage the choice of a target model to assess SQ to determine the composition of the estimated functional goals, which are then formally presented and evaluated using fuzzy logic [12]. There are the works in which researchers create a fuzzy SQ estimation model by integrating several existing models. Thus, a unique set of factors, criteria and sub criteria is formed [13].

The analysis of the approaches shows that in most works, the authors take the existing standard to assess SQ as the basis. The mathematical apparatus of fuzzy logic is mainly applied to the assessment of all quality indicators. In this paper, we propose the approach that allows SQ to be assessed at the trial operation stage, with the work divided into several stages and the subsequent aggregation of the resulting assessment. In this case, to evaluate some aspects of SQ, evaluated by an expert method, the mathematical apparatus of fuzzy logic will be used. The division of this process into separate stages is justified by the nature and specificity of quality indicator collection.

The work is based on GOST 28195-89 [14]. This standard describes SQ indicators and the ways of their evaluation. Integral quality assessment is formed by the convolution of all indicators. However, in addition to calculation methods that allow, for example, to evaluate the probability of failure-free operation indicator, the expert method is used when an expert must evaluate the indicator in the range [0..1].

In this paper, it is proposed to translate the process of indicator evaluation evaluated by the expert method into a language that is natural for a man using fuzzy logic. This will allow an expert to evaluate the quality in the usual terms (satisfactory, good, excellent) and, in the same turn, to formulate the result of the indicator evaluation in the range [0..1] through the stage of inference defuzzification.

It is also proposed to include the indicator of usability in the comprehensive assessment of SQ, obtained on the basis of user survey results, by the sociological method [15].

IMPLEMENTATION

According to GOST 28195-89 [14], SQ is evaluated by a set of 6 factors, each of which is evaluated by its own set of criteria [Fig. 1]. Moreover, each criterion is evaluated by a metric or a set of metrics. The metric is formed by evaluative elements. Thus, 4 levels of SQ score are formed. At all levels of the hierarchy (evaluation elements, metrics, criteria and factors), a single rating scale is adopted in the range [0..1].

The analysis of SQ assessment standards allows us to describe this process from the point of view of system analysis. The SQ (Sq) estimation system is a tuple (1):

$$Sq =< LQ;\ R_{F\times QC};\ R_{QC\times M};\ R_{M\times EE};\ Op > \quad (1)$$

Tuple (1) is described by the following set of elements:

LQ – many available quality indicators that evaluate all aspects of PT quality (2);

$$LQ = \{EE, M, QC, F\} \quad (2)$$

EE – the subset of the evaluation elements;

M – the subset of metrics;

QC – the subset of quality criteria;

F – the subset of factors.

$R_{F\times QC}$ – the binary relation between the elements of a multitude of factors and quality criteria;

$R_{QC\times M}$ – the binary relation between the elements of quality criterion and metric sets;

$R_{M\times EE}$ – the binary relation between the elements of metrics and evaluation elements;

Op – the set of operations to determine the value elements or metrics (provided that the metric is evaluated in a unique way). Moreover, this set includes operations for hierarchy indicator evaluation [Fig. 1], which are evaluated by several values.

The binary relations ($R_{F\times QC}, R_{QC\times M}, R_{M\times EE}$) are filled with the coefficients W_{ij}, that establish the relationship between the elements of two sets. Moreover, the values W_{ij} are weight coefficients reflecting the influence of the quality indicator j on i.

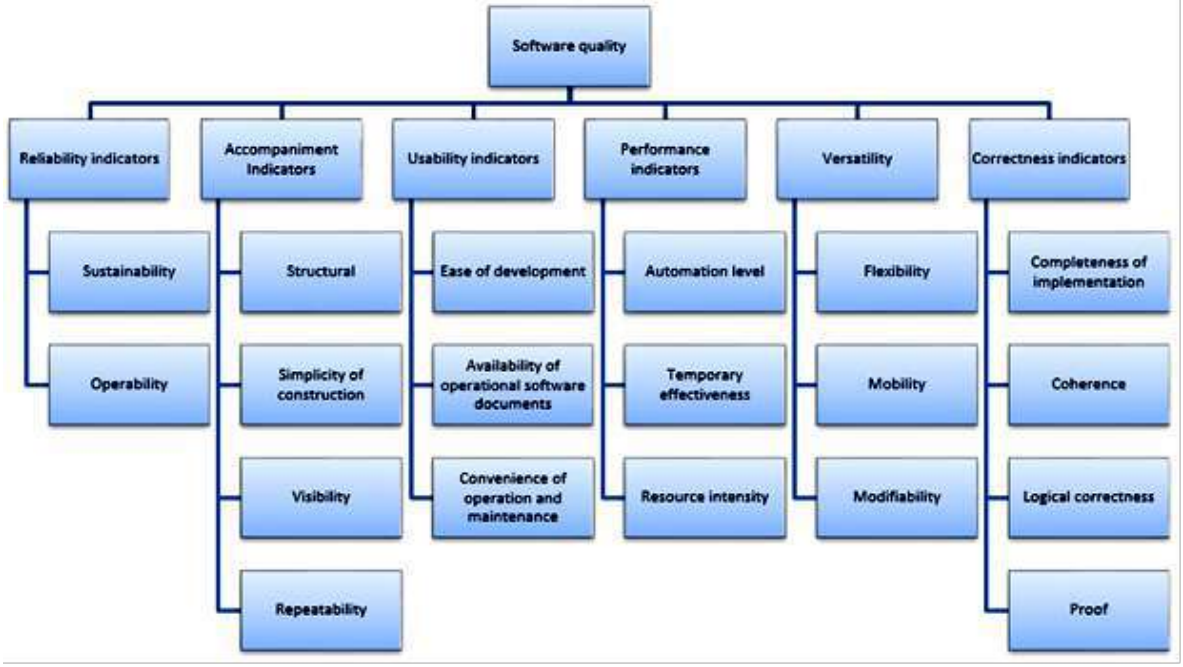

Fig. 1: Software quality assessment hierarchy.

According to GOST 28195-89 [14], the integral estimate of SQ is determined by the expression, which is the arithmetic mean weighted sum of factors, criteria, metrics and evaluation elements. The average assessment of the evaluation element, taking into account several of its values is carried out according to the formula (3).

$$ee_{kq} = \frac{\sum_{t=1}^{n} ee_t}{n} \qquad (3)$$

where ee_{kq} – the average estimate of the evaluation element q for the metric k;

n– the number of values q of the evaluation element ee_{kq} .

Based on the evaluation elements, an estimate of the metric (4) is formed.

$$m_{jk} = \frac{\sum_{i=1}^{Q} ee_{kq}}{Q} \qquad (4)$$

where m_{jk} – the average estimate of the metric for the criterion;

Q– the number of evaluation elements in the metric.

The calculated metrics are the basis for the quality criterion determination (5).

$$C_{ij} = \frac{\sum_{k=1}^{S}(m_{jk} * v_{jk})}{C_{ij}^{base}} \qquad (5)$$

where C_{ij} is the relative indicator of the criterion j for the factor i ;

S– the number of metrics that evaluate the criterion j ;

C_{ij}^{base} - The basic value of the criterion corresponding to the world level;

v_{jk} – the metric k weight coefficients for the criterion j . At that $\sum_{k=1}^{S} v_{jk} = 1$

Thus, each factor is evaluated (6).

$$F_i = \sum_{i=1}^{J}(C_{ij} * w_{ij}) \qquad (6)$$

where F_i – the factor i, where $i = \overline{1,6}$;

J – the number of criteria that evaluate the factor i;

w_{ij} – the weighting factors of the criterion j for the factor i. At that $\sum_{j=1}^{J} w_{ij} = 1$.

Thus, each factor is evaluated and compared with the value of the base indicator, which is determined at the design stage. Evaluation elements are calculated by various methods, depending on the availability of analytical information processing. [Fig. 1] shows the hierarchy of SQ indicators, which reflects the relationships between factors and quality criteria, which are further decomposed into metrics and evaluation elements.

However, from the position of direct assessment of each element, it is necessary to group the evaluation elements and metrics by the method of evaluation. In this regard, it is proposed to group according to the method of calculation into analytical methods and expert ones using fuzzy logic, which in turn are decomposed into an expert assessment and user assessment. In particular, end-to-end tests can be developed for users to simulate real user scenarios.

Formalization of expert assessments using fuzzy logic is necessary. It is proposed to evaluate the elements and metrics in the form of linguistic variables LV_{name}. The variable name corresponds to the name of the evaluation element. Each linguistic variable will be evaluated by a variety of terms $\{low, average, high\}$, that characterize the level of evaluation elements for the metric. The semantics of terms is defined on the interval [0, 1] and formalized by membership functions of the type $z, s, gauss$.

For example, according to [14], to evaluate the metric of PT work mastering (LV_{SD}) it is recommended to use assessment elements: the possibility of mastering PT according to the documentation (LV_{PDSTD}), the possibility of mastering PT using a test example (LV_{PMSTE}) and the possibility of phased mastering of PT (LV_{PPDS}). A fuzzy assessment of the metric mastering the work of PT is formed by the assessment of three linguistic variables through the logical conclusion of Mamdani [Fig. 2]. After the stage of defuzzification, the assessment of the variable LV_{SD} takes a numerical value, which is used in the assessment of the quality criterion according to the formula (5).

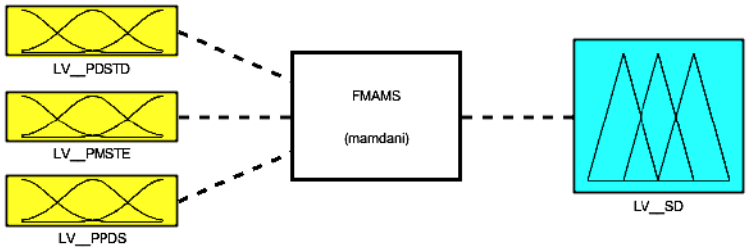

Fig. 2: Fuzzy metric assessment mastering software.

Each input linguistic variable for the inference system [Fig. 2] is evaluated on the basis of three terms. [Fig. 3] shows the semantics of the of the linguistic variable terms LV_{PDSTD}.

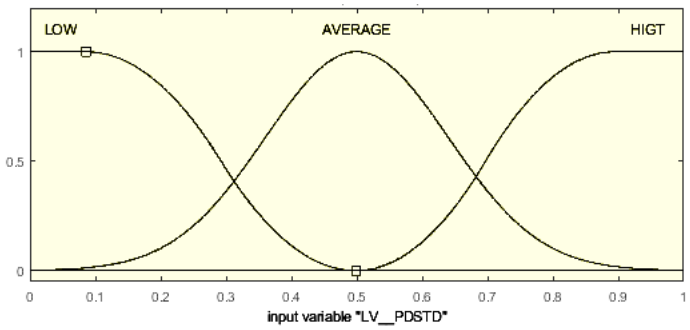

Fig. 3: The semantics of the terms {low, average, high} for the linguistic variable LV_{PDSTD}.

IIOAB JOURNAL
ISSN: 0976-3104

Thus, the proposed SQ estimation approach is reduced to the following sequence of actions:

1. The structuring of quality indicators according to (1) based on recommendations of regulatory documents and taking into account the specifics of PT development.
2. Determination of baseline indicators for each SQ factor.
3. Separation of the set of evaluation elements and metrics according to the evaluation method into subsets of analytical estimates and expert estimates using fuzzy logic, which in its turn are decomposed into an expert assessment and user assessment.
4. The implementation of a fuzzy inference to assess quality metrics based on linguistic variables that reflect the semantics of the evaluation elements, and obtaining the numerical value of the estimates based on defuzzification results.
5. Calculation of quality criteria based on metric estimates by the formula (5).
6. Calculation of factors based on the assessments of quality criteria by the formula (6).
7. The analysis of the obtained factor estimates for comparison with the values of the basic indicators and the development of a decision about the integral assessment of SQ.

SUMMARY

The developed approach allows us to estimate SQ based on the indicators decomposed into 4 levels of the hierarchy. A set-theoretic description of the SQ assessment process allows you to structure the hierarchy of indicators and the relationships between them, taking into account the degree of the lower level component hierarchy influence on the upper ones. The lower two levels (metrics and evaluation elements) are evaluated by analytical and expert methods using the mathematical apparatus of fuzzy logic. The semantics of linguistic variable terms are described by three terms. Quality criteria are evaluated as the weighted average of the metrics. Factors are estimated as a weighted sum of quality criteria.

CONCLUSION

The advisory nature of SQ standards allows developers to take them as a basis and modify them taking into account the specifics and needs of the developed PT. The proposed approach to the assessment of SQ using fuzzy parametric characteristics allows us to systematize the SQ process taking into account individual aspects of the project. The proposed set-theoretic description of the SQ system structures the elements of the system and establishes the relationships between them. This allows you to modify the standard SQ easily taken as the basis and formalize the required representation of the process.

The introduction of fuzzy parametric characteristics simplifies an expert's work during evaluation of a multitude of parameters evaluated by qualitative features. Linguistic variables are described by three terms corresponding to a high, low and average value of the indicator. Defuzzification of fuzzy estimates allows you to use them in the future for the numerical evaluation of quality indicators of a higher hierarchy level.

In SQ approach, they proposed to take into account the expert assessments of the developed system users. This will make it possible to form the estimates of real consumers of the system and evaluate the effectiveness of some aspects of the system practical use. The vision of users, as a rule, differs from the vision of developers, which may cause difficulties in the process of the system mastering.

Successful implementation of the proposed approach requires laborious work of experts on the formation of weighting factors for metrics and quality criteria. Weighting factors individually reflect the significance of the parameter when they develop the parameter estimate of the highest hierarchy level.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
Completed as the part of a comprehensive project implementation to create high-tech production "Development of a methodology and tools for applied applications, supporting the life cycle of information technology support and decision making for an effective implementation of administrative processes within the established authority", code 2017-218-09-187; Decree of the Russian Federation Government No. 218 issued on April 9, 2010.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Florea R, Stray V. [2019] The skills that employers look for in software testers. Software Quality Journal. 27(4):1449–1479.
[2] ISO/IEC 25000:2014 Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE. Date Views 10.09.2019 https://www.iso.org/obp/ui/#iso:std:iso-iec:25000:ed-2:v1:en

[3] Meng Y, Xin X, Xiaohong Z, Ling X, Dan Y, Shanping L. [2019] Software quality assessment model: a systematic mapping study. Science China Information Sciences, 62:191101.
[4] Senthil MC, Prakasam S. [2013] A Literal Review of Software Quality Assurance. International Journal of Computer Applications, 8(78): 25-30.
[5] Singh B, Kannojia SP. [2012] A Model for Software Product Quality Prediction. Journal of Software Engineering and Applications, 6(5): 395-401.
[6] State industry standard 28195-89 Quality control of software systems. General principles. Date Views 10.09.2019 http://docs.cntd.ru/document/1200009135
[7] Schramme M, Macias JA. [2019] Analysis and measurement of internal usability metrics through code annotations. Software Quality Journal, 27(4): 1505–1530.
[8] Haller K, Grotz R. [2015] Software Quality Beyond Testing In-house Code. The Tester, 54. Date Views 10.09.2019 https://cdn.bcs.org/bcs-org-media/2741/tester-2015.pdf
[9] Pasrija V, Kumar S, Srivastava PR. [2012] Assessment of Software Quality: Choquet Integral Approach. Procedia Technology, 6 (2012): 153 – 162
[10] Mittal H, Bhatia P. [2008] Software Quality Assessment Based on Fuzzy Logic Technique. International Journal of Soft Computing Applications. SIGSOFT Software Engineering Notes, 34(3):1-5.
[11] Challa JS, Paul A, Dada Y, Nerell V, Srivastava PR, Singh AP. [2011] Integrated Software Quality Evaluation: A Fuzzy Multi-Criteria Approach. Journal of Information Processing Systems, 7(3):473-518.
[12] Mansoor A, Streitferdt D, Fubi FF. [2015] Fuzzy Based Evaluation of Software Quality Using Quality Models and Goal Models" International Journal of Advanced Computer Science and Applications, 6(9): 2015.
[13] Kara M, Lamouchi O, Ramdane-Cherif A. [2016] Ontology Software Quality Model for Fuzzy Logic Evaluation Approach. Procedia Computer Science, 83(2016): 637-641.
[14] State industry standard 28195-89 Quality control of software systems. General principles. Date Views 10.09.2019 http://docs.cntd.ru/document/1200009135
[15] Lomazov AV, Lomazova VI, Lomakin VV, Asadullaev RG. [2019] Estimation of usability of the corporate applications integrated development platform based on the results of users' questionnaire surveys. Scientific and Technical Bulletin of the Volga Region, 5:45-49.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

SPECIAL ARTICLE

MATHEMATICAL APPROACH TOWARDS RECENT INNOVATION IN COMPUTATION AND ENGINEERING SYSTEM (MATRICS)

PREDICTION OF SOCIAL MEDIA BASED ON ARIMA AND ANN MODEL

Priyadarshini E.[1]*, Manjula T.[2], Krishnareddy Deepika[3], Anuhya Bandi[4]

[1,2]*Department of Mathematics, Sathyabama Institute of Science & Technology, Chennai, INDIA*

[3,4]*Department of ECE, Sathyabama Institute of Science & Technology, Chennai, INDIA*

ABSTRACT

One of the popular approaches for prediction of time series is Auto Regressive Integrated Moving Average abbreviated as ARIMA. The main objective of this paper is to predict the percentage of Facebook users with the time series approach. Methods: To obtain optimum accuracy, statistical support like autocorrelation and partial autocorrelation has been used. Standard statistical techniques are used to verify the validity of the model. The prediction power of ARIMA model was used to predict the percentage of Facebook users for succeeding month and to estimate the mean absolute percentile error (MAPE). Results: The performance of Artificial Neural Network using Multi Layer Perceptron (MLP) in the prediction of percentage of Facebook users was evaluated. In this study, a total of 122 data points (monthwise) of past 11 years from April 2009 to April 2019 has been taken to explore and predict Facebook users based on statistical and computational techniques. Conclusions: The Mean Absolute Percentage Error (MAPE) was used to evaluate the accuracy of the models.

KEY WORDS
Forecast, Auto Correlation, Partial Auto Correlation, Hidden layers, Multi Layer Perceptron

Received: 25 Oct 2019
Accepted: 24 Feb 2020
Published: 3 Mar 2020

***Corresponding Author**
Email:
priyaeb@gmail.com

INTRODUCTION

In today's advanced world, social media play a humungous part. Social media are the interactive technologies that are mediated by computer. They facilitate sharing and creation of information via virtual communities and networks. One such huge platform of social media is Facebook. It was founded in 2004 by Mark Zuckerberg and his fellow Harvard students. It is considered as one among "Big Four" technology companies which are Amazon, Apple and Google. Though founders initially confined websites membership to students of Harvard, later it was extended to IVY LEAGUE schools, MIT and higher education institutions in Boston area. In 2006 began the outgrowth when anyone who claimed to be at least 13 years old were allowed to register as a member. Facebook can be accessed from devices which have internet connectivity. By 2007, Facebook had 100,000 pages on which various companies promoted themselves. Pages of organizations began rolling out in the year 2009. The company had announced 500 million users in the month of July in 2010. Half of the people on Facebook used it daily for an average of 34 minutes. In 2015 Facebook announces that it had reached 2 million active advertisers with most gain coming from small businesses. In 2016, the number had reached 3 million with more percentage coming from outside United States. As they say that anything overused harms, Facebook's over usage did have multiple drawbacks. Spreading of fake messages was easy using such platform. In 2015, Facebook's algorithm was revised trying to filter out fake news stories and hoaxes. In addition to spreading fake news, the perpetrator of March 2019 New Zealand attacks used Facebook for live streaming of footage. Though Facebook is advantageous, it also has drawbacks of vulnerability. The outcome depends on proper and purpose of usage.

Peng et al. [1] applied the ARIMA model to forecast the crime which is helpful for the local police stations and municipal government. Hong et al. [2] predicted the traffic flow using ARIMA models. Using ARIMA models Weng Dongdong [3] had predicted the consumer price index (CPI). In the paper [4] Chinese mobile user forecast was discussed using ARIMA models. Babu and Reddy [5] in their paper analysed the average the global temperature using ARIMA models. Hussain et.al [6] made three types of forecasts such as historical, ex-post and ex-ante. Amin et al. [7] made analysis for potato prices and forecasts the nine future points. Priyadarshini et al. made forecasts of foreign exchange rates, silver rates and crude oil rates using ARIMA models [8, 9,]. A comparative analysis has been done using Artificial Neural Network and Auto Regressive Integrated Moving by Priyadarshini et al. [10].

ARIMA METHODS

ARIMA models are most commonly used to predict time series that can be stated by transformation lags of differenced series that appear in forecasting equation are called auto regressive terms . The lags of forecast errors are called moving average terms and a time series that needs to be differenced to be made stationary is called an integrated version of a stationary series. Random trend models, exponential smoothing model, auto regressive models and random walk are special cases of ARIMA models. A non-seasonal ARIMA model is classified as an "ARIMA (p,d,q)" model where:

- p is the number of auto regressive terms
- d is the number of non-seasonal difference

- q is the number of moving average terms

The following is how simple regression model can be expressed

$$Y_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + \cdots + b_p Y_{t-p} + e_t \quad [1]$$

Where Yt is the forecast variable yt-1,.....,Yt-p are explanatory variables and et is the error term. The name auto regression is used to denote the above equation due to the time-lagged values of the explanatory variable. The moving average model uses the past errors as the explanatory variables. A simple moving average model is represented as follows:

$$Y_t = b_0 + b_1 e_{t-1} + b_2 e_{t-2} + \cdots + b_p e_{t-p} + e_t \quad [2]$$

Similarly, a seasonal model can be represented as ARIMA (p,d,q) (P,D,Q). Basically, this method has three phases: model identification, parameters estimation and diagnostic checking. To identify the appropriate ARIMA model for a time series, the order(s) of differencing the series and remove the gross features of seasonality.

Artificial neural network (ANN)

Artificial Neural Network (ANN) uses a variety of optimization tools to learn from past experiences and uses this to predict and find new patterns. Using training algorithms, ANN systems are made to learn the percentage of Facebook users. Learning involves the extraction of rules or pattern from the historic data. In this neural network, using Multilayer Perceptron (MLP), models have been used for the prediction of Facebook users.

Predictors for the ANN

The five predictors for ANN are the technical indicators namely three monthly moving averages, six monthly moving averages, one yearly moving averages and two yearly moving averages respectively [Table 4].

Experimental results of multi-layer perceptron (MLP) neutral network

The MLP computes the new value of the weights and biases using the gradient descent. It quickly adjusts the network weights giving good performance. The space denoting the error for every combination of weights and biases is called as error space. The feed forward neural network architecture used in this experiment consists of two hidden layer along with one input and output layer respectively. The transfer function which is used in the hidden layer and output layer are hyperbolic tangent and identity [Table 5].

Backward pass:

i. Compute the error
ii. Compute
iii. Compute
iv. Keep updating the weights connecting the input layer to the hidden layer using the rule. After determining all the μ factors, the weights are adjusted for all layers simultaneously.

x : p X1 input vector
h : Weighted sum of input stimuli
r : m X1 output vector of hidden layer
g : Weighted sum of vj
Y : n X1 output vector of output layer
wij : weight connecting ith unit of output layer and jth unit of hidden layer
wjk : Weight connecting jth unit of hidden layer to kth unit of input layer
y : Actual output
yd : Desired output.

where i, j and k indices refers to the neurons in the output, hidden and input layers respectively, p, m and k are the number of neurons in input, hidden and output layer respectively.

Data analysis using arima

The analysis is taken for the data from the monthly percentage of Facebook users for a period of 10 years from 2009 to 2019 [Fig 1]. X-axis denotes the months from April 2009 to April 2019. The percentage of Facebook users was forecasted for the month of May 2019 using ARIMA modelling [Fig. 3].

a) Model identification:

The variable is transformed into a time series under forecasting. The value which varies over time around a constant variance and mean is the stationary series. The approach is to test stationary through checking the time plot. Using appropriate differencing non stationary is corrected. To obtain the stationarity

difference of order one was enough. Then the values of p and q are identified. The autocorrelation and partial autocorrelation are calculated for different orders [Fig 2]. The order p and q can be one. Four ARIMA models were calculated and the one which has minimum Bayesian Information Criterion (BIC) was chosen.

b) Model estimation and verification:

SPSS package is used to estimate the model parameters. Results are given in table. The model examination is concerned with testing the residuals then observe whether they have any systematic pattern which can be removed to improve. This can be done by verifying the autocorrelation and partial autocorrelation of various orders. Many correlation of upto 16 lags are calculated and this significance us verified by Box-Ljung test. This shows us that chosen ARIMA model is appropriate model.

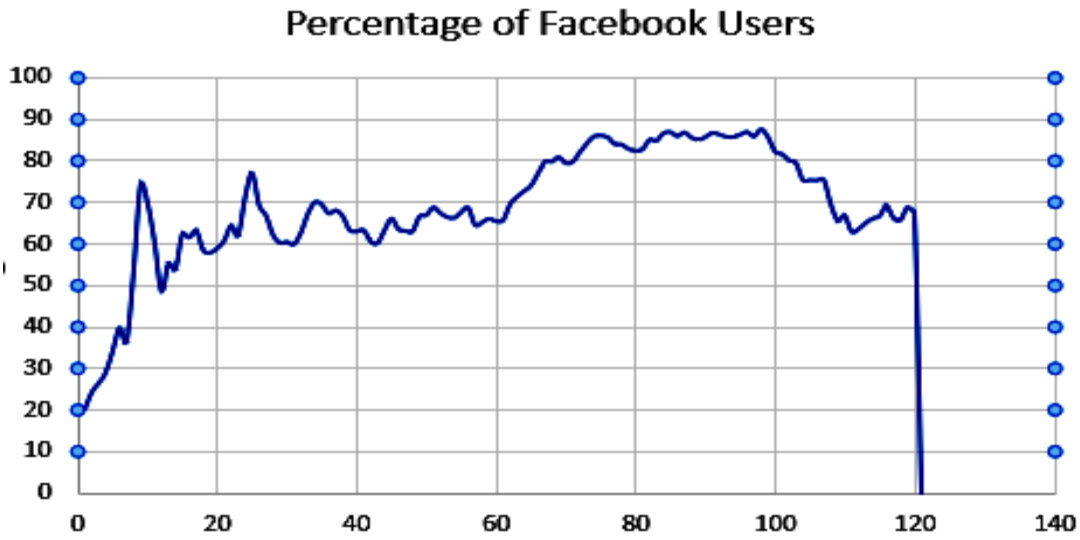

Fig. 1: Chart showing the percentage of Facebook users from April 2009 to March 2019.

Table 1: Model Description

FACE BOOK USERS	Model ID	Model Type
	Model 1	ARIMA(0,1,1)

Table 2: Model Fit

Fit Statistic	Mean	Min	Max	Percentile						
				5	10	25	50	75	90	95
Stationary R-squared	.043	.043	.043	.043	.043	.043	.043	.043	.043	.043
R-squared	.916	.916	.916	.916	.916	.916	.916	.916	.916	.916
RMSE	3.796	3.796	3.796	3.796	3.796	3.796	3.796	3.796	3.796	3.79
MAPE	3.931	3.931	3.931	3.931	3.931	3.931	3.931	3.931	3.931	3.93
MaxAPE	33.436	33.436	33.436	33.436	33.436	33.436	33.436	33.436	33.436	33.4
MAE	2.406	2.406	2.406	2.406	2.406	2.406	2.406	2.406	2.406	2.40
MaxAE	17.851	17.851	17.851	17.851	17.851	17.851	17.851	17.851	17.851	17.8
Normalized BIC	2.708	2.708	2.708	2.708	2.708	2.708	2.708	2.708	2.708	2.70

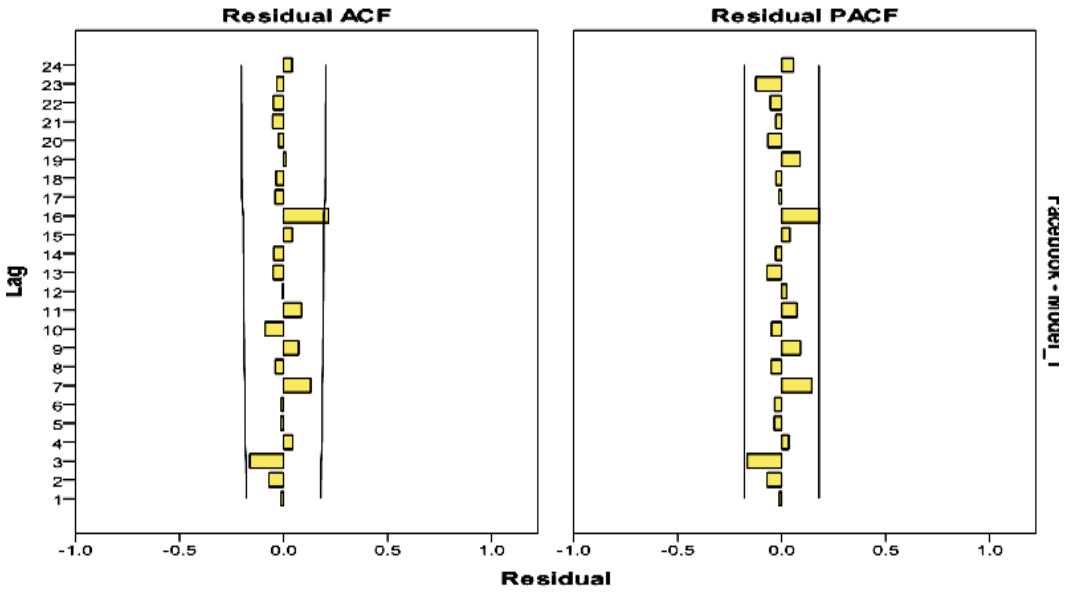

Fig. 2: Chart showing resudual ACF and PACF for the model ARIMA (0,1,1).

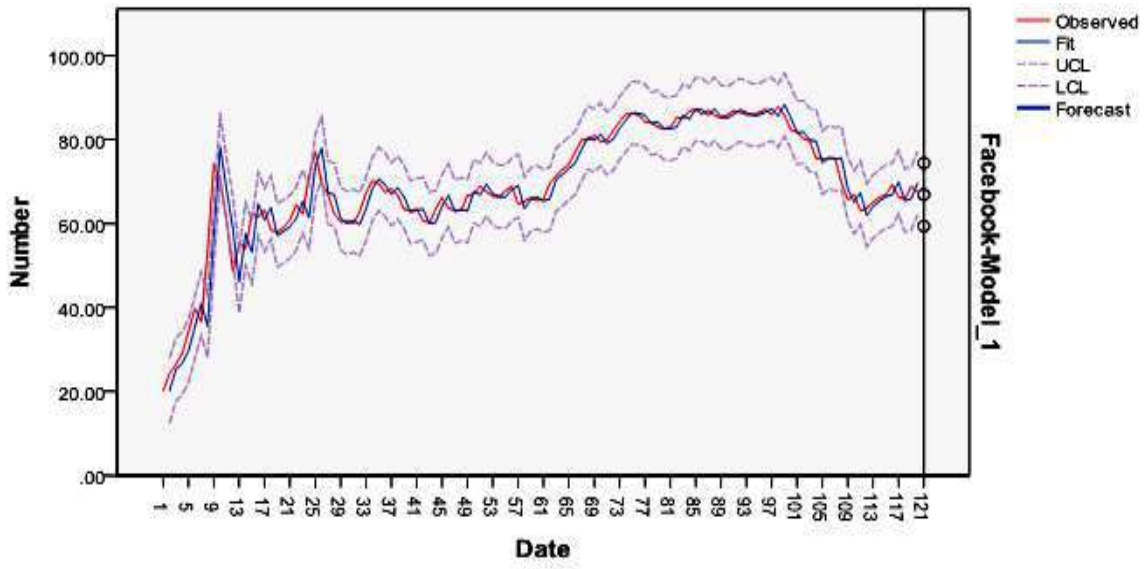

Fig. 3: Chart showing the actual and predicted values of the percentage of Facebook users.

Table 3: Case Processing Summary

		N	Percent
Sample	Training	67	69.1%
	Testing	30	30.9%
Valid		97	100.0%
Excluded		25	
Total		122	

RESULTS

In 2018, recent time's researchers conducted a survey on social media users, in which Facebook and YouTube dominate the social media landscape. Facebook remains the primary pattern for most of the Americans. It has been reported that roughly two thirds of U.S adults use Facebook and among them three quarters of those use Facebook on a daily basis. A majority of Americans with exception of those 65 and older now use Facebook. Prediction is made about the outcome of a future event based on a pattern of evidence. Using ARIMA methodology, we have estimated the percentage of number of users using Facebook for the month of June. In this study, 122 data points were taken from April 2009 to April 2019 and have been analyzed with help of ARIMA (0,1,1)[Table1]. From the analysis carried out the Mean

Absolute Error (MAE) was found to be 2.406, Root Mean Square Error (RMSE) was found to be 3.796 and Mean Absolute Percentage Error (MAPE) as 3.931[Table 2].

Table 4: Network Information

Input Layer	**Covariates**	1	2yr MA
		2	1yr MA
		3	6mnth MA
		4	3mnth MA
Hidden Layer(s)	Number of Units[a]		4
	Rescaling Method for Covariates		Standardized
	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		3
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Facebook
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares

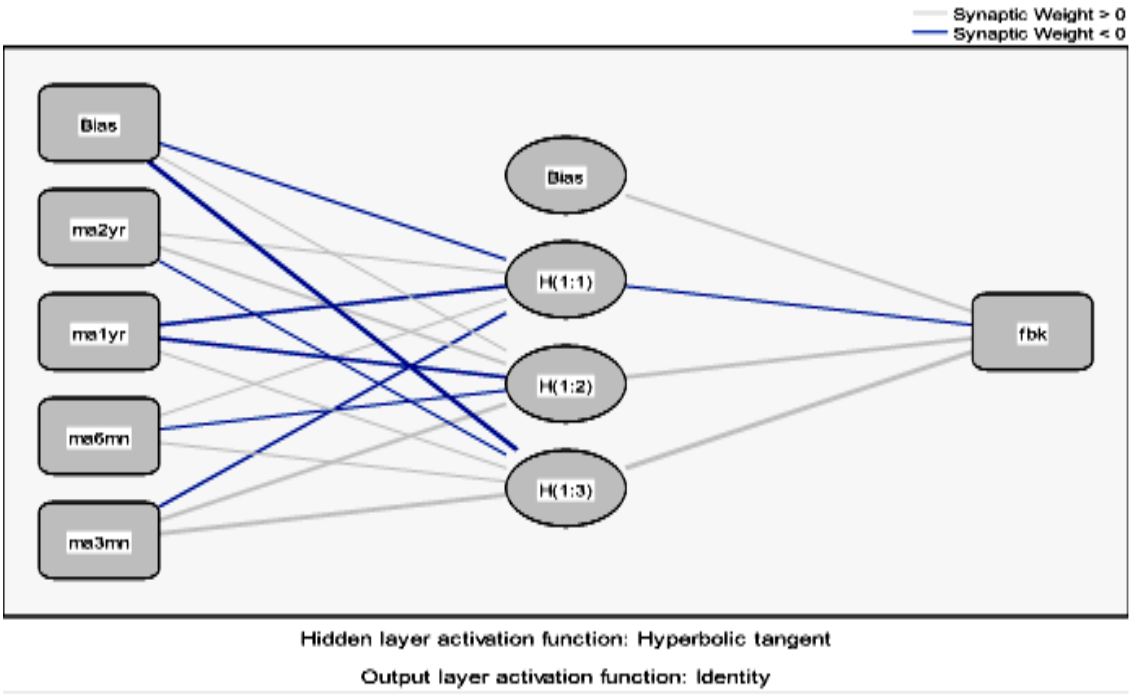

Fig. 4: Architecture of the ANN Model.

Table 5: Model Summary

Training	Sum of Squares Error	2.807
	Relative Error	.085
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	00:00:00.029
Testing	Sum of Squares Error	1.832
	Relative Error	.146

Table 6: Parameter Estimates

Since the MAPE, RMSE and MAE are very less, this model can be used for future prediction of Facebook users. The Artificial Neural Network model for training the data the sum of the squares of the error (SSE) was found to be 2.807[Table 3] and the relative error was 0.085 and for testing the data, the sum of the squares of the error was found to be 1.832 and the relative error was found to be 0.146[Table 6].The architecture of ANN consists of 5 input layers, hidden layers and output layer,the hidden layer activation function was hyperbolic function and output layer function was identity function.

Predictor		Predicted			
		Hidden Layer 1			Output Layer
		H(1:1)	H(1:2)	H(1:3)	Facebook
Input Layer	(Bias)	-.274	.139	-.638	
	ma2yr	.266	.280	-.145	
	ma1yr	-.427	-.440	.031	
	ma6mn	.050	-.042	.020	
	ma3mn	-.276	.609	.870	
Hidden Layer 1	(Bias)				.296
	H(1:1)				-.191
	H(1:2)				.752
	H(1:3)				1.070

Table 7: Prediction of Facebook users for the month of May, 2019

Model	ARIMA(0,1,1)	ANN	ACTUAL VALUE
Percentage of FACE BOOK Users	66.93	67.8	69.52

CONCLUSIONS

Using Arima model, the percentage of Facebook users for the month of May 2019 was found to be 66.9 and using ANN model ,it was found to be 67.8[Table 7].This type of analysis enables us to understand how a time series model can predict and provide us with appropriate result to help us modify the technology suitably and approach the possible upcoming challenges and scenarios. It also helps us to give the future forecast of such time series.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None.

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] Peng C, Hongyong Y, Xueming S. [2008] Forecasting Crime using the ARIMA model, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 5: 627- 630.

[2] H. Dong, L. Jia, X. Sun, C. Li and Y. Qin. [2009] Road Traffic Flow Prediction with a Time-Oriented ARIMA Model, Fifth International Joint Conference on INC, IMS and IDC, 2009:1649-1652.

[3] Dongdong W. [2010] The Consumer Price Index Forecast Based On ARIMA Model, WASE International Conference on Information Engineering , Beidaihe, Hebei, 1: 307-310.

[4] Xu Y. [2010] The Application of ARIMA Model in Chinese Mobile User Prediction, IEEE International Conference on Granular Computing, San Jose, CA, 2010:586-591.

[5] Babu NC, Reddy EB. [2012] Predictive Data Mining On Average Global Temperature Using Variants of ARIMA models, IEEE- International Conference On Advances In Engineering, Science And Management, 259-263

[6] Hossain Z, Abdus QS, Ali Z. [2006] ARIMA Model and Forecasting with Three Types Of Pulse Prices In Bangladesh: A Case Study, International Journal of Social Economics, 33(4):344-353.

[7] Amin F, Razzaque MA, [2000] Autoregressive Integrated Moving Average (ARIMA) Modelling for Monthly Potato Prices in Bangladesh, Journal of Financial Management and Analysis, 13(1):74-80.

[8] Priyadarshini E, Chandrababu A. [2011] Modeling And Forecasting Of Foreign Exchange Rates Using ARIMA Models, Proceedings of the National Conference on Recent Developments in Mathematics and its Applications at, SRM University, Chennai, Excel India Publishers:379 -383.

[9] Priyadarshini E, Chandrababu A. [2011] Forecasting of Crude Oil Rates using ARIMA Models, International Journal of Statistics and Systems(IJSS), Research India Publications, 6(3):287-29.

[10] Priyadarshini E. [2015] A Comparative Analysis Of Prediction Using Artificial Neural Network and Auto Regressive Integrated Moving, ARPN Journal of Engineering and Applied Sciences, 10(7):3078-3081.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

SPECIAL ARTICLE

MATHEMATICAL APPROACH TOWARDS RECENT INNOVATION IN COMPUTATION AND ENGINEERING SYSTEM (MATRICS)

AN APPLICATION OF RADIAL BASIS NEURAL NETWORK FUNCTION FOR RAINFALL PREDICTION

Nirmala Muthu*

Department of Mathematics, Sathyabama Institute of Science and Technology, Tamilnadu, INDIA

ABSTRACT

This research work deals with the prediction of annual rainfall of Chennai city, India using Neural Network algorithm. Chennai is one of the metropolitan city in India where disaster management is in need of knowing the extreme precipitation in advance for the control of floods and droughts. For this purpose, the monthly rainfall of Chennai is modeled using the Neural Network algorithm. The experimental result shows that the proposed neural network algorithm gives better accuracy. The error measures indicate the proposed model's performance accuracy.

KEY WORDS
Chennai, rainfall, prediction, neural network algorithm, error measure

Received: 04 Nov 2019
Accepted: 28 Feb 2020
Published: 4 Mar 2020

***Corresponding Author**
Email:
mnirmalamaths@gmail.com

INTRODUCTION

Rainfall prediction is one of the most important and challenging task carried over by meteorologists all over the world. Rainfall prediction is also useful for disaster management in tackling the extreme precipitations. There are many forecasters doing lot of research to obtain more accuracy in the model they construct. There are many factors which affect rainfall prediction such as temperature, humidity, wind speed, pressure, dew point etc. and hence prediction becomes more complicated for them since the real life data are nonlinear in nature. Nowadays Artificial Neural Network algorithms are applied to time series analysis.

Artificial Neural Network plays a major role in learning and capturing the characteristics of any nonlinear natural phenomena. MoniraSumi et al [9] made an attempt for deriving forecasting models for average daily and monthly rainfall of the Fukuoka city in Japan using the artificial neural network, multivariate adaptive regression splines, the k-nearest neighbor and radial basis support vector regression together with preprocessing techniques like moving average method and principal component analysis. The K-nearest neighbor, artificial neural network, and extreme learning machine were applied for the seasonal forecasting of summer monsoon (June-September) and post-monsoon (October-December) rainfall for the Kerala state of India by Yejnaseni Dash et al and proved that extreme machine learning performed better when compared to the other two models [12]. FhiraNhita et al constructed a forecasting system with evolving neural network algorithm to forecast the monthly rainfall of Bandung Regency in Indonesia [4] and were successful by arriving a forecast accuracy of 70%. Duong TranAnh et al [3] combined two preprocessing methods seasonal decomposition and Discrete Wavelet Transform to decompose the monthly rainfall series in Vietnam and constructed the prediction models based on Artificial Neural Network and Seasonal Artificial Neural Network. Kavitha Rani et al developed Artificial Neural Network based rainfall prediction model with teaching learning optimization algorithm to improve the accuracy of the model [7]. A feed forward neural network with back propagation and Levenberg-Marquardt algorithms based prediction models were constructed by Neelam Mishra et al for forecasting one-month and two-month ahead rainfall of Northern India [10].

In this study, a prediction model for annual rainfall in Chennai is constructed using Radial Basis Function (RBF) algorithm of Neural Network structure. For this purpose, the various parameters for the prediction of rainfall used are ElNino 3.4, Maximum Temperature, Minimum Temperature, Vapour Pressure and Cloud Cover as these factors play major role particular area.

MATERIALS AND METHODS

Chennai is the capital city of Tamil Nadu state in India. Its geographic location is 13°04'N latitude and 80°17'E longitude [5]. According to the 2011 Indian census, it is the sixth-most populous city and fourth-most populous urban agglomeration in India.Chennai has a tropical wet and dry climate.. The hottest part of the year is late May to early June with maximum temperatures around 35 °C – 40 °C (95–104 °F). The coolest part of the year is January, with minimum temperatures around 19 °C –25 °C (66–77 °F). The lowest recorded temperature was 13.9 °C (57.0 °F) on 11th December 1895 and 29th January 1905. The highest recorded temperature was 45 °C (113 °F) on 31 May 2003. The average annual rainfall is about 140 cm (55 in). The city gets most of its seasonal rainfall from the north–east monsoon winds, from mid–October to mid–December. Cyclones in the Bay of Bengal sometimes hit the city. The highest annual rainfall recorded is 257 cm (101 in) in 2005 (Table 1). Prevailing winds in Chennai are usually southwesterly between April and October and north-easterly during the rest of the year.

IIOAB JOURNAL
ISSN: 0976-3104

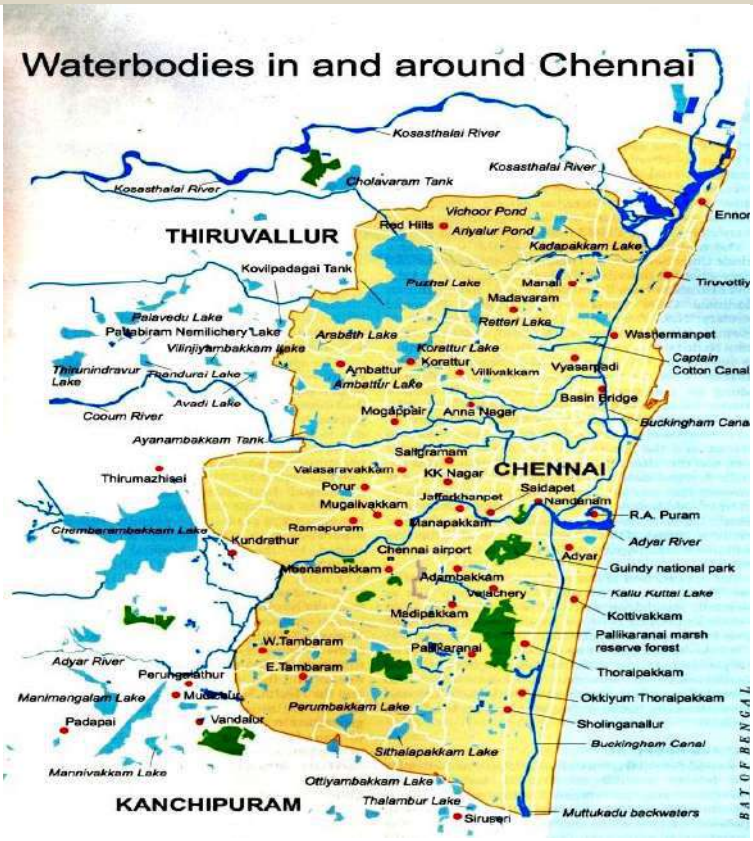

Fig. 1: Boundaries and water bodies of Chennai. (Source Maps Chennai.com)

Chennai has relied on the annual rains to replenish water reservoirs, as no major rivers flow through the area. Chennai has a water table at 2 metres for 60 percent of the year. Two major rivers flow through Chennai, the CooumRiver through thecentre and the Adyar River to the south (Fig.1). A third river, the Kortalaiyar, travels through the northern fringes of the city before draining into the Bay of Bengal, at Ennore. All the three rivers are heavily polluted by the industrial wastages and hence the water is not suitable for the use of public.

As Chennai is one among the metropolitan city in India, it depends entirely on rainfall to fulfil the daily needs of water for its people. Due to extreme precipitation during December 2015, the people in Chennai suffered for many days without food and drinking water. Many lost their lives, their belongings during the flood. Hence the knowledge of flood or drought is necessary for disaster management to take necessary precautionary actions to manage both.

This research article develops a prediction model based on Neural Network for predicting the annual rainfall in Chennai using the parameters like ElNino 3.4, Maximum Temperature, Minimum Temperature, Vapour Pressure and Cloud Cover. Table 1 gives the details of the parameters for predicting the annual rainfall in Chennai. For this purpose, a dataset containing the monthly rainfall of chennai, maximum temperature of the year, minimum temperature of the year, vapour pressure and cloud cover from the year 1901 to 2017 were obtained from India Meteorological Department, Pune, India and India Water Portal. The Sea Surface Temperature of Nino 3.4 (120°W-170°W and 5°S- 5°N) indices were obtained fromNational Oceanic and Atmospheric Administration, US for the same years.

Table 1: Chennai climate data (1981-2010)(source: India Meteorological Department)

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec	Year
Record High Temp (^{0}c)	34.4	36.7	40.6	42.8	45	43.3	41.1	40	38.9	39.4	35.4	33	45
RecordLow Temp (^{0}c)	13.9	15	16.7	20	21.1	20.6	21	20.6	20.6	16.7	15	13.9	13.9
AverageHigh Temp (^{0}c)	29.3	30.9	32.9	34.5	37.1	37	35.3	34.7	34.2	32.1	29.9	28.9	33.1
AverageLow Temp (^{0}c)	21.2	22.2	24.2	26.6	28	27.5	26.4	25.9	25.6	24.6	23.1	21.9	24.8
AverageRainfall (mm)	25.9	3.4	3.5	14.4	34.2	86.2	275.45	210.5	183.5	375.25	400.56	368.57	1981.43
Average Rainy days	1.4	0.8	0.3	0.8	1.8	4	10	11.5	10.7	17.8	21.5	19.5	100.1
Average Relative Humidity (%)	73	72	70	69	62	57	64	66	72	77	78	77	70
Mean Monthly Sunshine (Hrs)	268.3	268.1	293.6	290.2	279.9	202.6	185.2	193.6	198.6	194.6	182.7	204.3	2761.7

THE IIOAB JOURNAL
EDITED BY: Prof. Dr. S. Vaithyasubramanian

IIOAB JOURNAL
ISSN: 0976-3104

Artificial Neural Network (ANN)

An Artificial Neural Network is a computational model inspired by biological neural networks both structurally and functionally [11]. Many Neural Networks has been constructed by researchers for predicting the time series. It models the time series with various forms of activation functions by considering the underlying complex mathematical relationships between the input and output time series. Neurons are the group of interconnected computation units.

The two major types of Artificial Neural Networks are Multi-Layer Perceptron (MLP) networks and Radial Basis Function (RBF) networks which are applied by many researchers for predicting complex real life time series since they do not have any stationary constraint on the time series to be learned and predicted. The main aim of applying a neural network is to train the system to produce appropriate output patterns for the corresponding input patterns. The training algorithm has a set of training rules which varies as the algorithm varies. The learning algorithm and the neural network architecture varies according to their applications.Recent studies focus on the problem of weather forecasting using Radial Basis Function Network. The complex nature of the system can be modeled by RBF due to its nonlinear approximations. Generally RBF is applied in function approximation and time series prediction.

The design of RBF network is an approximation problem in a high dimensional space and the learning is equivalent to finding a surface in a multidimensional space that provides a best fit to the training data. The normalized Gaussian radial basis function network is applied to model the nonlineartime series. The data is fed into the hidden layer from the input layer.In the hidden layer the data is processed and transported to the output layer. The weights between the hidden layer and the output layer are modified during training. Each hidden layer neuron represents a basis function of the output space, with respect to a particular centre in the input space. The Gaussian function is used as the transform function in the hidden layer which is represented as

$$\phi_j(x) = \exp\left[\frac{-\|x - c_j\|^2}{2\rho^2}\right] \qquad (1)$$

where x is the training data, ρ is the width of the Gaussian function. This kernel is centered at the point in the input space specified by the weight vector. The closer the input signal is to the current weight vector, the higher the output of the neuron will be. Therefore the k^{th}output of the network $y_k$is represented as

$$y_k = w_0 + \sum_{1}^{M} w_{jk}\phi_j(x) \qquad (2)$$

Where$\phi_j(x)$ is the response of the j^{th} hidden unit and w_{jk} is the connecting weight between j^{th} hidden unit and the k^{th} output unit and w_0 is the bias term.

RESULTS

A time series plot (Fig.2) is used to display the time variation against the annual rainfall series. The dataset series related to Chennai was splitted into two:
1. Training Data consists of 100 years (1901 - 2000)
2. Testing Data consists of 17 years (2001 - 2017)

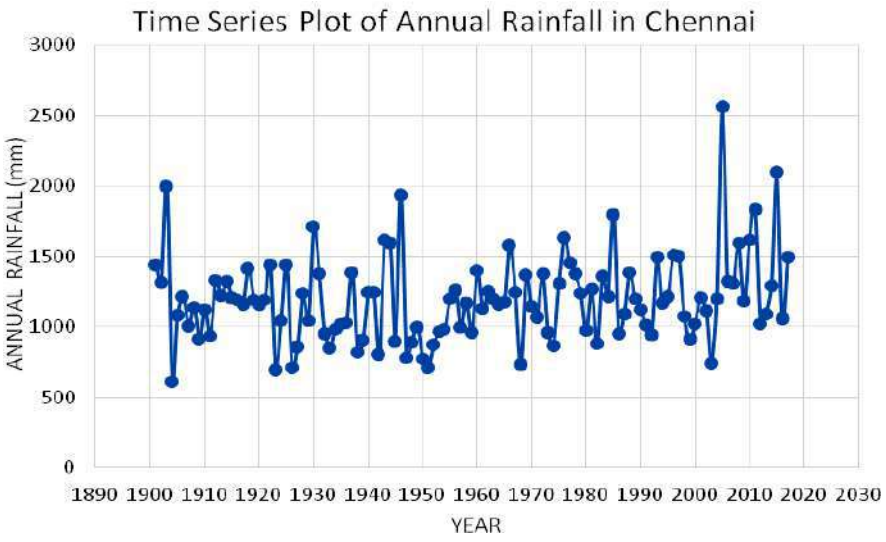

Fig. 2: Time series plot of annual rainfall in Chennai

Table 2: RBF network information

Layer	No of units	Activation function	Rescaling method
Input	6	-	Normalized
Hidden	1	Softmax	-
Output	1	Identity	Normalized

The RBF neural network architecture considered for this application was a single hidden layer with Gaussian Radial Basis Function (Table 2). Gaussian type Radial Basis Function is chosen here because it is similar with the Euclidean distance and it gives better smoothing and interpolation when compared to the other basis functions. The fully supervised learning algorithm is presented for the parametric estimation of the Radial Basis Neural Network.

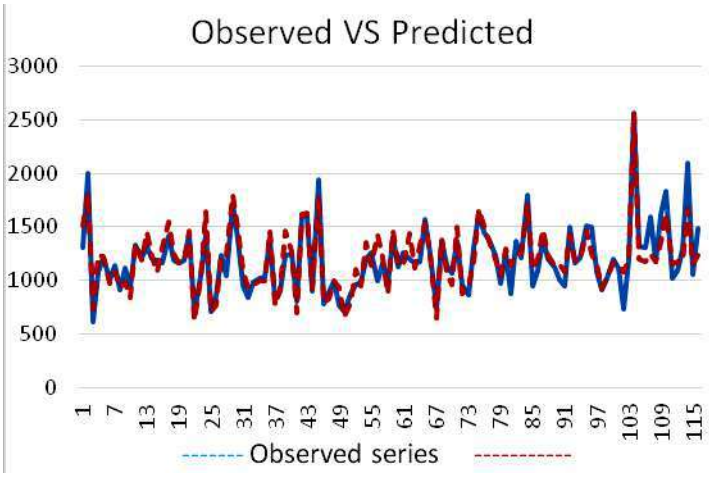

Fig. 3: Observed vs predicted annual rainfall series.

The RBF architecture consists of six units in the input layer, one unit in the hidden layer and one unit in the output layer [2]. After training, the network is tested with the test data set. For fitting a best model among the available models for a time series data, error measures are used. The Mean Average Percentage Error (MAPE) was used as forecasting accuracy measure in this analysis. The MAPE values of training data and test data are 0.0706 and 0.1266. The observed and predicted data is plotted (Fig.3) for visual inspection.

CONCLUSION

Traditional statistical time series forecasting methods, including moving average, exponential smoothing, and Auto-Regressive Moving Average, all assume stationarity and linearity of the time series. Artificial Neural Networks does not require any stationarity constraint on the time series to be learned and predicted. The highly flexible nonlinear regressive structure of Artificial Neural Network fit the target pattern space. In this research article, a prediction model based on rainfall parameters using RBF algorithm was constructed and the error measure shows the performance of the model. For some real world problems, Artificial Neural Network will never replace the existing conventional techniques but because of the fast growing applications it can be an alternative to those existing techniques.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Dariush R, Sadat H. [2017] Analysis water Quality by Artificial Neural Network in Bazoft River (Iran). J ChemPharma Res, 9(3):115-121.

[2] Dawson CW, Wilby RL. [2001] Hydrological Modeling using Artificial Neural Networks. ProgPhy Geography. 25(1):80-108.

[3] Duong TA, Thanh DD, Song PV. [2019] Improved Rainfall Prediction using combined Pre-Processing methods and Feed-Forward Neural Networks. J, 2:65–83.

[4] Fhira N, Adiwijaya, W, Izzatul UN. [2015] Planting Calendar

Forecasting System using Evolving Neural Network. Far East J. Elect Commu. 14(2):81-92.
[5] Geography of Chennai, https://en.wikipedia.org/wiki/Geography_of_Chennai
[6] Hung NQ, Babel MS, Weesakul S, Tripathi NK. [2009] An Artificial Neural Network model for rainfall forecasting in Bangkok, Thailand. Hydrol Earth SystSci, 13:1413-1425.
[7] Kavitha RB, Srinivas K, Govardhan A. [2014] Rainfall Prediction with TLBO Optimized ANN, 73:643-647.
[8] Mehnaza A. [2017] Application of ANN for the Hydrological Modeling. I J Research ApplSciEngg Tech, 5(7):203-213.
[9] Monira SS, Faisal Z, Hideo MH. [2012] A Rainfall Forecasting Method using Machine Learning Models and its Application to the Fukuoka City case. Int J Appl Math ComputSci, 22(4):841-854.
[10] Neelam M, Hemant SK, Sanjiv S, Upadhyay AK. [2018] Development and Analysis of Artificial neural Network Models for Rainfall Prediction by using Time-Series Data. I J Intelligent Systems and Applications, 1:16-23.
[11] Nirmala M. [2015] Computational Models for Forecasting Annual Rainfall in Tamilnadu. Appl Math Sci, 9(13):617-621.
[12] Yajnaseni D, Saroj KM, Bijaya K, et al. [2018] Rainfall Prediction for the Kerala state of India using Artificial Intelligence approaches. Cop and ElecEngg, 70:66 -73.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

SPECIAL ARTICLE

MATHEMATICAL APPROACH TOWARDS RECENT INNOVATION IN COMPUTATION AND ENGINEERING SYSTEM (MATRICS)

EVALUATION OF AQUIFER PARAMETER BY USING TRACER DYE

Priyadharshini B.[1*], Kavisri M.[2], Nandini K.[3]

[1,2]*Dept of Civil Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, INDIA*

[3]*Dept of Civil Engineering, Jerusalem College of Engineering Chennai, Tamil Nadu, INDIA*

ABSTRACT

Tracer dye can be used to transported through geologic media laterally with water. Three different dyes were selected for the laboratory study i.e. Rhodamine B, Sulphorhodamine, and Fluorescein. Initially column study was done to determine the adsorption of the dyes with the soil. In the batch study the fluorescein dye and sulphorhodamine shows good resistance to the adsorption of soil, but the concentration of fluorescein dye was varying with the temperature. Thiruvanchery Village, Kanchipuram district, Tamilnadu was selected for the field study and in field study Sulphorhodamine B and NACl were used as tracer. Single and two well dilution techniques had been used for the field to find out the parameters of the aquifer.16 open wells and 7 bore holes were identified. The dispersion coefficient estimated was 0.4 to 0.6 m2/day and dispersivity estimated was 13.5 to 15.5 m by single and two well dilution techniques.

KEY WORDS
Tracer, Dyes, Adsorption, Spectrophotometer, dilution techniques, dispersion coefficient

Received: 4 Oct 2019
Accepted: 24 Feb 2020
Published: 5 Mar 2020

INTRODUCTION

A tracer is a material that could be carried by the subsurface water which will give the indication about the groundwater measure and their corresponding path, the velocity of the flow in groundwater and polluted contaminants which might be transported by the water. Similarly, tracers dye support to identify the movement paths of water through a system. If enough information is collected about the geological media, then the dyes also help to determine the porosity, hydraulic conductivity, dispersivity, hydrogeological parameters and chemical distribution coefficients, [1-3]. The usage of tracers could improve our understanding of groundwater flow [4, 5]. The dye should travel through the same velocity and path as the water and should not interact with soil material and should be nontoxic [6]. The tracer would be present in a concentration of well above the background concentration of the equal component in the natural system [7][8]. Adsorption coefficient (k) by batch technique helps to explain the results of leaching studies, and given the pore water velocity, the leaching experiments can also yield k [11, 12].

The elementary theory and experimental specifics for the well tracer dilution technique are firm [14]. The single well dilution technique includes the outline of a identified quantity of tracer in a exact well and monitoring of the tracer concentration over a time period [15]. The velocity of filtration (V_f) is given by,

$$V_t = \frac{\pi r}{2\alpha t} ln \frac{C_o}{C} \quad (1)$$

The two well dilution techniques were injecting tracer in one well and monitoring in the well at downstream well by natural and forced gradient. The tracer travels in its usual gradient and not by means of pumping. The break through curve was obtained from this and analyses of the resulting tracer break through curve dispersion coefficient and dispersivity is estimated by equation [15].

$$C(r,t) = \frac{M}{4\pi BV\sqrt{\alpha_L}} e^{\left[-\frac{(r-vt)^2}{4D_L t}\right]} \quad (2)$$

The one-dimensional Dispersion equation is given below,

$$C(r,t) = \frac{\Delta M}{2Q\sqrt{\pi\alpha_L v t^3}} e^{\left[-\frac{(r-vt)^2}{4D_L t}\right]} \quad (3)$$

$$D_L = \alpha_L v \quad (4)$$

Where, DL - Longitudinal dispersion co efficient in m2/s, Q - Pumping rate of the well in m3/s, ΔM - Injected mass of tracer per unit section in kg, V - Velocity in m/s, αL- Longitudinal dispersivity in meter, r - Radial distance between two wells in meter.

***Corresponding Author**
Email:
dharspriya@gmail.com

IIOAB JOURNAL
ISSN: 0976-3104

MATERIALS AND METHODS

Dye selection

For the investigational study the three types of tracer dye were used namely Rhodamine B, Sulphorhodamine B and Fluorescein to determine the suitable dye for water study. The graph was plotted by using the known concentration's intensity value of the above tracer dyes. Standard curve for the dyes was shown [Table 1, Fig. 1].

Table 1: Intensity value for Sulphorhodamine B, Rhodamine B and Fluorescein

Concentration ppb	Fluorescent Dye SRB	Rhodamine	Fluorescein
1	0.02	0.088	0.63
2	0.034	0.156	0.711
4	0.053	0.294	1.006
6	0.075	0.456	1.116
8	0.094	0.619	1.261
10	0.116	0.736	1.902
20	0.325	1.283	2.723
40	0.557	2.049	5.518
60	0.915	2.844	6.938
80	1.119	3.639	9.456
100	1.426	4.26	11.35
200	2.123	6.845	20.56
400	3.345	7.452	32.91
600	4.523	8.125	46.38
800	5.359	9.025	54.95
1000	6.113	9.971	62.86

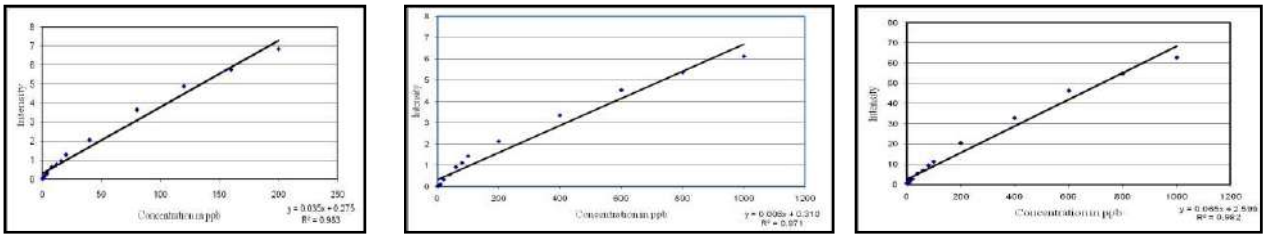

Fig. 1: Standard Graph for Rhodamine B, Sulphorhodamine and Fluorescein.

The dye was prepared at a known concentration level and transferred in to the column at different time break and the discharges of samples were collected. The tracer adsorption was measured from the discharge concentration and the proportion of adsorption was calculated from the standard graph. The column set up made for the laboratory experimental study [Fig. 2].

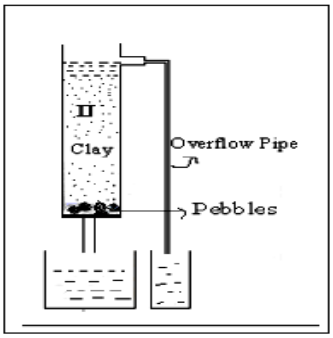

Fig. 2: Column study.

Study area

Thiruvanchery Village was selected for the field study. The field study was located in Kanchipuram District, Thambaram Taluk, Tamil Nadu, India. The study area was about 2 km2. The soil type in the study area was clayey soil and the cropping pattern followed by the farmer was paddy. Unconfined aquifer was identified

in the field. Totally 16 open wells were present in the study area and the depth of the wells of about 10 m. The water level in the well were measured for every 8 months and also it was noted that the water level was increasing during the monsoon season and in non-monsoon season and in summer season the water level had been decreased. During pre-monsoon and post monsoon season the fluctuation of the water level was about 1 m. The zone of aeration in the study area ranges from 0.5 to 1m.

RESULTS

Experimental batch study

A small investigation study was for the dye, to identify the dye loss due to sunlight, and due to minerals present in the water and also due to the impurities present in the water and soil, the sediment been varied to determine adsorbent of the dye loss. The loss of dye due to above factors were determined and shown [Fig. 3] [Fig. 4].

It shows that the loss of dye was more in surface water because of the variation in water temperature due to sunlight. Fluorescein dye was affected by the temperature and show resistance to chemical decay and adsorption of soil. When the tracer dyes were used in marine location or in contaminated water, the performance of dye would be disturbed. Smart and Laidlaw (1977) found a noticeable drop in the sum of dye loss to the adsorbing materials with increase in the dye concentration. The loss of dye during these test was shown in the [Table 1]

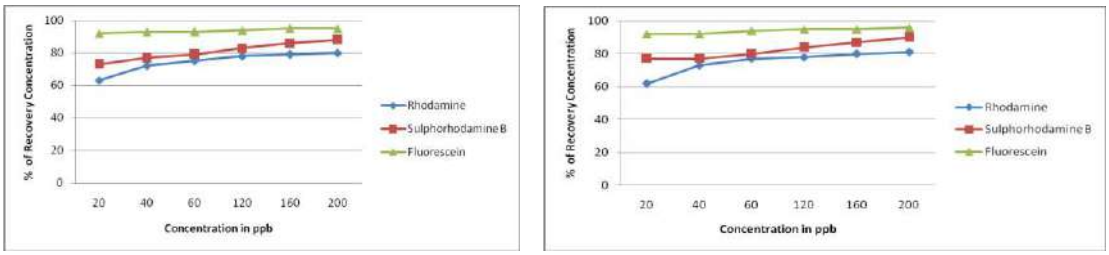

Fig. 3: Batch study for the different quality of water.

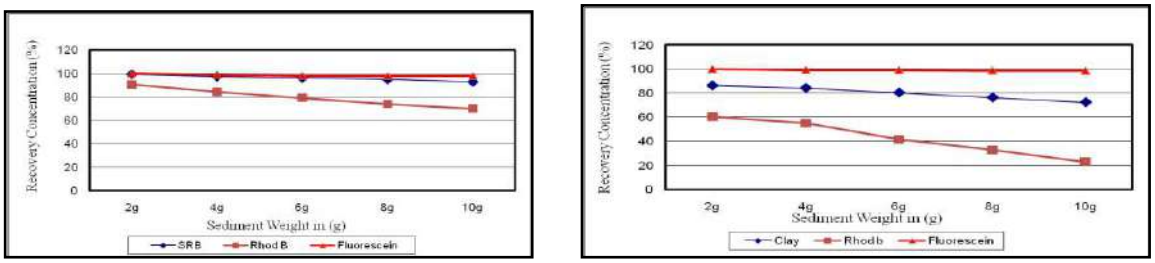

Fig. 4: Batch study for the different types of soil.

Finally, when the adsorption test was conducted by varying the sediment weight for all the three dyes, there is adsorption was very less for fluorescein dye. The result was plotted and tabulated [fig.5][Table 2]. it shows, when the sediment weight was increasing, the adsorption was more for other two dyes, fluorescein show resistance for soil adsorption.

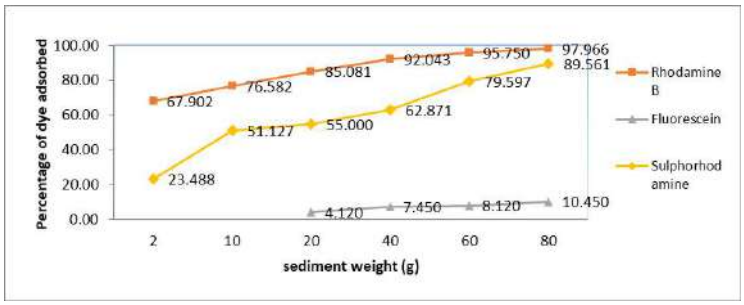

Fig. 5: Dye adsorption by varying sediment weight

Table 2: Batch study result for the dyes

Dyes	Percentage of dye loss					
	Loss due to sunlight	Chemical Decay	Contaminated water	Marine water	clay soil	sandy soil
Rhodamine B	11%	8%	20%	20%	75%	30%
Sulphorhodamine B	15%	3%	10%	11%	20%	4%
Fluorescein	20%	2%	6%	4%	4%	2%

Soil column test

For the column study, the column was initially filled with pebbles and gravel to a depth and after that clay soil was filled. The column was saturated with normal water. After saturation of soil column, the dye concentration of 200 ppb was passed through the soil. the tracer sulphorhodamine dye was tested under column test, the result showed that the dye was adsorbed to the soil. The recovery of dye had been reduced with increase duration of time period [Fig. 6].

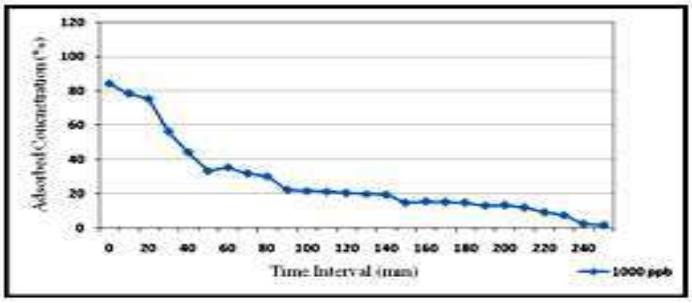

Fig. 6: Concentration Recovery of Sulphorhodamine B.

Fluorescein shows good resistance to adsorption when compare to Sulphorhodamine B [Fig. 7]. The main problem that occurred while using this dye was background fluorescence which present naturally in water. Therefore, it was mandatory to measure the background fluorescence before applying these dyes.

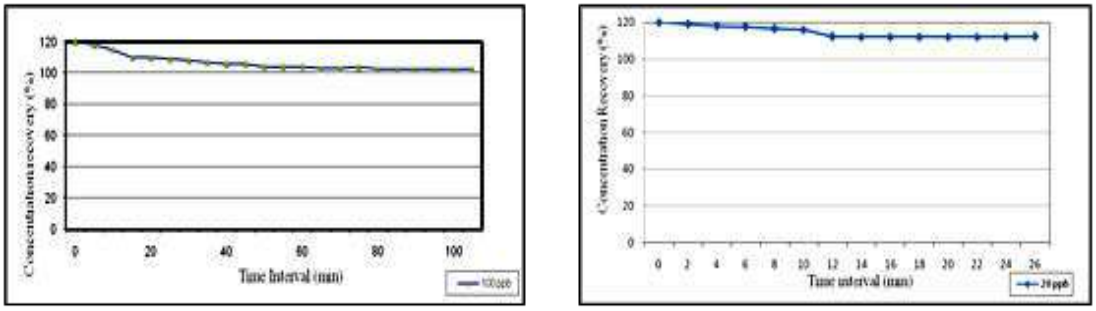

Fig. 7: Concentration Recovery of Fluorescein.

With this laboratory study the sulphorhodamine was selected for the field study, because of its easy availability and resistance to soil, where the Fluorescein was subjected to high background fluorescence[16]. Additionally, NACL was also used for a part of the field study.

Field study - well dilution technique

The well dilution technique was carried out with groundwater tracer such as organic tracer (Sulphorhodamine B) and chemical tracer (NaCl). The bore hole in the study area was plotted [Fig. 8]. The solution of 250 ppb concentration of one litre is poured instantaneously in the borehole L4. The water sample from the borehole L4 is collected from the time t=0 to the time t days until the background concentration attained. The concentration of the water samples collected from the borehole L4 is measured using the Fluorescent Spectrometer. The solution of 250 ppb concentration of 100 ml is poured instantaneously in the borehole B1. The water sample from the borehole B1 is collected from the time t=0 to the time t days until the background concentration in attained.

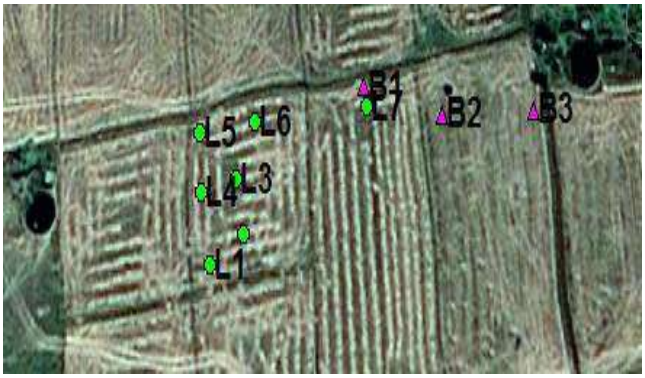

Fig. 8: Borehole Location in the Study Area.

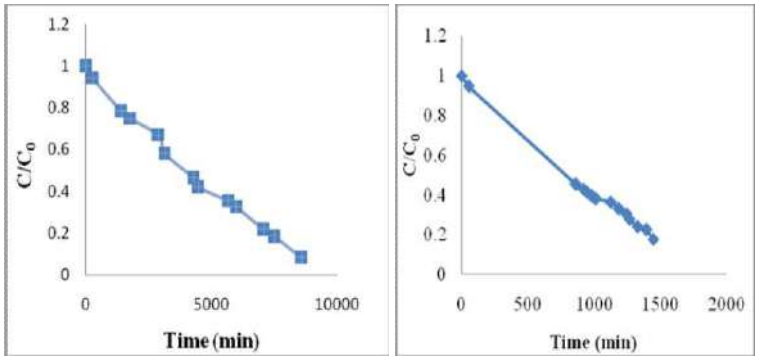

Fig. 9: Single Well Dilutions in Borehole L4 and B1 – Sulphorhodamine B.

400 gram of NaCl is diluted in the water sample from the study area and it is poured instantaneously in the borehole L1. The water samples are collected from the borehole L1 at different times. The concentration of NaCl in L1 was reduced in 10865 min. 200 gram of NaCl is diluted in the water sample from the study area and it is poured instantaneously in the borehole L6 [Fig. 9][Fig. 10].

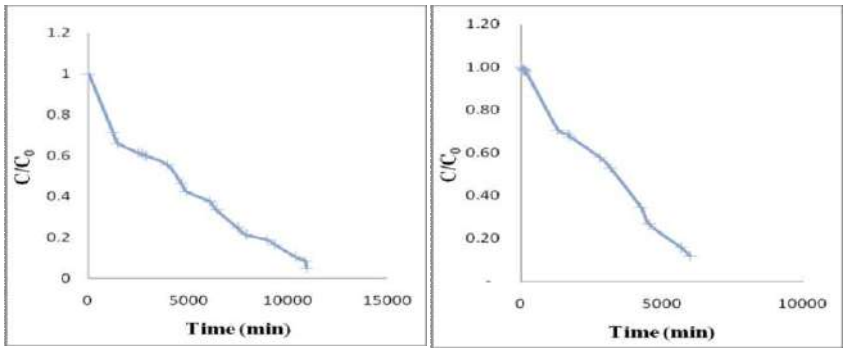

Fig. 10: Single Well Dilutions in Borehole L1 and L6 –NaCl.

From the above experiment the velocity and permeability was assessed by the given equation (1) and(2). The dispersion co efficient and dispersivity was estimated by equation (3) and (4) [Table 3] [Table 4].

Table 3: Estimated parameters by Single Well Dilution Technique

Groundwater Tracer	Sulphorhodamine B	NaCl
Filtration Velocity m/day	0.05	0.04
Permeability m/day	4.76	3.82
Dispersivity in m	11.5	10.6
Dispersion coefficient m^2/day	0.45	0.31

Table 4: Estimated parameters by Two Well Dilution Technique

Groundwater Tracer Breakthrough	Natural Gradient	Forced Gradient
Velocity m/day	10.05	108
Dispersivity in m	13.55	15.31
Dispersion coefficient m^2/day	0.43	0.61

CONCLUSION

The above study exposes that the Rhodamine B dye was seriously adsorbed to soil when compare to the other two dyes Sulphorhodamine B and Fluorescein. The amount of dye absorbance increases, if the adsorbent quantity and time increases. The batch study result shows that the fluorescein dye was affected by sunlight, therefore the change in temperature would affect the recovery dye concentration. It shows Fluorescein was greatly affected by photochemical decay. Also, it shows the good resistance to adsorption of soil. The adsorption of dye was less in clay soil, and the recovery of dye concentration was more and that was shown by soil column test. In column test the fluorescein dye recovery concentration was more when compare to the other dye. For field study the organic tracer dye sulphorhodamine and NACL was taken and used for well dilution techniques. From the single well dilution field study the filtration velocity assessed was 0.04 to 0.05 m/day, permeability estimated was 3.8 to 4.8 m/day, dispersion coefficient estimated was 0.3 to 0.45 m2/day and dispersivity estimated was 10.5 to 11.5 m. By the two well dilution studies the dispersion coefficient estimated was 0.4 to 0.6 m2/day and dispersivity estimated was 13.5 to 25m.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Shiau BJ, Sabatini DA. [1993] Influence of Rhodamine WT properties on Sorption and Transport in Subsurface Media, Journal of Ground Water, 31(5):913 – 920.

[2] David A. [2000] Sabatini, Sorption and Intraparticle Diffusion of Fluorescent Dyes with Consolidated Aquifer Media, Journal of Ground water, 38 (1):651-656.

[3] J.Fabryka-Martin D.O.Whittemore S.N.Davis P.W.Kubik P.Sharma [1991], Geochemistry of halogens in the Milk River aquifer, Alberta, Canada, 6 (4), 447 - 464

[4] Serge Brouyère, JordiBatlle,, AguilarPascal, Goderniaux, [2008], A new tracer technique for monitoring groundwater fluxes: The Finite Volume Point Dilution Method,28 , 121-140.

[5] Smart PL, Laidlaw MS, [1977] An evaluation of some Fluorescent dye for Water Tracing, Journal of Water Research, 13(3):15-33.

[6] Sumanjit, TejinderPal Singh W, Ishu K. [2008] Removal of Rhodamine B by adsorption on Walnut shell charcoal, Journal of Surface Sci Technol, 24:179-193.

[7] Torez K. [1999] Fluorescent dye and media properties affecting sorption and tracer selection, Journal of Ground water, 37:376 – 381.

[8] Priyadharshini B, Kavisri M. [2017] Behaviour Of Hrdrogeological Tracer Dyes, Rasayan journal of Chemistry, 10:1492– 1499.

[9] Priyadharshini B, Nandini K. [2015] Estimation of Aquifer Parameter By Well Dilution Technique International Journal of Applied Engineering Research, 10:11777-11786.

[10] Thomas S, Soerens, Ana G, et al. [2003] Characterizing DNAPL in Ground Water Using Partitioning Fluorescent Dyes, Ground water, 38:651 – 656.

[11] Markus F, Nu Nu W, Dyes. [2003] As Tracer for Vadose Zone Hydrology, Review of Geophysics, 41:1-31.

[12] Donald L, Et al. [1963] Fluorescent tracer for dispersion measurement, J. Sanitary Engineering Division, 89:1-22.

[13] Kumar B, Nachiappan P. [2000] 'Estimation of alluvial aquifer parameter by a single well dilution technique using isotopic and chemical tracer: a comparison' Proceedings of the conference Tracer and Modeling in Hydrology, IAHS, 53 - 56.

[14] Todd CR, Flower LA, Holmbeck SA. [2000] Regulation of Injected Groundwater Tracer Groundwater, 38:541-549.

[15] Tonder G, Riemann K, Dennis I. [2002] Interpretation of single well tracer tested using fractional-flow dimension part1: Theory and Mathematical models, Hydrogeology Journal, 10:351-356.

[16] Barnett MO, Jardine PM, Brooks SC, Selim HM. [2000] Adsorption and transport of U(VI) in subsurface media. Soil Sci Soc Am J, 64: 908–917.

[17] Bond WJ, Phillips IR. [1990 Cation exchange isotherms obtained with batch and miscible-displacement techniques. Soil Sci. Soc. Am. J. 54: 722–728.

[18] Schweich D, Sardin M, Gaudet JP. [1983] Measurement of a cation exchange isotherm from elution curves obtained in a soil column: Preliminary results. Soil Sci Soc Am J, 47: 32–37.

[19] Tomáš W, Martin S, Ji ří B. [2018] Use of sodium fluorescein dye to visualize the vaporization plane within porous media Journal of Hydrology, 565:331-340.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

SPECIAL ARTICLE

MATHEMATICAL APPROACH TOWARDS RECENT INNOVATION IN COMPUTATION AND ENGINEERING SYSTEM (MATRICS)

A CHARACTERIZATION ON GRACEFUL LABELING OF COMPLEMENT OF CAYLEY DIGRAPH

Rajeswari R.*, Manjula T., Parameswari R.

Sathyabama Institute of Science and Technology, Chennai-119, INDIA

ABSTRACT

A digraph is said to be edge graceful if there exists a bijection f: E → { 1, 2, ..., |E|} such that the induced mapping f: V → { 1, 2, ..., |V|-1} given by f*(vi) = ∑ f (eij) (mod |V|) taken over all the outgoing arcs of vi is a bijection where eij is the jth outgoing arc from the vertex vi. In this paper the graceful labeling of complement of Cayley digraph with n vertices and m generators and its line digraph were studied and obtained a characterization for the above graph to admit graceful labeling with respect to antimagic labeling .*

KEY WORDS
Cayley Digraph, Complement of a Graph, Edge Graceful labeling, Antimagic labeling

INTRODUCTION

Rosa introduced β valuation for a graph and later it was termed as graceful labeling. Various research papers were published on the topic of graceful labelings. If the communication grid is a graceful graph, we would then be able to label the connections between each centres such that each connection would have a distinct label.

Bloom and Hsu extended the concept of graceful labeling for digraphs and characterized algebraic structures based on the results. Further the relationship of graceful digraphs with other algebraic structures and the applications of graceful digraphs in networking were also discussed in Bloom and Hsu"s paper.

Being the next development, labelings on vertex symmetric digraphs were studied by Thirusangu et.al. [1]. labelings of special class of digraphs called quadratic residue digraph was studied in [2, 6, 7]. The necessary condition for a digraph to be edge graceful is q(q+1) ≡ 0 or (p/2) (mod p). We always know that the Cayley digraph has mp arcs and hence q(q+1) = mp(mp+1) ≡ 0 (mod p). Some theories of graceful labeling and magic labeling on Cayley digraphs were studied in [3][4][5]. The complement of Cayley digraph C.Cay(G,S) is a digraph on the same vertices such that two distinct vertices of C.Cay(G,S) are adjacent if and only if they are not adjacent in Cay(G,S)[8].

Line digraphs are fascinating structures in the study of dense digraphs.This paper characterizes some graceful technologies on the Complement of Cayley digraph and its line digraph.

Received: 14 Oct 2019
Accepted: 28 Feb 2020
Published: 5 Mar 2020

CASE STUDIES

Edge graceful labeling

A digraph is said to be edge graceful if there exists a bijection bijection f: E → { 1, 2, 3, ..., |E|} such that the induced mapping f*:V→{0, 1, 2, 3, ..., |V|-1} given by by $f^*(v_i) = \sum_{i=1}^{q} f(e_{ij}) (\mathrm{mod}|V|)$ taken over all the outgoing arcs of v_i, $1 \le i \le |V|$ is a bijection where e_{ij} is the jth outgoing arc of the vertex v_i.

Theorem

The Complement of Cayley digraph C.Cay(G,S) with generating set S admits Edge graceful labeling if | S | ≡ 0(mod2).

Proof: Consider the Cayley digraph with n vertices, m generators, m ≡ 0(mod2). The complement of Cayley digraph has n vertices and nq arcs where q = n - m-1. Denote the vertex set of C.Cay(G,S) as V= $\{v_1, v_2, v_3, \ldots v_n\}$ and the edge set of C.Cay(G,S) as E= $\{e_{ij}, 1 \le i \le n, 1 \le j \le q\}$. To prove that the complement of Cayley digraph admits edge graceful labeling, we have to show there exists a bijection f : E→{1, 2, 3, ..., |E|} such that the induced mapping f*:V→{0, 1, 2, 3, ..., |V|-1}given $f^*(v_i) = \sum_{i=1}^{q} f(e_{ij}) (\mathrm{mod}|V|)$) is a bijection.

Define f : E →{ 1, 2, 3, ..., |E|} as

***Corresponding Author**
Email:
rajeswarivel1998@gmail.com
Tel.: 9444126406

$$f(e_{ij}) = \begin{cases} (j-1)n+i & \text{if } j \text{ is odd} \\ jn+1-i & \text{if } j \text{ is even} \end{cases}$$

Then the induced function

$f^*(v_i) = \sum_{j=1}^{q} f(e_{ij})$

$= i + (q-1)/2 + (n(q^2-1))/2$

$= i + (q-1)/2 \pmod n$

Since q is constant, $f^*(v_i)$ is distinct for every i, $1 \leq i \leq n$. Hence the C.Cay (G,S) is edge graceful if $|S| \equiv 0 \pmod 2$.

Corollary

The line digraph of Complement of Cayley digraph C.Cay (G,S) is edge graceful if $|S| \equiv 0 \pmod 2$.

Proof

Consider the C.Cay (G,S) with $|S| = m$. It has n vertices and nq arcs, where q = n-m-1. Then by the definition of the line digraph, line digraph of Complement of Cayley digraph is again a q regular digraph where q is odd. That is every vertex of the line digraph has q incoming and q outgoing arcs. If the digraph of a group contains n vertices and nq edges, then the corresponding line digraph contains nq vertices and nq^2 arcs. Let us denote the vertex set of L(C.Cay(G,S)) as $V = \{ v_1, v_2, v_3, \ldots, v_{nq} \}$ and the edge set of C.Cay(G,S) as E = { eij, $1 \leq i \leq nq$, $1 \leq j \leq q$}.Then by the above theorem, if we define the labels for every q outgoing arcs of each vertex, we will get the edge graceful labeling of the line digraph of complement of Cayley digraph.

Antimagic labeling

A graph with q edges is called antimagic if its edges can be labeled with the set {1,2,3,...,q} such that the sums of the labels of the edges incident to each vertex are distinct.

Characterization

The complement of Cayley digraph C.Cay(G,S) with $|S| \equiv 0 \pmod 2$ admits edge graceful iff it is an antimagic.

Proof

From the above theorem we know that the regular digraph with odd number of outgoing arcs is edge graceful. Now we have to prove that it is antimagic too. Define f and f* as in the above theorem.
We have,

$$f^*(v_i) = \sum_{j=1}^{q} f(e_{ij})$$

$= i + (q-1)/2 + (n(q^2-1))/2$

Since n and q are constants, $f^*(v_i)$ is distinct for every i, $1 \leq i \leq n$. Hence the edge graceful C.Cayley digraph is always anti magic and vice versa.

Corollary

A graph G is called bi-edge - graceful if both G and its line graph L(G) are edge - graceful. Therefore the Complement of Cayley digraph C. Cay (G,S) with $|S| \equiv 0 \pmod 2$ is bi - edge – graceful.

Corollary

Every cayley digraph admits edge graceful labeling only when $|S| \equiv 1 \pmod 2$ but its complement is edge graceful only when $|S| \equiv 0 \pmod 2$.

IIOAB JOURNAL
ISSN: 0976-3104

Strong edge graceful labeling

Here we define a labeling called strong edge graceful labeling for digraphs by extending its range through which we can get edge graceful labeling for some family of digraphs which is not in the former case. We investigate the strong edge graceful labeling for the Complement of Cayley digraph and its line digraph.
A digraph is said to be strong edge graceful if there is an injection
f : E →{1,2,3,...,3 |E|/2} in such a way that the induced mapping $f^*(v_i) = \sum_{i=1}^{q} f(e_{ij}) \pmod{2|V|}$ taken over all the outgoing arcs of v_i is an injection, where e_{ij} is the j^{th} outgoing arc of the vertex v_i.

Theorem

The vertex symmetric digraph, complement of Cayley digraph C.Cay(G,S) admits strong edge graceful labeling.

Proof

Let C.Cay (G,S) be the complement of Cayley digraph with n vertices, nq arcs. Denote the vertex set of C. Cay(G,S) as V= $\{v_1, v_2, v_3, \ldots, v_n\}$ and the edge set of C.Cay(G,S) as E= $\{e_{ij}, 1 \le i \le n, 1 \le j \le q\}$ .To prove that the complement of Cayley digraph admits strong edge graceful labelling, we have to show that there exists an injection f : E →{1,2,3,...,3 |E|/2} such that the induced function $f^*(v_i) = \sum_{i=1}^{q} f(e_{ij}) \pmod{2|V|}$ is an injection. We prove this theorem in two cases.

Case(i):

When q ≡ 1(mod2)

$$f(e_{ij}) = \begin{cases} (j-1)n+i & \text{if } j \text{ is odd} \\ jn+1-i & \text{if } j \text{ is even} \end{cases}$$

Then the induced function

$$f^*(v_i) = \sum_{j=1}^{q} f(e_{ij})$$

$$= i + \frac{q-1}{2} + \frac{2n(q^2-1)}{4}$$

$$= i + \frac{q-1}{2} \pmod{n}$$

Since q is constant, f*(v_i) is distinct for every i, 1≤ i ≤ n. Hence the Complement of Cayley digraph C.Cay(G,S) is strong edge graceful if q ≡ 1(mod2).

Case(ii): When q ≡ 0(mod 2)

Define f : E →{1,2,3,...,3 |E|/2} as

$f(e_{ij})$ = (j-1)n+i if j=1,3,...(q-1)and q
jn+1-i if j=2,4,...,(q-2)

Then the induced function

$$f^*(v_i) = \sum_{i=1}^{q} f(e_{ij})$$

= 2 i +(q-2)/2 +(q-1)n (mod 2n)

Since q is constant, f*(v_i) is distinct for every i, 1≤ i ≤ n. Hence the Complement of Cayley digraph C.Cay (G,S) is strong edge graceful.

Corollary

The line digraph of Complement of Cayley digraph C.Cay (G,S) is strong edge graceful.

Proof

Consider the C.Cay (G, S) with | S | = m generators. It has n vertices and nq arcs, where q = n-m-1. Then by the definition of the line digraph, line digraph of Complement of Cayley digraph is again a q regular digraph.

That is every vertex of the line digraph has q incoming and q outgoing arcs. If the digraph of a group contains n vertices and nq edges, then the corresponding line digraph contains nq vertices and nq2 arcs. Let us denote the vertex set of L(C.Cay(G,S)) as $V = \{v_1, v_2, v_3, \ldots, v_q\}$ and the edge set of C.Cay(G,S) as $E = \{e_{ij}, 1 \le i \le n, 1 \le j \le q\}$. Then by the above theorem, if we define the labels for every q outgoing arcs of each vertex, we will get the edge graceful labeling of the line digraph of complement of Cayley digraph.

Proposition

Every edge graceful regular digraph is strongly edge graceful. But the converse need not be true.

CONCLUSSIONS

We have investigated graceful labelings on Complement of cayley digraphs and its line digraphs. We noticed an important upshot in the above process that the number of generators of Cayley digraph plays an important role in determining the existence of labelings of the above digraph. So this offers new horizons for future work that one can examine and analyze the various other labeling on the same architectures of cayley digraphs.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Thirusangu K, Atulya K, Nagar R, Rajeswari R. [2011] Labeling of Cayley digraphs, European Journal of Combinatorics Science. 32(1):133- 139.

[2] Parameswari R, Rajeswari R. [2014] Labeling of Paley Digraphs in International Electronic, Journal of Pure and Applied Mathematics, 7(3):127-135.

[3] Thamizharasi R, Rajeswari R. [2015] Labelings of Cayley Digraphs and its Line Digraphs, International Journal of Pure and Applied Mathematics, 101(5):681-690.

[4] Thamizharasi R, Rajeswari R. [2015] Graceful and Magic Labelings on Cayley Digraphs, International Journal of Mathematical Analysis, 9(19):947-954.

[5] Thamizharasi R, Rajeswari R. [2016] Labeling on Line Digraphs, Indian Journal of Science and Technology, 9(44): 1-5.

[6] Parameswari R, Rajeswari R. [2016] Labeling of Quadratic Residue Digraphs over Finite Field. Smart Innovation Systems and Technologies, 50(1): 387 – 396.

[7] Parameswari R, Rajeswari R. [2017] On Cordial and Anti Magicness of Quadratic Residue Digraph. Journal of Computational and Theoretical Nanoscience, 14:4553-4555,

[8] Rajeswari R, Udhayashree R, Nirmala M. [2019] Domination of Cayley digraph and its complement. International Journal of Recent Technology and Engineering, 8(2S11): 4005-4008.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

A NOVEL DESIGN OF 1-BIT COMPARATOR USING QUANTUM-DOT CELLULAR AUTOMATA

Kshitija Save, Siddhi Gudhekar, Debashrita Panicker, Sankit R. Kassa*

Dept. of Electronic Engineering, Usha Mittal Institute of Technology, SNDT Women's University Mumbai, INDIA

ABSTRACT

Background: *Quantum-dot Cellular Automata (QCA) is a transistor-less computation approach introduced as a renewal key to the fundamental limits faced by CMOS technology ensuring small size, high speed operation, high integration density, capacity and ultra-low power consumption. Quantum Dot cell is an artificial nano-scale molecule which does not require current flow to pass the information. Any digital circuit can be fully realized using QCA cells. The comparator is a vital digital circuit required to perform number of arithmetic operations in applications, like a PC.* ***Methods:*** *Comparison of the proposed approach with the previous approaches is realized, evaluated and verified by utilizing QCA Designer tool, a simulation tool for QCA circuits, Version 2.0.3.* ***Results:*** *In this paper, a novel design of a comparator in Quantum Dot Cellular Automata (QCA) technique has been proposed which is a promising design significantly declined in terms of effective area, latency and cell complexity, compared to other layouts, and its clock cycle is confined to bare minimum.* ***Conclusions:*** *Simulation of the proposed 1-bit comparator shows the advantage of the proposed approach over the previous approaches as it possesses approximately 73% reduction in total area and 74% reduction in total number of cell counts.*

KEY WORDS
QCADesigner tool version 2.0.3, Quantum-dot Cellular Automata, comparator; Quantum-dot Cell

INTRODUCTION

In 1965, the Gordon Moore law predicted that the capacity of computer chips would be doubled in approximately every 18 months. Till now, it has governed the development and performance of microprocessors very well [1]. The scaling of CMOS Technology has almost reached its physical limit, which has led to an extensive research for developing future generation ICs and discovery of new technologies. Quantum dot Cellular Automata is one of the emerging technologies at nano-scale level that has overcome the barrier of scaling. Lent et al proposed a physical implementation of an automaton using quantum-dot cells in 1993 [1]. It was first fabricated in the year 1997 [2]. Among other nano-scale level techniques, QCA has proved to be a better alternative because of its attractive features such as high-speed operation, low power consumption and small dimension. QCA does not use transistors [2]. Basically, QCA are array or wire of cells, which store information using electrons [3]. A digital comparator takes two numbers as inputs in binary form and determines whether one number is greater than, less than or equal to the other number [4]. Comparators are extensively used in central processing units and microcontrollers and thus have been researched upon and optimized in CMOS technology [4]. This paper deals with QCA based comparator designs as we have compared our proposed design with previous layouts.

Received: 24 Jan 2020
Accepted: 2 Mar 2020
Published: 10 Mar 2020

A 1-bit magnitude comparator design consisting of 73 cells and 0.06μm2 area was proposed in [4]. Another design of a reversible 1-bit comparator without wire crossing was proposed in [5]. It is constructed by using the Single Feynman and TR circuit and consists of 117 cells and 0.182μm2 area. In the paper [6], a 1-bit comparator using the most recent EX-OR entryway has been introduced. The QCA implementation of the proposed design utilizes 180° clock phase shift wire crossing and consists of 60 cells. An optimized 1-bit comparator design having 50% enhancement in the cell count and 59% enhancement in occupied area is presented in [7]. An efficient layout entailing the lowest cell count, area and clock cycle compared to the previous works is introduced in [8].

MATERIALS AND METHODS

QCA BASICS

QCA cell

A QCA cell is made up of four quantum dots with two electrons occupying opposite location in two different quantum dots. The electrons occupy opposite or diagonal corners as they are governed by the principle of Coulomb's repulsive force. [Fig. 1(a)] shows QCA cell in which binary 1 and binary 0 are indicated by polarity +1 and -1 respectively.

***Corresponding Author**
Email:
rel1356@mnnit.ac.in

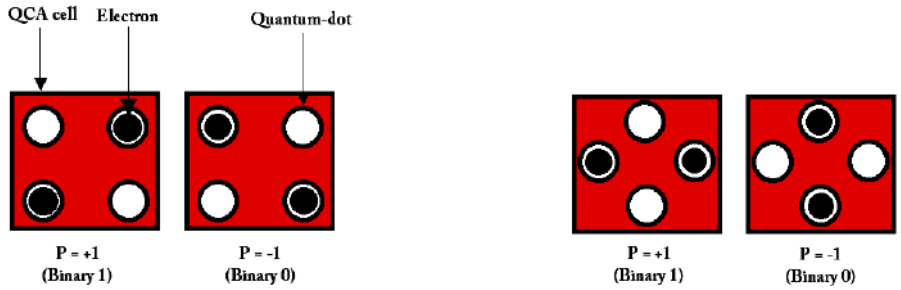

Fig. 1(a): QCA Cells

QCA Logic gate

The QCA Inverter and QCA Majority gate are the fundamental elements of any QCA circuit. A QCA Inverter is made up of minimum 2 QCA Cells in which one is an input cell and the other is output cell. On the other hand, a QCA Majority gate is made up of five QCA cells consisting of three input cells, a decision cell and an output cell. As a result of Coulomb's repulsive force in this structure, stable state of the output is dependant on the input. Hence electrons are placed at maximum distance.

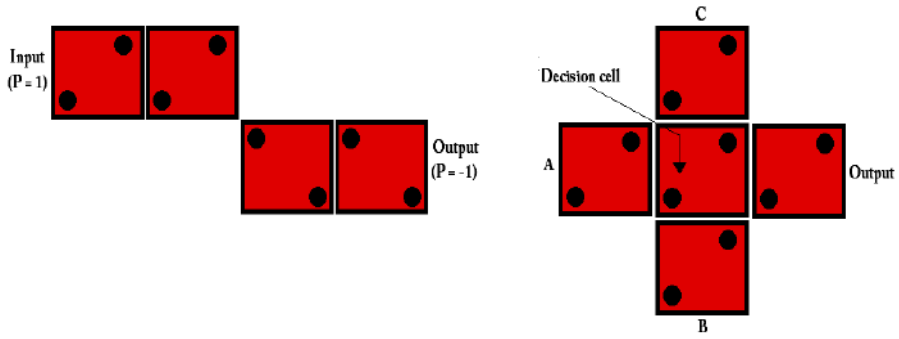

Fig. 1(b): QCA Inverter and Majority gate

The QCA Inverter and QCA Majority gate are the fundamental elements of any QCA circuit. A QCA Inverter is made up of minimum 2 QCA Cells in which one is an input cell and the other is output cell. On the other hand, a QCA Majority gate is made up of five QCA cells consisting of three input cells, a decision cell and an output cell. As a result of Coulomb's repulsive force in this structure, stable state of the output is dependant on the input. Hence electrons are placed at maximum distance.

Wire crossing

The simplest arrangement of cells is given by placing quantum-dot cells in series. [Fig. 1(c)] shows such arrangement of four quantum-dot cells. The bounding boxes in the figure do not represent physical implementation, but it is shown to identify individual cells. If polarization of any cells in the arrangement shown in [Fig. 1(c-a)] is changed then rest of the cells are immediately synchronized to the new polarization due to coulombic interactions between them. There are two types of wires possible in QCA. A simple binary wire shown in [Fig. 1(c-a)] and an inverter chain, which has 45-degree inverted QCA cells side by side shown in [Fig. 1(c-b)].

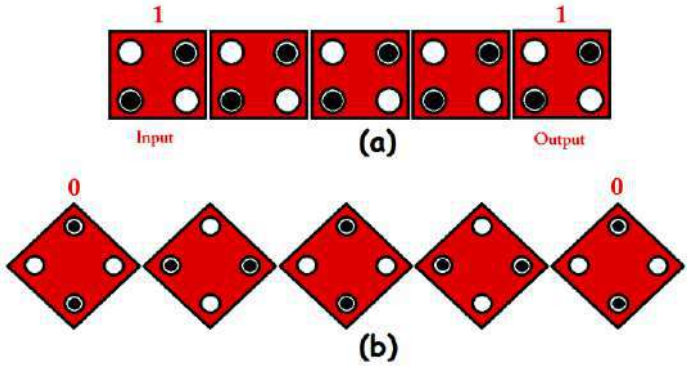

Fig. 1(c): (a) Series arrangement of QCA cells. **(b)** Series arrangement of 45-degree inverted QCA cells.

IIOAB JOURNAL
ISSN: 0976-3104

QCA clocking

QCA clocking is used to determine data flow direction and also provides power to the circuit. It has four zones depending on the phases. There is a phase shift of 90° in each clocking zone. As shown in [Fig. 1(d)] there is four phases namely switch, hold, release and relax. In switch phase, electrons present in the quantum dots contain minimum energy after which the clock signal amplitude increases, potential energy of electrons begin to rise and finally the electrons gain the highest potential energy leading to the end of this phase. In the hold phase, also known as the high phase, electrons become effectively energized to exceed the tunneling barrier and the cells obtain null phase. In other words, the barriers are held at high value and the cell is now acting as an input to the next stage [9] .In the release phase, the high to low phase, actual computation is performed and electrons start to dissipate potential energy. [10] In the relax phase, the low phase, electrons bear minimum energy and are confined into the quantum dots. [10]

In QCA cells, having different colours means that they are under different clocks and having same colour means that they are under same clock. In QCA, Green refers to clock 0, Violet refers to clock 1, Blue refers to clock 2 and White refers to clock 3. [11]

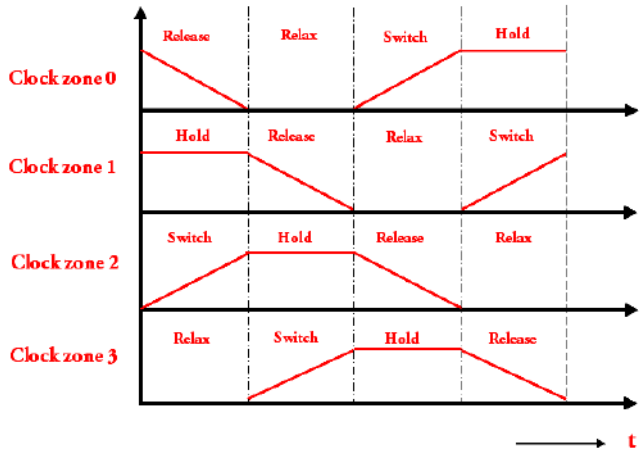

Fig. 1(d): Clocking in QCA

RESULTS

The proposed comparator design using QCA technology has two 1-bit inputs and three 1-bit outputs.

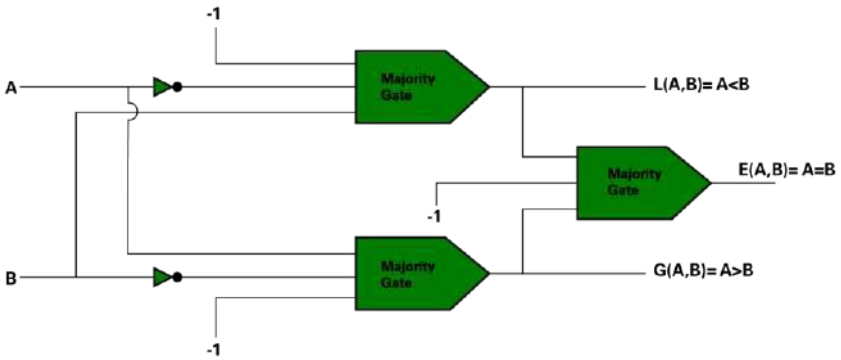

Fig. 2(a): Proposed 1-bit QCA comparator circuit

The inputs are indicated by A and B, and the outputs are indicated by L (A, B), E (A, B), and G (A, B). The relation between outputs and inputs are defined as follows.

L (A, B) =$\bar{A}$. B where A<B
E (A, B) = A$\bar{\odot}$B where A=B
G (A, B) = A. $\bar{B}$ where A>B

From the above equations it is observed that when input A is less than the input B, the output L (A, B) is "1" and it is "0" for E (A, B) and G (A, B). But if input A is greater than the input B, the output G (A, B) is "1" and other outputs remain "0". When both the inputs A and B are equal, only the output E (A, B) is "1". [8]

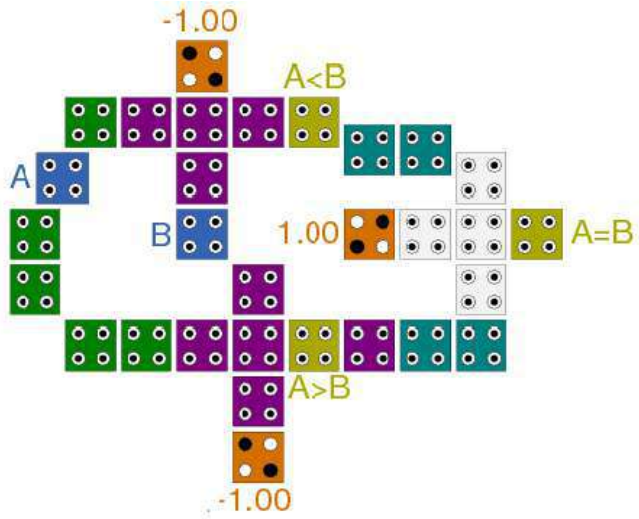

Fig. 2(b): Proposed 1-bit QCA comparator circuit

The layout design of proposed 1-bit comparator as shown in [Fig. 2(b)] consists of two inverters and three majority gates. The majority gates are used here for implementation of AND gates [8]. This layout of 1-bit QCA comparator circuit requires 30 number of QCA cells.

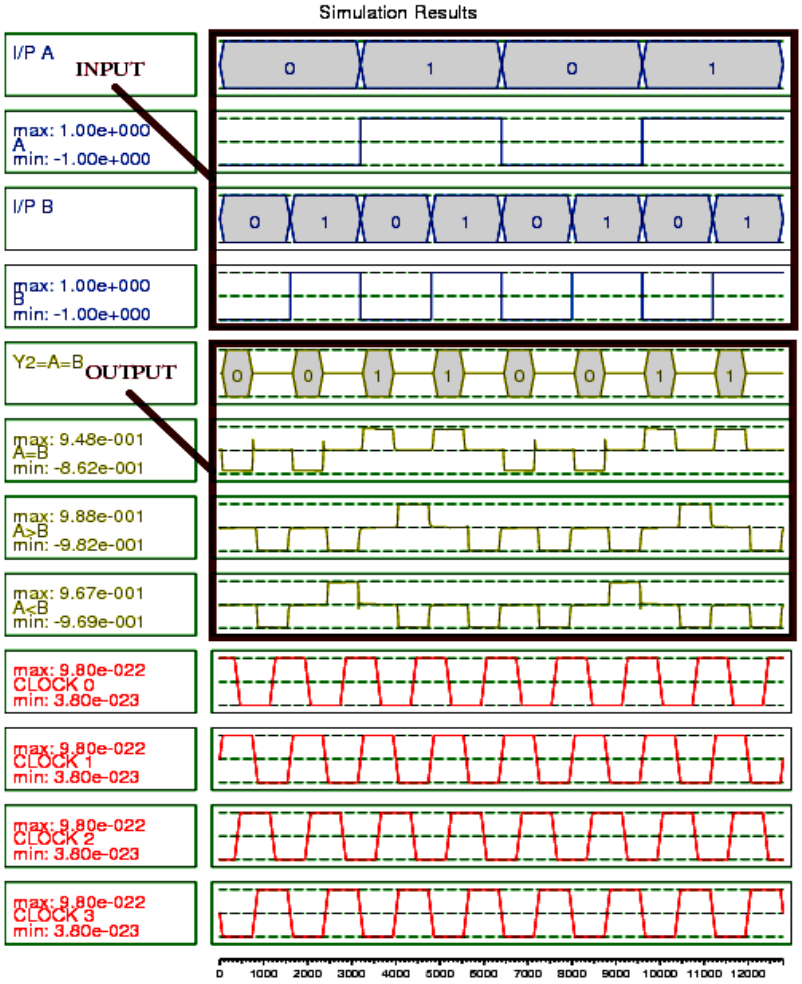

Fig. 2(c): Simulation for proposed QCA comparator

The outputs of the proposed 1-bit comparator circuit are correctly obtained after 1 clock cycle delay. It requires 0.05 µm2 area and 30 QCA cells.

DISCUSSION

Table 1 shows the comparison of 1-bit Comparator with other comparators.

Various QCA comparator circuits mentioned in [Table 1] have shown good performance but improvements have been entailed in the proposed comparator.

Table 1: Comparison table for 1-bit comparators

Sr no.	Reported designs	Number of QCA Cells	Area (μm^2)	Time Delay (Clock Cycle)	Cost (Area× Clock cycles)
1	[1]	73	0.06	1	0.06
2	[4]	117	0.182	1	0.182
3	[3]	60	0.10	0.5	0.05
4	Proposed design	30	0.05	1	0.05

Area and delay are shown in terms μm2 and clock cycle respectively. To determine the cost, following equation can used [8]

Cost=Area × Delay

The proposed design of 1-bit Comparator has major advantages in terms of cost and area as compared to other designs [8].

Following is the graph denoting cell count, area occupied and cost of various comparators in comparison with the proposed layout.

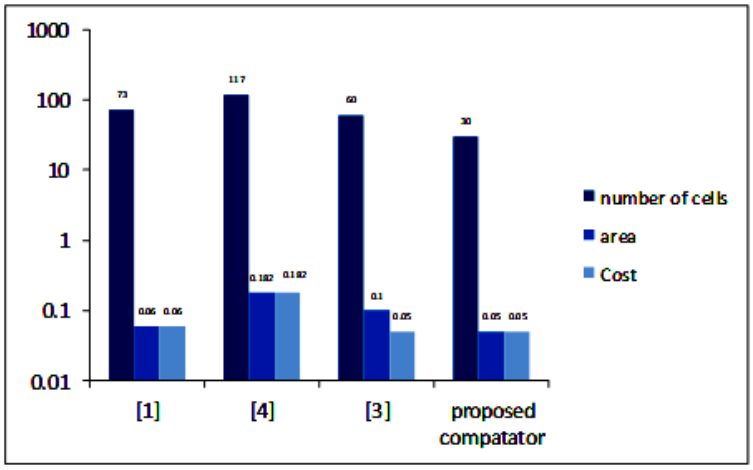

Fig. 3: Comparative analysis for number of cells, area and cost

The proposed comparator is seen to have lesser cell complexity, lesser area and lower cost compared to the previous designs.

CONCLUSION

In this paper, a new design of 1-bit comparator has been proposed, which is robust and efficient than the previous designs. It surpasses its counterparts in terms of area as well as complexity. The 1-bit QCA comparator circuit is carefully constructed using Majority gates, XNOR gates and Inverter gate. The functionality and efficiency of the proposed designs has been verified using QCA Designer 2.0.3 simulation tool. The obtained results indicate that the designed 1-bit comparator circuit requires 0.05 μm2 area, which is 73% less in total area compared to the previous designs and 74% less cell count compared to the previous designs, as only 30 QCA cells have been utilized. It has 1 clock cycle delay. In future, this design can be implemented in several calculative applications, which may perform a vital function of a general-purpose nano-processor as well as image processing applications.

IIOAB JOURNAL
ISSN: 0976-3104

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to our Principal Dr. Sanjay Pawar and our Head of Department Dr. Shikha Nema for their kind co-operation and encouragement which helped us in the completion of this project in timely manner.

FINANCIAL DISCLOSURE

No finance was provided for this study

REFERENCES

[1] Rafiq Beigh SM, Mustafa M, Ahmad F. [2013] Performance Evaluation of Efficient XOR Structures in Quantum-Dot Cellular Automata (QCA), Circuits and Systems, 4: 147-156.
[2] Sankit R. Kassa R, Nagaria K, Karthik R. [2018] Energy efficient neoteric design of a 3-input Majority Gate with its implementation and physical proof in Quantum dot Cellular Automata, Nano Communication Networks, 15(1): 28–40.
[3] Sankit R, Kassa R, Nagaria K. [2016] An Innovative Low Power Full Adder Design in Nano Technology Based Quantum Dot Cellular Automata, Journal of Low Power Electronics, 12(2):107-111.
[4] Bahniman G, Shoubhik G, Smriti K. [2012] Quantum Dot Cellular Automata Magnitude Comparators, 2012 IEEE International Conference on Electron Devices and Solid State Circuit (EDSSC), 1-2.
[5] Abdullah-Al-Shafi M, Bahar AN. [2016] Optimized design and performance analysis of novel comparator and full adder in nanoscale. Cogent Engineering, 3: 1237864.
[6] AMallaiah A, Swamy GN, Padmapriya K. [2017] 1-bit and 2-bit comparator designs and analysis for quantum-dot cellular automata, Nanosystems: Physics, Chemistry, Mathematics, 8(6): 709–716.
[7] Krishna R, Asaduzzaman Md, Bhuiyan MMR, Bahar AN. [2017] Design and Implementation of 1-bit Comparator in Quantum-dot Cellular Automata (QCA), Cumhuriyet Universitesi Fen Fakultesi Fen Bilimleri Dergisi (CFD), 38(1): 146-152.
[8] Ahmadreza S, Abdalhossein R, Hamid M. [2019] Design of efficient coplanar 1-bit comparator Circuit in QCA Technology, Facta Universitatis Series: Electronics and Energetics, 32(1): 119-128.
[9] Usha M, Vaishali D. [2017] Quantum-dot Cellular Automata (QCA): A Survey, Institute of Technology, Nirma University, Ahmedabad, India.
[10] Angona S, Alam MMB, Bahar AN. [2017] Design of 1-bit Comparator using 2 Dot 1 Electron Quantum-Dot Cellular Automata, International Journal of Advanced Computer Science and Application, 8(3): 481-485.
[11] Ratna C, Sreyashi D, Rajib G, Maitreyee RM, Sagar SR, Pallabi M. [2017] Nano-Calculator Using Quantum Dot Cellular Automata (QCA), Institute of Engineering & Management, Kolkata, 978-1-5386-1703-8/17/$31.00 IEEE.
[12] Karthik R, Jyothi K. [2019] Novel Design of Full Adders using QCA Approach, International Journal of Advanced Trends in Computer Science and Engineering, 8(3):501-506.
[13] Aishwarya T, Snehal B, Sankit K. [2018] Design and analysis of (2x1) and (4x1) Multiplexer circuit in Quantum dot Cellular Automata approach, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(6S3): 277-281.
[14] Sankit R, Kassa, R, Nagaria K. [2016] A novel design of quantum dot cellular automata 5-input majority gate with some physical proofs, Journal of Computational Electronics, Springer Publication, 15(1): 324-334.
[15] Sankit R, Kassa R, Nagaria K. A Novel Design for 4-Bit Code Converters in Quantum Dot Cellular Automata, Journal of Low Power Electronics, 13(3): 482-489.
[16] Karthik R, Sankit RK. [2018] Implementation of flip flops using QCA tool, Journal of Fundamental and Applied Sciences, 10(6S): 2332-2341.
[17] Walus K, Dysart T, Jullien G, Budiman R. [2004] QCADesigner: A rapid design and simulation tool for quantum-dot cellular automata, IEEE Trans. Nano, 3(1): 26-29.
[18] Gin S. Williams H. Meng P, Tougaw D. [1999] Hierarchical design of quantum-dot cellular automata devices, Applied Physics, 85(7): 3713-3720.
[19] Amlani A, Orlov G, Toth GH. Bernstein C, Lent S, Snider GL. [1999] Digital Logic Gate Using Quantum-dot Cellular Automata, Science, 284(5412): 289-291.
[20] Tougaw PD, Lent CS. [1994] Logical devices implemented using quantum cellular automata, Applied Physics, 75(3): 1818-1825.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

HYBRID METAHEURISTIC ALGORITHM TUNED BACK PROPAGATION NEURAL NETWORK FOR INTRUSION DETECTION IN CLOUD ENVIRONMENT

Ayyappan Thirumalairaj[1*], Mohan Jeyakarthic[2]

[1]*Dept. Of Computer Science, Kunthavai Naacchiyaar Govt. Arts College for Women, Thanjavur, Tamil Nadu, INDIA*

[2]*Tamil Virtual Academy, Kottur, Chennai, Tamil Nadu, INDIA*

ABSTRACT

Background: *Cloud computing (CC) being the representation of the technology utilizes the infrastructure for computing in a proficient way. This kind of computing provides massive amount of significance in enhancing the productivity that minimizes the cost and verifies the risk handling management. Intrusion detection systems (IDS) are commonly employed to detect malicious activities in the network of communication and also its host.* ***Methods:*** *This paper presents new clustering based hybrid metaheuristic algorithm tuned back propagation neural network (BPNN) based IDS, called C-HMT-BPNN for effective identification of intrusion in the network. The proposed model involves two stages namely threshold based K-means clustering technique and back propagation neural network (BPNN) based classification. To optimize the weights and biases of BPNN, Particle Swarm Optimization (PSO) with Gravitational Search Algorithm (GSA) called PSO-GSA model has been developed.* ***Results:*** *The proposed model has been tested using a set of two IDS dataset namely, NSL-KDD 2015 and CICIDS 2017 dataset. The obtained experimental outcome clearly ensured the superior characteristics of the proposed model over the compared methods in a significant way.* ***Conclusions:*** *The proposed HMT-BPNN model resulted to a maximum accuracy of 99.51% which is increased to 99.75% by the inclusion of clustering techniques.*

KEY WORDS
Clustering, Cloud computing, Metaheuristic, Intrusion detection

Received: 1 Mar 2020
Accepted: 24 Mar 2020
Published: 30 Mar 2020

***Corresponding Author**
Email:
a.thirumalairaj@gmail.com

INTRODUCTION

At present times, cloud computing (CC) becomes familiar and offers on-demand computing services ranging from applications to storage and processing power using the internet and on a pay-as-you-go basis. [1]. The cloud satisfies the requirement of users for trustworthy access to data and the corresponding resources. More number of businesses has been applying CC through the features of on-demand self-service, broad network application, resource pooling, quick elasticity and valid facilities. These metrics enables several clients to concentrate on business processes and controlling the computational resources using Cloud Service Provider (CSP) [2]. This cloud method tends to minimize the business cost with respect to simple installation of hardware and its procedures with software as well as the hardware updates by ensuring suitability and accessibility of diverse computational resources [3]. Generally, the deployment of CC takes place in three ways, namely Private cloud, Public cloud and Hybrid cloud. Public cloud can be used by common people and owned by a private or academic, government or combined organizations. This private cloud infrastructure has been maintained by a broad as well as individual organization. The Hybrid cloud is referred as an appropriate integration of different architectures which might be private, public which becomes as an individual cloud [4].

In recent times [5], Google, Amazon and the Salesforce.com are the important providers of cloud services (CS) and expand the facilities of storing the applications as well as processing enhancement for every year. The data non-availability of services and applications have been induced interms of denial of service (DOS) or distributed DoS (DDOS) where the attackers use CS [6]. The IDS becomes a required unit of defensive metrics which is capable of protecting the systems from a dangerous attack. Furthermore, safeguarding a system is considered to be the most important potion of CC platform. The main theme of IDS is to find and response to the events of intrusion as emerged from selfish nodes [7]. The IDS is defined as a model applied for detecting and responding the intrusion events. Also, it is referred as a method applied for predicting the intrusions of a network. IDS is a process of detecting actions which happens in a network and tries to satisfy the confidentiality, security or network accessibility to apply the trust procedures [8, 9]. Misuse prediction relied method is assumed to be the analysis of intrusions that exhibits forecasting intrusions uses an effective intrusion strategy [10]. These models are more efficient in detecting predefined attacks. Alternatively, the anomaly based prediction shows the performance conducted by an investigation of modified patterns in a system. These modifications are applied for detecting the difference in patterns of predetermined and unknown attacks. Furthermore, the abnormal nature can be identified. Also [11], IDS has 2 features which are based on the host network. The IDS is found at defense system that helps to monitor the harsh events existing in a system. CC has 2 various modules that are knowledge-based IDS as well as behavior-based IDS to analyze the IDS of CC platform.

Several types of meta-heuristic frameworks are used in solving the issues related with scheduling. [12] deployed a model in-depth analysis of PSO and the application of workflow scheduling techniques has presented for CC environment in this study. Also, it offers the classifications of developed system according to PSO that has been used in research directions. A method for PSO-optimized Back Propagation (BP) which is assumed to be an NN based MapReduce on a Hadoop platform along with PSO as well as corresponding model was presented by Cao et al. [13]. It is applied for optimizing BP NN from the initial

weights and thresholds helps in developing classification methodologies as well as the accuracy. The MapReduce based parallel programming technique is applied to attain the simultaneous process of BP methods to deal hardware and communication problems while addressing BP and NN datasets. Furthermore, the system depicts a maximum accuracy of classification and enhanced efficiency of time which denotes an increment that is attained from parallel processing to smart models of big data. Hyper-heuristic methodologies helps in finding essential solutions to schedule in CC systems and further extension of allocating the simulation outcome to enhance the network lifetime. [14] deployed an alternate new Multi-Objective PSO (MOPSO) and Genetic Algorithm (GA) based hyper-heuristic technologies to schedule the resource in the form hybrid model. The working function of this method is estimated under the application of Cloud Sim toolkit. Various researchers made a comparison over hybrid scheduling technique by using recent heuristic as well as scheduled models [15]. The attained outcome has exhibited an optimal performance when compared with current approaches with respect to cost reduction and enhanced network lifetime. Consequently, proposed system has implemented a maximum resources application with span and throughput.

This paper presents new clustering based hybrid metaheuristic algorithm tuned back propagation neural network (C-HMT-BPNN) based IDS for effective identification of intrusion in the network. The proposed model involves two stages namely threshold based K-means clustering technique and back propagation neural network (BPNN) based classification. To optimize the weights and biases of BPNN, Particle Swarm Optimization (PSO) with Gravitational Search Algorithm (GSA) called PSO-GSA model has been developed. The proposed model has been tested using a set of two IDS dataset namely, NSL-KDD 2015 and CICIDS 2017 dataset.

MATERIALS AND METHODS

Figure-1 shows the process involved in C-HMT-BPNN model. The proposed model involves two main stages, namely clustering and classification. For clustering process, thresholding based K-means clustering technique is applied to cluster the data prior to classification. Next to that, the HMT-BPNN based classification model is applied to classify the clustered data to identify the existence of intrusions exist in the network data.

Threshold based K-means clustering

K-means is defined as a simple as well as effective unsupervised classification technique. K-means is a popular division based clustering models which tries to explore a user with definite clusters given by the centroids. It is said to be a common distance-based clustering method where the distance has been applied as a value of similarity, in which a minimum distance objects exhibits a higher affinity.

- Initiate $k = 2$ as the expected variable has 2 feasible results namely, normal and anomaly.
- Compute the input data for every nearby cluster center with the application of Eq. (1).

$$S_i^{(z)} = \left\{ x_p : \left\| x_p - m_i^{(z)} \right\|^2 \leq \left\| x_p - m_i^{(z)} \right\|^2 \forall_j, 1 \leq j \leq k \right\} \quad (1)$$

- Using Eq. (2), upgrade the cluster centers by re-evaluating the mean of all input data allocated to every cluster.

$$m_i^{(z+1)} = \frac{1}{|S_i^{(z)}|} \sum_{x_j \in S_i^{(z)}} x_j \quad (2)$$

- In order to implement the k-means cluster to a terminate step, it has looped through step (b) and (c) till a convergence of a mean value is attained.

As a result, k-means cluster provides the result by eliminating unwanted clustered data as well as to make a decision of exploring a novel dataset to classify using Eq. (3). When a novel data size is maximum then, supervised classification is applied, otherwise, it repeats k-means clustering technique till obtaining an adoptable cluster size.

$$newsize = \frac{leftdata}{totalsum} \quad (3)$$

In this paper, a thresholding mechanism is incorporated, which enables the clustering technique to group at least 70% of entire data. The clustering process gets iterated till 70% of clustered data is attained.

Classification Methods

Once the clustering process gets completed, HMT-BPNN model is applied for data classification. The HMT-BPNN is actually a BPNN based classifier which undergoes tuning by the use of hybrid metaheurisitc algorithm, named PSO-GSA. The PSO-GSA is applied to tune the weights and biases of BPNN. Generally, PSO is referred to be an optimizing model which depends upon the foraging behavior of birds, and random initialization of population as well as regular extension of searching task. When exploring an optimized solution, every bird has been assumed to be a particle with no mass and volume.

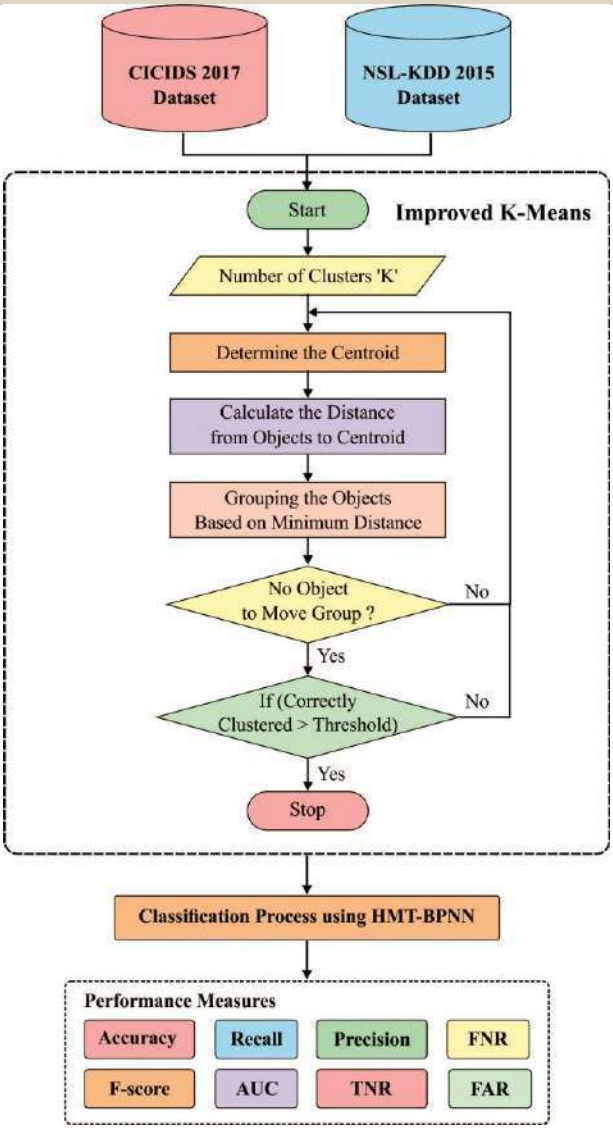

Fig. 1: Overall Process of Proposed Method

At the time of processing a search operation, the particles are capable of recording the recent best position (*pbest*) as well as global best position (g*best*). Velocity and position of every particle are estimated as given below:

$$v_i^{z+1} = w \cdot v_i^z + c_1 \cdot \text{rand} \cdot (\text{pbest} - x_i^z) + \text{c}_2 \cdot rand \cdot (gbest - \text{x}_\text{i}^\text{z}) \quad (4)$$

$$x_i^{z+1} = x_i^z + v_i^{z+1} \quad (5)$$

where v_i^z and x_i^z are referred as current velocity and position of ith particle at jth iteration, c_1 and c_2 are said to be acceleration coefficients which helps to manage the influence of $pbest$ and $gbest$ in searching task, correspondingly, rand denotes a random value in [0, 1], $pbest(z)$ represents the recent best position of every particles at tth iteration, $gbest$ implies a best position over the compared particles, and w signifies inertia weight.

GSA model has been named as a novel strategy which depends upon the law of gravity. The agents present in GSA are assumed as objects along with masses. Agents are inspired with one another by using a gravity force. If the quality becomes higher then, the gravity is stronger. Hence, position of an agent with higher mass is termed as best solution. Let N agents have d-dimension where the position of ith agent is given by,

$$X_\text{i} = \left(x_i^1, x_i^2, \ldots, x_i^d\right)(\text{i} = 1,2,\ldots,N) \quad (6)$$

In a tth time, the force is working on ith agent from jth agent which is written as,

$$F_{ij}^d = G(\text{z})\frac{M_i(\text{z})M_j(\text{z})}{R_{ij}(\text{z}) + \varepsilon}\left(\text{x}_j^d(\text{z}) - \text{x}_i^d(\text{z})\right) \quad (7)$$

where $M_i(z)$ and $M_j(z)$ are signified as masses of ith agent as well as jth agent, correspondingly , $G(z)$ denotes a gravitational constant at zth time, ε refers a lower constant, and R_{ij} depicts the Euclidian distance from ith and jth agent. In zth time, overall force has been used on ith agent as defined in the following:

$$F_i^d(z) = \sum_{j=1, j\neq i}^{N} \text{rand} \cdot F_{ij}^d(z) \tag{8}$$

where $rand$ represents a uniform random variable in $[0, 1]$. Based on the law of motion, the acceleration of an agent in zth time could be expressed as follows:

$$a_i^d(z) = \frac{F_i^d}{M_i(z)} \tag{9}$$

For every iteration process, velocity and position of ith agent gets updated under the application of 2 given functions:

$$v_i^d(z+1) = \text{rand} \times v_i^d(z) + a_i^d(z) \tag{10}$$
$$x_i^d(z+1) = x_i^d + v_i^d(z+1) \tag{11}$$

where $rand$ implies a uniform random variable from the interval $[0, 1]$ and $x_i^d(z)$ and $v_i^d(z)$ are the current position and velocity, correspondingly.

In GSA, an agent does not distribute the population details by one another and contains a vulnerable ability of developing. BY exploiting the global optimum searching potential of PSO as well as local searching capability of GSA, every agent are upgraded with using the velocity of PSO as well as the acceleration of GSA. This technique is named as PSO-GSA [16]. Hence, exploration and exploitation ability has been integrated with modified variables. The velocity and position of ith agent are extended by given 2 equations:

$$v_i^{z+1} = w \cdot v_i^z + c_1' \cdot \text{rand} \cdot ac_i + c_2' \cdot rand \cdot (gbest - x_i^z) \tag{12}$$
$$x_i^{z+1} = x_i^z + v_i^{z+1} \tag{13}$$

where w denotes the inertia weight, v_i^z , x_i^z , and $ac_i(z)$ are said to be the velocity, position, and acceleration of ith particle at tth iteration, whereas c_1' and c_2' are constant acceleration coefficients, correspondingly. In this study, it is assumed with that c_1' and c_2' is exponential functions expressed as:

$$c_i' = c_{start} \cdot \left(\frac{c_{end}}{C_{Start}}\right)^{1/(1+k/Z_{max})} \tag{14}$$

where c_{start} denotes an initial value, c_{end} implies the final value, Z_{max} signifies the higher iteration value, and k represents a current iteration value. To differentiate from GSA-PSO, GSA-PSO with functional acceleration coefficients (14) it is termed as I-PSO-GSA.

The PSO-GSA model is used for optimizing the weights and biases of BPNN as well as Mean Square Error (MSE) which is employed as Fitness Function (FF) of HPT-BPNN model. The FF of kth training sample is described as:

$$MSE = \frac{1}{q}\sum_{k=1}^{q}\sum_{i=1}^{m}\left(o_i^k - d_i^k\right)^2 \tag{15}$$

where q refers a count of training samples, d_i^k shows an required outcome of ith input unit from a kth training instance, and o_i^k reflects the original result of ith input unit from kth training sample.

When a structure of BPNN is a $r - s - z$ structure, where r indicates the node count present in an input layer, s depicts the value of nodes from a hidden layer, and z shows a number of the nodes from a resultant layer. Then,N agents of population has $L_i(i = 1,2,\ldots,N)$ as d-dimension vector$\left(l_{i,1}, l_{i,2}, \ldots, l_{i,d}\right)$, where $d = rs + s + sz + z$. By mapping L_i as the weights on the basis of BPNN, the components $l_{i,1}, l_{i,2}, \ldots, l_{i,rs}$ of $L_i$is assumed to be the weights acquired from input and hidden layer, the components $l_{i,rs+1}, \ldots, l_{i,rs+s}$ of L_i is based on hidden layer, and the components $l_{i,rs+s+1}, \ldots, l_{i,rs+s+sz}$ of L_i are the weights among hidden layer as well as output layer, and components $l_{i,rs+s+st+1}, \ldots, l_{i,D}$ depends upon the output layer.

RESULTS

Dataset Description

For assessing the effective performance of the presented C-HMT-BPNN model, an experimental validation takes place using a set of two benchmark dataset namely NSL-KDD 2015 [17] and CICIDS 2017 [18]. The first NSL-KDD 2015 dataset holds a sum of 125973 instances with 41 attributes. Among the 125973 instances, around 67343 and 58630 instances fall into the Normal and Anomaly categories respectively. The second CICIDS 2017 dataset comprises 2830743 instances with the occurrence of 80 features.

Among the total number of instances, 2273097 instances comes under Normal class and rest of the instances comes under Anomaly class.

Results Analysis

Fig. 2 illustrates the confusion matrix generated by the proposed model on the applied dataset. Fig. 2a shows the confusion matrix offered by the proposed model before clustering on the applied NSL-KDD 2015 dataset. The figure clearly stated the proposed model offers a maximum of 67138 instances as Normal and 58219 instances as Anomaly. Similarly, Fig. 2b shows the confusion matrix offered by the proposed model after clustering on the applied NSL-KDD 2015 dataset. The figure clearly stated the proposed model offers a maximum of 50724 instances as Normal and 45897 instances as Anomaly. Fig. 2c shows the confusion matrix offered by the proposed model before clustering on the applied CICIDS 2017 dataset. The figure clearly stated the proposed model offers a maximum of 2223538 instances as Normal and 553235 instances as Anomaly. Similarly, Fig. 2d shows the confusion matrix offered by the proposed model after clustering on the applied CICIDS 2017 dataset. The figure clearly stated the proposed model offers a maximum of 2189257 instances as Normal and 491513 instances as Anomaly.

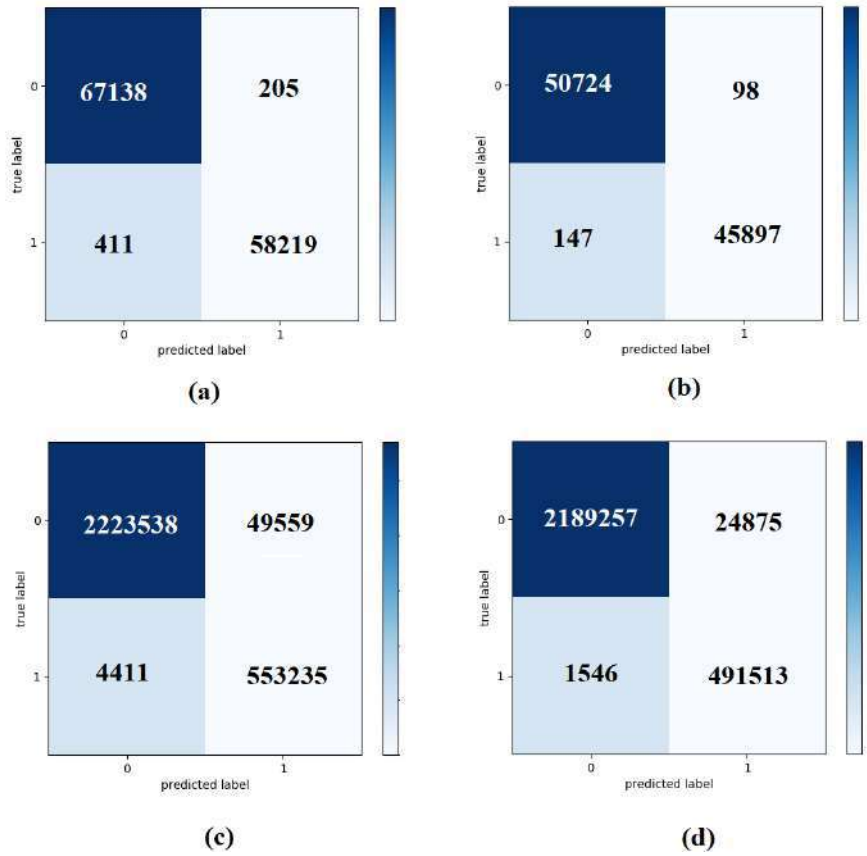

Fig. 2: Confusion Matrix Generated at the Time of Execution a) Before Clustering NSL-KDD 2015 b) After Clustering NSL-KDD 2015 c) Before Clustering CICIDS d) After Clustering CICIDS

Figures - 3 and -4 show the classification outcome of the proposed model on the applied dataset, with and without clustering interms of different measures namely false acceptance rate (FAR), true negative rate (TNR), false negative rate (FNR), area under curve (AUC), precision, recall, accuracy and F-score are employed. On measuring the results on the applied NSL-KDD 2015 dataset, it is noted that the C-HMT-BPNN model before clustering offers effective results with minimum FAR value of 0.04 and is further reduced to 0.002 after clustering. Similarly, the C-HMT-BPNN model attains minimum FNR values of 0.007 and 0.003 under before and after clustering respectively. At the same time, the C-HMT-BPNN model provides a higher TNR rate of 99.65% before clustering and gets increased to 99.79% after clustering. The proposed C-HMT-BPNN model also demonstrated maximum AUC value of 99.52% before clustering and is raised to 99.75% after clustering. In the same way, the C-HMT-BPNN model achieves a maximum precision value of 99.70% before clustering and is increased to 99.81% by the inclusion of clustering process. Afterwards, the C-HMT-BPNN model shows its effective results by offering maximum recall values of 99.39% and 99.71% before and after clustering respectively. In these lines, the C-HMT-BPNN model exhibits maximum classification accuracy of 99.51% before clustering and the clustering process

enhances it to 99.75%. At last, the C-HMT-BPNN model reaches to an F-score value of 99.54% and is raised to 99.76% by the use of clustering process.

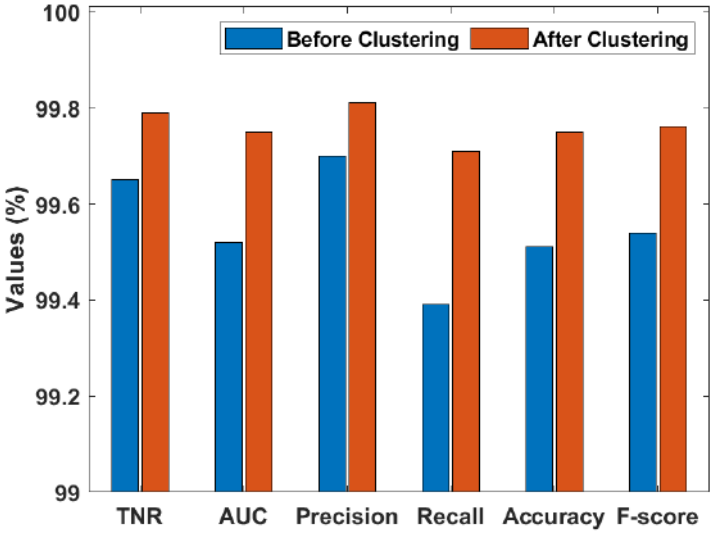

Fig. 3: Classifier results analysis of C-HMT-BPNN model on NSL-KDD 2015 dataset

To validate the classifier outcome on the tested CICIDS 2017 dataset, it is observed that the C-HMT-BPNN model before clustering provides supreme outcome with a lower FAR value of 0.082 and is further reduced to 0.048 after clustering. Likewise, the C-HMT-BPNN model reaches to minimum FNR values of 0.008 and 0.003 under before and after clustering respectively. Simultaneously, the C-HMT-BPNN model attains a high TNR of 91.78% before clustering and gets increased to 95.18% after clustering. The proposed C-HMT-BPNN model additionally exhibited higher AUC value of 95.79% before clustering and is raised to 97.56% after clustering. Similarly, the C-HMT-BPNN model resulted to a maximum precision value of 97.82% before clustering and is increased to 98.88% by the inclusion of clustering process. Afterwards, the C-HMT-BPNN model shows its effective results by offering maximum recall values of 99.80% and 99.93% before and after clustering respectively. In these lines, the C-HMT-BPNN model provides a maximum classification accuracy of 98.09% before clustering and the clustering process enhances it to 99.02%. Finally, the C-HMT-BPNN model leads to a maximum F-score value of 98.80% and is raised to 99.40% by the use of clustering process.

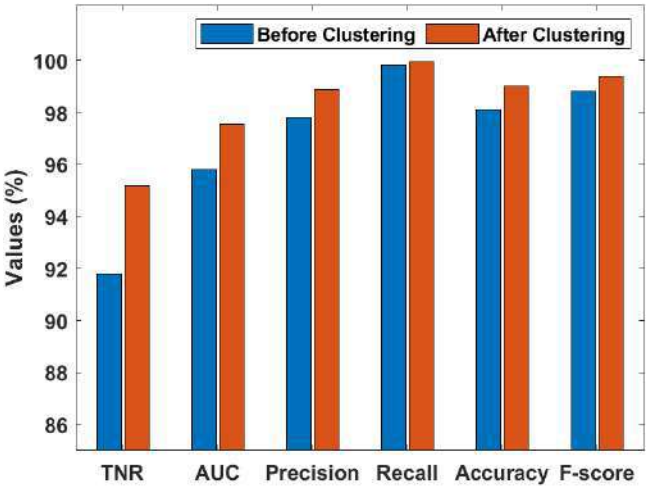

Fig. 4: Classifier results analysis of C-HMT-BPNN model on CICIDS 2017 dataset

DISCUSSION

To validate the consistent results of the presented C-HMT-BPNN model on the IDS dataset, an extensive results analysis is made with the lately presented IDS models [19] namely Cuckoo optimization, cuckoo search with PSO (CS-PSO), PSO-SVM, Behaviour Based IDS, Gaussian Process, Deep Neural Network with SVM, GA+Fuzzy, Fuzzy C-means and Gradient Boosting models interms of accuracy. The resultant values are shown in Fig. 5. After observing the values exist in the table, it is evident that the CS-PSO model exhibited poor outcome with a least accuracy of 75.51%. Then, it is apparent that the Gradient Boosting

model has reached to an accuracy of 84.25%, which is superior to the accuracy provided by the CS-PSO algorithm. On the other hand, the Gaussian Process and the DNN+SVM models show better results over the existing models by the attainment of near identical accuracy values of 91.06% and 92.03% respectively. Besides, even higher detection performance is showed by Fuzzy C-means model by offering an accuracy value of 95.30%. Concurrently, the GA+ Fuzzy and Cuckoo Optimization algorithms have accomplished manageable and identical detection results by offering accuracy values of 96.53% and 96.888% respectively. In line with, even higher detection outcome is achieved by Behaviour Based IDS model which can be noticed from the accuracy value of 98.89% whereas competitive results of 99.10% and 99.36% of accuracy are provided by the existing PSO-SVM and IPSO-NN models. But, it is interesting that the HMT-BPNN model has outperformed all the existing methods and achieved a higher accuracy of 99.51% on the applied dataset. Furthermore, it is noted that the C-HMT-BPNN model has shown superior performance and offered a maximum accuracy of 99.75%.

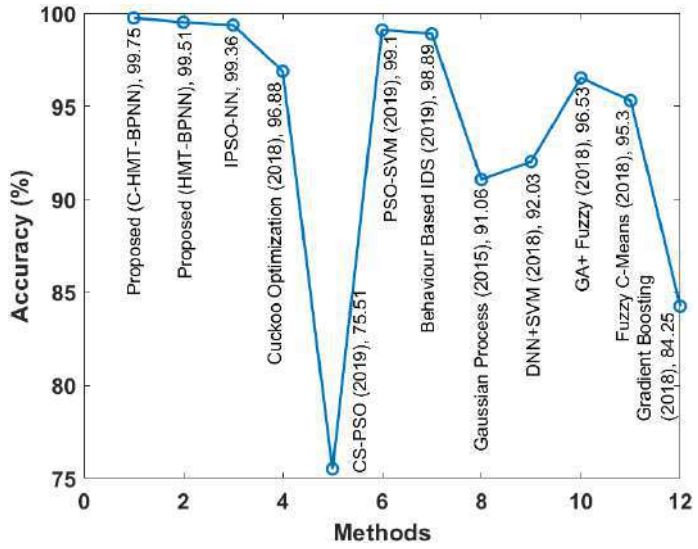

Fig. 5: Accuracy analysis of C-HMT-BPNN model with state of art models

CONCLUSION

This paper presents new C-HMT-BPNN model for effective identification of intrusion in the network. The proposed model involves two main stages, namely clustering and classification. For clustering process, thresholding based K-means clustering technique is applied to cluster the data prior to classification. Next to that, the HMT-BPNN based classification model is applied to classify the clustered data to identify the existence of intrusions exist in the network data. The proposed model has been tested using a set of two IDS dataset namely, NSL-KDD 2015 and CICIDS 2017 dataset. The obtained experimental outcome clearly ensured the superior characteristics of the proposed model over the compared methods in a significant way. The proposed HMT-BPNN model resulted to a maximum accuracy of 99.51% which is increased to 99.75% by the inclusion of clustering techniques.

CONFLICT OF INTEREST
The authors have expressed no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Shiraz M, Gani A, Khokhar RH, Buyya R. [2012] A review on distributed application processing frameworks in smart mobile devices for mobile cloud computing. IEEE Communications surveys & tutorials, 15(3):1294-313.

[2] Popović K, Hocenski Ž. [2010] Cloud computing security issues and challenges. In The 33rd international convention mipro, IEEE, 344-349.

[3] Carlin A, Hammoudeh M, Aldabbas O. [2015] Defence for distributed denial of service attacks in cloud computing. Procedia computer science,1;73:490-497.

[4] Murugan S, Jeyakarthic M. [2019] An Efficient Bio-Inspired Algorithm Based Data Classification Model For Intrusion Detection In Mobile Adhoc Networks. The International journal of analytical and experimental modal analysis, 11(11): 834-848.

[5] Carlin A, Hammoudeh M, Aldabbas O. [2015] Intrusion detection and countermeasure of virtual cloud systems-state of the art and current challenges. International Journal of Advanced Computer Science and Applications, 6(6): doi: 10.14569/IJACSA.2015.060601
[6] Shelke MP, Sontakke MS, Gawande AD. [2012] Intrusion detection system for cloud computing. International Journal of Scientific & Technology Research, 1(4):67-71.
[7] Butun I, Morgera SD, Sankar R. [2013] A survey of intrusion detection systems in wireless sensor networks. IEEE communications surveys & tutorials,16(1):266-82.
[8] Peddabachigari S, Abraham A, Grosan C, Thomas J. [2007] Modeling intrusion detection system using hybrid intelligent systems. Journal of network and computer applications, 30(1):114-32.
[9] Mohod AG, Alaspurkar SJ. [2013] Analysis of IDS for cloud computing. International Journal of Application or Innovation in Engineering & Management, 2:344-9.
[10] Rowland CH, Psionic Software Inc [2002]. Intrusion detection system. US Patent, 6405318.
[11] Liao HJ, Lin CH, Lin YC, Tung KY. [2013] Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1):16-24.
[12] Masdari M, Salehi F, Jalali M, Bidaki M. [2017] A survey of PSO-based scheduling algorithms in cloud computing. Journal of Network and Systems Management, 25(1):122-58.
[13] Cao J, Cui H, Shi H, Jiao L. [2016] Big data: A parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. PloS one, 11(6):e0157551.
[14] Kumari KR, Sengottuvelan P, Shanthini J. [2017] A hybrid approach of genetic algorithm and multi objective PSO task scheduling in cloud computing. Asian Journal of Research in Social Sciences and Humanities, 7(3):1260-71.
[15] Zaman S, El-Abed M, Karray F. [2013] Features selection approaches for intrusion detection systems based on evolution algorithms. InProceedings of the 7th International Conference on Ubiquitous Information Management and Communication, 1-5.
[16] Hu H, Cui X, Bai Y. [2017] Two Kinds of Classifications Based on Improved Gravitational Search Algorithm and Particle Swarm Optimization Algorithm. Advances in Mathematical Physics. 2017.
[17] CICIDS2017 data set. [2019] https://www.unb.ca/cic/datasets/ids-2017.html (Accessed on 14 Jan, 2020)
[18] NSL-KDD Dataset of NSL-KDD University of new Brunswick. [2019] http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html. (Accessed on 14 Jan, 2020)
[19] Chiba Z, Abghour N, Moussaid K, Rida M. [2019] Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms. Computers & Security, 86:291-317.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

AN ENSEMBLE OF OPTIMAL DEEP LEARNING ARCHITECTURE WITH RANDOM FOREST CLASSIFIER FOR CONTENT BASED IMAGE RETRIEVAL SYSTEM

Purushothaman Anandababu*, Mari Kamarasan

Department of Computer and Information Science, Annamalai University, Chidambaran, INDIA

ABSTRACT

Content based image retrieval (CBIR) extract the details from the images depending upon the content exist in the image as feature descriptors. It intends to the process of retrieving images with maximum resemblance among the visual content exist in the massive databases. Feature extraction and similarity measurement are the two essential steps involved in CBIR. This paper presents an ensemble of optimal AlexNet architecture with random forest (RF) classifier called EOANA-RF model to effective retrieve the images from the databases. Since AlexNet model does not offer superior results on large databases, the AlexNet architecture is optimized in three ways. Firstly, the average pooling undergoes replacement with max-ave pooling, Maxout is employed in fully connected (FC) layers and hidden layer is included to map high-dimension features to binary codes. Then, similarity measurement and RF based classification process takes place to retrieve the images related to the query image (QI) from the databases and classifies it. The performance of the proposed model undergoes validation using Corel10K dataset. The obtained simulation outcome verified the enhanced performance of the proposed model on the applied dataset.

KEY WORDS
Image retrieval, CBIR, Corel 10K, AlexNet, Similarity measurement

INTRODUCTION

In general, information retrieval as become an essential objective along with the requirements of multimedia data processing for detecting the practical information. Hence, image retrieval (IR) is been a typical and most organized tool. It is more essential to execute and enhance the tools used for IR to explore images present on internet in a productive manner. Many types of traditional IR system depends upon a keyword search which tends to several disadvantages like, maximum requirement of man power as well as dependency on personal aspect that generates improper outcome. In order to manage these limitations, a novel mechanism has been established named as CBIR method [1]. This model is comprised with collection of techniques which mainly focus in minimum level image features, for instance texture values, structure as well as color signature for retrieving images from database which are based on QI provided by a customer [2]. Previous CBIR models perform in an inconvenient form while using higher level methods that mainly aims in minimum and maximum level visual features that are not contributed in retrieval function. Hence, 2 modules has been enhanced initially, Region based image retrieval (RBIR) which is based on image representation to be divided as regions features that depends upon image perception by an user. Alternatively, Relevance feedback (RF) helps to assure the user inclination [3]. The core theme of CBIR technique is to retrieve the images which are relevant to QI [4].

CBIR uses the model of query by example that helps in retrieving same images of input image using a definition regarding a QI missed by a user, also CBIR method has been operated by QI features extraction, once the system starts for extracting a feature. Feature vector has been estimated for the obtained features, CBIR shows all images present in a database along with a vector, after inputting the QI, CBIR model determines the feature vector to compare with alternate vectors saved for each image exist in a database, images which are comprised with higher features are same as QI has to be retrieved. To improvise the IR process the Region-based visual signatures [5] has been employed in image segmentation. According to the performance of optical system of a human, images should be differentiated as features of region by image similarity. Such models determine the segmented region features from the object level and implement at the granularity region where previous models are applied to image representation by applying global features. [6] implied as model to extract features by applying image binarization to improve the images retrieval as well as exploration with the application of CBIR. Many developers sampled the model by applying 2 public datasets which has limited the feature size of an image. Some of the statistical values which are relied on precision as well as recall measures applied for estimation purpose. The main demerit of this technique is to misclassify the QI that affects the function of image than other conventional techniques. [7] projected a method image representation as well as feature extraction under the application of bandelet transform, which is comprised with the information in the form of image. The artificial neural network (ANN) has been applied for used for IR process where the system functions were computed by the application of 3 public data sets such as: Coil, Corel, and Caltech 101. Here, the precision as well as recall measures were employed for estimating the retrieval efficiency. [8] projected a technique to conduct the IR process with help of statistical tests, like Welch's t-tests and F-ratio. These modules are structured input QI. In this approach, the whole image has been assumed to be textured image, whereas in structured image, the shape is divided as different regions according to the behavior. Initially, F-ratio test has been employed to pass the images to energy spectrum testing. Followed by, when images are effective in 2 tests then it is decided that images are identical. Otherwise, it is dissimilar. In order to calculate the performance validation Mean Average Precision score have been employed.

Received: 5 Mar 2020
Accepted: 28 Mar 2020
Published: 30 Mar 2020

***Corresponding Author**
Email:
drpabcud@gmail.com

IIOAB JOURNAL
ISSN: 0976-3104

[9] deployed an image descriptor for extracting texture and color as it has similar impact on CBIR. In [10], a local structure descriptor is deployed to process the IR. Local structure descriptor has been developed on the basis of local structures of colors where the combination of color, shape and texture are presented as a single unit to perform IR process. Also, a technique used for feature extraction that is capable of extracting local structure histogram with the application of local structure descriptor. [11] implied a model termed as IR using interactive genetic algorithm (GA) to calculate the maximum number of desired features and compares with the QI images. This method has been sampled under a set of 10,000 images to evident the effectiveness of a presented technique. [12] applied a CBIR mechanism by employing the combination of Speeded-Up Robust Features (SURF) as well as Scale Invariant Feature Transform (SIFT). Therefore, presentations of such local features are utilized to retrieve images as SIFT is robust in rotation and scale changing, while SURF is highly stringer for illumination difference. The combination of SURF and SIFT improves the CBIR efficiency. GA algorithm is used for optimization [13]. GA is very productive in identifying an optimal solution from a search space. To improve the function of meta-heuristic like GA, a local search mechanism is essential to assist the GA in searching the solution space instead of finding a search space. The best local search technique is great deluge algorithm (GDA). This GDA were established by [14] in the form of local search model. The main aim of this technique has been evolved from analogy of hill climbing and trying to travel in a direction of finding a way to maintain dry feet and water level is increasing by a GDA. Also, GDA has been included in GA when it is more efficient to produce best solution rather applying GA model [15].

This paper presents an ensemble of optimal AlexNet architecture with RF classifier called EOANA-RF model to effective retrieve the images from the databases. Since AlexNet model does not offer superior results on large databases, the AlexNet architecture is optimized in three ways. Firstly, the average pooling undergoes replacement with max-ave pooling, Maxout is employed in FC layers and hidden layer is included to map high-dimension features to binary codes. Then, similarity measurement and RF based classification process takes place to retrieve the images related to the QI from the databases and classifies it. The performance of the proposed model undergoes validation using Corel10K dataset. The obtained simulation outcome verified the enhanced performance of the proposed model on the applied dataset.S

MATERIALS AND METHODS

The working principle of the proposed EOANA-RF model is shown in Figure-1. Initially, the feature extraction process takes place on the images present in the database by the use of optimal AlexNet architecture. Then, the extracted features are stored in a repository. Upon providing the query image, the feature extraction process again takes place and similarity measurement is carried out to determine the resemblance between the extracted feature vectors if the QI and the feature vectors exist in the database. The images with higher resemblance will be retrieved. Afterwards, the retrieved images will undergo classification by the use of RF classifier where the retrieved images will be grouped into respective classes. The processes involved in the EOANA-RF model will be explained in the following subsections.

Optimized AlexNet model

Basic architecture: The presented system is based on AlexNet, which is said to be a traditional deep convolution neural network (DCNN). AlexNet is comprised with 5 convolution layers, 3 pooling layers and 3 FC layers. The convolution layers as well as pooling layers were applied for extracting image features while FC layers applies the convolution layers and pooling layers, that match a 2D feature vectors as ID feature vectors. Though it is existed with a semantic space which can be limited by including a depth of network, it improves the processing time simultaneously. The main goal is to reduce the semantic gap with the application of few optimizing process applied on the structure of AlexNet in case of accurate and compact image representation.

AlexNet can be optimized using convolutional layers and FC layers to attain particular middle-level feature descriptors.

- The max-ave pooling criterion is applied in pooling layers for the representation of local feature.
- Max-out activation has been completely employed in FC layers to fit the global feature.
- The hidden layer was included in FC layer to transform the global feature vectors as binary codes.

Top K images were ranged under the application of Hamming distance which is referred as retrieval outcome. This mechanism is named as Optimized AlexNet for Image Retrieval (OANIR).
Feature extraction: In case of semantic IR, exploring a best feature extraction as well as representation has been considered as a crucial procedure. Even though a conventional AlexNet are capable of performing best feature extraction, it has only minimum network depth which is assumed to be a major challenging operation. In OANIR system, few optimizations were developed with no improvement of network depth.

Max -Ave pooling for local features: The main objective for pooling is to obtain the major features from combined features at the time of provisioning significant features and eliminating the unwanted features. Pooling is comprised with merits for optimal feature representation. It is considered to be more compact representation which is assumed to be an invariant to image conversion and robust for noise. To obtain a better image features, state-of-the-art deep learning (DL) techniques were added such as max pooling, spatial pyramid model as well as average pooling. In case of convolved matrix image features along with the size of p×k, for all p-dimensional feature vector v_i, it can be defined as 2 pooling types, namely, max pooling and average pooling as given in Eqs. (1) and (2),

$$f_m(v) = \max v_i \tag{1}$$

$$f_a(v) = \frac{1}{p}\Sigma_{i=1}^{p} y_i \tag{2}$$

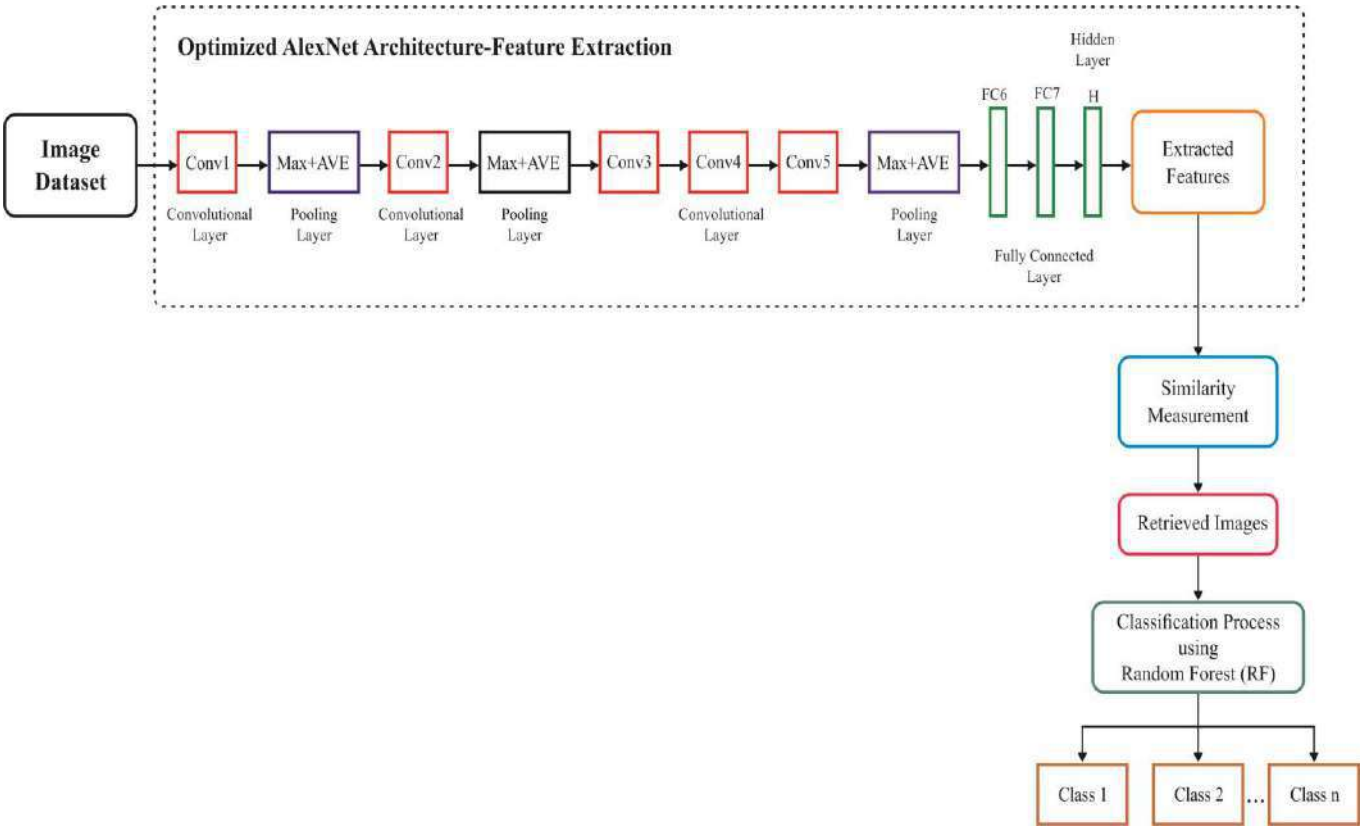

Fig. 1: Overall process of EOANA-RF model

Once the convolutional task is completed, distribution of image features for every patch could be named as exponential distribution along with a mean $E(X) = 1/\lambda$ and variance $D(X) = Var(x) = 1/\lambda^2$. Hence, the adjacent cumulative distribution function is $F(x;\lambda) = 1 - e^{(-\lambda x)}$. The maximum kurtosis from a provided exponential distribution is capable of modelling visual feature. The main feature acquired by a salient region is corresponding to maximum kurtosis present in a data distribution.

Let P implies a cardinality of pooling. The cumulative distribution of a max-pooled feature is given by

$$F(P) = \left(1 - e^{(-\lambda x)}\right)^P \tag{3}$$

The mean isolation is expressed as,

$$\mu_{m=}\left(H(P)\right)/\lambda \tag{4}$$

and variance is

$$\sigma_m^2 = \frac{1}{\lambda^2}\Sigma_{1=1}^{p}\frac{1}{l}\left(2H(l) - H(P)\right), \tag{5}$$

where $H(k) = \Sigma_{i=1}^{k}\frac{1}{i}$ is a harmonic series.Hence, for each P,

$$\frac{\mu_1}{\mu_2} = \frac{\delta_1}{\delta_2} = \frac{\lambda_1}{\lambda_2} \tag{6}$$

The distribution might be optimally isolated when a scaling factor of a mean value is higher than scaling factor of SD. Moreover, as $H(p) = \log(P) + v + o(1)$, if $P \to \infty$, where v refers anEuler's constant, which can be represented as:

$$\Sigma_{1=1}^{P} \frac{1}{l}\left(2H(l) - H(p)\right) = \log(P) + O(1) \tag{7}$$

The distance from mean values are developed rapidly when compared with SD. Therefore, a crucial result could be attained if there is no smoothing performance, higher cardinalities gives best signal-to-noise ratio. A feature available in convolutional layer has related data, like position as well as relative position. If the distributions of image features are smooth and flat, max-pooling operation removes the associated local spatial data. Then, it influences the feature extraction and representation to a great extent. At this point, ave-pooling has the responsibility to manage locally correlated data. Thus, the pooling function can be written as,

$$f(v) = \alpha_1 \max v_i + \alpha_2 \frac{1}{p} \Sigma_{=1}^{P} y_i \tag{8}$$

where $\alpha_1 + \alpha_2 = 1$. For the training phase, it is pointed that the learned method is capable of providing best image features with all types of images. Fix $\alpha_1 = \alpha_2 = 0.5$ to improvise the robustness of a technique of input data while in a testing stage, 2 solutions are offered to resolve with diverse cases. If the input images are comprised with maximum quality pixels and only minimum noises, then it is fixed as $\alpha_1 = \alpha_2 = 0.5$ to validate the saliency feature and the average feature. While the input images are constrained with lower quality pixels and with a higher noises, then it is fixed as $\alpha_1 = 0.3, \alpha_2 = 0.7$ to focus on average features when compared with maximum kurtosis noise features that minimize the influence of image noises in extracting features. The MNIST and CIFAR−10 are existed with lower quality pixels whereas SUN397 and ILSVRC2012 are present with higher quality pixels.

Non-linear activation Maxout: The FC layers perform matrix multiplication, assumed to be more similar to feature space conversion. It is also used in data extraction as well as combination. Features in FC layers indicate global information. When merged with a non-linear mapping activation function, the FC layers triggers a non-linear conversion. The limitation of this function is that, it has lower data for spatial structure. These shortcomings can be resolved by using average pooling. Previous activation functions, such as Sigmod and ReLu, fits the 2D functions, and it shows that Max-out function can fit diverse dimensional function along with global approximation. It reaches a tremendous performance with dropout training. The Max-out technique is referred as a forward propagation structure with a largest resultant active form. In a provided input, the result for Max-out can be expressed as:

$$h_i(x) = \max_{j \in [1,k]} z_{ij} \tag{9}$$

where $z_{ij} = x^T W \ldots ij + b_{ij}, W \in \Re^{d \times k}$, and $b \in \Re^{m \times k}$ are learned parameters.The inclusion of Max-out parameters of a network tends to acquire maximum computation time for feature extraction. To attain major retrieval efficiency, the neural nodes present in FC6 and FC7 layers are reduced to 2048, and minimize the dropout value from 0.7 to 0.5. Finally, the sparseness image feature representation maintains the representation accuracy simultaneously. Binary code for large scale data: Here, the hidden layer pattern has been applied to reach the productive IR in large scale database. It undergoes mapping with high dimension features as compact binary codes as well as removes the unwanted features concurrently. Therefore, the activation function could be shown as:

$$a_n^H = \sigma(a_n^7 W^H + b^H) \tag{10}$$

where $\sigma(.)$ denotes a Sigmod logistic regression that manages the result from (0,1). a_n^7denotes the resultant feature vectors in FC7 layer, W^H represents weights, and b^H implies bias parameters of hidden layer. The binary code function is written as:

$$b_n = \begin{cases} 1 & a_n^H > 0.5 \\ 0 & a_n^H \leq 0.5 \end{cases} \tag{11}$$

Under the application of binary operation from a hidden layer, feature vectors undergoes mapping with binary codes. For IR process, OANIR network tends to extract image features for QI. Followed by, the result of hidden layer and binary codes were filtered with an activation function. Then, a same image has been ranged using Hamming distance of binary codes from QI as well as database images.

Query matching

Feature vector for QI as Q is denoted as $f_Q = \left(f_{Q_1}, f_{Q_2}, \ldots\ldots f_{Q_{Lg}}\right)$ has been attained after completing the feature extraction process. Likewise, every image present in a database has been presented with feature vector $f_{DB_j} = \left(f_{DB_{j1}}, f_{DB_{j2}}, \ldots\ldots f_{DB_{jLg}}\right); j = 1,2, \ldots\ldots, |DB|$. The main aim of this process is to select n best images which are same as QI. It involves in selecting n top images by estimating the distance from a QI and image in a database $|DB|$. To match these images, it is employed with 4 diverse similarity distance measures as given in the following.

$$Manhattan\ distance\ value: \mathrm{D(Q, I_1)} = \sum_{\mathrm{i=1}}^{\mathrm{Lg}} |\,\mathrm{f_{DB_{ji}}} - \mathrm{f_{Q,i}}| \tag{12}$$

$$Euclidean\ distance\ value: \mathrm{D(Q, I_1)} = \left(\sum_{\mathrm{i=1}}^{\mathrm{Lg}} \left(\mathrm{f_{DB_{ji}}} - \mathrm{f_{Q,i}}\right)^2\right)^{1/2} \tag{13}$$

$$Canberra\ distance\ value: D(Q, I_1) = \sum_{i=1}^{Lg} \frac{|f_{DB_{ji}} - f_{Q,i},|}{|f_{DB_{ji}}| + |f_{Qi}|} \tag{14}$$

$$\mathrm{d_1}\ distance\ value: D(\mathrm{Q}, I_1) = \sum_{i=1}^{Lg} |\frac{f_{DB_{ji}} - f_{Q,i}}{1 + f_{DB_{ji}} + f_{Q,i}}| \tag{15}$$

where $f_{DB_{ji}}$ denotes the i^{rh} feature of j^{rh} image present in database $|DB|$.

RF classifier

The RF classification method is a well-known approach for linear and non-linear classification issues which are relatively novel. Hence, it comes under the category of ML methods named as ensemble models. Ensemble learning is defined as a learning approach that contributes in various mechanisms that is applied for resolving an individual predictor. It is operated by producing several classification methods that learns and develops autonomous detection. Such prediction has been integrated as single prediction which has to be an optimal one when compared with prediction done by a classifier. RF is a considered to be an ensemble learning that applies ensemble of DTs. Gradient boosting employs a collection of weak learners and provides enhanced prediction accuracy. Therefore, the attained simulation outcome from a sample depends upon previously obtained result. In GD, the limitations of predictions are referred as negative gradients. For all steps, a novel tree has been fit to the negative gradients of existing trees.The RF classification model is comprised with an integration of tree classifiers in which every classifier has been produced under the application of random vector that is tested in an autonomous input vector, and every tree casts a single vote for well-known class that tends to classify the input vector. Hence, an RF classifier applied in this work has arbitrarily chosen features at all nodes to be used for tree development. Bagging is a technique used in generating training data set by randomly obtaining N examples, where N denotes the size of the actual training set which has been deployed for all selected feature combination. The constantly employed attribute selection values in DT establishment are IG, Ratio criterion and Gini Index. The given training set T, selecting in a random manner that belongs to few class C_i, the Gini index might be expressed as:

$$\sum\sum_{j \neq i} (f(C_i, T)/|T|)(f(C_j, T)/|T|) \tag{16}$$

where $f(C_i, T)/|T|$ implies the possibility of selected classC_i.

RESULTS

In this section, the experimental validation of the proposed EOANA-RF model takes place. The dataset used, results offered by the EOANA-RF model and the comparative analysis is carried out in the following subsections.

Dataset used

The important points of proposed EOANA-RF methods are applied for testing in contrast to a standard Corel10K dataset [16]. The used dataset is constrained with a collection of 10,908 distinct images. A group of 100 classes are present in an employed dataset and 100 images have been exploited in all classes. This model utilizes a group of ten classes where everyone holds a set of100 images. The images

are divided into different groups from animals, sports, food, etc. Only few testing images from the dataset is depicted in Figure-2.

Fig. 2: Sample Test Image.

Fig. 3: (a) Query Image (b) Retrieved Images.

Figure-3 depicts the qualitative results analysis of the EOANA-RF model on the applied Corel10K dataset. It is shown that the EOANA-RF method effectively retrieves a set of images for the applied QI.

Result analysis

Figure-4 offer the outcomes obtained by the EOANA-RF method interms of precision and recall under a set of ten classes. On the applied set of images under Buses class, the EOANA-RF model attains a maximum precision and recall of 95.67% and 86.13% respectively.

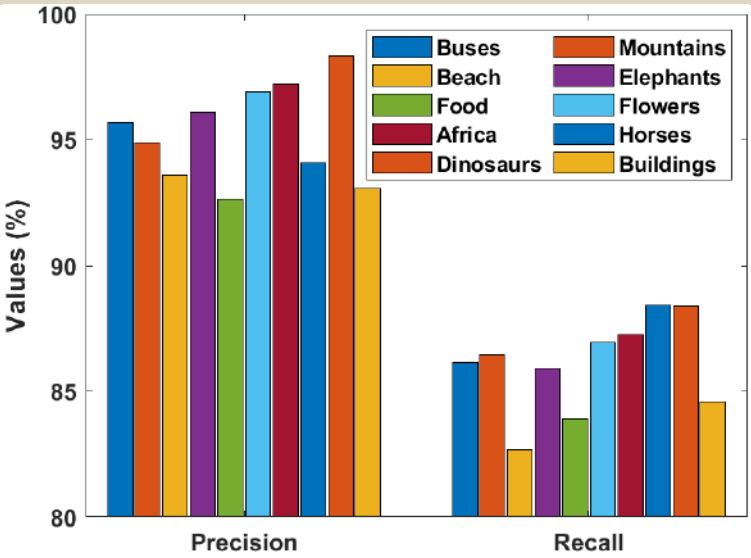

Fig. 4: Results of Proposed EOANA-RF Method in terms of Precision and Recall

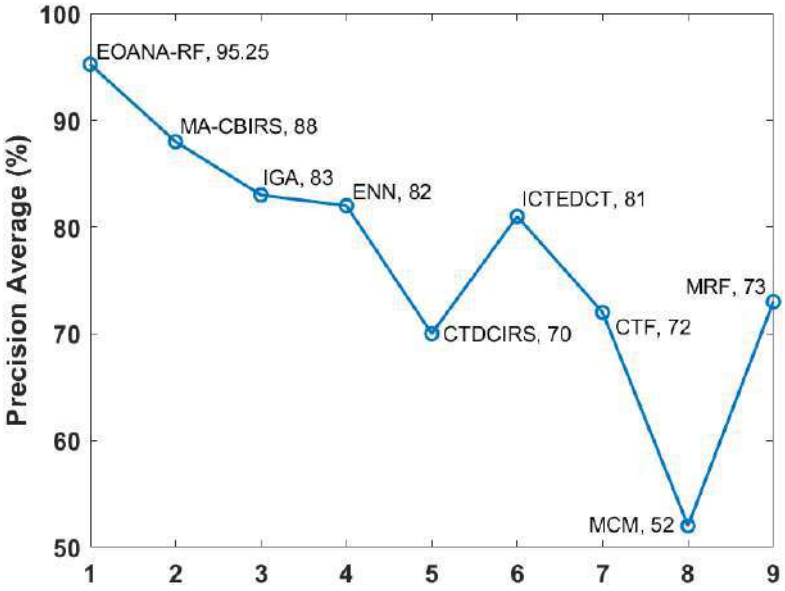

Fig. 5: Average precision analysis

Simultaneously, the MA-CBIRS model attempts in showing competing simulation outcome by attaining maximum average precision of 88%.On the same way, the previous IGA model seeks to average precision rate of 83% and the ICTEDCT techniques refers slightly better results with average precision of 81%.Similarly; the traditional MRF model demonstrates worst retrieval process by reaching very less average precision value of 73%. Therefore, poor retrieval outcome is produced by used CTF, CTDCIRS and MCM techniques by accomplishing minimum average precision rate of 72%, 70% and 52% correspondingly. The maximum average precision value of 95.25% is provided by EOANA-RF model which illustrates the stable retrieval outcome on the applied images. In order to ensure the uniform retrieval performance on the used test images, an average recall measure can be estimated for each applied methods [Figure-5].

Discussion

Figure-6 defines the investigation of IR process of various methods with respect to average recall. The figure clearly points that the projected EOANA-RF mechanism gives maximum retrieval final outcome with the average recall of 86.07%.Concurrently, the MA-CBIRS model attempts in showing reasonable result than other techniques with average recall measure of 70%. In line with this, the classical IGA approach seeks for an average recall value of 69%. The ENN and CTDCIRS models providemoderate outcomeswith same average recall measure of 16%. Similarly, the previous ICTEDCT and MCM methods depictan impractical retrieving performance by reaching a least average recall of 14%. But, poor retrieval results are produced with the application of CTF technique by achieving minimum average recall value of 10%. The

highest average recall value of 50.73 by a deployed RCM model displays its reliable retrieving outcome on the applied images.

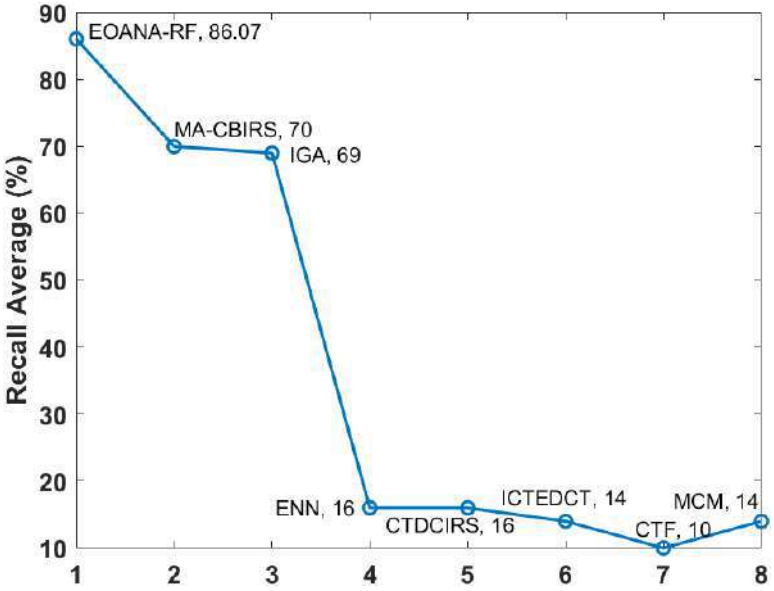

Fig. 6: Average recall analysis.

After observing the above-mentioned figures, it is obviously clear that the proposed model provides supreme retrieval and classification results over the compared methods in a significant way. It is verified from the maximum average precision and recall values of 95.25% and 86.07% respectively.

CONCLUSION

This paper has presented an effective EOANA-RF model to attain efficient retrieval of images from the databases. As AlexNet method is not capable in providing qualified results in case of large databases, the AlexNet architecture has been optimized in 3 ways. Initially, the feature extraction process takes place on the images present in the database by the use of optimal AlexNet architecture. Then, the extracted features are stored in a repository. Upon providing the query image, the feature extraction process again takes place and similarity measurement is carried out to determine the resemblance between the extracted feature vectors if the QI and the feature vectors exist in the database. The images with higher resemblance will be retrieved. Afterwards, the retrieved images will undergo classification by the use of RF classifier where the retrieved images will be grouped into respective classes. The performance of the proposed model is validated by applying Corel10K dataset. Hence, experimental values shows the improved results of the projected system has reached to a higher average precision and recall values of 95.25% and 86.07% correspondingly. In future, the performance of the proposed system can be improvised under the application of hyper parameter tuning models.

CONFLICT OF INTEREST
The authors have expressed no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Tseng VS, Su JH, Huang JH, Chen CJ [2008] Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation. IEEE Transactions on Multimedia, 10(2):260-7.

[2] Kanimozhi T, Latha K, [2013] A Meta-heuristic optimization approach for content based image retrieval using relevance feedback method. In: Proceedings of the World Congress on Engineering 2013 London, U.K

[3] Datta R, Joshi D, Li J, Wang, J.Z, [2008] Image retrieval: ideas, influences, and trends of the new age. ACM Comput, Surveys (Csur) 40: 5.

[4] Carneiro G, Chan AB, Moreno PJ, Vasconcelos N [2007] Supervised learning of semantic classes for image annotation and retrieval, Pattern Analysis and Machine Intelligence. IEEE Transactions on, 29: 394–410.

[5] Yuvaraj D, Hariharan S. [2016] Content-based image retrieval based on integrating region segmentation and colour histogram. The International Arab Journal of Information Technology, 13(1A):203-207.

[6] Das R, Thepade S, Bhattacharya S, Ghosh S. [2016] Retrieval architecture with classified query for content based image

recognition. Applied Computational Intelligence and Soft Computing, 2016(1)1-9.
[7] Ashraf R, Bashir K, Irtaza A, Mahmood, MT [2015] Content based image retrieval using embedded neural networks with bandletized regions. Entropy, 17(6): 3552-3580.
[8] Seetharaman K, Selvaraj S [2016] Statistical tests of hypothesis based color image retrieval. Journal of Data Analysis and Information Processing, 4(2): 90.
[9] Feng L, Wu J, Liu S, Zhang H [2015] Global correlation descriptor: a novel image representation for image retrieval. Journal of Visual Communication and Image Representation, 33:104-114.
[10] Zeng Z [2016] A novel local structure descriptor for color image retrieval. Information, 7(1): 9.
[11] Madhavi KV, Tamilkodi R, Sudha KJ [2016]. An innovative method for retrieving relevant images by getting the top-ranked images first using interactive genetic algorithm. Procedia Computer Science, 79: 254-261.
[12] Ali N, Bajwa KB, Sablatnig R, Chatzichristofis SA, Iqbal Z, Rashid M, Habib HA [2016]. A novel image retrieval based on visual words integration of SIFT and SURF. PloS one, 11(6): e0157428.
[13] Goldberg DE [1989]. Genetic algorithms in search, optimization, and machine learning, Addison-Wesley Longman Publishing Co., Inc.75 Arlington Street, Suite 300 Boston, MA, United States.
[14] Dueck G [1993] New optimization heuristics: The great deluge algorithm and the record-to-record travel. Journal of Computational physics, 104(1): 86-92.
[15] Foutsitzi GA, Gogos CG, Hadjigeorgiou EP, Stavroulakis, GE [2013] Actuator location and voltages optimization for shape control of smart beams using genetic algorithms. In Actuators. Multidisciplinary Digital Publishing Institute. 2(4): 111-128.
[16] Corel 10K dataset, available at http://www.ci.gxnu.edu.cn/cbir/Dataset.aspx

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

EFFICIENT QUERY KEYWORD INTERPRETATION FOR SEMANTIC INFORMATION RETRIEVAL

Sonia Setia*, Jyoti Verma, Neelam Duhan

Department of Computer Engineering, J. C. Bose University of Science and Technology, YMCA, Faridabad, INDIA

ABSTRACT

Due to vast information available over the World Wide Web and its unstructured nature it is becoming more difficult to get relevant information. Currently, information retrieval techniques are keyword based. They do not finally capture the semantic meaning of a query keyword. To overcome this problem, conceptual knowledge to information retrieval has been introduced by means of taxonomy. In this paper, a semantic information retrieval mechanism has been presented which translates query keywords to categories belonging to a taxonomy. A hybrid similarity method has been proposed which finds the closest category corresponding to a keyword using a thesaurus like WordNet. Evaluation of the proposed approach has been done for semantic based declarative querying process which shows better results in terms of precision, recall and F-measure.

KEY WORDS
Information Retrieval, Semantic-based IR, Taxonomy, categorization, keyword-based IR

Received: 15 Jan 2020
Accepted: 21 Apr 2020
Published: 3 May 2020

***Corresponding Author**
Email:
setiasonia53@gmail.com
Tel.: +91-8383007704

INTRODUCTION

In past few years, keyword-based information retrieval systems have been used to retrieve information over World Wide Web. Currently, most search engines are purely based on keyword based information retrieval. They accept query in the form of keywords and output those documents which contain the given keywords. But these search engines do not consider the semantic meaning of those provided keywords. Therefore, they provide number of false links of documents and users are not able to find relevant information. The main aim of an information retrieval process is to retrieve the relevant information corresponding to the given query. In particular, this requires understanding users' needs precisely enough to allow for retrieving a precise answer using some semantic technologies. Taxonomies appear to be useful method to allow for more semantics based search. Therefore, there is a need of translating keyword-based information retrieval to category-based information retrieval. In this paper, an approach has been presented to interpret query keywords using knowledgebase available through taxonomies. Based on presumptions about how individuals portray their data needs, proposed approach translates a keyword based query into category-based query. The evaluation of the proposed methodology has been done on queries given by few users at our institute. It uses the knowledge base of the semantic portal available at http://www.dmoz.org.in/ and displays better results in terms of precision, recall and F-measure.

A, significant work has been performed in literature to find the similarity between two elements. While these approaches claim for remarkable results, but the approach is not clear enough that how this has achieved. In fact, it is observed that users are more comfortable in keyword based search. But, it also seems important to design an approach for interpretation of keywords such that more meaningful and relevant information can be retrieved.

Chahal et al. [1] proposed a technique to compute similarity for semantic web documents that is based upon conceptual instances found between the keywords and their relationships. Authors explored all relevant relations that may exist between the keywords which explores the user's interest and based upon that determine the similarity between documents.

Formica [2] proposed a similarity measure for Fuzzy Formal Concept Analysis (FFCA), which is a general form of Formal Concept Analysis (FCA) which is used for modelling of uncertainty information. Although FFCA became very popular for semantic web development. But the problem with the given work is that manual development of ontologies is a time consuming process. Further, for constructing the fuzzy ontologies Zhang et al. [3] proposed an automated approach by using Fuzzy Object Oriented Database (FOOD) model. This way it supported the automated process for retrieval of information.

De Maio et al. [4] proposed new retrieval approach which is based on ontologies. By supporting data organization and visualization, it provides a friendly navigation model. The major challenges faced by researchers are to find the efficient techniques of sharing and searching the information with the rapid growth of web. By using the concept of Fuzzy, Kohli and Gupta [5] solved the challenges of information retrieval system. Aloui et al. [6] proposed a semiautomatic method to design and extract ontology which is based on clustering, fuzzy logic, and formal concept analysis (FCA). Authors represented the ontology as a set of fuzzy rules. Prot´eg´e 4.3 has been used to evaluate the proposed approach. Results shows that by using ontology mapping, more relevant information can be retrieved. Kandpal et al. [7] proposed a new methodology for ontology alignment. Ontology alignment is done by retrieving the similar concepts of two different ontologies. If directly concepts are not matched of two different ontologies, then similarity can be calculated of expanded terms. Major challenge is to provide accurate information of user's uncertain query words.

Rani et al. [8] proposed a hybrid retrieval system which integrates ontology and fuzzy logic concept to find information. Fuzzy type 1 has been used for documents and fuzzy type 2 has been used for words to prioritize the retrieved list.

A critical look at the literature motivates us to propose an approach to retrieve more relevant information by considering the semantics of user's query. Here, in this work, we have proposed a method which interprets the query keywords semantically in the form of taxonomy. To achieve this task, keyword to category (terms belonging to taxonomy) mapping has been done by using proposed hybrid similarity matching method. Currently, no real automatic solution has been found using knowledge-base for keyword to category mapping. So, the surveying of literature work motivates the semantic mapping of keywords by using the concept of taxonomy to retrieve more relevant information.

MATERIALS AND METHODS

This paper proposed a taxonomy-based approach for query interpretation which is on the ambition of producing more precise query from a given keyword so that more relevant information can be retrieved. Domain taxonomy has been used to retrieve more precise query in the form of categories belonging to taxonomy. Therefore, a keyword to category mapping approach has been proposed which is depicted in [Fig. 1].

1. At the first stage, users put queries for retrieving information, these queries are parsed into keywords or phrases, typically n-grams (n-gram is a n word sequence).
2. To retrieve more relevant information, these n-grams are mapped to the categories T = {c1, ..., c¬k.} of a domain taxonomy. This mapping is performed using a similarity matching method which is based on domain specific taxonomy. It uses thesaurus to find closest category corresponding to a keyword. Moreover, we used WordNet as thesaurus.
3. To better characterize the objects, weights are assigned to the keywords according to the frequency of queries corresponding to an object. Therefore, the taxonomy categories' weights are also updated. And finally, the resulted category based query is passed to Search engine for more relevant information retrieval in terms of this precise query.

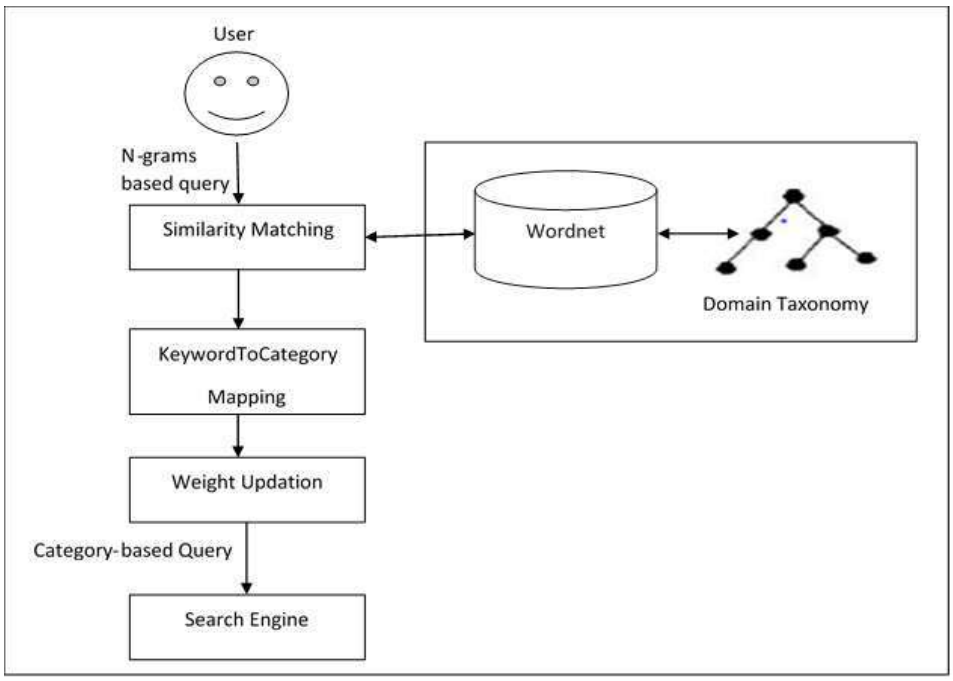

Fig. 1: Translation of keyword-based query to category-based query

Wordnet

WordNet is an online lexical reference system where, English nouns, verbs, adjectives, and adverbs are organized into Synsets. Each one is representing an underlying lexical concept. "Synsets" is a set of synonyms which represent a concept or a knowledge of a set of terms. Synsets make diverse semantic relations for instance synonymy (similar) and antonymy (opposite), hypernymy (super concept)/hyponymy (sub concept) (also known as a hierarchy/taxonomy), meronymy (part-of), and holonymy (has-a). For each keyword in WordNet, we can have a set of senses. For example, the word "wind" has eight verb senses and eight noun senses. The first sense of "wind" as a noun gives the following path:

wind –>weather; weather condition,
atmospheric condition→
atmospheric phenomenon-->
physical phenomenon--> natural phenomenon,
nature--> phenomenon

Similarity matching method

It is the major component for our semantic retrieval mechanism. Query interpretation is done by using similarity matching method. In order to find the closest category in the taxonomy T for a keyword k, we calculate the similarity through the mechanisms provided by the thesaurus.

If the keyword belongs to the taxonomy, then it is included as it is. Otherwise, most similar category is found corresponding to the keyword by using proposed Hybrid similarity method which is the integration of Type-based similarity and Path-based similarity.

Type-based similarity: If a keyword k has been defined as synonym of a category c it means keyword is directly related to this category i.e. keyword is a type of this category. Then this category is assigned to the corresponding keyword with similarity value 1 and no need to compute similarity with other categories.

Path-based similarity: If no direct synonym is found in that case, path based similarity will be computed for keyword to category mapping. Wu & Palmer similarity measure has been used to compute similarity between senses of k, Sn(k) and the categories c in T, that measures similarity between two terms. We select the pair (k,c) which is having maximum similarity and map keyword k to the taxonomy category c.

Using Wu & Palmer similarity we can compute the path-based similarity between two nodes a, b of the given taxonomy by using following formula:

$$S(a,b) = \frac{2 * depth(LCS(a,b))}{depth(a) + depth(b)}$$

where LCS is Least Common Sequence of a and b.

Breadth First Search traversal algorithm has been used to traverse the taxonomy while comparing the keywords with categories to reduce the search space. First it compares the keywords with the categories at top level of taxonomy. The category having highest similarity is explored further to find relevant sub-category. This process is repeated until most relevant category is found. Finally, keyword and category pair i.e. (k,c) pair that gives maximum similarity s has been selected. After this complete process, each keyword is mapped to a category with a similarity s respectively. Once a query has been augmented with appropriate categories it can be handed over to a search engine that is designed to pinpoint information. The challenge of the algorithm is to be able to select the right category corresponding to keywords in order to improve the information retrieval. The algorithm follows these steps:

```
Algorithm keyword Category Mapping(k, t)
1. For all sns ϵ Sn(k) do
2. For all c ϵ T do
3. sim ← max(WPsim(sns, c));
4. done
5. snscsim= max({sim});
6. cmax = c ϵ O, for which (sim == snstsim);
7. done
8. kcsim = max({snscsim});
9. category = c ϵ {cmax}, for which (sim == kcsim);
10. return(category, sim);
11. done
```

Algorithm analysis are broadly classified into three types such as.

- Best Case: If keyword is available in taxonomy then it is considered as it is for information retrieval.
- Average Case: Otherwise best possible match is found for a category which has the maximum similarity with keyword.
- Worst Case: If there is no category found for a keyword then ignore that keyword and we assume that keyword doesn't belong to our specific domain.

v) Weightage of the Resulted query. The normalized weight has been assigned to every mapped category derived from the similarity matching algorithm which can be calculated by given formula
Weight of category= w*s

Where, w represents weight of the keyword
s represents similarity value between keyword and mapped category

RESULTS

The proposed approach for the translation of query keywords with respect to a domain specific taxonomy is incorporated in our prediction system framework called Semantic Prefetching System [9] which has been intended to help a blend of search and investigation in information bases. We will presently depict a potential interaction of a user with the proposed framework. For the evaluation of the proposed approach, we have asked our colleagues at our institute to provide queries. It uses the knowledge base of the semantic portal of Dmoz [10]. Few of them were expelled which were out of scope of our domain specific knowledge base. For the evaluation, users manually allotted conjunctive queries corresponding to the natural user queries. A query produced by our approach is considered as accurate if it recovered indistinguishable answers from the hand crafted query. Few examples of the queries given by our users are shown in [Table 1].

Table 1: Translation of query to conjunctive query

User Query	Corresponding Conjunctive Query
Guitar	Stringed instrument
Techno	Dance
Karaoke	Music equipment
Veena	Stringed instrument
Flute, Sitar	Wind Instrument, Stringed instrument
Piano	Keyboard instrument
Vocoders	Electronic instrument

We evaluated the proposed approach in terms of precision, recall and F-Measure. Precision P is calculated by the number of accurately interpreted query keywords divided by the total query keywords interpreted by system. Recall R is calculated by the number of accurately interpreted query keywords divided by all the query keywords. F- measure is harmonic mean between precision and recall. In case, the query is interpreted automatically by our system, our system obtains a precision 84%, a recall 72% and F-Measure 77% as depicted in [Fig.2].

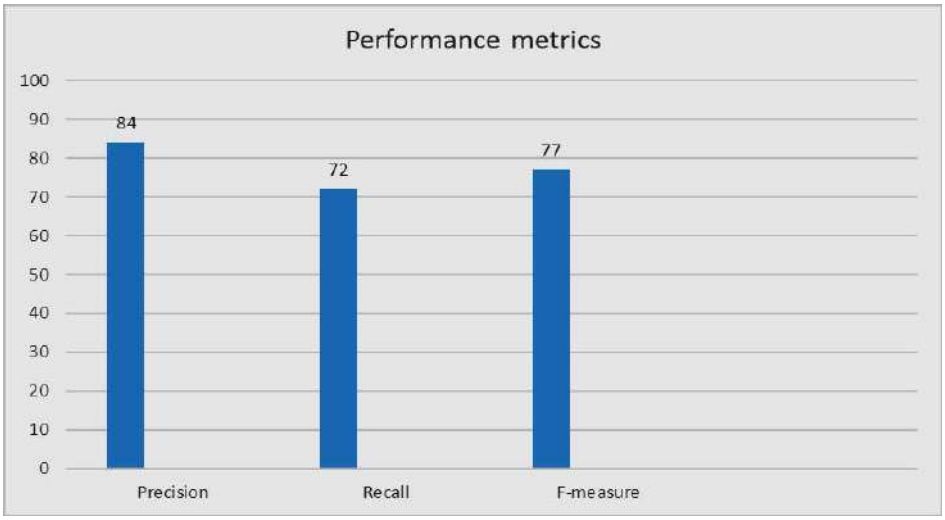

Fig. 2: Performance metrics for our proposed approach

DISCUSSION

The proposed approach has been evaluated with the knowledgebase of music domain [10]. Evaluation has been performed by our colleagues who do not have any knowledge of considered domain. They have been asked to provide general queries related to music domain. Some of the queries were obviously removed which were out of scope of music domain. Noticeably, this is a problem as this will affect the recall measure. But, here we have achieved 72% recall which means 72 out of 100 user given queries has been mapped to domain taxonomy terms which is a great achievement as compare to paper presented in [11], where, authors claimed 50% recall for their proposed approach. Precision shows that the generated conjunctive queries by our approach is correct in most of the cases which is approximately 84% which is also a large percent as compare to [11] where, authors claimed 69% precision. In short This paper proposed a novel approach for keyword to category mapping so that more precise query can be retrieved for better results. A novel hybrid similarity matching method has also been proposed which has been evaluated against few users given queries. Results shows that our approach gives 84% precision, 72% recall and 77 % F-measure. Novelty of the work has been covered in the following points:

- A novel similarity matching method between two words has been proposed, which uses Thesaurus like WordNet to find similarity. Breadth First search traversal algorithm has been used to traverse the taxonomy while calculating similarity. It is a hybrid approach which integrates Type based and path based similarity.
- A novel keyword to category mapping technique has been proposed which exploits the proposed similarity matching method to find similarity between user given query keyword and category belonging to taxonomy and finally, finds the best matched category corresponding to users given query

CONCLUSION

This paper proposed an approach for interpretation of query keywords in more precise manner. It supports the Information Retrieval system to overcome the limitations of traditional Information retrieval system so that users can retrieve more relevant information respective to their queries. It also helps to improve Hit-Miss ratio. By using domain specific taxonomy, proposed system will support information retrieval system semantically. This system can handle the semantic issues for information retrieval.
Based on the proposed approach, retrieval system can be extended to support different domains. It also can be extended to other local languages.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
I would like to express my special thanks of gratitude to my supervisors Dr. Jyoti and Dr. Neelam Duhan, it is truly an honor. Thank you for all the advice, ideas, moral support and patience in guiding me through this project.

FINANCIAL DISCLOSURE
There are no financial conflicts of interest to disclose.

REFERENCES

[1] Chahal P, Singh M, Kumar S. [2013] An ontology based approach for finding semantic similarity between web documents. International Journal of Current Engineering and Technology, 3(5): 1925–1931.
[2] Formica A. [2013] Similarity reasoning for the semantic web based on fuzzy concept lattices: an informal approach. Information Systems Frontiers, 15(3): 511–520.
[3] Zhang F, Ma Z M, Fan G, Wang X. [2010] Automatic fuzzy semantic web ontology learning from fuzzy object-oriented database model. Database and Expert Systems Applications, 6261: 16–30.
[4] DeMaio C, Fenza G, Loia V, Senatore S. [2012] Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis. Information Processing & Management, 48(3): 399– 418.
[5] Kohli S, Gupta A. [2014] A survey on web information retrieval inside fuzzy framework. In Proceedings of the Third International Conference on Soft Computing for Problem Solving, 259: 433– 445.
[6] Aloui A, Ayadi A, Grissa-Touzi A. [2014] A semi-automatic method to fuzzy-ontology design by using clustering and formal concept analysis. In Proceedings of the 6th International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA '14): 19–25.
[7] Kandpal A, Goudar R, Chauhan R, Garg S, Joshi K. [2014] Effective ontology alignment: an approach for resolving the ontology heterogeneity problem for semantic information retrieval. Intelligent Computing, Networking, and Informatics, 243: 1077–1087.
[8] Rani M, Muyeba M, Vyas O. [2014] A hybrid approach using ontology similarity and fuzzy logic for semantic question answering. Advanced Computing, Networking and Informatics, 1: 601–609.
[9] Setia S, Jyoti, Duhan N. [2019] Semantic Prefetching Based Hybrid Prediction Model. International Journal of Scientific & Technology Research, 8(12): 3936-3941.
[10] http://www.dmoz.org.in
[11] Tran T, Cimiano P, Rudolph S, Studer R. [2007] Ontology-Based Interpretation of Keywords for Semantic Search. ISWC/ASWC, LNCS 4825: 523–536.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

MATHEMATICAL APPROACH TOWARDS RECENT INNOVATION IN COMPUTATION AND ENGINEERING SYSTEM (MATRICS)

SMART DRIP IRRIGATION AND REAL TIME MONITORING SYSTEM USING IOT AND DATA ENCRYPTION ALGORITHM

Antony Vigil, W. Abelwin Pereira*, Divya Jaikanth Naicker, J. Anto Levin

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamilnadu, INDIA

ABSTRACT

Irrigation systems are a significant factor with regards to modern cultivating techniques. This paper describes an implementation that embeds all the required technologies which are necessary for agriculture. Several parameters such as temperature, humidity, moisture associated with plants, these are measured with the help of sensors. Motion sensor is used to trigger the camera .i.e. it helps in switching the camera on/off and also gives the ability to record and capture images. This model helps us to predict the crop yields. The results are taken into account to switch the motor. We use hashing techniques to encrypt and secure the data collected from the sensors and transferred through Message Queueing Telemetry Transport (MQTT). The camera is operated via open CV, The data is stored in a cloud server and encrypted with advanced encryption standard (AES). This whole system is integrated together and the stored data is used for future analysis.

KEY WORDS
Internet of things, Sensors, Open CV, MQTT, Real time monitoring, Precision agriculture

Received: 28 April 2020
Accepted: 10 June 2020
Published: 12 June 2020

***Corresponding Author**
Email:
abelwin.w@gmail.com
Tel.: +91 9080905584

INTRODUCTION

Irrigation is a very important factor when it comes to industrial farming. It helps in growing agricultural crops as well as in maintaining landscapes Water is a valuable common asset, however it covers ¾ of the earth its accessibility for human use is moderately less. Be that as it may, in contrast to power and fuel, the significance of water use and improvement in the worldwide perspective isn't that perceived [1]. Smart irrigation and other modern agricultural practices help in using the water resources in an efficient manner. Absence of motorization is viewed as one reason for the ruin of agribusiness fundamentally in India. Since the greater part of the farmlands being used are kept up by physical work of the ranchers and the labour that is as of now doing it is evaluated to diminish extraordinarily in the years to come.

IOT in straightforward terms is the aggregate system of interconnected gadgets which are implanted with or incorporate sensors, programming, organize network and fundamental hardware that help and empower them to gather and trade information making them responsive and brilliant in nature. It is one of the quickly developing fields and has been useful in conveying social and financial advantages for both the created just as creating economies [2]. The expansion in worldwide populace has brought about issues like constrained accessibility of land for agribusiness, increment in flighty climate conditions likewise constrained us to move towards savvy agrarian practices. Therefore, the utilization of IOT just as investigation of information are both utilized to upgrade the operational proficiency and produce in the farming sector [3].

Precision agriculture is a mechanical move from the deep rooted cultivating methodologies which targets advancing the results by watching a horticultural framework via automatizing, estimating and reacting stages while keeping the general control of the framework effective as far as assets utilized or vital for agriculture. It for sure opens up a wide point of view for utilization of innovations that have been effectively utilized like remote detecting [4]. Real-time sensing systems is one of the major components that are needed for precision agriculture to be widely implemented and used [5].

Smart agriculture includes integration of IOT and sensor, wireless networks. The sensor technology here helps in recording the data from different places and parameters related to plants. The parameters include moisture of the soil, temperature etc. Soil moisture is measured with the help of moisture sensor [6]. Sensor node also helps in communication. The sensor technology coupled with internet of things gives a smart and varied approach to the nature of agriculture carried out, although there are instances when the interface is a bit slow and the implementation is a bit costly [7]. Real time monitoring is done with the help of camera [8]. Precision agriculture involves real time monitoring of the field and checking parameters like soil moisture, pH of the soil, humidity level and temperature level. It involves routing protocols, wireless sensor networks [9]. A system was designed to forecast and predict the crop yield and a possibility of an attack like pest attack and virus attack. Here the significant spotlight is on the most proficient method to improve the facility and continuous observing of the rural field. The crop simulation system only helps in studying the dependencies between the environment factors and the grain yield. The work is not enhanced to analysis end to end delay and end to end output [10].

IIOAB JOURNAL
ISSN: 0976-3104

MATERIALS AND METHODS

Proposed methodology

The proposed system here aims to do precision agriculture. We collect the required data from the different sensors which help in measuring different aspects associated with the plants. The collected data is stored and is mailed to the authorized personnel using MQTT [11]. The data is analyzed and the yield is predicted accordingly, also the stored data can be used in the future as a reference [12]. Smart irrigation helps in saving water. The setup is fully automatic in nature and also integrates all the systems making it one whole system hence making it smart in nature [13]. The complexity of the existing system has been reduced by replacing complex algorithms with simpler ones. The usage of sensors and camera for monitoring the parameters helps in making a smarter and an easy to understand setup. This not only enhances the age-old farming method but also paves a path towards smart irrigation as well as farming in a smarter and simpler manner [14]. Sensors have been employed to collect data. The sensors used here are temperature, humidity and moisture sensors. Temperature is estimated in Celsius and moistness and dampness is estimated in percentage. The gathered information is put away. The data from both the sensors as well as the camera is collected and stored. The data from the sensors is hashed using the hashing technique. The hashed data is thereafter sent via MQTT, That is the data is mailed to authorised personnel [15]. Camera is used with the help of open CV for the purpose of real time monitoring. The data collected through the camera is encrypted using the data encryption algorithm. The data collected is then understood, analysed and manipulated accordingly [16]. The analysis result is then used for precision agriculture. The data thereafter is mailed and stored in the cloud for future reference or analysis (refer [Fig. 1]).

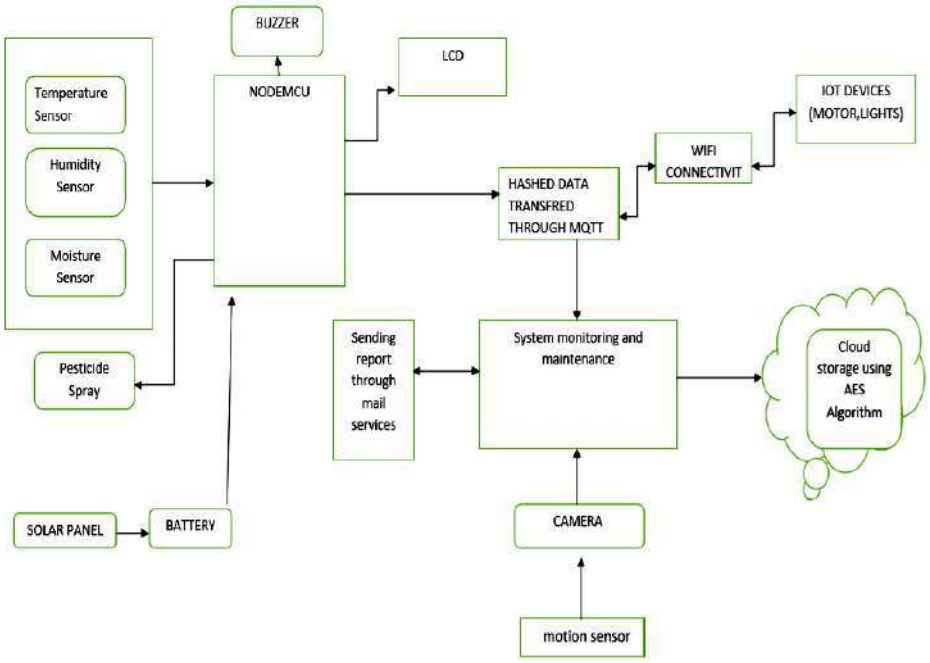

Fig. 1: overall architecture of system and design

Hardware requirements

a) Humidity sensor: This sensor is used here to monitor the humidity variation of the environment where the crops are cultivated. This is a digital sensor and measures the humidity value in percentage format.
b) Dampness sensor: The dampness sensor is utilized to gauge the water substance or dampness level in the dirt. At the point when the water levels in the dirt are low, the module yield is at elevated level, else the yield is at low level.
c) Temperature sensor: The temperature sensor measures the temperature of the environment based on the obtained value the controller controls the pump.
d) Motion sensor: This sensor is used to detect the presence of any motion i.e. to check whether any animals or humans are in the field and after that it triggers the camera. It has a 360-degree viewing angle [17].
e) Camera: High resolution camera with night vision is required for capturing images and recording the videos [18].
f) IOT devices such as smart motor pumps and smart lights. A buzzer is used to intimate about the water level or presence of any animals by producing sounds [19].

IIOAB JOURNAL
ISSN: 0976-3104

Modules description

a) **Data Collection:** Data is collected from sensor which holds parameters such as humidity, temperature, moisture. The pictorial data is captured from high definition cameras. These data is used for various operations. The temperature sensor gives information about the temperature it is measured in degree Celsius or Fahrenheit. This gives us information about the temperature in the field. The humidity sensor gives the data in percentage which provides the information about the water vapour content present in the field; this estimates the amount of water to be sprinkled. The moisture sensor gives the data about the moisture content present in the soil, which helps in concluding the amount of water to be sprayed. Through open CV we are able to run the camera which helps us to monitor and obtain the pictorial data from the camera in the field all the time, here the camera is triggered with the help of motion sensor [20].

b) **Preprocessing**: The data from the Node MCU is hashed with simple hashing methods to provide a minimal level of security to the data. This data is further transferred through MQTT which help to control IOT devices such as motors and lights. There is an option to add more IOT devices for future purpose. Here the major emphasis is on precision agriculture and monitoring plant parameters with the help of sensors. The maintenance and operational computer are connected with the circuit to take control over the system which could be useful in case of emergency. The levels of water to be fed is decided and predicted by the data collected from the humidity and moisture sensors the reason to use humidity alone with moisture is to predict the weather and reduce the water usage. The data from camera is processed, the camera here runs on open CV. All the data collected from the sensors is transformed into a graph using an algorithm which runs on Mat lab and a report is generated and sent to the mail id provided. Only unauthorized-personnel can access the stored data hence ensuring data security. Adding a feature like buzzer to intimate the famer about the water level and LCD to check the working of the system (refer [Fig. 2]) [21].

c) **Data storage and analysis**: The processed pictorial data is further stored in cloud with which it is further stored in the cloud and secured by Advanced Encryption Standards (AES) but in this project we use only level 1 because in agriculture irrigation and monitoring is important than security and unwanted access is also restricted. Here we use asymmetric data encryption standards as it is more viable (data encryption). These data can be retrieved later for research purpose and analysis, the data from the sensor is stored in the mail with appropriate information or metadata [22].

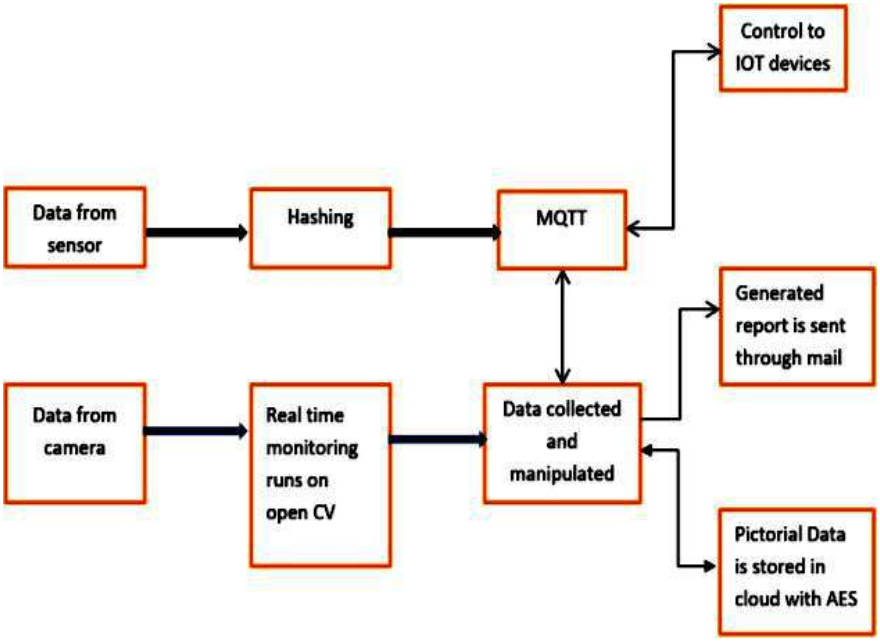

Fig. 2: data flow diagram

RESULTS

The graph in [Table 1] depicts the readings recorded by the three sensors namely humidity, temperature and moisture sensor. The graph is the result analysis of recordings recorded over the days. The x axis represents the time and the y axis represents the readings [Fig. 3]. The temperature is recorded in celsius. Dampness and moisture are represented in rate. The temperature was around 31 degrees it rose by 5 degrees and reached about 36 degrees on the mid time of the day. Humidity on the other hand decreased and reached to 34.48% as compared to beginning of the day. Moisture percentage differs by the amount of water fed into the fields and by the dew collected by the soil in the morning [23].

Table 1: Sensor readings in tabular format

S.NO	TIME	Temperature (in degree celsius)	Humidity (in percentage %)	Moisture (in percentage%)
1	1:00am	27	56	50
2	5:00am	26	54	55
3	9:00am	29	46	45
4	1:00pm	32	50	35
5	5.00pm	29	52	60
6	9:00pm	28	54	55
	AVERAGE	28.5	52	50

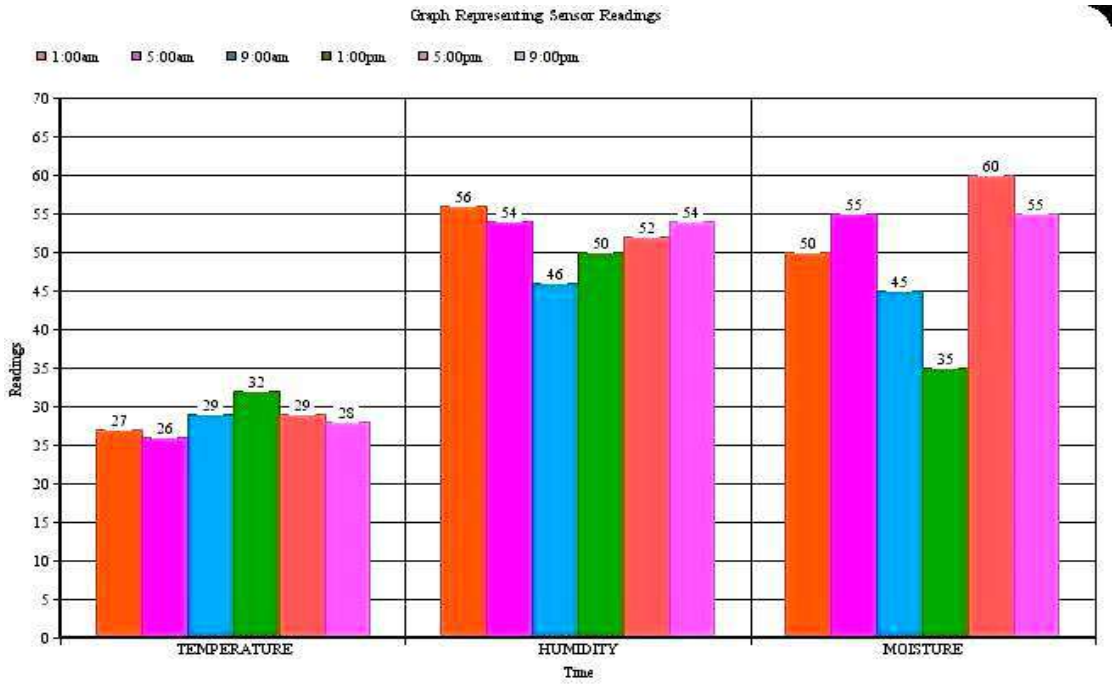

Fig. 3: Graph representing sensor readings

CONCLUSION

Developing an android application which in turn can be used for real time monitoring as well as for remote accessing and other uses. The android application can be designed in such a manner that the controls can be done from a faraway place like the house of the farmer. The monitoring system can also in turn be added for monitoring the plants in other fields. Advanced algorithms, like neural artificial intelligence algorithms can be implemented [24,25]. Hence forth, embracing computerization and more brilliant approaches to improve the manner in which we develop crops as well as harvests can keep the nation from confronting a dim time of emergency when the manual strategies will in general miss the mark out to nowhere. At this moment, if there should arise an occurrence of little and minimal homesteads, the squandered human work is extremely high and the yields are seen as low per capita work power. This makes the circumstance of the rancher right now more terrible than before since the cash produced using the effectively low yield is currently part to be given to the workers, henceforth, leaving nothing to the rancher. With keen and propelled strategies, for example, the model proposed here, the work power can be cut by an incredible sum without diminishing the yield, however expanding it enormously. This proposed model can be utilized for different purposes joined with the correct parts and coded with the necessary programming dependent on anything the rancher requires.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
There are no financial conflicts of interest to disclose.

REFERENCES

[1] Chahal NS, Vinit V, Aakanksha G, Darshan S. [2017] Optimization of water consumption using dynamic quota based smart water management system, 2017 IEEE Region 10 Symposium (TENSYMP), IEEE,, doi:10.1109/TENCONSpring.2017.8070075.
[2] Raghu K, Sitaramaraju G, Harish KV. [2017] Knowledge based real time monitoring system for aquaculture using IOT, doi: 10.1109/IACC.2017.0075.
[3] Olakunleelijad. [2018] Student member, IEEE, Tharek abdul rahman, Member, IEEE, Iqbafeorikumhi, Member, IEEE, Chee yen leow, member, IEEE, MHD nourh india, member, IEEE, an overview of internet of things and data analytics in agriculture: benefits and challenges, doi: 10.1109/JIOT.2018.2844296.
[4] Dmitriishadrin, Alexander M, Maxim F, et al. [2019] Enabling precision agriculture through embedded sensing with artificial intelligence, IEEE Transactions on Instrumentation and Measurement, IEEE, doi: 10.1109/TIM.2019.2947125.
[5] So PM, Julius G. [2015] Precision agriculture: challenges in sensor and electronics for real time soil and plant monitoring, 2017 IEEE Biomedical Circuits and Systems Conference (BioCAS), IEEE, doi: 10.1109/BIOCAS.2017.8325180.
[6] Mancuso M, Bustaffa F. [2006] A Wireless Sensors Network for Monitoring Environmental Variables in a Tomato Greenhouse. In Proceedings of the IEEE International Workshop on Factory Communication Systems, Torino, Italy, 27(30):107-110.
[7] Georgiou O, Raza U. [2017] Low Power Wide Area Network Analysis: Can Lo Ra Scale?, in IEEE Wireless Communications Letters, 6(2);162-165.
[8] Balamurali K, Kathiravan K. [2015] Analysis of various routing protocols for precision agriculture using wireless sensor networks, 2015 IEEE Technological Innovation in ICT for Agriculture and Rural Development (TIAR), IEEE, doi: 10.1109/TIAR.2015.7358549.
[9] Zhuang J, Xu S, Li Z, Chen W, Wang D. Application of intelligence information fusion technology in agriculture monitoring and early- waiting research, doi:10.1109/ICCAR.2015.7166013.
[10] Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M. [2015] Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications, in IEEE Communications Surveys & Tutorials, 17(4):2347-2376.
[11] Chiyurl Y, Miyoung H, Shin-Gak K, et al. Implement smart farm with IOT technology on Information and communication network technology. doi: 10.23919/ICACT.2018.8323908.
[12] Estrin D, et al. [2001] Instrumenting the World with Wireless Sensor Networks, Proc Int'l Conf. Acoustics, Speech, and Signal Processng (ICASSP 2001), 4, IEEE Press, doi:10.1109/ICASSP.2001.940390.
[13] Jose A. [2004] Gutierrez, Ed. Callaway, and Raymond. Barrett, Low-Rate Wireless Personal Area Networks: enabling wireless sensors with IEEE 802.15.4, IEEE Press.
[14] Beckwith R, Teibel D, Bowen Jones P. Report from the Field: Results from an Agricultural Wireless Sensor Network, proceedings of 29th IEEE LCN'04, Tampa, Florida, doi: 10.1109/LCN.2004.105.
[15] Burrell J, Brooke T, Beckwith R. [2004] Vineyard computing: sensor networks in agricultural production. IEEE Pervasive Computing, 3(1):38–45.
[16] Gomez C, Paradells J. [2010] Wireless home automation networks: A survey of architectures and technologies, IEEE Commun Mag, 48(6):92–101.
[17] Wang N, Zhang NQ, Wang MH. [2006] Wireless sensors in agriculture and food industry-Rencent development and future perspective, Computers and Electronics in Agriculture, 50(1):1-14.
[18] Zhou Q. [2004] status and tendency for development in remote sensing of agriculture situation, Journal of China Agricultural Resources and Regional Planning, 25(5):9-14.
[19] Guo Z, Chen P, Zhang H, Jiang M, Li C. [2012] IMA: An integrated monitoring architecture with sensor networks, IEEE Trans. Instrument. Meas, 61(5):1287–1295.
[20] Corbellini S, Parvis M. [2016] Wireless sensor network architecture for remote non-invasive museum monitoring, in Proc Int Symp Syst Eng (ISSE), Edinburgh, UK, 34– 40.
[21] Nakutis et al., [2015] Remote Agriculture Automation Using Wireless Link and IOT Gateway Infrastructure, 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), Valencia, 99-103. doi: 10.1109/DEXA.2015.37.
[22] Xu J, Zhang J, Zheng X, Wei X, Han J. [2015] Wireless Sensors in Farmland Environmental Monitoring, 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Xi'an, 372-379. doi: 10.1109/CyberC.2015.17.
[23] Meonghun L, Jeonghwan H, Hyun Y. [2013] Agricultural Protection System Based on IOT, IEEE 16th International Conference on Computational Science and Engineering, doi: 10.1109/CSE.2013.126.
[24] Monika J, Ashwani K, Rushikesh B. [2013] Image Processing for Smart Farming: Detection of Disease and Fruit Grading, IEEE Second International Conference on Image Information Processing (ICIIP), doi: 10.1109/ICIIP.2013.6707647.
[25] Liu C, Ren W, Zhang B, Lv C. [2011] The application of soil temperature measurement by lm35 temperature sensors, International Conference on Electronic and Mechanical Engineering and Information Technology, 88(1):1825– 1828.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

EXPERT OPINION

ANALYTICS: FUTURE OPERATING MODEL

Aakash Chhablani*

Manager, Deloitte Consulting, 38921 Snapdragon Pl, Newark, CA 94560, USA

ABSTRACT

Today's fastest-growing companies have one thing in common—they harness technology and innovation to their advantage. Cloud, analytics, artificial intelligence (AI), machine learning (ML), and faster connectivity fuel business model disruption and industry transformation. Yet for every shining star, we see a struggling one. Newspapers, travel agencies, taxis, film photography, landline phone service, cable television, video rentals, they too invested in data management and technology but failed to capitalize on enormous volumes of data. Hence, the advent of analytics and requirements for data insights have become important - organizations are investing heavily on data analytics to use analytics as an advantage in the future by identifying and shaping customer preferences. These organizations expect to achieve greater business results by spending on data-focused initiatives like other core business functions. In addition, role of analytics is quickly transforming and emerging as leaders are tasked with enhanced executive decision making, improving operational efficiency and empowering innovation. These leaders are looking to use analytics to evolve from classical data management to a data advantage perspective. This paper highlights the key trends emerging in analytics space and how companies can evolve their current modus operandi when it comes to data analytics and use it to their competitive advantage to stay ahead in the marketplace.

KEY WORDS
Analytics, data, services operating model; data management;

Received: 16 May 2020
Accepted: 19 June 2020
Published: 21 June 2020

***Corresponding Author**
Email:
achhablani@deloitte.com
Tel.: +1-650-804-3426

INTRODUCTION

In Organizations across the world are noticing that data has emerged as a key asset and are investing in analytics to glean customer and operational insights from their data. Fifty nine percent of respondents to Deloitte's 2018 Global CIO Survey identified data as their organization's top focus area [1]. Regardless of the industry, data analytics can enhance and amplify what the organizations can see and do. In the era where waiting to see how things shake out is no longer a viable option, organizations are increasingly relying on their current data to predict future. There are three emerging trends driving the increasing need for data analytics. First and foremost is the Data growth and proliferation over last few years [2]. Rapid digitization of processes, the physical-digital-physical loop of data, and digital exhaust from intelligent products are creating high volumes of siloed data.(IDC predicts the total volume of data stored electronically in 2020 to be around 44 zettabytes; the same value was 33 zettabytes in 2018 [3]. This data, proliferating across the value chain in customer relationship management (CRM); configure, price, and quote (CPQ); enterprise resource planning (ERP); and other on- premise and cloud enterprise systems, creates both opportunities and risks for organizations. Combing through it enables companies to develop more targeted products and services, enhance feature sets, offer rich customer service, and more. However, the data's sheer volume and siloed nature often leads to multiple sources of truth, lack of trust in data and metrics, and significant security risk [4].

Second trend driving the need for advanced analytics is cloud and flexible consumption. Deloitte CIO survey respondents stated that they plan to dedicate nearly half of their IT spending to cloud, which represents an increase of more than 20 percent from the previous year [5]. CIOs also noted that organizations are shifting from on premise models to flexible consumption models in almost all cloud technology use cases to enable a growth agenda, improve business agility, and increase scalability [6]. Cloud computing is emerging to be a force multiplier in the data and analytics space to create more opportunities for enterprises [7].

Finally, future of work is changing. Accelerating connectivity and increasingly powerful cognitive tools are changing the nature and future of work. Cloud, automation, and AI are making it easier for business professionals to scout, implement, and maintain technology without intervention from IT staff and, in the process, redefine organizational roles and responsibilities. For example, by using prebuilt cloud solutions, business professionals in multiple industries are managing real-time exceptions of order price with minimal IT staff assistance. The advancement of borderless technology is also changing the perception of IT as the owners of an organization's data [8]. As analytics continues to be pushed out from behind the IT curtain into the business, technology SMEs' primary role is transitioning from developing reports to facilitating provision of clean, accurate, and secure data "on the fly" to business SMEs to accelerate decision-making.

EVOLUTION TOWARD AN ANALYTICS SERVICES OPERATING MODEL

Harnessing is the emerging trends of data growth and proliferation, cloud and flexible consumption, and future of work thank can generate rapid changes in organizations' structure, operations and processes. This also enables organizations to transition towards the analytics-as-a-service model by shifting from a product to a services operating model [Fig. 1]

IIOAB JOURNAL
ISSN: 0976-3104

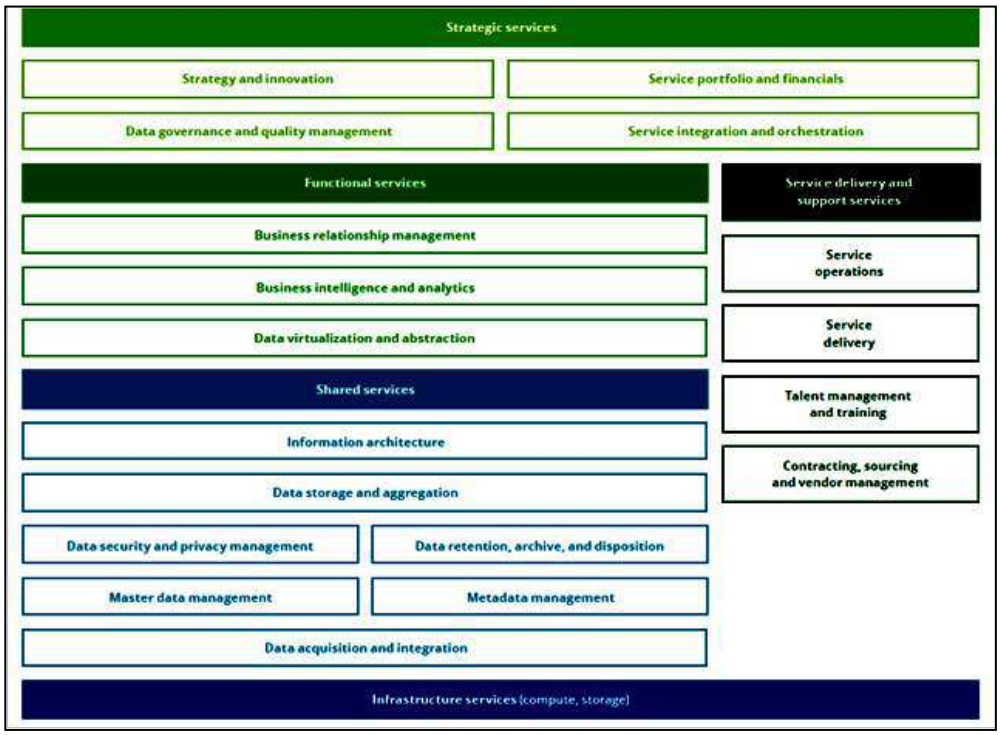

Fig. 1: Proposed analytics services operating model of the future

The analytics operating model of the future spans strategic, functional, support, shared, and infrastructure services. Strategic services include developing and refining the strategy for data management and analytics and conduct research on emerging trends in the market, technology, process, industry, and people. In addition, these services can help manage initiatives and opportunities as a formal investment portfolio. Moreover, this functional will maintain an up-to-date service catalog and manage budget, chargeback, SLAs and KPIs. Next functional services arm will work jointly with business functions to understand their strategy and priorities. This arm also assists business functions to triage, prioritize and achieve their full analytics goals. Operationally, this arm can provide reports, self-serve analytics, visualization processes and tools to automate the generation of analytics and insights. Service delivery and support services provide service design and development and release management execution. This function also ensures services operations and support in terms of event such as Covid-19 [9]. Additionally, this services function procures vendors to support delivery of services within the organization, recruit, retain and manage performance of talent within the organization. Shared services ensure data quality, security and privacy. This pillar also helps support information architecture and data modeling. Finally, the infrastructure services act as a backbone of the operating model and includes computing, storage and routine maintenance activities.

POTENTIAL BENEFITS

According to Deloitte's Global CIO Survey, organizations are using digital technologies and capabilities to transform business operations (69 percent) and drive top-line growth through improved customer experiences [10]. Cloud-based data platforms, coupled with service-based analytics operating models, can support these objectives by:

Enabling data democratization

To help businesses shift their value proposition from products to ongoing, data-driven services - From R&D and sales to account management and aftermarket services, cloud-based AI/ML capabilities create opportunities to improve efficiency and enhance customer experiences, helping organizations attract and retain customers, as well as drive significant, service-driven value [Fig. 2]. For example, one $4 billion storage client shifted to outcome-based services via digital adoption and to an agile mindset through a common data platform, enabling shared business alignment on "metrics that matter."

Supporting cloud-driven customer centricity

By eliminating functional silos to establish a frictionless customer experience - Customer journey–stage KPIs built on a cloud-enabled, unified data platform can facilitate a 360-degree view and a culture of customer-centricity. For example, a global communications network company started its shift to a customer data

strategy as-a-service model by using a defined customer data management platform to foster personalized and contextual engagement, with the aim of driving growth in customer acquisition and lifetime value [11].

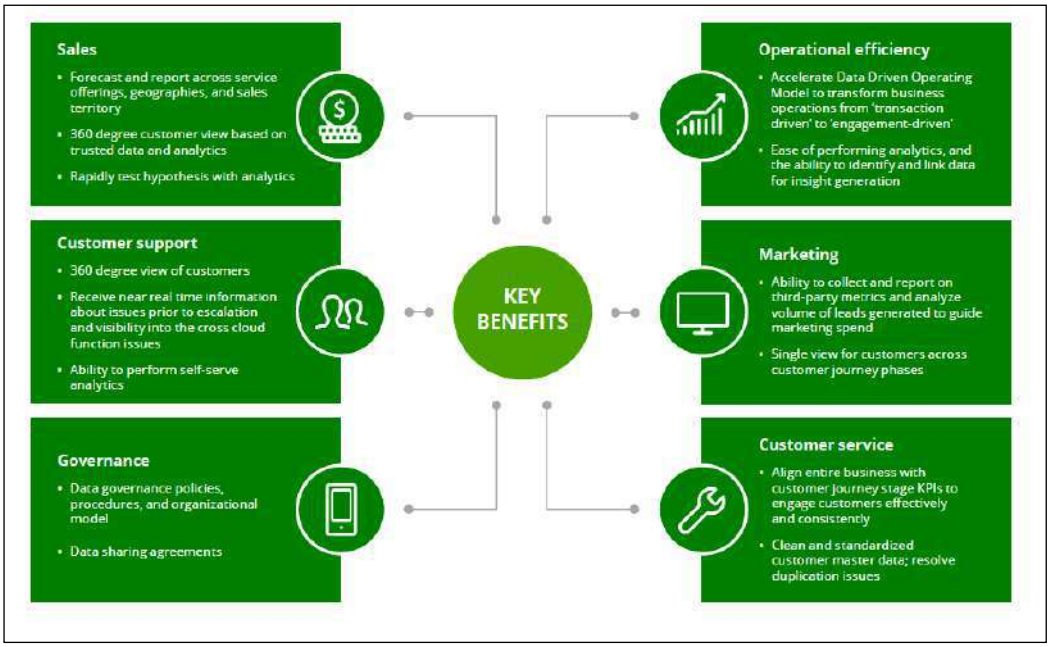

Fig. 2: Customer and operational benefits

CONCLUSION

The boundary between business and technology issues is blurring, accelerating organizations' move toward an analytics operating model through cloud adoption and an evolving business-IT construct. This model is a key foundational element to help their organizations harness emerging trends, develop actionable insights, and deliver results and value more quickly to business and IT stakeholders.

As with any major change, the transition to the analytics operating model of the future requires a shared vision among key leaders, early identification and engagement of the right sponsors, and setting bold yet achievable short-term goals. This transition will help organizations realize their long-term vision of using data-as-a-service and data-as-a-strategic asset for competitive advantage in tomorrow's world. Leaders will find plethora of use cases in their business functions where change in move to predictive analytics operating models will yield significant benefits and reducing risks in the long run. Moreover, the redefined operating model for analytics will create a predictive environment to have the foresight to identify and respond to any potential crisis and challenging situations such as Covid-19 ensuring minimal disruption in operations [12]. Finally, this journey of analytics modernization will require embedding analytics and data driven culture in all aspects of business decision making. This stage will be the ultimate nirvana of data advantage where leaders will be able to use data to shape their organizations future. The vision they establish and day-to-day work practices they instill and reinforce can determine how the organization's culture evolves and whether it supports or prevents the operating model from delivering on the promise of the joint business-technology strategy and data advantage.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
There are no financial conflicts of interest to disclose.

REFERENCES

[1] Deloitte Survey Insights [2018]: Global CIO Survey: Manifesting Legacy

[2] Stainslaw K. [2018] Beyond databases, architectures and structures

[3] Article: International Data Corporation (IDC), The digitization of the World: From Ede to Core November [2018]: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[4] Evan W. [2011] Security Risk Management, 40-53

[5] Deloitte Press box Khalid K. [2020] https://www2.deloitte.com/us/en/insights/topics/leadership/global-technology-leadership-study.html
[6] Todd H. [2011] Consumption Economics: The new rules of the tech
[7] Article: Cloud computing paradigm Sandeep M. [2014] https://www.forbes.com/sites/oracle/2014/01/09/cloud-computing-is-a-force-multiplier-for-emerging-markets/#6a80dcf81d18
[8] Article: Technology strategy in Borderless world [1999] https://www.inderscienceonline.com/doi/abs/10.1504/IJTM.1999.002741
[9] Forbes Insights: Louis C. [2020] Covid-19 and analytics spending https://www.forbes.com/sites/louiscolumbus/2020/05/10/how-covid-19-is-changing-analytics-spending/#61a6163011cf
[10] Deloitte Article Analytical Insights: Richard S. Navin W, Anuj S, Payal S. [2020] Operational Models www2.deloitte.com/global/pages/analytics/insights/2020/11
[11] Deloitte Article Analytical Insights: Richard S. Operational Models Richard S, Navin W, Anuj S, Payal S. [2020] www2.deloitte.com/global/pages/analytics/insights/2020/11
[12] Article Alteryx: Andy D. Use of analytics to minimize disruption. [2020] https://www.alteryx.com/input/beyond-covid-19-how-to-leverage-ai-and-analytics-to-minimize-supply-chain-disruption-in-the

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

AN ENHANCED OPTIMIZATION APPROACH FOR IMPROVING CLASSIFICATION ACCURACY IN DATA MINING

Chidambaranathan Krubakaran[1*], Kaliyappan Venkatachalapathy[2]

[1]*Dept. Of Computer Science Engineering, Annamalai University, Annamalai Nagar – 608002, Chidambaram, Tamilnadu, INDIA*

[2]*Dept Of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar - 608002, Chidambaram, Tamilnadu, INDIA*

ABSTRACT

This work presents an enhanced version of JAYA minimization algorithm to solve the issues in data mining classification approached with different objective functions that reflect in accuracy, reliability, cost, and efficiency improvement. The proposed enhanced JAYA algorithm was obtained by modifying the equations that is used in obtaining the best and worst solutions. Formulation of issues with respect to reliability and efficiency is highlighted in this paper to convert the multi-level objective into mono level objective function. To prove the advantages of proposed EJAYA algorithm two different test sets are used and results are obtained for different scenarios. Conventional genetic algorithm is compared with the obtained results to show the superiority of the proposed EJAYA algorithm with different complexities.

KEY WORDS
Data mining, EJAYA, Classification accuracy, Reliability

Received: 5 June 2020
Accepted: 15 July 2020
Published: 17 July 2020

INTRODUCTION

Data mining is the process of extract information from the implicit unknown information source through classification and learning process [1]. Using computer programmes the similarities and dissimilarities and its patterns are classified and organized automatically to form a data set. This useful information helps in research to obtain better results which are applicable in many fields such as big data, medical data processing and other applications. Most of the data classification process depends on the learning process to obtain the data automatically. Using general concept learning the concept learning task is obtained in machine learning process. It categorizes the instances into positive class and negative class by train the instances and then groups the information. Using Boolean valued function these two classes are obtained [2]. The general format of concept learning deals with more than two classes of instances to obtain the information from the training instances. Based on the classified results the models are selected. Precisely based on the positive and negative instances the new unknown is compared to identified and grouped into that respective instance. This kind of learning process is given as supervised learning as the class membership of the instances are known. In unsupervised learning the training instances doesn't know the classes so the instances are grouped through data analysis [3]. Unsupervised learning is derived from the supervised learning to make use of information class and the two-step strategy is followed to obtain the class information. Fig. 1 gives an illustration about data mining process [4].

To evaluate the precision of classified data, performance evaluation through classification algorithms is generally used. In this the real data is split into two data sets such as test samples and train samples [5]. In this training samples are used to obtain the learning model and test samples are used to evaluate the accuracy of the designed application [6]. In this process the test samples are given to the model with hidden classes and then it predicts the class labels. Once it predicted the classes then it is compared with respective class labels. This evaluates the prediction accuracy of the designed application. In the comparison if two labels are same then the prediction result is considered as success as positive otherwise it is considered as error or negative [7]. Calculating the error rate is an essential factor in performance evaluation since it defines the proportion of errors obtained for the whole set [8]. This error rate on test samples are meaningful and it is used to evaluate the model similarly the error rate on training samples are useful to know since the model is derived from the same. Fig. 2 depicts the data classification process as two steps which include train data and test data sets [9].

In the unsupervised learning technique data clustering is recognized as most prominent process. In this a given dataset is categorized into likeness and disparity metric to group the useful information from the raw data set [10]. A conventional clustering algorithm requires assumptions that include the cluster structure group and adaptable objective function [11]. The natural paradigm to accommodate the data in the feature space and obtaining the exact number of partitions for a single objective function through clustering algorithm is essential. Also, it is required to estimate a combined solution which is stable and lower sensitive to noise. Similarly, multi objective clustering through multi objective optimization aims to provide several trades-off with numerous objectives [12]. It aims to cluster the data set into comparable groups to obtain the multi objective function. But it has limitations under specific conditions to apply. Conventional and metaheuristic techniques are available to optimize the required functions [13].

***Corresponding Author**
Email:
kirubabcet2018@gmail.com

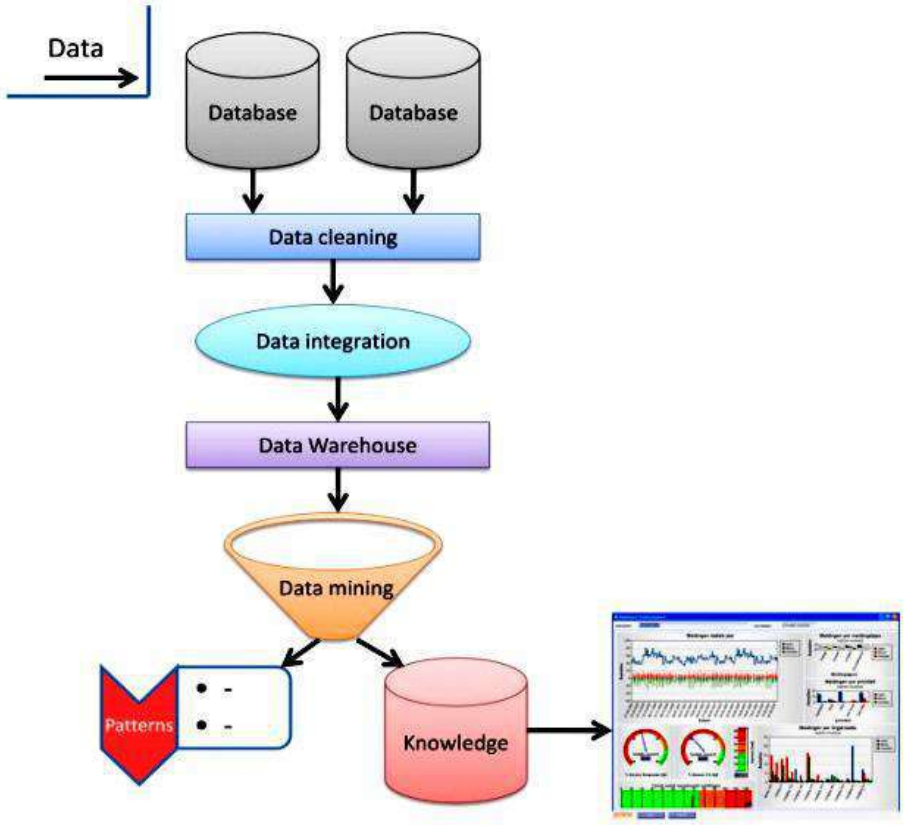

Fig. 1: Data mining process flow

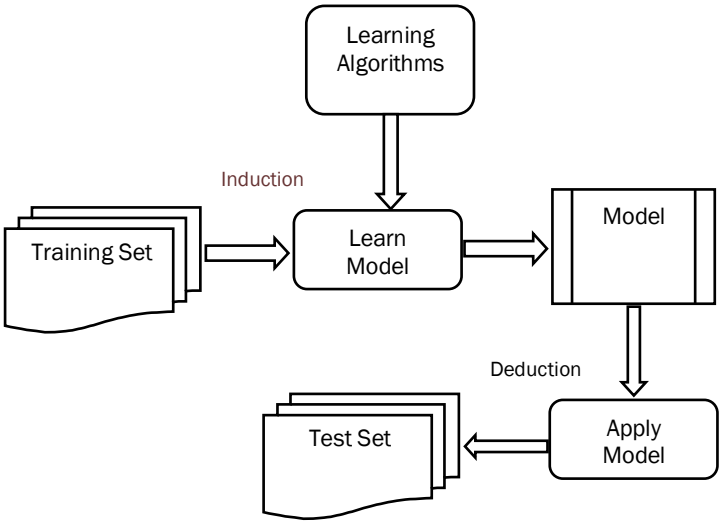

Fig. 2: Data Classification Process

Recent research interest is moved towards to optimization to improve the performance of classification results either it may be supervised or unsupervised model. Majority of the issues require this optimization strategy to improve the multiple objective and computation methodologies. Due to its simplicity and effective calculation to obtain precise results optimization involved in most of the applications now a days. Advantage of this optimization models are based on the problem concept we can modify add and remove the components to achieve better results. Also, these are robust and powerful search procedures which portray the solutions by selecting, segmenting and rearranging the set of various solutions to obtain new solution. This ideal solution increases the research interest in the field of data mining with multi objective optimization which helps in resolving many complex issues. This research article aims to provide an optimized clustering algorithm through Jaya minimization algorithm in an enhanced version by modifying the intra cluster likeness and linter cluster likeness functions. Using large set of data and k determination which is suitable for data sets and cluster validity indices, convergence solution is obtained.

METHODS

Real time data set was created based on the data array observed from four different classes of people which include 50 males and 45 females out of which 25 are elderly in the age group above 50 and 70 are in the age group of 25 – 50. Every two hours the instances for the subjects are observed for activities like sitting, walking, standing, jogging and running. The proposed method has been compared with a sequential model [2] for performance comparison.

The performance of data mining depends on the classification accuracy and clustering parameters. This proposed model provides an enhance version to improve the clustering accuracy and efficiency of the data mining approach. The mathematical model of JAYA minimization is presented in this section for better understanding of optimization in data mining model. JAYA minimization algorithm is developed to resolve optimization issues under constrained and unconstrained problems. It avoids the worst result scenario by obtaining optimal solution for the issues in data mining applications. By achieving accomplishments while obtaining optimal solution it try to attempt long way from the worst solution. This helps to move the genuine solution for the issues. It is mainly application perspective approach and the advantage over conventional optimization model is its unrestricted parameter selection. So that by selecting any two common parameters this algorithm operates. Population size and number of iterations are the two important parameters highlighted in JAYA optimization models.

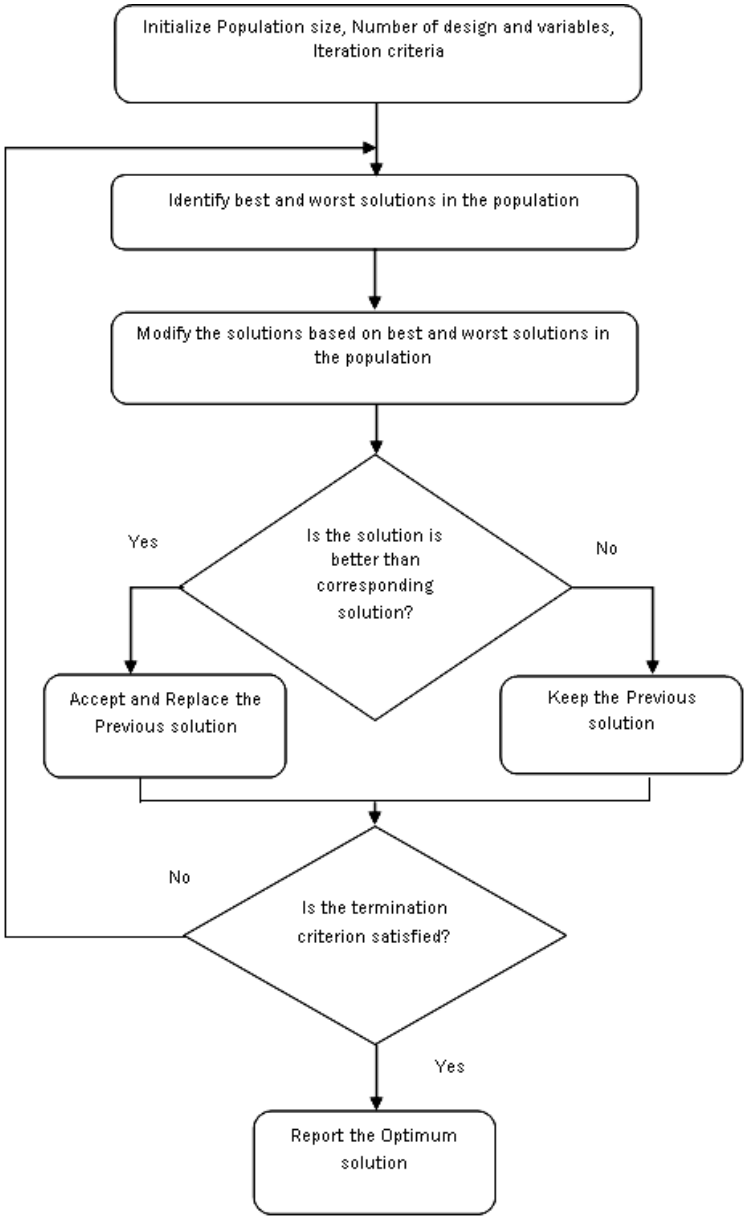

Fig. 3: JAYA minimization algorithm process flow

Since the conventional models differs by selecting different parameters for each application the advantage of JAYA implies in its calculation abilities by ignoring the changing environment constrains and reduces the optimization issues. This helps to implement JAYA minimization model in many real time applications which has complex and large heterogeneous datasets. It doesn't need any specific algorithmic parameters and even it doesn't need tuning parameter before the optimization process. Accepting the best solution and rejecting the worst solution based on the design criteria is main factor in JAYA algorithm. Fig. 3 gives the general illustration of JAYA minimization model flow process which is used in the proposed model.

Let $f(x)$ is the objective function which requires minimization or maximization. Let us assume the iteration as i and the maximum number of variables are n in numbers. The candidate solution is defined as m which is the actual population size. Let the variables be $j = 1,2,3 \ldots n$ and the population size be $k = 1,2,3 \ldots m$. For obtaining the $best$ solution from the entire data it is given as $f(x)_{best}$ and the worst solution is given as $f(x)_{worst}$. If the $X_{j,ki}$ is the value of j^{th}variable for the k^{th} set while it performs the i^{th}iteration then the value of modified function is obtained as

$$X'_{j,ki} = X_{j,ki} + r_{1j,i}(X_{j,best,i} - |X_{j,ki}| - r_{2j,i}(X_{j,worst,i} - |X_{j,ki}|)) \quad (1)$$

Where $X_{j,best,i}$ is the value of $best$ solution for the j variable and $X_{j,worst,i}$ is the worst solution for the j variable. The update value is given as $X'_{j,ki}$ which includes two random functions for the j^{th} variable when i^{th} iteration is in progress. The range of iteration is [0,1] and the random variables are given as $r_{1j,i}$ and $r_{2j,i}$.

The tendency to obtain the solution which is closer to best solution is obtained from the difference values $r_{1j,i}(X_{j,best,i} - |X_{j,ki}|)$ and the tendency to avoid the worst solution is obtained from the other values $r_{2j,i}(X_{j,worst,i} - |X_{j,ki}|)$. Once the best value is obtained the factor $X'_{j,ki}$ accepts the best solution and this becomes the input for the next iteration.

The absolute value for the solution helps to enhance the ability of the algorithm and this operation considers the value of $X'_{j,ki}$ to reflect the allowable values. If the value is exceeding the corresponding upper and lower boundaries then the desired operation is represented as

$$X'_{j,ki} = \begin{cases} 2x'_j - x'_{j,ki} & if\, x'_{j,ki} < x'_j \\ 2x''_j - x'_{j,ki} & if\, x'_{j,ki} < x''_j \\ x'_{j,ki} & otherwise \end{cases} \quad (2)$$

Based on the value of objective function the counter past of individual target and the vector is compared to obtain the best functional value. Otherwise the target vector is retained with the same population as

$$X'_{j,ki} = \begin{cases} x'_j & if\ f(x'_j) \le f(x_j) \\ x_j & otherwise \end{cases} \quad (3)$$

Where, f is the cost function which used to minimize the functional parameters.

The pseudo code for the proposed EJAYA algorithm is given as follows in a summarized manner

---

Initializepopulation size,
Initialize number of designs, variables and meeting criteria,
Initialize number of fitnessfunction evaluations
Analyse the fitness function value for each candidate;
Categorize into best and worst solution
Fitness function =population;
While fitness function <Max_fitness function
do
Select the best candidate xbest and the worst candidate xworst from the population;
For i = 1 to population
do
Select the fitness function value for the updated candidate;
Fitness function = population + 1;
Accept the new solution if it is better than the old one
End for
End while.

RESULTS

The proposed model optimization behavior is experimented and compared with sequential algorithm. Experimentation is performed by implementing C language in GCC v.4.8.5 compiler. The platform is composed of two nodes and each node is comprised of x5660 processors which has processing core frequency of 2.8 Ghz with infinite communication band. The simulation parameters are given in Table 1.

Table 1: Simulation parameters

S.No	Parameter	Value
1	Population	512 and 1024
2	Iteration parameter	25000
3	Number of runs for each node	30 and 40

The speedup function for the proposed model and sequential model is given in Fig. 4 and Fig. 5 with different number of runs. From the results it is observed that proposed model achieves better speedup function in both run values compared to another model. The function values are used to calculate the speedup values and it gradually increases for each run and reaches a maximum of 10 for the last value. While the proposed model achieves more speedup and another model achieves 8.56 for their last function value.

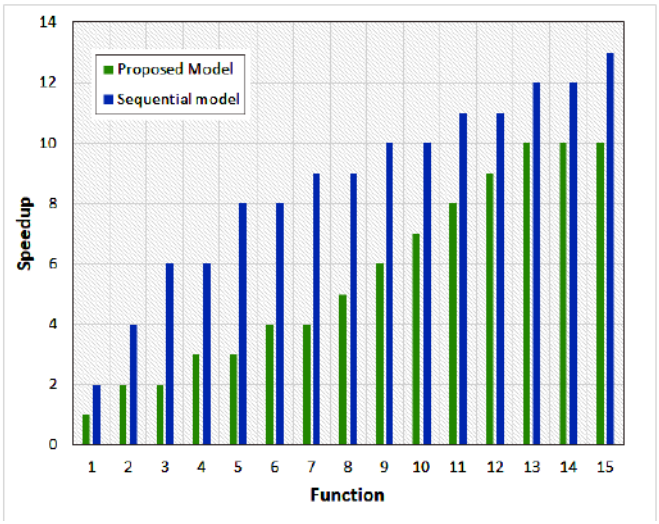

Fig. 4: Speedup comparison for 30 Runs

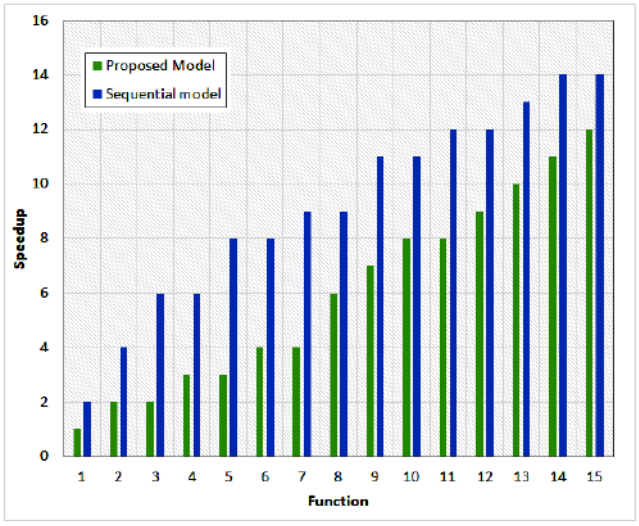

Fig. 5: Speedup comparison for 40 Runs

Fig. 6 gives a comparison of fitness function for proposed model and sequential model through the mean value and obtained value. It is observed that proposed model achieves better fitness function by providing

values near to mean value. While the sequential model deviates in fitness function with respect to mean value. Maximum of 30000 iterations are considered in the comparison process.

The efficiency comparison of each model is given in Fig. 7 and Fig. 8 for different iteration values and population size of 512. It is observed that the number of iterations doesn't affect the performance in proposed model while it is important parameter in sequential model which affects the performance. Fig. 7 gives the efficiency comparison for sequential model and Fig. 8 gives an efficiency comparison of proposed model. 30 runs are used for both population sizes.

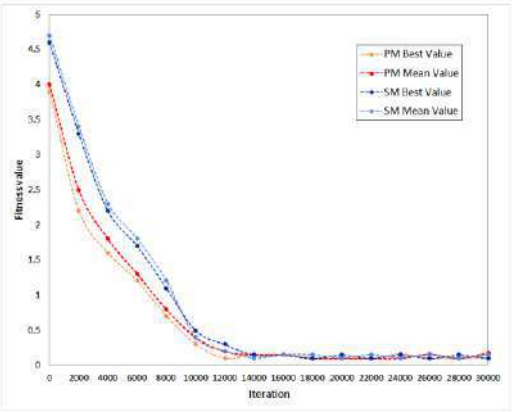

Fig. 6: Fitness function comparison for proposed model and sequential model with best and mean value for 30000 iterations

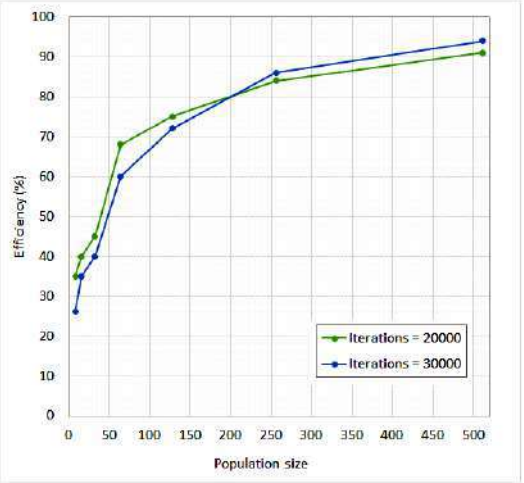

Fig. 7: Efficiency Performance of Sequential model for 30 Runs

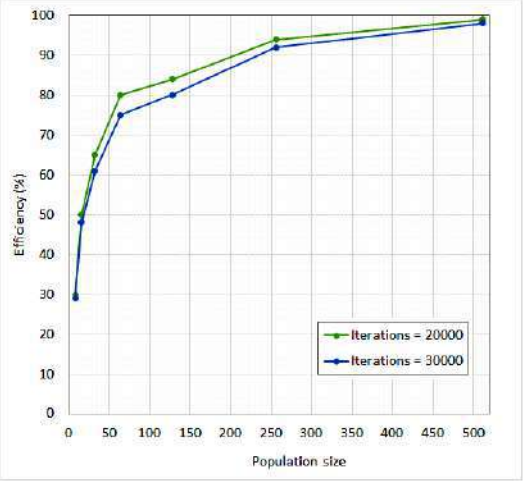

Fig. 8: Efficiency Performance of proposed model for 30 Runs

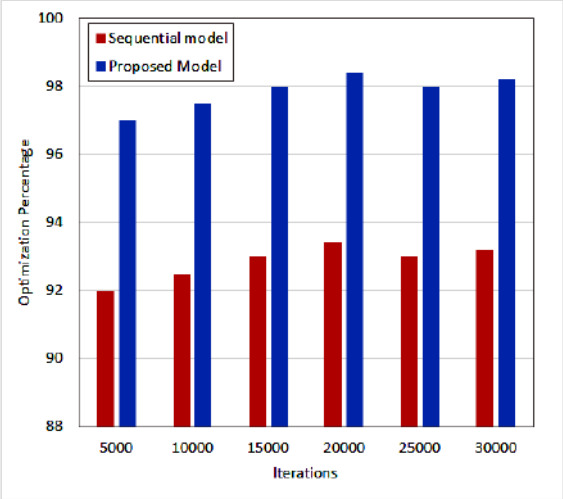

Fig. 9: Optimization Performance Comparison

Fig. 9 gives a comparison of optimization ratio for obtaining best solution with respect to proposed model and sequential model. The percentage of obtaining solution varies with 5% when compared to sequential model which makes a huge difference in data mining approach.

CONCLUSION

This research work provides an enhanced JAYA optimization model to overcome the issues in conventional data mining models and solves the large data set objectives. Using this enhanced optimization model the performance of obtaining best solution was increased. Conventional sequential algorithm is used in experimentation for comparing the proposed model results and it is validated. Through this proposed model we achieve an efficiency of 98% (as observed in Fig-7) and an optimization ratio of 98.2% (as observed in Fig- 9) which are much better than conventional models.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] dos Santos BS, Steiner MT, Fenerich AT, Lima RH. [2019] Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. Computers & Industrial Engineering, 138:106-120.

[2] Lee G, Yun U. [2018] A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives. Future Generation Computer Systems, 68:89-110.

[3] Salehi H, Das S, Biswas S, Burgueño R. [2019] Data mining methodology employing artificial intelligence and a probabilistic approach for energy-efficient structural health monitoring with noisy and delayed signals. Expert Systems with Applications, 135:259-72.

[4] Arslan AK, Colak C, Sarihan ME. [2016] Different medical data mining approaches based prediction of ischemic stroke. Computer methods and programs in biomedicine, 130:87-92.

[5] Tsai CF, Lin WC, Ke SW. [2016] Big data mining with parallel computing: A comparison of distributed and Map Reduce methodologies. Journal of Systems and Software, 122:83-92.

[6] Apiletti D, Baralis E, Cerquitelli T, Garza P, Pulvirenti F, Michiardi P. [2017] A parallel mapreduce algorithm to efficiently support item set mining on high dimensional data. Big Data Research, 10:53-69.

[7] Gürbüz F, Turna F. [2018] Rule extraction for tram faults via data mining for safe transportation. Transportation research part A: policy and practice, 116:568-79.

[8] Ryang H, Yun U. [2016] High utility pattern mining over data streams with sliding window technique. Expert Systems with Applications, 57:214-31.

[9] Lin HY, Yang SY. [2019] A cloud-based energy data mining information agent system based on big data analysis technology. Microelectronics Reliability. 97:66-78.

[10] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. [2017] Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15:104-16.

[11] Kavakiotis I, Samaras P, Triantafyllidis A, Vlahavas I. [2017] FIFS: A data mining method for informative marker selection in high dimensional population genomic data. Computers in biology and medicine, 90:146-54.

[12] Zhang J, Williams SO, Wang H. [2018] Intelligent computing system based on pattern recognition and data mining algorithms. Sustainable Computing: Informatics and Systems, 20:192-202.

[13] Sekar K, Mohanty NK. [2017] Combined mathematical morphology and data mining based high impedance fault detection. Energy Procedia, 117:417-23.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

INTELLIGENT FEATURE SELECTION WITH SOCIAL SPIDER OPTIMIZATION BASED ARTIFICIAL NEURAL NETWORK MODEL FOR CREDIT CARD FRAUD DETECTION

Gurumurthy Krishnamurthy Arun*, Kaliyappan Venkatachalapathy

Department of Computer and Information Science, Annamalai University, Chidambaram, TN, INDIA

ABSTRACT

In recent days, credit card fraud remains as an essential problem for theft and fraud commitment by the use of payment cards like credit or debit cards. For resolving this problem, financial industries have started to develop fraud detection algorithms. Data mining and machine learning (ML) approaches can be used to investigate the normal and abnormal patterns along with individual transactions in order to raise an alarm for possible frauds. In this view, this study develops an intelligent Feature Selection (FS) with social spider optimization (SSO) algorithm based artificial neural network (ANN) model called SSO-ANN for credit card fraud detection. The proposed SSO-ANN model involves preprocessing, ant colony optimization (ACO) algorithm based FS and SSO-ANN based classification. The proposed SSO-ANN model has the ability to detect the existence of frauds in credit card payments. The performance of the SSO-ANN model has been tested against two benchmark dataset namely German Credit dataset and Kaggle's Credit Card Fraud Detection dataset. The experimental outcome pointed out that the SSO-ANN model has shown superior results with the maximum classifier accuracy of 93.20% and 92.82% on the German Credit dataset and Kaggle's Credit Card Fraud Detection dataset.

KEY WORDS
Credit card fraud, Data mining, Classification, Machine learning, Feature Selection

INTRODUCTION

Generally, fraud is a mischief and criminal activity which is performed to gain the economical and personal applications. The fraud events can be reduced under the application of two approaches such as, Fraud prevention as well as Fraud detection. Initially, Fraud prevention is defined as proactive model that eliminates the occurrence of fraud. Secondly, fraud detection is required while an illegal transaction is made by a criminal. Specifically, credit card fraud is a general crime that happens often which is carried out by stealing the details of a person. The credit card transactions can be processed in a digital and physical manner [1].

With the improved credit card users, several crime actions were also improved. Even though there is massive number of authentic models, some of the mischief actions are still in progress which is highly complex to investigate. Using the internet, Fraudsters conceals the position and identity details. The credit card fraud has major impact on economic sector. The overall credit card fraud has attained staggering of USD $21.84 billion. The credit card loss, especially for merchants leads in major deviations such as losing the cost details, administrative charges, and so on [2]. As the merchants should tolerate the loss, and few products might have increased price, or offers and incentives were minimized. Hence, it is optional for loss reduction, and well-defined fraud detection model is applied to avoid the fraud cases. Diverse works are proposed on credit card fraud detection.

Received: 2 June 2020
Accepted: 10 July 2020
Published: 20 July 2020

ML and relevant approaches are utilized prominently such as ANN, rule-induction system, Decision Trees (DT), Logistic Regression (LR), and Support Vector Machines (SVM). Such technologies were applied either uniquely or by integrating various models for developing hybrid approaches. In this method, credit card fraud detection is processed using Random Forest (RF), SVM and LR. An Artificial Immune Recognition System (AIRS) is applied for credit card fraud detection as projected in [3]. AIRS is an extended version of reputed AIS scheme, in which negative selection is employed for reaching maximum precision. Hence, the accuracy is improved and minimizes the system response time to greater extent. A tailored Fisher Discriminant function has been utilized for detecting the credit card fraud [4]. The alteration of conventional functions makes highly sensitive instances.

In order to enhance the prediction of credit card frauds, an effective model is presented in [5]. A data set derived from a Turkish bank has been employed. Every transaction was measured as a fraudulent. The misclassification rates are limited with the help of Genetic Algorithm (GA) as well as scatter exploration. The newly presented technique is highly beneficial when compared with existing outcome. An alternate economical loss is a financial statement fraud. The methodologies like SVM, LR, Genetic Programming (GP) and Probabilistic NN (PNN) were applied to find financial statement fraud. A data set with 202 Chinese industries is employed. The t-statistic is applied for feature subset selection, in which 18 and 10 features are decided in 2 phases. The final outcome shows that the PNN performs quite well than GP. A fraud detection technique depends upon the user account's visualization and threshold-type examination as projected in [6]. Then, Self-Organizing Map (SOM) has been utilized as a visualization framework.

***Corresponding Author**
Email:
arunnura2370@gmail.com

Hybrid approaches are the integration of several technologies. A hybrid scheme is composed of consisting of the Multilayer Perceptron (MLP), NN, SVM, LR, and Harmony Search (HS) optimization have been applied in [7] for detecting corporate tax evasion. HS is applicable in identifying optimal parameters for

classification models. Under the application of data acquired from food and textile applications in Iran, MLP with HS optimization has attained maximum accuracy. The hybrid clustering mechanism with noise prediction ability was utilized in [8] for detecting fraud in lottery and internet games. The training data set undergoes compression over main memory at the time of increasing stored data-cubes. This model has reached maximum detection value with minimum false alarm rate.

The economic crisis can be handled using clustering and classifier ensemble methodologies which forms hybrid approaches finally [9]. The integration of SOM and k-means models are applied in clustering, whereas LR, whereas classification applies MLP, and DT approaches. The SOM and MLP classifier are said to be a remarkable combination, which produces standard accuracy. The concatenation of various models like RF, DR, Roush Set Theory (RST), and Back propagation NN (BPNN) were utilized [10] for developing fraud detection approach in case of corporate financial statements. Organizational financial statements were employed as data set. The final outcome with RF and RST is capable of reaching best classification accuracy.

The model which finds automobile insurance fraud as presented in [11]. A Principal Component Analysis (PCA) based RF method is combined with capable nearest neighbour technique. The traditional classical majority voting in RF has been substituted with effective nearest neighbour module. Overall data sets are employed in the experimental study. The PCA dependent method has generated a maximum classification accuracy and minimum variance, as related with RF and DT methodologies. The concatenation of GA and Fuzzy C-Means (FCM) is effectively applied [12] for exploring the fraud activity involved in automobile insurance sector. The sample records are divided into normal and suspicious classes according to the developed clusters. By removing the original and fraud records, the malicious cases are examined in future with the application of DT, SVM, MLP, and Group Method of Data Handling (GMDH). Hence, SVM performs well by reaching a resourceful specificity and sensitivity rates than other approaches.

The contribution of the study is given as follows. This study introduces a new FS with SSO algorithm based ANN model called SSO-ANN for credit card fraud detection. The proposed SSO-ANN model operates on three different stages namely preprocessing, ACO algorithm based FS(ACO-FS) and SSO-ANN based classification. The proposed SSO-ANN model has the capability of detecting the presence of frauds in credit card payments.

METHODS

The simulation outcome of the SSO-ANN model has been tested against two benchmark dataset namely German Credit dataset (https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)) and Kaggle's Credit Card Fraud Detection dataset (https://www.kaggle.com/mlg-ulb/creditcardfraud).

The block diagram of the proposed SSO-ANN model is demonstrated in Fig. 1. As shown, initially, the input data undergo pre-processing and converts the data into a compatible format. Then, ACO-FS algorithm is executed for selecting the useful count of features from the pre-processed data. Finally, SSO-ANN based classification process takes place to determine the existence of credit card frauds or not.

Preprocessing: Initially, preprocessing of input credit data takes place in two stages namely format conversion and data transformation. During format conversion, the input data in .csv format is converted into .arff format. Next, in the data transformation stage, the numerical values are converted into corresponding categorical values, i.e. values in 0's and 1's are converted into good and bad credits. Once the data is preprocessing, it is sent to the FS process for selecting the desired number of features.

ACO-FS model: In this section, ACO-FS model has been discussed the way of selecting features from the preprocessed data. The provided feature set of size n, the FS issues is mainly applied to find the least size of s feature subset (where s<n), at the time of retaining maximum accuracy while presenting the actual feature subset. An incomplete does not indicate the ordering among the features of solution. Simultaneously, the next feature has been selected and it is not affected by existing feature which appends the partial solution. Hence, there is no requirement equal size for FS problem [13]. The matching of FS problem to ACO method is comprised with following procedures:

- Graph depiction
- Heuristic popularity
- Pheromone extension
- Solution development

SSO-ANN based Classification Model: In this framework, the processes involved in the SSO-ANN based classification model have been discussed in the following subsections.

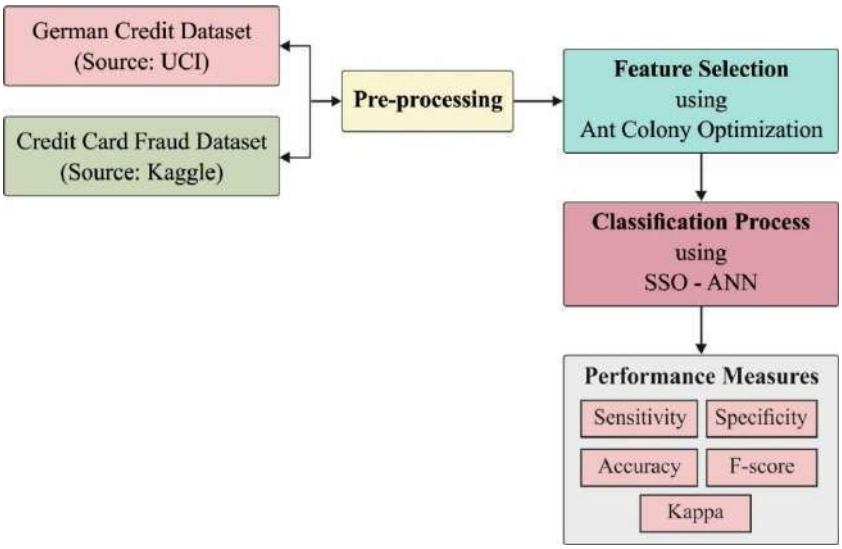

Fig. 1: Block diagram of SSO-ANN model

Structure of ANN: The ANN is defined as a soft computing model which is extensively applied for data processing. It has been evolved from biological nervous systems, like working function of human brain. It is illustrated with respect to weighted directed graphs where node is treated as artificial neurons as well as directed edges among neurons defined weights. It is classified into 2 classes namely, Feed forward and Recurrent networks. Initially, Feed forward networks are defined as static based while recurrent networks are dynamic. It generates a single set of resultant values rather producing a series of values from provided input. The Feedforward networks are generally memory-less and autonomous of existing network. When presenting a novel input pattern, the neuron outcomes are determined [14].

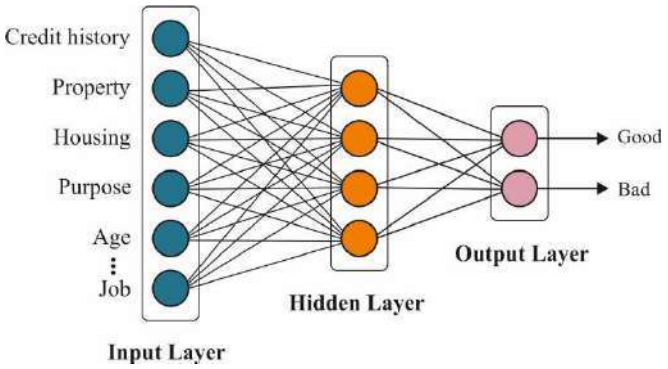

Fig. 2: Structure of ANN

Every neuron in the input and hidden layers are correlated with other neurons of subsequent layer. The neurons present in the hidden layers were utilized for calculating weighted sums of inputs and a threshold. The architecture of MLP includes input layer, hidden layer, and output layer, as depicted in Fig. 2. The input layer shows the parameters of datasets. The performance of hidden layer means the attributes of datasets that is not linearly separated while output layer offers the essential outcome. The final results show the function of neurons using a sigmoid activation function which is depicted by Eq. (1).

$$p_v = \sum_{u=1}^{n} w_{uv} xu + \theta v, m_v = f_v(p_v) \qquad (1)$$

where p_v denotes linear integration of inputs $x_1, x_2, \ldots, x_n$, and thethresholdθ_v, w_{vu} is the association weight among the input x_u, neuron v, and f_v represents activation function of v^{th} neuron, and m_v means the final outcome. A sigmoid function is assumed to be an activation function as provided in Eq. (2).

$$f(z) = \frac{1}{1 + e^{-z}} \qquad (2)$$

The MLP can be trained using BP learning technique that is said to be a Gradient Descent (GD) model for accessing the weights. Every weight vector (w) was loaded using minimum random measures from

pseudorandom sequence generator. Hence, it consumes maximum procedures for network training, and modified weights are processed at every step. The mentioned problems can be resolved by using SSO algorithm for computing the optimal value of weight and threshold functions, as SSO have the ability of calculating parallel weight and finds viable solutions.

SSO Algorithm: The SSO model depends upon the cooperative nature of social-spiders as projected by Cuevas [15].Here, search space is referred as communal web and spider's location is the best solution. The most important feature of social-spiders is a female-based population. Here, male spiders are minimum when compared with female in the overall community. The count of females N_f is selected randomly with least proportion$N$, which is determined as:

$$N_f = [(0.9 - rand * 0.25) * N], \quad (3)$$

where $rand$implies random values among [0, 1]. The value of male spiders N_m is measured using:

$$N_m = N - N_f . \quad (4)$$

All spiders gain a weight according to the fitness rate of attained solution:

$$w_u = \frac{fitness_u - worst}{best - worst}, \quad (5)$$

where $fitness_u$ refers the fitness value accomplished by estimating u^tspider's position $u = 1,2,\ldots,N$. The $worst$ and $best$ shows the inferior and superior fitness value of whole population, correspondingly.The communal web is mainly applied for transmission between the colony members. The data undergoes encoding as tiny vibrations are based on the weight as well as distance of spider which has been generated by the given expression:

$$Vb_{uv} = w_v e^{-d_{uv}^2}, \quad (6)$$

where $d_{u,v}$ means the Euclidean Distance among the spider u and v. There are 3 kinds of relationships namely,

- Vibrations Vb_{uc} which are perceived by spider u by transmitting the data by the member c, and it is closer to u with maximum weight, such as $w_c > w_u$;
- Vibrations Vb_{ub} which are perceived by spider u which is forwarded by spider b with optimal weight of whole population; and
- Vibrations Vb_{uf} are perceived by spider u as data is sent by closer female f.

It is clear that weight (w) and bias (b) attributes are constrained with higher influence on ANN function [14]. In this work, the SSO technique is employed for parameter optimization of ANN. The ANN model undergoes training with the parameters present in the social spider. The 10-fold cross-validation (CV) approach is employed for evaluating the fitness function (FF). The FF can be represented as follows.

$$Fitness = 1 - CA_{validation} \quad (7)$$

$$CA_{validation} = 1 - \frac{1}{10}\sum_{i=1}^{10}\left|\frac{TP}{TP+TN}\right| \times 100 \quad (8)$$

where, TP and TN represent the number of true as well as false classifications correspondingly.
By using derived fitness function, Eq. (5) becomes

$$w_u = \frac{\left[1 - \left(1 - \frac{1}{10}\sum_{i=1}^{10}\left|\frac{TP}{TP+TN}\right| \times 100\right)_u\right] - worst}{best - worst} \quad (9)$$

Finally, the above equation is used to determine the weight according to the fitness rate.

RESULTS

This section describes the function of the SSO-ANN approach on two benchmark dataset. The dataset applied, performance metrics and the results are explained in the following sections.

Dataset used: The dataset used for assessing the experimental values of the SSO-ANN model are German Credit [16] and Credit card fraud detection [17] dataset. Firstly, the number of instances in the German Credit dataset is 1000 credit applicants with 20 attributes. The number of classes is two including good and bad credits. The number of instances in good credit is 700 and the remaining 300 instances come under bad credit. Secondly, the credit fraud detection dataset includes the credit card transactions

by cardholders in Europe in September, 2013. This dataset includes a total of 2,84,315 transactions with 30 attributes. This dataset also comprises the instances under good (1) and bad (0) credit classes. The information related to the dataset is given in Table 1.

Table 1: Dataset Description

Descriptions	German Credit Dataset	Credit Fraud Detection Dataset
Source	UCI	Kaggle
# of instances	1000	284807
# of attributes	20	30
# of class	2	2
Classes: Good/Bad	700/300	284315/492

Table 2 provides the FS results offered by the ACO algorithm on two dataset along with its best cost. The table values pointed out that 4features were chosen by ACO algorithm on the German Credit dataset with the best cost of 0.14. Besides, the number of features chosen in Credit Fraud Detection dataset is 13 with the best cost of 0.832.

Table 2: ACO based Selected Features and its Best Cost

Dataset	Selected Features	Best Cost
German Credit	9,2,1,4,7	0.140
Credit Fraud Detection	1,2,3,5,6,8,9,11,16,19,21,24,28	0.832

Table 3 provides the comparative study of the classifier outcome offered by different classification models on German Credit dataset in terms of several measures. The table depicts that the DT model has demonstrated lower classifier outcome with the sensitivity and specificity of 76.75% and 52.61% respectively. Simultaneously, the RBF Network has surpassed DT model by attaining the sensitivity and specificity of 78.58% and 59.55% respectively. In the same way, the MLP model has outperformed the earlier models with the sensitivity and specificity of 79.72% and 55% respectively. Along with that, the LR model has obtained even better outcome with the sensitivity and specificity of 79.82% and 60.74% respectively. On continuing with, the ACO-DC model has shown moderate results with the sensitivity and specificity of 77.93% and 69.87% respectively. Besides, the NGSAII has tried to show acceptable classifier outcome with the sensitivity and specificity of 89% and 89% respectively. At the same time, the SMOPSO model has showcased near optimal sensitivity and specificity of 90% and 90% respectively. However, the proposed SSO-ANN model has outperformed all the compared methods with the maximum sensitivity and specificity of 93.88% and 91.43% respectively.

Table 3: Performance Evaluation of Various Classifiers on German Credit Dataset

Classifier	Sensitivity	Specificity	Accuracy	F-score	Kappa
SSO-ANN	93.88	91.43	93.20	95.21	83.49
MLP	79.72	55.00	72.80	80.84	33.98
RBFNetwork	78.58	59.55	74.30	82.58	34.10
LR	79.82	60.74	75.20	82.99	37.50
DT	76.75	52.61	71.20	80.41	26.93
ACO-DC	77.93	69.87	76.60	84.74	36.13
NSGA II	89.00	89.00	85.10	-	-
SMOPSO	90.00	90.00	92.30	-	-
PSO-SVM	-	-	81.50	-	-
GA-SVM	-	-	80.50	-	-
SVM	-	-	77.50	-	-

ABC-SVM	-	-	84.00	-	-
SVM-IGDFS	-	-	82.80	-	-
SVM-GAW	-	-	80.40	-	-
KNN-IGDFS	-	-	70.20	-	-
KNN-GAW	-	-	75.80	-	-
NB-IGDFS	-	-	77.30	-	-
NB-GAW	-	-	76.80	-	-

Fig. 3 shows the analysis of the results attained by SSO-ANN model in terms of accuracy. The figure showed that the proposed SSO-ANN model has showcased effective classifier outcome by offering a maximum accuracy of 93.20% on the applied German Credit dataset.

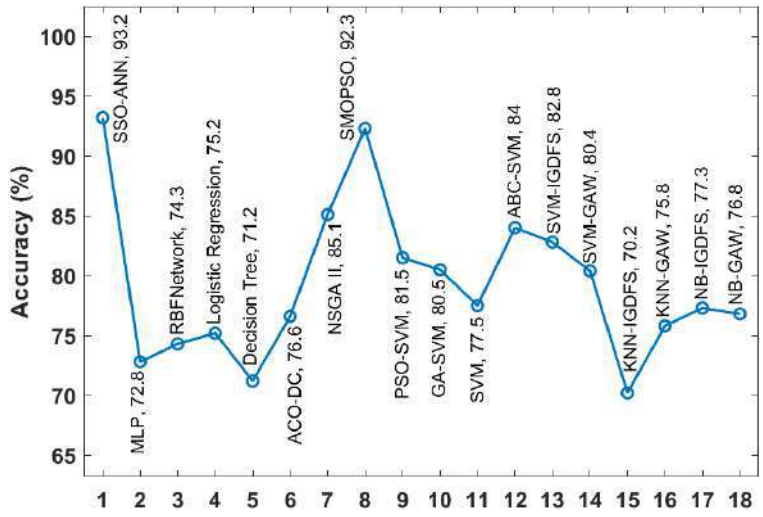

Fig. 3: Accuracy analysis of various models on German Credit dataset

A comparative study of the classifier results provided by diverse classification methods on Credit Card Fraud Detection dataset with respect to various measures is given in [Table 4]. The table showcases the sensitivity and specificity analysis of the SSO-ANN technique on the sampled Credit Card Fraud Detection dataset. The table implies that the NB model has illustrated least classifier outcome with the sensitivity of 62.40%. Concurrently, the DT has surpassed NB method by accomplishing the sensitivity and specificity of 67.83% and 68.51% correspondingly. Along with that, the RF model has an outstanding performance when compared with alternate approaches with the sensitivity of 67.89%. In line with this, the LR model has attained manageable outcome with the sensitivity and specificity of 76.52% and 78.70% respectively. On the same way, the RBF Network model has showcased gradual outcome with the sensitivity and specificity of 79.41% and 80.49% respectively. Meantime, the MLP model has demonstrated closer optimal sensitivity and specificity of 81.39% and 82.40% correspondingly. Therefore, the presented SSO-ANN model has performed effectively than compared methods with the greater sensitivity and specificity of 90.47% and 92.57% respectively.

Table 4: Performance Evaluation of Various Classifiers on Credit Card Fraud Detection Dataset

Classifier	Sensitivity	Specificity	Accuracy	F-score	Kappa
SSO-ANN	90.47	92.57	92.82	89.20	85.20
MLP	81.39	82.40	82.53	80.13	73.50
RBFNetwork	79.41	80.49	80.86	79.68	68.71
Logistic Regression	76.52	78.70	78.92	76.01	65.89
Decision Tree	67.83	68.51	69.93	68.82	63.60
Random Forest	67.89	-	91.96	78.11	-
Naïve Bayes	62.40	-	83.00	74.20	-

After examining the above-mentioned figures and tables, it is evident that the SSO-ANN algorithm has found to be an effective tool for credit card fraud detection.

CONCLUSION

This study has introduced a new FS based classification model called SSO-ANN for credit card fraud detection. Initially, the input data undergo preprocessing to transform the data into a compatible format. Then, ACO-FS algorithm is executed to select the useful number of features from the preprocessed data. Finally, SSO-ANN based classification process takes place to determine the existence of credit card frauds or not. The performance of the SSO-ANN model has been tested against two benchmark dataset namely German Credit dataset and Kaggle's Credit Card Fraud Detection dataset. The experimental outcome pointed out that the SSO-ANN model has shown superior results with the maximum classifier accuracy of 93.20% and 92.82% on the German Credit dataset and Kaggle's Credit Card Fraud Detection dataset. In future, the performance of the SSO-ANN model can be improved by the use of clustering techniques.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Adewumi AO, Akinyelu AA. [2017] A survey of machine-learning and nature-inspired based credit card fraud detection techniques, International Journal of System Assurance Engineering and Management, 8:937–953.

[2] Quah JT, Sriganesh M. [2008] Real-time credit card fraud detection using computational intelligence, Expert Systems with Applications, 35(4):1721–1732.

[3] Halvaiee NS, Akbari MK. [2014] A novel model for credit card fraud detection using Artificial Immune Systems," Applied Soft Computing, 24:40–49.

[4] Mahmoudi N, Duman E. [2015] Detecting credit card fraud by modified Fisher discriminant analysis, Expert Systems with Applications, 42(5):2510–2516.

[5] Duman E, Ozcelik MH. [2011] Detecting credit card fraud by genetic algorithm and scatter search, Expert Systems with Applications, 38(10):13057–13063.

[6] Olszewski D. [2014] Fraud detection using self-organizing map visualizing the user profiles, Knowledge-Based Systems, 70:324–334.

[7] Rahimikia E, Mohammadi S, Rahmani T, Ghazanfari M. [2017] Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran, International Journal of Accounting Information Systems, 25:1–17.

[8] Christou IT, Bakopoulos M, Dimitriou T, Amolochitis E, Tsekeridou S, Dimitriadis C. [2011] Detecting fraud in online games of chance and lotteries, Expert Systems with Applications, 38(10):13158–13169.

[9] Tsai CF. [2014] Combining cluster analysis with classifier ensembles to predict financial distress Information Fusion, 16:46–58.

[10] Chen FH, Chi DJ, Zhu JY. [2014] Application of Random Forest, Rough Set Theory, Decision Tree and Neural Network to Detect Financial Statement Fraud–Taking Corporate Governance into Consideration, In International Conference on Intelligent Computing, 221–234.

[11] Li Y, Yan C, Liu W, Li M. [2017] A principle component analysis based random forest with the potential nearest neighbor method for automobile insurance fraud identification, Applied Soft Computing, to be published, DOI: 10.1016/j.asoc.2017.07.027.

[12] Subudhi S, Panigrahi S. [2017] Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," Journal of King Saud University-Computer and Information Sciences, to be published, DOI: 10.1016/j.jksuci.2017.09.010.

[13] Uthayakumar J, Metawa N, Shankar K, Lakshmanaprabu SK. [2020] Financial crisis prediction model using ant colony optimization. International Journal of Information Management, 50:538-556.

[14] Gambhir S, Malik SK, Kumar Y. [2017] PSO-ANN based diagnostic model for the early detection of dengue disease. New Horizons in Translational Medicine, 4(1-4):1-8.

[15] Cuevas E, Cienfuegos M, ZaldíVar D, Pérez-Cisneros M. [2013] A swarm optimization algorithm inspired in the behavior of the social-spider. Expert Systems with Applications, 40(16):6374-6384.

[16] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[17] https://www.kaggle.com/mlg-ulb/creditcardfraud

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

MATHEMATICAL APPROACH TOWARDS RECENT INNOVATION IN COMPUTATION AND ENGINEERING SYSTEM (MATRICS)

OPTIMAL WHALE OPTIMIZATION ALGORITHM BASED ENERGY EFFICIENT RESOURCE ALLOCATION IN CLOUD COMPUTING ENVIRONMENT

Natarajan Subalakshmi[1], Mohan Jeyakarthic[2*]

[1]*Computer Science and Engineering Wing, Annamalai University, Chidambaram, TN, INDIA*

[2]*Tamil Virtual Academy, Chennai, TN, INDIA*

ABSTRACT

In cloud, proper allocation of resources improves the exploitation of resources as well as energy efficiency, provider's profit and client's fulfilment. Whale Optimization Algorithm (WOA) is a new bio-inspired meta-heuristic technique inspired from the social hunting nature of humpback whales. WOA suffers premature convergence that causes it to trap in local optima. To resolve it, this paper proposes a new energy efficient resource allocation in cloud computing environment using optimal whale optimization algorithm with tumbling effective called WOA-TRA model. Here, the WOA is hybridized with tumbling effect which has good exploration ability for function optimization problems to derive an energy aware solution. The proposed WOA-TRA model methodology attempts to optimize allocation of resources for improving the energy efficiency of the cloud platform by fulfilling the quality of service (QoS) requirements of the end user. To effectively utilize the energy and QoS requirements, the WOA-TRA technique is utilized in two levels. In the first level, WOA-TRA technique assigns Virtual Machines (VMs) resources to jobs, whereas in second level, WOA-TRA technique assigns Physical Machines (PMs) resources to VMs. The presented model is simulated in CloudSim platform and a detailed comparative analysis is made with the state of art resource allocation techniques. The experimental analysis states that the presented WOA-TRA technique offered desired QoS and enhanced energy efficiency by effectively utilizing the available resources over the compared methods.

KEY WORDS
Resource allocation, Energy aware, Whale optimization, Local optima

INTRODUCTION

Infection Cloud Computing (CC) is a model which consists of massive ability in trading and business. It contains maximum number of certain resources that could be obtained and utilized whenever the demand arouses [1]. The obtained resources could be used across the network. Cloud presents every source as service and is comprised of three services namely Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The IaaS performs basic operations such as computation, memory, networking and so on. They are offered to the customers in order to deploy as well as to implement software randomly [2]. Here, the resources are planned definitely and assigned on the basis of customer demands. Therefore, the purpose of allocating resources ensured the requirement of every application processing. Since the above operations are performed, resource allocation in CC is a mandatory issue. Regardless of allocating resources, allocating enough resource for user demand to convince quality of service (QoS) attributes is an alternate challenge for an organization to reduce the power consumption as well and carbon footprints.

Based on the statement of [3], entire data centre energy conservation from servers, memory, interactions, resolving, and power-distributing tools considers 1.7–2.2% of total power applied in US. Using the numerous amount of energy, present data centre release maximum carbon dioxide (CO_2) thought Argentina. When it is left on present path, data centre carbon-dioxide outcome would increase widely by 2020 [4]. Since there is rapid development of cloud, industries, and research institutes identify the feasible paths to minimize the power utilization. Data centres construction is normally prepared for retaining resources in peak time [5]. Energy efficiency could be enhanced by optimally balancing the resources. In order to reduce maximum power utilization, resource allocation should be efficient. Massive amount of energy could be stored by consolidating server as well as switching off unique servers. But, consolidation of servers are possible financially due to some limitations namely cost of migrating, violating QoS as several disturbances occurs while performing final process of physical machines (PM) using enough resources, so that virtual machine (VM) could be transferred easily. At this point, energy conserved by PM could be stored by varying the corresponding voltage.

Beloglazov et al., 2012 [6] presented an energy effective as well as QoS aware resource allocating technique absed on heuristics. A technique for reducing number of VM has also been deployed. Lower and upper threshold consumption has been fixed in order to predict the underweight and overloaded systems. If any server drops below the specified threshold level, then each VM implemented on the system is transferred to alternate machine whereas the resource conservation is above the threshold value then more than single values are transformed among the values. Practically cloud platforms identical multi-core machines are utilized. The VM is placed for reducing the resource wastage as well as energy conservation. The first pheromone value is declared to VM-host transformation. This value denotes the possibility of a host which is to be chosen to allocate VM with certain constraints. This technique operates using central processing unit (CPU) operation, speed, storage and so on.

Received: 26 June 2020
Accepted: 18 July 2020
Published: 22 July 2020

***Corresponding Author**
Email:
Jeya_karthic@yahoo.com
Tel.: 8667859294

EDITED BY: Prof. Dr. S. Vaithyasubramanian

IIOAB JOURNAL
ISSN: 0976-3104

In Kinger et al., 2014 [7], an event driven detection model has been proposed to maintain the temperature of a server with particular threshold values. Temperature predictor often observes the temperature of external system. In this model "unified list" has been applied for storing electricity and threshold temperature for all nodes. It is usually upgraded once the duration is fixed that leads to network congestion, degrading performance, restricted scalability and so on. Quarati et al., 2013 [8] projected 2 stages of brokering algorithm for hybrid cloud along with the intention of increasing brokers economy rate as well as customer convenience. Initial stage is to schedule the requested facilities on private or public cloud which is done on the basis of predefined resources. Further division from first stage of cloud namely possibility, reservation should be static and effective. The next stage is to apply lower resources in order to declare the resource for every service. Hence, requested services are executed on the physical system which consists of improved assessable sources. The deployed model leads to dissimilar distribution of overhead between servers and the machines that are operated with high efficiency. Overload tends to cause hot spot issue and raise the level of failure.

Lee et al., 2014 [9] applied the computation model that relies on resource allocating principle for green cloud. All PMs of a data centre are allocated with a performance measure that depends upon the processing speed of CPU, count of cores, storage ability and so on. Any PM could be assigned to VM, when the performance value is accurate to VM necessities. Improper sharing of overhead between servers might leads to power wastage. Raycroft et al., 2014 [10] examined the impact of VM allocating that is based on consumption of energy. Here, simulating operation is carried out for similar kind of applications whereas realistic clouds conduct different types of application. The communication cost between VM and QoS is not considered. Therefore, the movement of VM between areas becomes impractical when it comes to massive size of VM. Feller et al., 2011 [11] employed multi-dimensional ant colony optimization (ACO) that is relied on workload consolidating technique. This technique applies the resource consumption in order to detect the upcoming demands of resource.

Gao et al., 2013 [12] signified the multi-objective ACO model for virtual system to be fixed. Using this placement, it reduces the wastage of resource and energy application. It has the objective of utilizing server to complete ability that results in creating hot spots as well as maximizes service-level agreement (SLA) violations. With the application of complete usage of server it causes major heat disperse that tends to minimize the servers consistency. Nathani et al., 2012 [13] introduced alternate and latest consumption methods for end sensitive lease of VM. The proposed model attempts for scheduling novel lease as deadline in more than one time slots. To create a new room for novel lease, the algorithm reallocates the previously scheduled deadline sensitive leases in case if it could not be declared to single or multiple time slots. If reallocation is carried out properly i.e., fails to produce limited deadline schedule and backfilling is used to accompany new lease. Proposed model is emerged with a demerit where it consumes maximum pre-emption value that improves overhead.

ACO model is applied for heterogeneous operators which are implied in Chen et al., 2011 [14]. Local search method is used for enhancing the efficiency of energy when there is possible declaration decision produced by new technique. This model is claimed with 15.8% energy than prototyped model of ACO. Huang et al., 2013 [15] adapted a sub-optimal resource managing approach. Here, wide resource scheduling module applies residual resource table as well as resource utilization measuring table in order to evaluate VM which is essential to offer defined level of services. Genetic Algorithm (GA) is developed to reallocate resources for attaining optimal computation. However the newly presented model meets the failure of single point. Therefore, centralized global resource scheduling method, residual resource table, as well as resource consumption table would create decreased performance while maximum requests arise in VM.

Garg et al., 2011 [16] presented green CC technology to minimize carbon footprint with inconvenient QoS. This research applies Green Offer Directory (GOD) and Carbon Emission Directory (CED) in order to provide green facilities for the customers. The CED balances information which is relevant to energy efficient cloud service. According to the data from 2 repositories, price as well as carbon footprint of particular leasing is estimated. Hence, providers have the responsibility of publishing footprint as well as effective energy for corresponding public directories. Thus service provider could publish the modified information to gain maximum standard in market. Xu and Fortes, 2010 [17] established different objectives for VM allocating technique. By the consumption of disk, inter VM interaction cost is not considered. Wu et al., 2014 [18] projected priority job scheduling in a effective manner for CC. CPU frequencies in terms of maximum and minimum are preferred as the essential features to perform this priority scheduling job. Each server is declared with few weights on the basis of computation obtained. In order to perform the job, server should be chosen on the basis of weight assigned as well as SLA needed by customers.

This paper introduces a whale optimization algorithm for resource allocation problem in cloud environment. WOA is a new bio-inspired meta-heuristic technique inspired from the social hunting nature of humpback whales. WOA suffers premature convergence that causes it to trap in local optima. To resolve this issue, this paper this limitation of WOA, in this paper, WOA is hybridized with tumbling effect which has good exploration ability for function optimization problems and derives a WOA with tumbling effect based energy aware solution called WOA-TRA model. The proposed WOA-TRA model methodology attempts to optimize allocation of resources in for improving the energy efficiency of the cloud platform by fulfilling the quality of service (QoS) requirements of the end user. To effectively utilize the energy and QoS requirements, the WOA-TRA technique is utilized in two levels. In the first level, WOA-TRA technique assigns

Virtual Machines (VMs) resources to jobs, whereas in second level, WOA-TRA technique assigns Physical Machines (PMs) resources to VMs. The presented model is simulated in CloudSim platform and a detailed comparative analysis is made with the state of art resource allocation techniques. The experimental analysis stated that the presented WOA-TRA technique offered desired QoS and enhanced energy efficiency by effectively utilizes the available resources over the compared methods.

The presented WOA-TRA model

The presented WOA-TRA framework assigns the resource for all jobs with the application of ACO. The resources are employed efficiently in order to store energy and satisfy the demands of every job. Every job is comprised with few resources as well as QoS necessities. The QoS parameter of a job is connected with weight value. WOA-TRA assigns resources for jobs based on the demand for resources and weight values of QoS attributes. Some of the significant features of WOA-TRA are listed as follows:

- It simply allocates the sources for job to enhance the application of resources which maximizes the energy efficiency of cloud structure.
- In order to save energy idle PM is turned to sleep mode.
- Observing the use of resources for processing unit, storage, internet bandwidth of a PM to attain effective resource allocation.
- Dynamic scaling operation is carried out to preserve energy.
- Consolidation of server helps in reducing active servers.

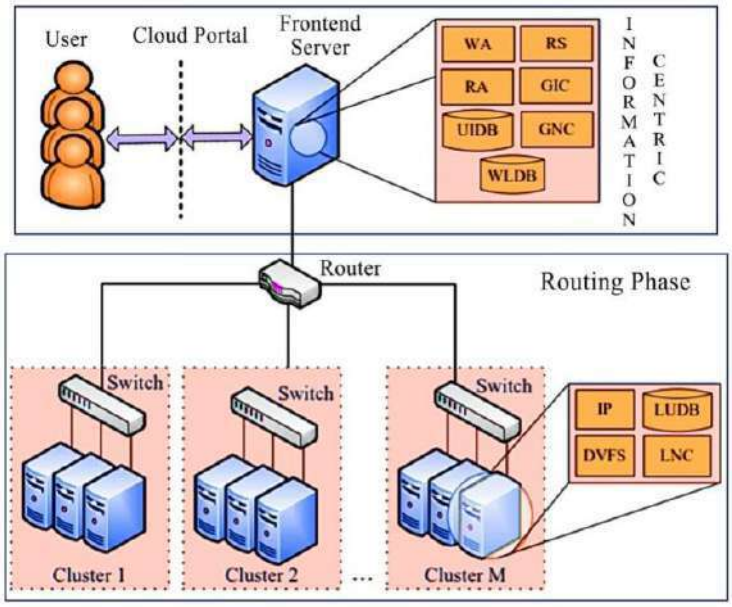

Fig. 1: Components in WOA-TRA model

The different number of components involved in WOA-TRA model is shown in [Fig. 1] and the basic steps involved in the resource allocation process is shown in [Fig. 2].

Cloud Portal: It offers an interface for cloud consumers in order to induce their corresponding job and fixed QoS.
Workload analyzer (WA): It examines essential QoS features of jobs and divides into various classes with the help of k-means clustering technique.
Resource scheduler (RS): It produces job scheduling execution process.
Resource Allocation (RA): This model employs ACO for allocating jobs to VM, where VM in turn assigns to PM. Resource allocation is performed on the basis of demands and values of QoS attributes which is linked with a task.
Global information collector (GIC): It obtains a resource application information form information probes (IP) of all PM as well as saves in Utilization Information Database (UIDB).
Utilization information database (UIDB): Data regarding resource application for each PM is recorded in UIDB, which could be acquired for future allocation and VM transforming solutions.
Global node controller (GNC): It starts current migration of VM executing on PM while resource utilization of PM violates Lower Green Threshold (LGT) or Upper Green Threshold (UGT) limit.
Workload database (WLDB): It is helpful in saving the data regarding every job.
Information probes (IP): It observes the use of resource by processors, memory, network bandwidth of PM and saves monitored values in Local Utilization Database (LUDB).
Local utilization database (LUDB): Application of resource of PM is stored in this feature.

IIOAB JOURNAL
ISSN: 0976-3104

Dynamic voltage frequency scaling (DVFS): It helps to adapt the voltage and frequency of PM for energy conservation as well as to minimize the heat dispersion. Voltage of PM could be modified based on the resource demand of VM implementation.

Local node controller (LNC): It turns PM to sleep mode when it is identified as unique for particular interval of duration.

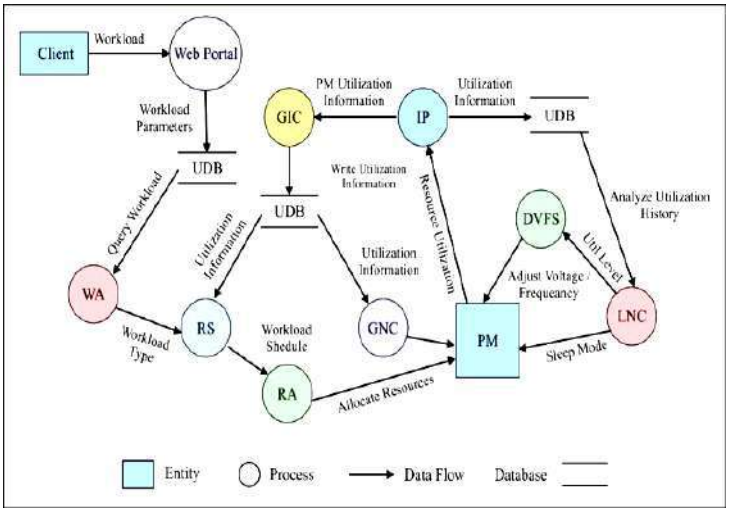

Fig. 2: Energy Based Resource Allocation Steps

The main intention of proposed model is to decrease energy conservation, entire implementation time, cost spent for execution by effective application of resources. Here, weighted addition is applied in order to scale different objectives into one resource.

Inspiration of WOA

Whale is treated as the largest mammals in the world. A mature whale could develop to a maximum of 30m long with180t weights [19]. There are seven various important types of huge mammal namely killer, Minke, Sei, humpback, right, finback, and blue. Whale is generally treated as killers. They do not sleep, as they should breathe from the surface of oceans. In reality, half of the whale's mind only sleeps. The most amazing factor is that they are extremely clever mammals with feeling. The exciting things on humpback whales are their individual hunting technique. This hunting process is termed as bubble-net feeding model. Humpback whales choose chasing school of krill's otherwise little fishes nearby surface. It has been monitored that hunting's are completed with generating particular bubbles beside encircle otherwise '9'-shaped way. An interesting factor regarding this technique is bubble-net feeding which is identified only from whales; also the enhanced model of feeding is spiral bubble-net method which is developed for determining better optimizing function.

Mathematical concept and optimization technique

In this segment, the mathematical concept of surrounding victim, circling bubble-net feeding scheme and initial exploration of prey were discussed. The WOA techniques are then presented.

Encircling prey

Humpback whales could identify the victim location as well as surrounded them. Because of the fact that the location of best position in the exploration spaces are not known before, the WOA technique considers the present optimal candidate results as the victim otherwise nearby optimization. Behind the optimal exploring agents as described, the further exploration agents gets resolved,; hence, attempts for informing their locations towards the optimal explore agents. This performance is signified with subsequent equations:

$$\vec{D} = \left|\vec{C}.\overrightarrow{P^*}(z) - P(z)\right| \quad (1)$$

$$\vec{P}(z+1) = \overrightarrow{P^*}(z) - \vec{L}\cdot\vec{D} \quad (2)$$

where z denotes the present iteration, $\vec{L}$ and $\vec{C}$ are co-efficient vectors, P^* is the location vector of optimal result achieved until now, $\vec{P}$ is the location vector, $\|$ is the total value, while $.$ is the element-via-element multiplication. It is value for declaring now that P^* must be informed in all iterations when there is an optimal solution. The vectors $\vec{L}$ and $\vec{C}$ are computed as follows:

$$\vec{L} = 2\vec{l} \cdot \vec{r} - \vec{l} \quad (3)$$
$$C = 2 \cdot \vec{r} \quad (4)$$

where $\vec{l}$ is linear reduced from two to zero above the way of iterations (together searching and utilization stages) while $\vec{r}$ is a arbitrary vectors in [0, 1].

The location (P, Q) of a explore agents could be informed give to the location of present optimal records (P^*, Q^*). Several places about the optimal agents could be obtained in terms of the present location with changing the value of $\vec{L}$ as well as $\vec{C}$ vectors. It must be noticeable with interpret the arbitrary vector $(\vec{r})$ it is probable for reaching some location in the explore space located among the key points. Consequently, Eq. (2) permits several explore agents for updating its location in the region of the present optimal result as well as reproduces surrounding the victim. The similar method could be continuing for exploring spaces by n dimension, while explore agents would progress in hypercube about the optimal result achieved till now. As declared in the before segment, the humpback whales as well harass the victim by the bubble-net approach. These techniques are mathematically created as follows:

Bubblenet harassing technique

To mathematically define the bubble-net performance of humpback whales, 2 strategies are employed and the bubblenet exploration mechanism is shown in [Fig. 3] [19]:

Shrinking surrounding mechanism: These performances are attained with reducing the rate of $\vec{l}$ in the Eq. (3). Noticeable the variation series of $\vec{L}$ is also reduced with$\vec{l}$. In further words $\vec{L}$ is an arbitrary rate of interval $[-l, l]$ where l is reduced from two to zero above the iteration way. Surroundings arbitrary rates to $\vec{L}$ in [-1, 1], the novel location of a explore agents could be described wherever among the unique location of the agents while the location of the present optimal agents.

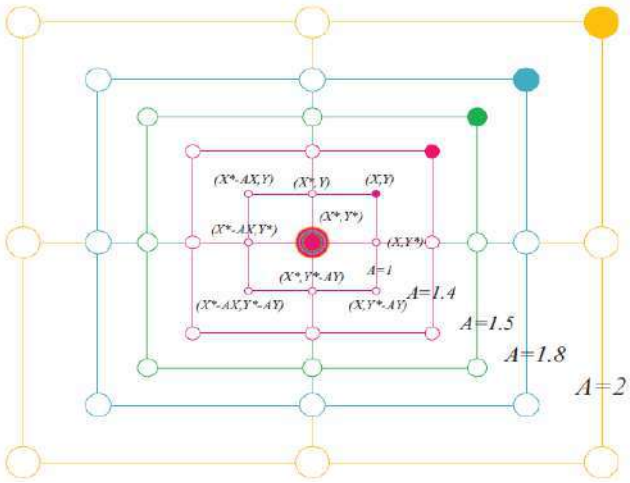

Fig. 3: BubbleNet explore mechanism

Spiral informing location: It computes the distance among the whale positioned at (P, Q) as well as victim positioned at (P^*, Q^*). A circling equations are then generated among the location of whale as well as victim for imitating the helix-shaped progress of humpback whales as pursues:

$$\vec{P}(z+1) = \overrightarrow{D^i} \cdot e^{b\alpha} \cdot \cos(2\pi l) + \overrightarrow{P^*}(z) \quad (5)$$

where $\overrightarrow{D^i} = \left|\overrightarrow{P^*}(z) - \vec{P}(z)\right|$ represents the distance of the ith whales to the victim (optimal result achieved till now), b is stability to define the form of the logarithmic circling, a is a arbitrary number in [-1,1], and . is an element-via-element multiplication.

Noticeably, humpback whales swims and encircle the victim in a decrease surround as well as beside a circling shape way at the same time. To concept this concurrent actions, we consider that there is a possibility of 50% of selecting among them either by decreasing the surrounding mechanism or the circling concept for updating the location of whales in optimization. The mathematical concept is as pursues:

$$P(z+1) = \begin{cases} \overrightarrow{P^*}(z) - \vec{L} \cdot \vec{D} & if\ x < 0.5 \\ \overrightarrow{D'} \cdot e^{ba} \cdot cos\ (2\pi l) + \overrightarrow{P^*}(z) & if\ x \geq 0.5 \end{cases} \quad (6)$$

where x is an arbitrary number in [0, 1]. Besides the bubble-net technique, the humpback whales explore the victim arbitrarily.

Search for prey (exploration phase)

The similar manner dependent on the difference of the $\vec{L}$ vector could be used for exploring for victim. In detail, humpback whales explore arbitrarily give to the location of every other. Consequently, we utilize $\vec{L}$ by arbitrary values larger than 1 or smaller than -1 to force explore agent for moving isolated from a mention whale. In difference for utilization stage, we inform the place of a explore agent in the searching stage provided for an arbitrarily selection explore agent rather than the optimal explore agent found until now. This mechanism and $|\vec{L}| > 1$ underline searching as well as permit the WOA technique for executing a global exploration. The mathematical concepts are pursues:

$$\vec{D} = \left| C \cdot \overrightarrow{P_{rand}} - \vec{P} \right| \quad (7)$$

$$\vec{P}(z+1) = \overrightarrow{P_{rand}} \rightarrow -\vec{L}.\vec{D} \quad (8)$$

where $\overrightarrow{P_{rand}}$ is an arbitrary location vector (a arbitrary whales) selected from the present population.

A few probable locations surrounding an exact result by $\vec{L} > 1$. The WOA technique can be found by a group of arbitrary results. At every iteration, exploring agents inform their locations in terms of an arbitrarily selected explore either agent or optimal solution achieved until now. The attributes are reduced from two to zero that gives searching as well as utilization, correspondingly. An arbitrary explore agents are selected when $|\vec{L}| > 1$; as the optimal solutions are chosen when $|\vec{L}| < 1$ to update the location of explore agents. Based upon the value of$x$, WOA is capable to switch among also a circling or around association. At last, the WOA techniques are ended with the pleasure of a completion reason.

From hypothetical location, WOA could be measured a global optimization as it contains searching/utilization capacity. Besides, the presented hypercube mechanism describes a explore space in the region of the optimal result while allocates further explore agents for searching the present optimal list within that field. Flexible difference from the explore vector L allocates the WOA technique for easily transferring among searching as well as utilization: with reducing$L$, a few iterations are dedicated for search $(|L| \geq 1)$ while the rest is devoted for searching purposes $(|L| < 1)$. Extraordinarily, WOA contains only 2 important internal attributes to be changed $(L\ and\ C)$. Even though mutation and other development functions may contains WOA formulation for complete replicating the performance of humpback whales, we determined for reducing the sum of heuristics while the amount of interior attributes thus executing a very fundamental versions of the WOA technique.

The limitation of WOA lies in the local trapping of optimum values which is resolved by the use of tumbling effect. The new solutions are created by the use of bacterial foraging algorithm (BFO). The motion of bacteria in the human intestine while searching the nutrient rich location away from harmful place takes place by the use of locomotory organelles called as flagella through the chemotactic motion through swimming or tumbling. In the WOA-TRA model, the choice of whale motion will be decided using the fitness function. When the whales shift toward the optimal fitness value, then the motion of whales is called as swimming. In other cases, every whale follows the chemotactic motion of bacterium.

Two level resource allocation

This paper comprises WOA-TRA technique which is applied in 2 stages: Allocating VM resources to jobs, Allocation of PM resources to VM.

Allocation of VM resources to Jobs

Every task of the client requires some resource necessitates and QoS needs. Every QoS variable is linked to few values which represents the priority over the other ones. The weights of QoS properties could be represented in three types namely absolute weighting, relative weighting, and arbitrary weighting. Here, relative weighting is applied to represent QoS variables. The issue of allocating VM resources to jobs undergo mapping into the construction graph $G_1 = (N_1, L_1)$. The node set N_1 comprises of every VM

and jobs. A set of L_1 edges completely links all the nodes of the graph G_1. Every individual edge (a, g) of the graph G_1 is allocated to the bubblenet and is shown below.

$$\tau_{ag} = \frac{1}{la} \quad (9)$$

where a and g indicates exclusive identification number of a job, and number of a VM, la is the length of the job a. Since the inverse of job length is employed as the bubblenet nature, more importance is provided to shorter jobs over the longer ones.

Allocation of PM resources to VM

Here, the process of allocating resources undergo mapping to a construction graph $G_2 = (N_2, L_2)$. The node set N_2, comprises VM and PMs. L_2 indicates a collection of edges which are completely linked nodes. Every edge (g, j) of the graph G_2 is linked to the bubblenet τ_{gj} and heuristic information η_{gj}, where g isthe identification number of a VM and j is identification number of a PM. It can be represented as

$$\tau_{gj} = \frac{1}{\frac{vm_g}{FM_j}} \quad (10)$$

where vm_g is memory requirements of VM g, and FM_j is available memoryspace of PM j.

PERFORMANCE VALIDATION

Simulation setup: The experimental validation of WOA-TRA technique takes place using CloudSim environment. To analyze the results, a set of methods used for comparative analysis are FFD, EARA and MGGA. A set of 5 data centres were generated with the specifications as provided in [Table 1]. At every data centre, PM fulfils with terms, as provided in [Table 2], were generated. For comparison of the experimental results, a set of 4 kinds of VMs are employed as given in [Table 3]. The intention of executing jobs with various QoS necessitates is to validate the goodness of the WOA-TRA with respect to energy efficiency, number of PM needed, and quality of services. The simulation results of the WOA-TRA technique is investigated under different job count. The experiments were iterated for 25 rounds.

Table 1: Details of Data Centres

Name	Processing Cost	Memory Cost	Storage Cost	Bandwidth Cost	Time Zone
Data Centre 1	3	0.05	0.10	0.10	3.0
Data Centre 2	3.5	0.07	0.10	0.11	5.0
Data Centre 3	4	0.09	0.10	0.07	5.5
Data Centre 4	5	0.10	0.10	0.13	8.0
Data Centre 5	5.25	0.12	0.10	0.15	10.0

Table 2: Details of PMs

PM Type	CPU	Cores	RAM	Storage (TB)	Bandwidth (gbps)
1	1000	4	8	2	10
2	1500	8	16	2	10
3	2000	12	32	2	10
4	3000	20	64	4	10
5	5000	36	64	4	10

Table 3: Details of VMs

VM Type	CPU	Number of Cores (PEs)	RAM	Bandwidth (gbps)
1	500	1	512	1
2	1000	2	1024	2
3	2000	4	2048	4
4	4000	8	4096	8

RESULTS AND DISCUSSION

Fig. 4 illustrates the comparative results of the presented WOA-TRA technique with respect to PMs utilized by various models for fulfilling the computational needs of particular job counts. The presented WOA-TRA model offers better results over all the compared three methods with respect to varying PMs utilized for deploying various jobs. [Fig. 5] portrays the comparative analysis of the total energy consumption takes place by various methods. The total energy utilization of the proposed method is lower than the energy

utilized by compared methods due to the fact that it utilizes few PMs for deploying provided jobs. It can be noted that the FFD model consumed maximum energy utilization and offered poor performance over the compared methods. At the same time, slightly lower and near identical energy consumption is achieved by MGGA and EARA methods. But, the presented WOA-TRA model offers minimum energy consumption over all the compared methods.

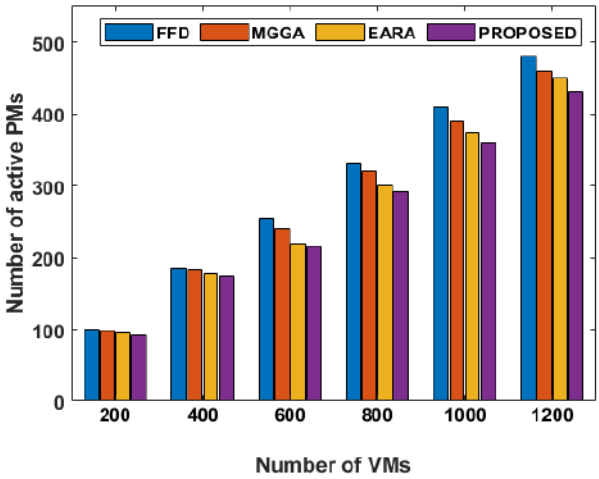

Fig. 4: Results analysis of diverse methods in terms of number of active PMs

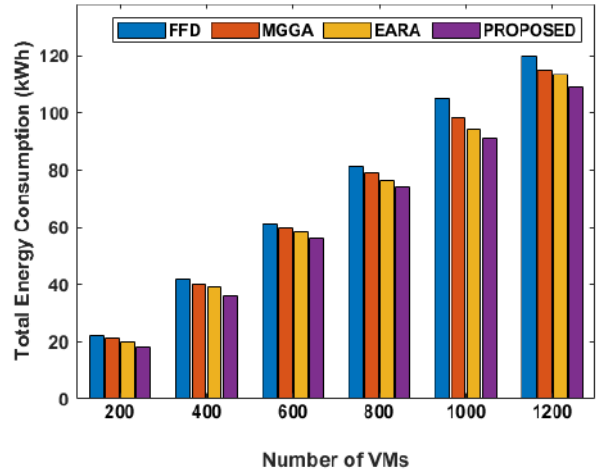

Fig. 5: Results analysis of diverse methods in terms of total energy consumption

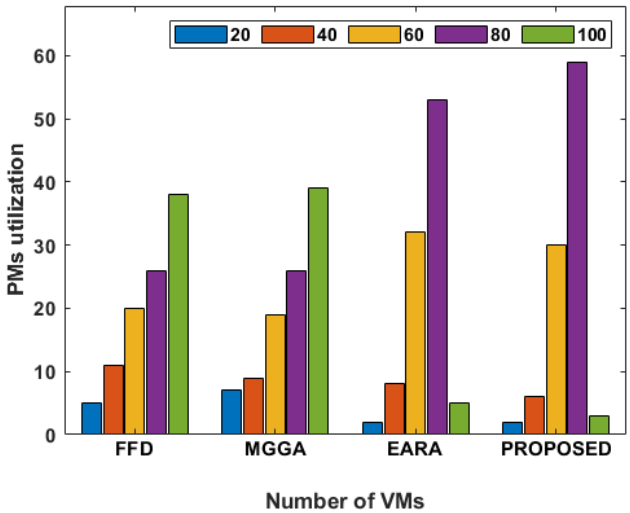

Fig. 6: Results analysis of diverse methods in terms of PMs utilization

Fig. 6 shows the comparison of the PM resource utilized by various methods. The figure clearly stated that the presented WOA-TRA technique has the capability to manage the resources in an effective way. [Fig. 7]

offers a comparative analysis of different methods with respect to average number of VM migrations. It is observed that the FFD method exhibits lower migrations over all the compared methods. It is also noted that maximum number of migrations are carried out using MGGA technique. Next to that, the EARA and presented WOA-TRA techniques offers moderate number of migrations under varying VMs. Since migration takes place to consolidate PMs, energy conservation can be achieved by switching them to sleep state. It is also noted that around 2 migrations takes place to complete 1200 jobs and it does not have any influence of the system performance. Besides, the WOA-TRA will balance the energy loss by migration through the switching off idle PMs to sleep mode.

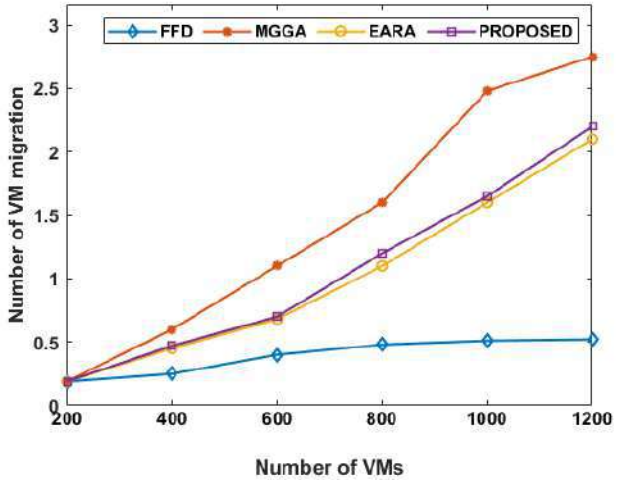

Fig. 7: Results analysis of diverse methods in terms of number of VM migrations

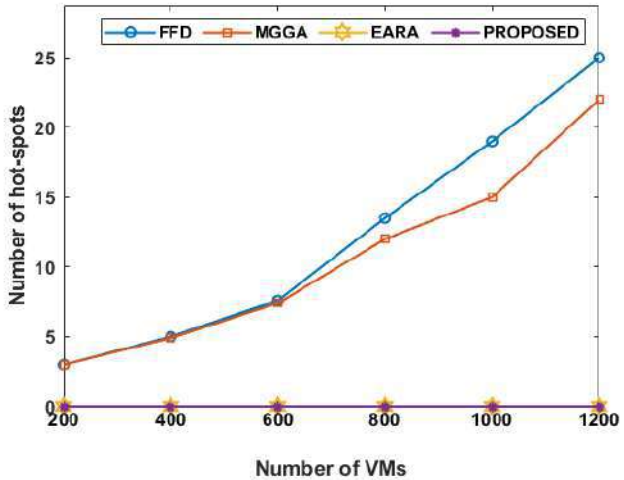

Fig. 8: Results analysis of diverse methods in terms of hot spots.

Fig. 8 displays the comparative results of various methods with respect to the creation of hot spot under varying jobs from 200 to 1200. A PM can be considered as a hot spot when it utilizes the resource up to 100 %. Here, more hot spots are generated by the FFD model due to the fact that it tried to make use of PM to its total capacity. But, the WOA-TRA model does not produce any hot spot which considerably degrades the performance and reliability of PM. Furthermore, the generation of hot spots requires more cooling systems and also raises the possibility of hardware failure. Therefore, WOA-TRA is considered to have more reliability and energy efficiency. [Fig. 9] offered a comparative analysis of diverse methods with respect to total energy utilization by CC platform. The figure indicated that the energy consumption of the presented WOA-TRA model is significantly lower than other methods.

Finally, the computational energy of the presented and compared methods takes place and is shown in Fig. 10. It indicates the total energy utilized usually determined in Watt hours (Wh), to find appropriate resource to every job. It is shown that minimum computation energy is required by the presented model resources for all the jobs.

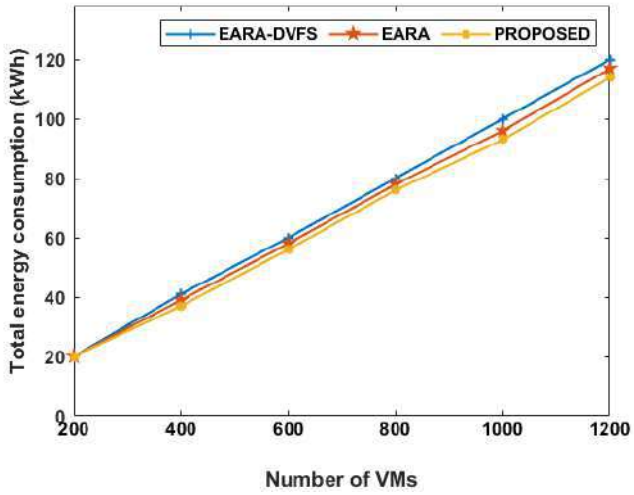

Fig. 9: Results analysis of diverse methods in terms of total energy consumption under varying VMs.

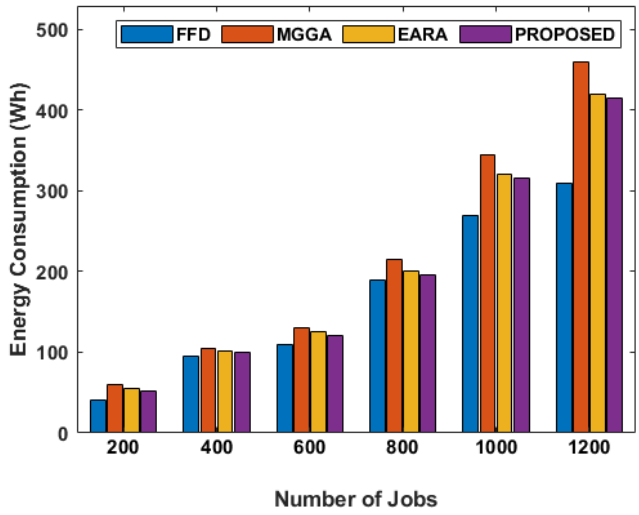

Fig. 10: Results analysis of diverse methods in terms of total energy consumption under varying jobs

This figure shows that the EARA consumes less energy in computation than other methods except FFD which achieves slightly lower computation energy. On observing the experimental results, it can be ensured that the presented model offers maximum performance by effectively allocates the resources in the cloud platform.

CONCLUSION

This paper has introduced a WOA-TRA for resource allocation problem in cloud environment. The WOA is hybridized with tumbling effect which has good exploration ability for function optimization problems and derives a WOA with tumbling effect based energy aware solution called WOA-TRA model. To effectively utilize the energy and QoS requirements, the WOA-TRA technique is utilized in two levels. In the first level, WOA-TRA technique assigns VMs resources to jobs, whereas in second level, WOA-TRA technique assigns PMs resources to VMs. The experimental analysis stated that the presented WOA-TRA technique offered desired QoS and enhanced energy efficiency by effectively utilizes the available resources over the compared methods.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] Fox A, Griffith R, Joseph A, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I. [2009] Above the clouds: A berkeley view of cloud computing. Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS, 28(13):2009.

[2] Kumar A, Kumar R, Sharma A. [2018] Energy Aware Resource Allocation for Clouds Using Two Level Ant Colony Optimization. Computing and Informatics, 37(1):76-108.

[3] Koomey J. [2011] Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times, 9(20):161.

[4] Kaplan JM, Forrest W, Kindler N. [2008] Revolutionizing Data Center Energy Efficiency. Technical report, McKinsy & Company.

[5] Vogels W. [2008] Beyond Server Consolidation. Queue, 6(1):20–26.

[6] Beloglazov A, Abawajy J, Buyya R. [2012] Energy-Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing. Future Generation Computer Systems, 28(5):755–768

[7] Kinger S, Kumar R, Sharma A. [2014] Prediction Based Proactive Thermal Virtual Machine Scheduling in Green Clouds. The Scientific World Journal, doi: 10.1155/2014/208983.

[8] Quarati A, Clematis A, Galizia A, D'Agostino D. [2013] Hybrid Clouds Brokering: Business Opportunities, QoS and Energy-Saving Issues. Simulation Modeling Practice and Theory, doi: 10.1016/j.simpat.2013.01.004

[9] Lee HM, Jeong YS, Jang HJ. [2014] Performance Analysis Based Resource Allocation for Green Cloud Computing. The Journal of Supercomputing, 69(3):1013–1026.

[10] Raycroft P, Jansen R, Jarus M, Brenner PR. [2014] Performance Bounded Energy Efficient Virtual Machine Allocation in the Global Cloud. Sustainable Computing: Informatics and Systems, 4:1–9.

[11] Feller E, Rilling L, Morin C. [2011] Energy-Aware Ant Colony Based Workload Placement in Clouds. 12th IEEE/ACM International Conference on Grid Computing (GRID), doi: 10.1109/Grid.2011.13.

[12] Gao Y, Guan H, Qi Z, Hou Y, Liu L. [2013] A Multi-Objective Ant Colony System Algorithm for Virtual Machine Placement in Cloud Computing. Journal of Computer and System Sciences, 79:1230–1242.

[13] Nathani A, Chaudhary S, Somani G. [2012] Policy Based Resource Allocation in IaaS Cloud. Future Generation Computer Systems, 28(1):94–103.

[14] Chen H, Cheng AMK, Kuo YW. [2011] Assigning Real-Time Tasks to Heterogeneous Processors by Applying Ant Colony Optimization. Journal of Parallel and Distributed Computing, 71(1):132–142.

[15] Huang CJ, Guan CT, Chen HM, Wang YW, Chang SC, Li CY, Weng CH. [2013] An Adaptive Resource Management Scheme in Cloud Computing. Engineering Applications of Artificial Intelligence, 26(1):382–389.

[16] Garg SK, Yeo CS, Buyya R. [2011] Green Cloud Framework for Improving Carbon Efficiency of Clouds. Euro-Par 2011 Parallel Processing. Lecture Notes in Computer Science, doi: 10.1007/978-3-642-23400-245.

[17] Xu J, Fortes JAB. [2010] Multi-Objective Virtual Machine Placement in Virtualized Data Center Environments. IEEE/ACM International Conference on Green Computing and Communications (Green Com), doi: 10.1109/GreenCom-CPSCom.2010.137.

[18] Wu CM, Chang RS, Chan HY. [2014] A Green Energy-Efficient Scheduling Algorithm Using the DVFS Technique for Cloud Datacenters. Future Generation Computer Systems, doi: 10.1016/j.future.2013.06.009.

[19] Mirjalili S, Lewis A. [2016] The whale optimization algorithm. Advances in engineering software, 95:51-67.

THE IIOAB JOURNAL

EDITED BY: Prof. Dr. S. Vaithyasubramanian

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

MATHEMATICAL APPROACH TOWARDS RECENT INNOVATION IN COMPUTATION AND ENGINEERING SYSTEM (MATRICS)

AN EFFECTIVE STOCK MARKET DIRECTION PREDICTION MODEL USING WATER WAVE OPTIMIZATION WITH MULTI-KERNEL EXTREME LEARNING MACHINE

Mohan Jeyakarthic[1*], Sachithanantham Punitha[2]
[1]*Tamil Virtual Academy, Chennai, Tamil Nadu, INDIA*
[2]*Dept. of Computer Science, DGGA College for Women, Tamil Nadu, INDIA*

ABSTRACT

Presently, forecasting of stock market return is commonly considered as a prediction problem. The intrinsic volatile characteristic of stock market all over the globe makes the forecasting procedure a difficult process. The reduction in predictive error rate would considerably decrease the risks in investment processes. This paper introduces a new hybridization of water wave optimization with multi-kernel extreme learning machine (WWO-MKELM) for stock market return prediction. The presented WWO-MKELM model comprises three major steps namely preprocessing, feature extraction and classification. At first, preprocessing is carried out by the use of exponential smoothing technique. Subsequently, the features will be extracted from the preprocessed dataset. Next, WWO-MKELM based model is employed to predict the stock prices. The presented WWO-MKELM model has the ability to predict whether the stock prices will be raised or reduced in advance. The WWO-MKELM technique is simulated by the use of Apple (APPL) and Facebook (FB) stocks. The attained experimental results defined that the WWO-MKELM model has provided better performance over the compared methods.

KEY WORDS
Stock market, Prediction, Classification, WWO algorithm, Machine learning

INTRODUCTION

In last decades, the effective differences in stock prices cannot be detected in previous days. Random Walk [1], as well as Efficient Market Hypothesis (EMH) revealed that a market would be considered as robust and efficient on the basis of recent information, if it is not feasible to detect the market flow due to the randomness of stock prices, then it is similar to the risks which can be recovered, economical gain cannot be enhanced. The task of stock market price detection is highly complex and it provides a maximum return from increasing measure for risk prediction. Followed by, the Wisdom of Crowd strategy pointed that diverse individuals can provide accurate estimation of information has been recovered appropriately. Still it becomes unknown and it is not applicable to detect the result of stock market; regardless, some of the individuals and organizational shareholders are capable of striking the market to make better profits. An incomplete detection is simplified due to diverse irregularities present and because of the presence of massive parameters that influences the market value per day. Consequently, stock markets are highly vulnerable for differences for prominent for making random conflicts in stock prices.

Various economical as well as statistical professionals have invoked to refer that the stock market prices could be predicted partially. Followed by, a novel economical expert has pointed the mental and behavioral elements of stock-price computation as well as volatility [2]. Followed by, only few professionals have stated the determining patterns would enable the shareholder to accomplish maximum gain from risk-based values. Recently, stock market detection methods are deployed using Machine Learning (ML) as well as Data Mining (DM) methodologies. Some of the related works are defined in the following. Predictive approaches have been applied to forecast the future patterns in stock market operations that offer a method to improve the determining capabilities and predefine efficient market strategy as well as diffusion schemes [3]. ML, besides, which is set of present approaches. Finally, diverse methods are applied to forecast stock prices such as Support Vector Machine (SVM), Deep Neural Network (DNN) [4], Random Forest (RF), naïve bayes (NB), and so on to accomplish the method of predictability with maximum efficiency. Simultaneously, DM methods are highly employed with regular stock data.

Generally, Autoregressive Integrated Moving Average (ARIMA) framework [5] has been applied for finding and identifying differences in time series. In Dai and Zhang, 2013 [6], closing prices were employed for analyzing the firm of 3M that have data from the interval of September 2008-August 2013. Massive methods have been used for detective approaches and applied for forecasting the dimension of stocks on effective day's data sample. It is also capable of predicting the amount for future n days. It is stated that US stock market is partially robust, which denotes technical as well as fundamental analysis cannot be employed and attain higher gain. However, the prolonged predictive model offers best accuracy that had surpassed if the time window is 44 days. Then, SVM has reached optimal accuracy.

In Di, 2014 [7], collection of 3 stocks has been applied with the help of symptoms like RSI, on balance volume, Williams's %R, and so forth. Over these features, highly randomized tree model, as defined by Geurts and Louppe, 2011 [8], for Feature Selection (FS) that is provided to SVM with Radial Basis Function (RBF) for training a method. It is identical which the stock market price values are highly non-static, non-parametric, chaotic and noisy by default that provides in threatening investment. Followed by, the trends of stock market prices are considered as a random task with alterations that is important shortly. It is

Received: 26 June 2020
Accepted: 23 July 2020
Published: 28 July 2020

***Corresponding Author**
Email:
jeya_karthic@yahoo.com

EDITED BY: Prof. Dr. S. Vaithyasubramanian

IIOAB JOURNAL
ISSN: 0976-3104

essential to reveal the latest aspects of upcoming stock price trends should help in limiting the risk. Dealers are extremely a stock in current time periods with the values in future days. Moreover, it is pointed that the exact detection of movements in stock market prices would enhance the profit as well as to limit the loss. Hence, it is essential to deploy a new method that detects the direction of trends in stock prices.

Devi et al., 2015 [9] projected a hybridization of Cuckoo Search (CS) with SVM approaches in which CS method is employed for tuning the SVM parameters. Giacomel et al., 2015 [10] projected a trading agent with the application of Neural Network (NN) ensemble which detects the particle stock when it is minimum or maximum. Boonpeng and Jeatrakul, 2016 [11] applied a one-against all (OAA-NN) and one-against-one neural network (OAO-NN) to classify the purchasing, capturing or selling data and they are compared with the outcomes provided by conventional NN. In Qiu and Song, 2016 [12], an optimized ANN by applying Genetic Algorithm (GA) is used to predict the stock market prices identically.

Alrasheedi and Alghamdi, 2014 [13] used diverse classes of classification methods for SPP in Saudi stock exchange from 2006-2013. Dow Jones dataset has been utilized with 5-fold cross validation method. Milosevic, 2016 [14] presented a ML scheme for investigating the security of future cost for longer period. It is highly applicable in finding exact improvement in organizational measures in last decades. The main aim of applying ML model is to train the previous data that is suitable for predicting the stock price and to search the trends for certain time period. Here, diverse classification approaches are related for SPP. Once the comparison is completed, RF accomplished better results interms of Precision, F-score and recall.

Leung et al., 2014 [15] employed a structural SVM (SSVM) to predict the SSP. Hence, the developed model activates SSVM for learning predictive approach while complicated graph input such as massive edges of nodes. Therefore, results of SPP have positive and negative class labels that represented the enhancement as well as minimization in stock prices. Additionally, it is used under 3-fold cross validation to explore original measure as well as SSVM feature C is set as. Finally, it accomplished the maximum accuracy that has ensured that this model undergoes training with no application of over-fitting. Thus, the working principle of ML approach is highly applicable in detecting the stock prices.

Qiu and Song, 2016 [16] introduced an Artificial Neural Network (ANN) for SPP in Japanese stock exchange. It focuses in identifying the upcoming stock prices. To improve the classification accuracy, ANN is integrated with GA that generates a GA-ANN approach for attaining efficient SPP. In this model, GA is elected to enhance the accuracy of ANN and to eliminate the converging issues from back-propagation (BP) method. Hence, processing analysis is allocated as hybrid GA-ANN that assists to attain a better hit value which has to be higher than earlier model. Guo et al., 2015 [17] Implemented a hybrid approach which combined 2-D Principal Component Analysis (2D) (PCA) and RBF NN for SPP at Shanghai stock market. It has been selected with 36 stock market attributes as input values, in which a sliding window is used for retrieving input data. Besides, 2D-PCA is suitable to limit the dimensions of data and filter of intrinsic parameters. Consequently, RBFNN applies the data processed using 2D-PCA for detecting the upcoming stock price. Therefore, the simulation outcome represented that the applied method performs quite-well than MLP model.

Alkhatib et al., 2013 [18] applied k-Nearest Neighbor (KNN) technique as well as non-linear regression scheme for SPP in major organizations that is listed in Jordanian stock exchange which is applicable for users and suppliers in effective decision making operation. Based on the attained result, kNN approach should be effective and applicable to reach the lower error while forecasting the outcome that is corresponding to actual stock prices. Guo et al., 2014 [19] developed an SPP scheme using PCA, canonical correlation analysis (CCA) and SVM methodologies. First, 2 parameters have been filtered from previous closing cost as well as 39 scientific features retrieved from independent component analysis. As a result, SVM is applied to detect upcoming stock price.

This paper introduces a new hybridization of water wave optimization with multi-kernel extreme learning machine (WWO-MKELM) for stock market return prediction. The presented WWO-MKELM model comprises three major steps namely preprocessing, feature extraction and classification. At first, preprocessing is carried out by the use of exponential smoothing technique. Subsequently, the features will be extracted from the preprocessed dataset. Next, WWO-MKELM based model is employed to forecast the stock prices. The presented WWO-MKELM model has the ability to predict whether the stock prices will be raised or reduced in advance. The WWO-MKELM technique is simulated by the use of Apple (APPL) and Facebook (FB) stocks. The attained experimental results defined that the WWO-MKELM model has provided better performance over the compared methods.

MATERIALS AND METHODS

The entire task of the projected WWO-MKELM approach is showcased in [Fig. 1]. The presented approach is enclosed with pre-processing, feature extraction, as well as classification. The above mentioned sub processes are defined in the upcoming sections.The entire task of the projected WWO-MKELM approach is showcased in [Fig. 1]. The presented approach is enclosed with pre-processing, feature extraction, as well as classification. The above mentioned subprocesses are defined in the upcoming sections.

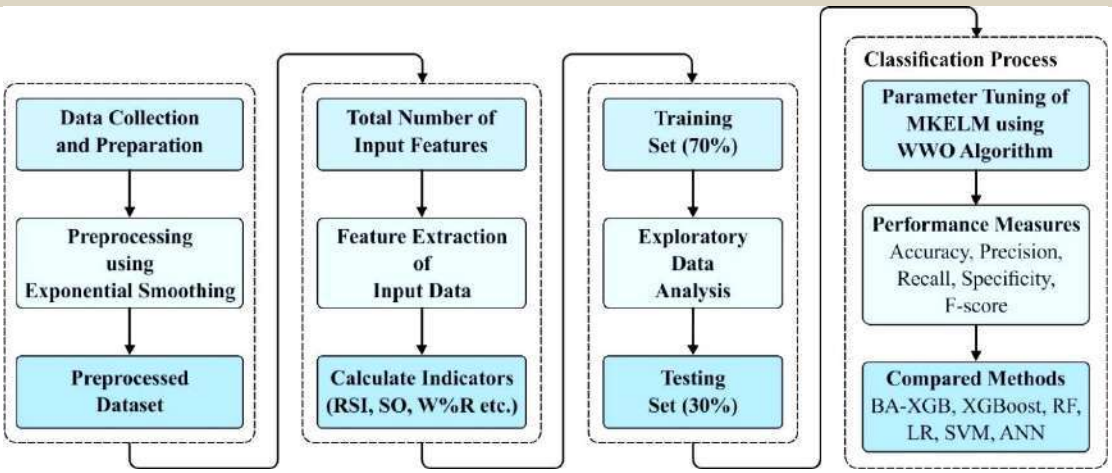

Fig. 1: Block diagram of WWO-MKELM Model.

Preprocessing: The upgraded smoothing offers maximum weights for recent observations and limited as same as existing observation. The statistic sequence of Y is determined in recursive manner using:

$$S_0 = Y_0$$

$$\text{for } z > 0, S_z = \alpha * Y_z + (1-\alpha) * S_z\text{-}1 \qquad (1)$$

where α refers the smoothing factor and $0 < \alpha < 1$. Higher value of α is applicable to reduce the smoothing phase. When $\alpha = 1$, smoothed statistic is same as actual processing. In addition, smoothed statistic S_z is determined rapidly as prominent observations are suitable. Hence, the smoothing assists in avoiding random variations from historical data by activating the model for discovering a prolonged stock price.

The technical detecting is determined from exponentially smoothed time period which is developed as feature matrix. Additionally, target is detected in u^{th} day that determines using the given below:

$$targ_u = sign(close_{u+d} - close_u) \qquad (2)$$

where d implies value of days. When the value of $targ_u$ is $+1$, it represents the existence of positive shift in price in advanced d days; while -1 denotes the presence of negative shift after d days, which provides clear definition for corresponding stock price. Additionally, $targ_u$ measures are allocated as labels for uth row in feature matrix.

Feature extraction process: Here, the closing price of stock is considered and collect these metrics in past decades. Hence, input data is described as $(date, price_{closing})$. Therefore, data is composed of some detectors which are determined [20].

Relative strength index (RSI)

In general, a stock is overbought while there is demand enhanced the money. Therefore, it interprets the stock overvalue and reduced price correspondingly. In addition, it is meant as oversold at the time of there is a limitation of price under a positive value. The result is projected due to the panic sold task.

Stochastic oscillator (SO)

SO utilizes the trend of price. Based on this strategy, momentum is changed. It determines the density of closing price which is related to minimum to maximum range over a specific time limit.

Williams percentage range (W%R)

It is named as Williams %R is a secondary momentum predictor, which is similar to SO. It shows the markets closing price level in combination of higher amount for look-back time period. Such value from -100 to 0. When the measure is higher than -20, it is a sell point while it belongs to -80.

Moving average convergence divergence (MACD)

It is described as a momentum predictor which combines 2 moving averages of stock. Primary average is around 26-day Exponential Moving Average (EMA) and then moving average is 12-day EMA.

IIOAB JOURNAL
ISSN: 0976-3104

Price rate of change (PROC)

It is referred as a scientific predictor which represents the ratio of modification in price between recently fixed price and existing price that is monitored for certain time period.

On balance volume (OBV)

It uses the alteration in quantity to estimate the modification in stock prices. The technical indicator is utilized for finding buying and selling moments of a stock, by considering the aggregating volume: as it encloses the volumes for days when the price is enhanced, and reduce the quantity of the price is limited, and correlated with the existing day price correspondingly.

WWO-MKELM based Classification Model: ELM is a single-hidden layer of feedforward neural network (FF-NN). It defines the input weights as well as biases in random manner and explanatory calculates the resultant weights in spite of tuning regularly. The ELM has incoming weights ω, bias b, hidden nodes L, and output weights α.

An effective ELM is represented in mathematical format as given in the following:

$$H\alpha = T \quad (3)$$

$$H = \begin{bmatrix} h_1(x_1) & \cdots & h_L(x_1) \\ \vdots & & \vdots \\ h_1(x_n) & & h_L(x_n) \end{bmatrix} \quad (4)$$

$$= \begin{bmatrix} G_1(\omega_{111}x + b_1) & \cdots & G_L(\omega_{1L} \cdot x_1 + b_L) \\ \vdots & & \vdots \\ G_1(\omega_{n1} \cdot x_n + b_1) & & G_L(\omega_{n1} \cdot x_n + b_L) \end{bmatrix}$$

where Himplies the hidden layer output matrix, α represents the final output weight matrix, Tdefines the matrix of target output, and $G_L$signifies the activation function applied in all hidden neurons. The final weight $(\alpha)$is attained by resolving the Eq. (3) under the application of Moore-Penrose normalized inverse of H:

$$\alpha = H^{\dagger}T \quad (5)$$

In order to enhance the ability of ELM, it reformed into (5) as:

$$\alpha = H^T\left(\frac{1}{C} + HH^T\right)^{-1}T \quad (6)$$

where C defines the regularization variable, and resultant function of ELM is provided in the following:

$$y = h(x)\alpha = h(x)H^T\left(\frac{1}{C} + HH^T\right)^{-1}T \quad (7)$$

The ELM function is enhanced under the application of sum of diverse activation functions and kernel functions.

A KELM is defined as an ELM with kernel functions from hidden layer nodes. A kernel function has been applied for data matching for high-dimensional feature space and transforms a nonlinear into linear issues. The kernel function is described under the application of Mercer's constraints for unstable feature mapping $h(x)$:

$$k(x_i, x_j) = h(x_i) \cdot h(x_j) \quad (8)$$

$$K = HH^T$$

where $k(x_i, x_j)$ defines the kernel function as well as K describes the kernel matrix. The resultant function of KELM is attained by replacing Eq. (8) into Eq. (7):

$$y = h(x)\alpha = \begin{bmatrix} k(x,x_1) \\ \vdots \\ k(x,x_n) \end{bmatrix}\left(\frac{1}{C} + K\right)^{-1}T \quad (9)$$

The normalization function as well as learning capability of KELM is based on the class of kernel function employed in hidden layer nodes. It is also composed of massive kernel functions that are classified as 2 major classes such as Global and Local. Initially, Global kernel functions like linear and polynomial kernel

functions contains massive generalization function; however, a vulnerable for learning ability and influenced by instances away from one another whereas local kernel functions like Gaussian and wavelet kernel functions are composed of robust learning capability, hence, the lower normalization function is influenced by samples nearby one another.

A linear integration of diverse kernel functions offers a multi-kernel function which meets the Mercer's conditions. An MKELM is attained by applying multi-kernel function from hidden nodes of KELM. It is present with a managing attribute (λ) among diverse kernels are represented as given below:

$$k(x_i, x_j) = \sum_{n=1}^{N} \lambda_n k_n(x_i, x_j) \qquad (10)$$

Here, it has been examined with the efficiency of MKELM for categorization as well as position of error in under the application of 2 kernel functions like, wavelet and polynomial:

$$k(x_i, x_j) = \lambda k_1(x_i, x_j) + (1 - \lambda) k_2(x_i, x_j) \qquad (11)$$

$$0 \leq \lambda \leq 1$$

where $k_1(x_i, x_j)$ defines the Morlet wavelet kernel function that is provided below:

$$k_1(x_i, x_j) = \cos\left(1.75\left(\frac{x_i - x_j}{\gamma}\right)\right) e^{\frac{(x_i - x_j)^2}{2\gamma^2}} \qquad (12)$$

and $k_2(x_i, x_j)$ refers the polynomial kernel function that is represented as:

$$k_2(x_i, x_j) = (1 + x_i \cdot x_j)^d \qquad (13)$$

Where, γ and d implies the dilation of wavelet kernel as well as polynomial degree, correspondingly. Such attributes should be normalized for MKELM by WWO algorithm to accomplish best function.

WWO algorithm is developed from shallow water wave techniques for solving optimization problems [21].

RESULTS

Dataset used: For experimentation, data from 2 firms such as Facebook (FB) and Apple (APPL) as well as other accessible data has been applied. These organizations are tested in random manner, without adverse considerations from the background or financial impact which offered in public. It is emphasized that the diversity of these firms are chosen for analyzing stock prices which is important to ensure the efficiency of this model. The actual values are obtained from data that has been attained from entry of data, closing price, volume, and so forth. From the actual format, size of parallel data to stocks of diverse organizations are varied from 10 kB to700 kB, under the application of rows which corresponds to closing prices and changes from 1180 and 10,700 [20].

Results Analysis: [Table 1] shows the results analysis of the WWO-MKELM model interms of distinct methods under different trading window sizes.

Table 1: Results of classification using WWO-MKELM

Company Name	Trading Window	Accuracy	Recall	Precision	Specificity	F-Score
AAPL Stock	3	67.90	74.00	70.00	60.00	70.00
	5	75.76	82.00	76.00	59.00	77.00
	10	80.15	85.00	85.00	79.00	84.00
	15	84.90	87.00	86.00	80.00	86.00
	30	87.83	90.00	89.00	83.00	89.00
	60	92.48	96.00	94.00	89.00	95.00
	90	96.23	98.00	96.00	93.00	96.00
FB Stock	3	69.44	76.00	70.00	66.00	75.00
	5	76.40	88.00	72.00	64.00	80.00

	10	83.50	94.00	84.00	72.00	87.00
	15	88.22	93.00	93.00	84.00	92.00
	30	90.47	98.00	95.00	85.00	96.00
	60	90.27	99.00	92.00	63.00	97.00
	90	97.14	99.00	99.00	75.00	99.00

DISCUSSION

A comparative analysis of the WWO-MKELM model with existing models [20] interms of accuracy is made, and the outcome is tabulated in [Table 2] and [Fig. 2]. The table values defined that the LR and SVM models have demonstrated ineffective prediction results, which has resulted to a minimum accuracy of 55% and 58%.

Table 2: Comparisons of Proposed with Existing Methods

Methods	Accuracy
WWO-MKELM	97.14
BA-XGB	96.42
XGBOOST	83.00
RF	92.00
LR	55.00
SVM	58.00
ANN	72.00

On the other side, the ANN model leads to a slightly higher accuracy value of 72%. Along with that, the XGBoost model has exhibited a moderate classifier accuracy of 83%. In line with that, the RF model has demonstrated somewhat higher accuracy of 92%. Furthermore, the competitive predictive accuracy of 96.42% has been achieved by the BA-XGB model whereas the maximum accuracy of 97.14% has been obtained by the presented WWO-MKELM model.

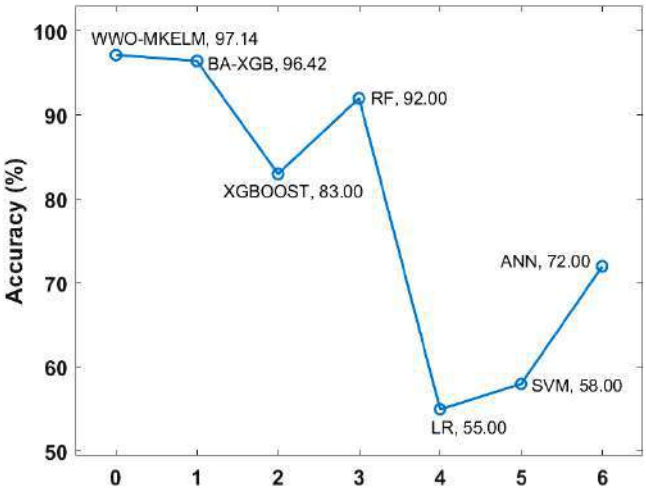

Fig. 2: Accuracy analysis ofWWO-MKELM model with existing models

CONCLUSION

This paper has developed a new WWO-MKELM model for stock market return prediction. The presented WWO-MKELM model initially performs preprocessing using exponential smoothing technique. Next, the features will be extracted from the preprocessed dataset. Afterwards, WWO-MKELM based model is employed to detect the stock prices. The presented WWO-MKELM model has the ability to predict whether

IIOAB JOURNAL
ISSN: 0976-3104

the stock prices will be raised or reduced in advance. The WWO-MKELM technique is simulated by the use of APPL and FB stocks. The simulation results attained that the WWO-MKELM model has showcased effective outcome with the maximum accuracy of 97.14%. Then, the function of WWO-MKELM model has been increased by the use of outlier detection techniques.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Malkiel BG. [2003] The efficient market hypothesis and its critics. The Journal of Economic Perspectives, 17(1): 59–82.

[2] Veeramanikandan V, Jeyakarthic M. [2019] Forecasting of Commodity Future Index using a Hybrid Regression Model based on Support Vector Machine and Grey Wolf Optimization Algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 10(10): 2278-3075.

[3] Veeramanikandan V, Jeyakarthic M. [2019] An Ensemble Model of Outlier Detection with Random Tree Data Classification for Financial Credit Scoring Prediction System. International Journal of Recent Technology and Engineering (IJRTE), 8(3): 2277-3878.

[4] Murugan S, Jeyakarthic. [2019] Optimal Deep Neural Network based Classification Model for Intrusion Detection in Mobile Adhoc Networks. Jour of Adv Research in Dynamical & Control Systems, 11(10):1374-1387.

[5] Pai PF, Lin CS. [2005] A hybrid arima and support vector machines model in stock price forecasting. Omega, 33(6):497–505.

[6] Dai Y, Zhang Y. [2013] Machine learning in stock price trend forecasting. Stanford University http://cs229.stanford.edu/proj2013/DaiZhangMachineLearningInStockPriceTrendForecasting.pdf.

[7] Di X. [2014] Stock trend prediction with technical indicators using SVM. Stanford University, DOI: 10.2991/ammsa-17.2017.45.

[8] Geurts P, Louppe G. [2011]. Learning to rank with extremely randomized tree. JMLR: Workshop and Conference Proceedings, 14:49–61

[9] Devi KN, Bhaskaran VM, Kumar GP. [2015] Cuckoo optimized SVM for stock market prediction. In: IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICJJECS), DOI: 10.1109/ICIIECS.2015.7192906.

[10] Giacomel F, Galante R, Pareira A. [2015]. An Algorithmic Trading Agent based on a Neural Network Ensemble: a Case of Study in North American and Brazilian Stock Markets. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, DOI: 10.1109/WI-IAT.2015.43.

[11] Boonpeng S, Jeatrakul P. [2016]. Decision support system for investing in stock market by using OAA-neural network. In: 8th International Conference on Advanced Computational Intelligence Chiang Mai, Thailand, DOI: 10.1109/ICACI.2016.7449794.

[12] Qiu M, Song U. [2016]. Predicting the direction of stock market index movement using an optimized artificial neural network model. PLoS One, 11(5): 155-133.

[13] Alrasheedi M, Alghamdi A. [2014] Comparison of Classification Methods for Predicting the Movement Direction of Saudi Stock Exchange Index, Journal of Applied Sciences, 14(16):1883-1888.

[14] Milosevic N. [2016] Equity forecast: Predicting long term stock price movement using machine learning. arXiv preprint arXiv:1603.00751.

[15] Leung CK, MacKinnon RK, Wang Y. [2014] A machine learning approach for stock price prediction. Proceedings of the 18th International Database Engineering & Applications Symposium. ACM, https://doi.org/10.1145/2628194.2628211

[16] Qiu M, Song Y. [2016] Predicting the direction of stock market index movement using an optimized artificial neural network model. PloS one, 11(5):155-133.

[17] Guo Z, Wang H, Yang J, Miller DJ. [2015] A stock market forecasting model combining two-directional two-dimensional principal component analysis and radial basis function neural network. PloS one, 10(4): 122-385.

[18] Alkhatib K, Najadat H, Hmeidi I, Shatnawi MK. [2013] Stock price prediction using k-nearest neighbor (knn) algorithm. International Journal of Business, Humanities and Technology, 3(3):32-44.

[19] Guo Z, Wang H, Liu Q, Yang J. [2014] A feature fusion based forecasting model for financial time series. PloS one, 9(6):101-113.

[20] Jeyakarthic M, Punitha S. [2020] Hybridization of Bat Algorithm with XGBOOST Model for Precise Prediction of Stock Market Directions, International Journal of Engineering and Advanced Technology (IJEAT), 9(3):3375-3382.

[21] Zheng YJ. [2015] Water wave optimization: a new nature-inspired metaheuristic. Computers & Operations Research, 55:1-11.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

HYBRIDIZATION OF MODIFIED GREY WOLF OPTIMIZATION WITH INERTIA PARTICLE SWARM OPTIMIZATION BASED CLUSTERING TECHNIQUE IN WIRELESS SENSOR NETWORKS

Arutchelvan Karunanidhi[1*], Sathiya Priya Ramalingam[1], Bhuvaneswari Chandran[2]
[1]*Department of Computer and Information Sciences, Annamalai University, TN, INDIA*
[2]*Department of Computer science, Government Arts and Science College, Thiruvennainallur, TN, INDIA*

ABSTRACT

Wireless Sensor Networks (WSN) become a hot research topic and is commonly employed for data gathering applications. Energy efficiency is treated as an important issue in the design of WSN. As the clustering technique helps to achieve energy efficiency in WSN, several works have been carried out on the proper selection of cluster heads (CHs). Clustering is assumed as an NP hard problem and metaheuristic algorithms are applied for resolving it. This paper presents a hybridization of modified grey wolf optimization algorithm and inertia particle swarm optimization with dynamic velocities (IPSO-DV) called HMGWOIPSO-DV based clustering technique in WSN. The proposed algorithm selects CHs in two levels namely IPSO-DV based tentative CH selection and MGWO based final CH selection. Besides, the presented model derived a fitness function by the use of residual energy, distance to base station (BS), and distance to nearby nodes. The proposed model has the capability of competently choosing the CHs, attains energy efficiency and maximum network lifetime. The performance of the presented model takes place under diverse aspects and the obtained experimental results ensured the goodness of the presented model.

KEY WORDS
Clustering, Energy efficiency, Metaheuristics, GWO algorithm, PSO algorithm

INTRODUCTION

Wireless sensor networks (WSNs) are composed of a massive collection of nodes to observe and save the physical parameters in the environment [1]. In WSN, the node distribution is arbitrary. It is placed in the unmanned regions where the batteries cannot be replaced and it cannot be recharged easily, and the nodes are distributed in a random fashion. The count of nodes is maximum and initial power has been employed for charging the battery. Hence, the power applications as well as network lifespan are the major constraints which affect the network [2]. The data gathered by nodes are forwarded to base station (BS) or sink for computation. Data transmission is processed in a single-hop or multi-hop fashion [3]. The main applications of WSNs are observing forest fires, managing the condition of serious patients, armed forces as well as traffic [4]. Generally, the major challenges of WSN are fault tolerance, scalability, costs, hardware constraints, consistency, WSN topology, transmission environment and power utilization. The 2 methods applied for enhancing the WSNs lifespan are Clustering and Routing. In clustering phase, a collection of sensor nodes are fixed in a class named a cluster on the basis of general parameters.

An effectively qualified node in a cluster is termed as Cluster Head (CH). The responsibility of CH is to gather the data collected from Cluster Members (CM) and send it to BS which depends upon the data transmission. The CHs send the received data to sink using single-hop transmission, while CHs forwards the obtained data to higher-level CHs while multi-hop transmission and higher-level CHs sends the data to sink. Multi-hop transmission is often applied in large-scale networks. Basically, CM is classified into 2 classes which are composed of general nodes and CHs. Thus, few meta-heuristic approaches and Computational intelligence (CI) methodologies like artificial bee colony (ABC), artificial immune systems (AIS), Reinforcement learning (RL), Evolutionary Algorithms (EA) has been employed for the process to resolve the NP-hard optimization problem. Transmitting data to BS from the sensor node is performed by best CH which is a major challenge for routing protocol. Best CH election model results in energy reduction, latency, distance, and so forth.

The combined particle swarm optimization (PSO) and Fuzzy-related CH election approach have been presented to serve an efficient clusters-aided routing process to expand the network lifetime [5]. The CH election model applies the advantages of PSO to accomplish clustering according to the data correlated with the scientific position. Followed by, it is composed of an enhanced PSO method to compute effective CH nodes from a hierarchical topology network. It is processed to enhance the networked lifestyle with a limited degree of sensor nodes' fatality. Besides, a combined Simulated Annealing (SA) and Differential Evolution (DE) related CH election method has been projected for choosing best count of CH nodes in the clustering process [6]. The unified method highly focuses on eliminating the premature death of CH nodes to accomplish a prolonged network lifecycle. LEACH, Harmony Search Algorithm (HSA), modified HSA, and DE methods are employed for analyzing purposes.

The Enhanced Artificial Fish Swarm Algorithm (EAFSA) has been projected to reduce the power application of a system by effective CH election process [7]. The EAFSA has the advantages of foraging characteristic of the fish swarm to extract a feasible count of features that contributes to frequent CH selection. It has the fitness to estimate and validate the ability of sensor nodes to CH conversion. Moreover, it is estimated over GA and variant LEACH aided CH election schemes. An ABC-related CH election process is applied to enhance the effectiveness of the developed cluster according to the evaluation of multi-objective fitness

Received: 24 Aug 2020
Accepted: 12 Oct 2020
Published: 15 Oct 2020

***Corresponding Author**
Email:
karutchelvan@yahoo.com

function [8]. The ABC-based model employed the condition of least hop-count to maintain capable data transmission. It is considered for reducing initial power consumption with improved throughput, packet delivery ratio (PDR) as well as network lifetime. A combined PSO and ACO-based clustering model has been presented to improve the data and energy dispersion effectively [9]. It applied the primitive parameter of Residual Energy (RE) and intra-cluster distance for developing FF which intends to accomplish data aggregation process. It utilized multi-dimensional features of sensors for determining the importance to find the responsibility of CH in a system. Furthermore, it ensured a phenomenal enhancement in network duration than swarm-intelligent (SI) models for examination.

This paper presents a hybridization of modified grey wolf optimization algorithm and inertia particle swarm optimization with dynamic velocities (IPSO-DV) called HMGWOIPSO-DV based clustering technique in WSN. The proposed algorithm selects CHs in two levels namely IPSO-DV based tentative CH selection and MGWO based final CH selection. Besides, the presented model derived a fitness function by the use of residual energy, distance to base station (BS), and distance to nearby nodes. The proposed model has the capability of competently choosing the CHs, attains energy efficiency and maximum network lifetime. The performance of the presented model takes place under diverse aspects and the obtained experimental results ensured the goodness of the presented model.

MATERIALS AND METHODS

The major theme of this model is to decide essential CHs among the ordinary sensors by assuming the power efficiency; thus the network lifecycle can be extended. In order to select the CH with power efficiency, remaining energy of the sensor nodes and many other distance parameters are considered with average intra-cluster distance among the sensors and the distance from BS. The CH process is carried out place under 2 stages like Tentative CH selection by applying IPSO-DV model as well as final CH election by using MGWO-LF model. [Fig. 1] shows the working process of proposed method.

Derivation of fitness function

Assume f_1 as a function of average intra-cluster and BS distance of CHs. It is mandatory to reduce f_1 for selecting best CH. Secondly, f_2 is a function that is an inverse of overall energy for elected CH. For normalizing objective functions the measures have to be from (0, 1). Such functions would be applied to retrieve Fitness Function (FF) for optimization approach as depicted in the function:

$$\min F = \alpha \times f_1 + (1-\alpha) \times f_2 \quad (1)$$

Subject to

$$dis(s_i, CH_j) \leq d_{max}, \quad \forall s_i \in S, \quad CH_j \in C \quad (2)$$

$$dis(CH_j, BS) \leq R_{max}, \quad \forall CH_j \in S \quad (3)$$

$$E_{CH_j} > T_H, \quad 1 \leq j \leq m \quad (4)$$

$$0 < \alpha < 1 \quad (5)$$

$$0 < f_1, f_1 < 1 \quad (6)$$

Average intra-cluster distance

It determines the average sum of distances of sensor nodes from the elected CH, where $\frac{1}{l_i}\sum_{i=1}^{l_i} dis\,(s_i, CH_j)$. For intra-cluster communication, the sensor nodes intake few energy while transferring data to CH. For consuming minimal energy, the average intra-cluster communication distance has to be reduced. It refers that the selected CH is closer to sensor nodes.

Average sink distance

It is defined as a ratio of distance among CH_j and BS to count of sensor nodes l_i in CH_j i.e. $\frac{1}{l_i} dis(CH_j, BS)$. In case of data routing phase, every CH has to route the collected data to sink. Hence, in order to minimize the power consumption, the distance of CHs from BS should be limited. Then, the objective f_1 for best selection is reducing average intra-cluster as well as BS distance of the CHs.

$$\min f_1 = \sum_{j=1}^{m} \frac{1}{l_j} \left(\sum_{i=1}^{l_j} dis(s_i, CH_j) + dis(CH_j, BS) \right) \quad (7)$$

Energy Parameter

E_{CH_j} refers to the power of CH_j, $1 \le j \le m$ that is elected from normal sensor nodes iteratively. $\sum_{j=1}^{m} E_{CH_j}$ would be the overall energy for elected CHs. Thus, while the optimal CH election is processed, it is better to limit the overall energy of decided CHs, where the reciprocal has to be reduced.

$$\min f_2 = \frac{1}{\sum_{j=1}^{m}\left(E_{CH_j}\right)} \quad (8)$$

In the presented model, it is limited to 2 objective functions and does not affect one another. Hence, the 2 functions are combined to offer best outcomes.

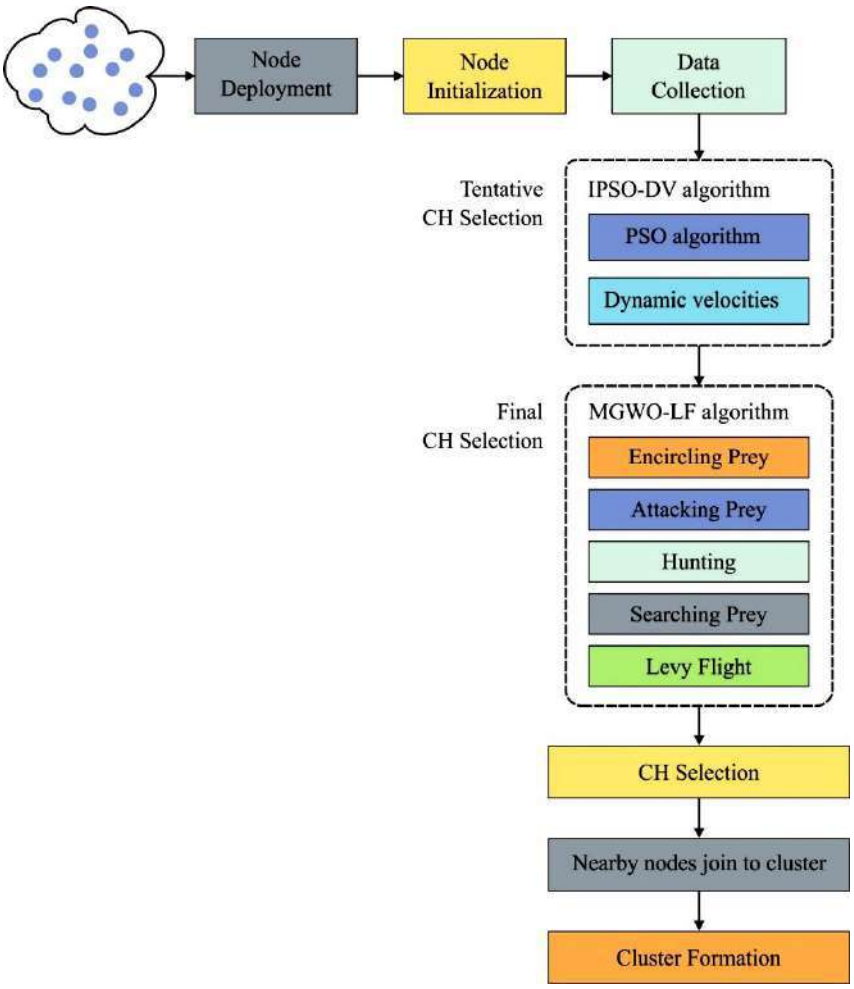

Fig. 1: Overall working process of HMGWOIPSO-DV model

IPSO-DV algorithm based Tentative CH Selection

In order to enhance the function of traditional PSO model, dynamic velocities has been used. The novel velocity of IPSO-DV is extended as given below:

$$v_{ij}(t+1) = \begin{cases} wv_{ij}(t) + c_1 r_1\left(pbest_{ij} - x_{ij}(t)\right) + c_2 r_2\left(gbest_j - x_{ij}(t)\right) & if r1 > 0.5 \\ \tau\left(v_{ij}(t) + c\left(P_{mj} - x_{ij}(t)\right)\right) & otherwise \end{cases} \quad (9)$$

where v_{ij} refers the velocity of particle i in a dimension j at t^{th} generations, r1 and r2 shows the uniform random values from [0,1], c1 and c2 showcases the cognitive as well as social coefficient parameters.

$$w = w_{max} - \frac{w_{max} - w_{min}}{iter_{max}} \times t \quad (10)$$

where w_{max} w_{min} defines the high and low values of inertia weight.

$$\tau = \frac{2}{\left|2 - c - \sqrt{c^2 - 4c}\right|} and c = c_1 + c_2, c > 4 \qquad (11)$$

and

$$P_{mj} = \frac{c_1 pbest_{ij} + c_2 gbest_j}{c} \qquad (12)$$

The place of a particle upgraded with the help of given function:

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \qquad (13)$$

A cluster is created by sink on the basis of centralized clustering. For clustering BS, data is broadcasted as messages for all sensor nodes. Behind receiving the message, it sends the node data such as node ID, position, energy-loss and energy loss ratio as well as current energy to send BS. The BS deploys clusters under the application of IPSO-DV method and telecasts a cluster- message for sensors where it has a cluster. All sensors store the message and invoke CH selection process. Then, sensor nodes retain "my cluster list." The iteration is invoked for selecting CH. In all rounds, Ibest solution is assumed as particle with maximum value than so called as global best (gbest) solution. Finally, particle with gbest solution is approved as best CH.

MGWO-LF based Final CH Selection

The traditional GWO model concentrates on hunters towards the prey on the basis of α, β, and δ (leader dominant wolves). Thus, population of GWO is limited towards stagnation in LO for many cases. Therefore, GWO's problem of inferior convergence is assumed to be the major issues. Then, effective GWO model is not suitable for computing better modification from exploration to exploitation phases. From the previous application, LF has been employed. The LF guides GWO searching operation on the basis of exploring patterns. Using this model, it makes sure that GWO is appropriate for handling global exploration efficiently. Besides, the stagnation problems are resolved. Additionally, the qualities of candidate solutions are improved in Lévy- based GWO for complete process. At the initial phase, portion of delta wolves in social behavior is operated by other wolves and incorporated LF method.

In LGWO, delta wolves are not employed since the hunting operation is performed under various conditions. Actually, updating and initialization of wolves affects the entire whole searching operation. It is highlighted that the dominant nature of wolves guides GWO for saving best solutions which guides remaining candidates of population. Followed by, social behavior becomes effective at the time of using 3 kinds of wolves like alpha, beta and omega. Thus, the place of wolves in LGWO is enhanced on the basis of β types using provided equation:

$$\vec{X}(t) = 0.5 \times (\vec{X} - \vec{AD} + \vec{X} - \vec{AD}) . \qquad (14)$$

In this approach, the actions are performed towards best minima across search space. Followed by, random walks may be suitable in developing the animal actions. Lévy motion is named as diverse non-Gaussian random principles in which random steps are processed on the basis of Lévy stable distribution. Lévy distribution is projected by clear power- law function as given below:

$$L(s) - |s|^{-1-\beta}, 0 < \beta \leq 2, \qquad (15)$$

where s implies the parameter and β refers the Lévy index to balance the scalability. The Lévy distribution is formulated as:

$$L(s, \gamma, \mu) = \begin{cases} \sqrt{\frac{\gamma}{2\pi}} \exp\left[-\frac{\gamma}{2(s-\mu)}\right] \frac{1}{(s-\mu)^{\frac{3}{2}}} & 0 < \mu < s < \infty \\ 0 & s \leq 0 \end{cases}, \qquad (16)$$

where μ refers a shifting parameter and $\gamma > 0$ denotes a reliable parameter. Lévy distribution is reformed according to Fourier transform (FT) as given in the following:

$$F(k) = \exp\left[-\alpha |k|^{\beta}\right], 0 < \beta \leq 2, \qquad (17)$$

Where α expressed the variable from [−1, 1], so called as skewness and scale factor. According to the literature, diverse values of β influence the architecture of LF distribution. While measures of β are limited, prolonged jumps are created; or else, LF can generate lesser jumps utilizing maximum values of β. Based on the Exploration factor, the searching is randomly distributed objects, while computing Lévy walk on LF-based path along with static velocity. Nevertheless, it is advantageous to imply hunting model of wolves in GWO based on the LF model. In LF assists LGWO in resembling wolves' hunting nature in practical time

IIOAB JOURNAL
ISSN: 0976-3104

than GWO. Moreover, it is used for LF as a secondary objective that resolves the stagnation problems of GWO. Thus, latter modification to GWO, novel places are defined as:

$$\vec{X}_{new}(t) = \begin{cases} 0.5\times\left(\vec{X}_\alpha - \vec{A}_1\vec{D}_\alpha + \vec{X}_\beta - \vec{A}_2\vec{D}_\beta\right) + \alpha \oplus Levi(\beta) & |A| > 0.5 \\ 0.5\times\left(\vec{X}_\alpha - \vec{A}_1\vec{D}_\alpha + \vec{X}_\beta - \vec{A}_2\vec{D}_\beta\right) & |A| < 0.5 \end{cases} \quad (18)$$

where it depicts the step size to relevant for scales of problems, β implies the Lévy index from $(0,2)$ and $\oplus$ projects the entry wise improvisation. According to the value of $|A|$, the novel operator reforms wolves to gain best management among exploration as well as exploitation due to the LF- based jumps. During this point, adjusted GWO is suitable for accomplishing best outcomes, there is a possible LGWO for utilizing and improves optimal-qualified decisions. Moreover, these operators improve the exploitive nature of wolves in last rounds. Followed by, x implies aarbitrary quantity to dimension of wolves. Thus, if $|A| > 0.5$, the operator is maximized as:

$$\vec{X}_{new}(t) = 0.5\times\left(\vec{X}_\alpha - \vec{A}_1\vec{D}_\alpha + \vec{X}_\beta - \vec{A}_2\vec{D}_\beta\right) + rand\ (size\ (D)) \oplus Levi(\beta), \quad (19)$$

where$D$ implies the dimension. Mantegna method is accurate model for providing stochastic variables in which the probability density is changed to Lévy steady distribution specifically balanced by a parameter $(\alpha(0.3 < \alpha < 1.99)$. So, Mantegna principle is used for accomplishing LF for complete exploration operation. Then, in Eq. (19), step size is computed by:

$$rand\ (size(Dim)) \oplus Levi(\beta) \sim 0.01\frac{u}{v^{-\beta}}(\vec{X}(t) - \overrightarrow{X^R}(t)), \quad (20)$$

where$u$ and v values can be attained on the basis of normal distributions;

$$u \sim N(0,\sigma_u^2), v \sim N(0,\sigma_v^2), \quad (21)$$

with

$$\sigma_u = \left[\frac{\Gamma(1+\beta)\sin(\frac{\pi\beta}{2})}{\Gamma(\frac{1+\beta}{2})\beta\times 2^{\frac{\beta-1}{2}}}\right]^{\frac{1}{\beta}}, \sigma_v = 1, \quad (22)$$

where Γ refers the classical gamma function. During this model, β variable is dynamic, hence an arbitrary value from $(0,2)$ interval must be chosen in iterations for LF operation. Followed by, LF offers lower and occasional long-distance jumps. The random β attribute would improvise exploitation and exploration movements between the iterations.

RESULTS AND DISCUSSION

This section validates the efficient performance of the proposed technique with respect to count of alive nodes, dead nodes, and network stability. For comparative purposes, different methods are utilized [10]. [Table 1] depicts the parameter settings.

Table 1: Parameter Settings

Parameter	Value
Network dimension	100*100m^2
Node count	100,300,500
Number of BS	1
Initial energy	0.5
Transmit/Receive Energy E_{elec}	50nJ/bit
Threshold distance (d_0)	80m
Packet size	2000 bits
Population size (P)	100

Fig. 2 depicts the detailed alive node analysis of the presented and existing models under a varying number of nodes. From the figure, it is apparent that the HPSOGWO and FBECS algorithms have exhibited poor network lifetime by attaining minimum number of alive nodes. Followed by, the IPSO-DV algorithm has tried to show better network lifetime over the previous models with slightly higher number of alive nodes. Simultaneously, the MGWO-LF algorithm has demonstrated somewhat better network lifetime over the compared methods. But the proposed method has showcased excellent results by attaining maximum network lifetime with the maximum number of alive nodes. For instance, under the varying node count of 100 with the execution round of 1000, the presented model has reached to a maximum number of 50 nodes whereas the MGWO-LF, IPSO-DV, HPSOGWO and FBECS algorithms have attained a minimum number of 28, 25, 6 and 6 nodes respectively. Besides, under the varying node count of 300 with the execution round of 1000, the presented model has reached to a higher number of 153 nodes whereas the

MGWO-LF, IPSO-DV, HPSOGWO, and FBECS algorithms have attained a minimum number of 120, 40, 6, and 0 nodes respectively. Also, under the varying node count of 500 with the execution round of 1000, the presented model has offered a superior performance with a higher number of 210 nodes whereas the MGWO-LF, IPSO-DV, HPSOGWO, and FBECS algorithms have attained a minimum number of 160, 140, 26 and 5 nodes respectively.

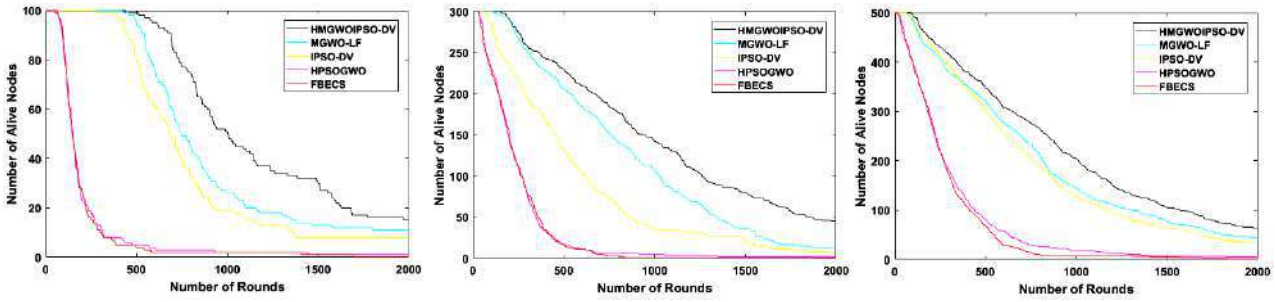

Fig. 2: Alive node analysis of proposed model under varying node count

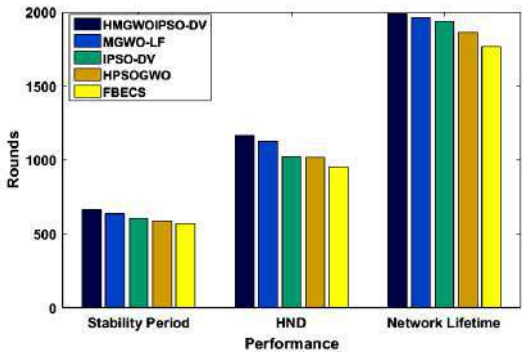

Fig. 3: Network lifetime analysis

Fig. 3 illustrated the further validation of network lifetime efficiency of the presented model; a comparison analysis is demonstrated with the previous method by means of stability period, half node die (HND) as well as network lif

Fig. 4 illustrates the analysis of the presented model with existing methods under varying node count interms of number of packets. The figure depicted that the proposed method has achieved a higher number of packets under varying number of nodes. At the same time, the MGWO-LF and IPSO-DV algorithms have attained a slightly lower number of packets. Likewise, the HPSOGWO and FBECS algorithms have reached a minimum number of packets.

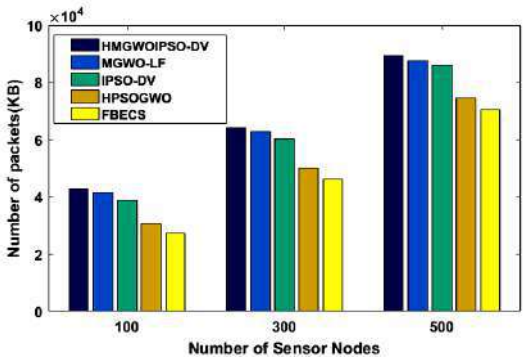

Fig. 4: Comparative results analysis of proposed and existing methods in terms of number of packets transmitted

CONCLUSION

This paper has developed an effective hybrid clustering technique called HMGWOIPSO-DV algorithm. The presented algorithm involves CH selection in two levels namely IPSO-DV based tentative CH selection and MGWO based final CH selection. Besides, the presented model derived a fitness function by the use of residual energy, distance to BS, and distance to nearby nodes. The proposed model has the capability of competently choosing the CHs, attains energy efficiency, and maximum network lifetime. A series of

IIOAB JOURNAL
ISSN: 0976-3104

experiments were carried out to ensure the energy efficient performance of the projected model. The inclusion of two algorithms for CH selection leads to the optimal CH selection and thereby network lifetime gets improved. The obtained simulation outcome ensured that the proposed model has outperformed the compared methods in a significant way. In future, the proposed model can be extended to the use of data aggregation techniques to reduce the quantity of data transmission in WSN.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None.

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] Yi D, Yang H. [2016] HEER–A delay-aware and energy-efficient routing protocol for wireless sensor networks. Computer Networks. 104:155-73.

[2] Kannan G, Raja TS. [2015] Energy efficient distributed cluster head scheduling scheme for two tiered wireless sensor network. Egyptian Informatics Journal. 16(2):167-74.

[3] Hari U, Ramachandran B, Johnson C. [2013] An unequally clustered multihop routing protocol for wireless sensor networks. In2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 1007-1011.

[4] Bhattacharjee S, Roy P, Ghosh S, Misra S, Obaidat MS. [2012] Wireless sensor network-based fire detection, alarming, monitoring and prevention system for Bord-and-Pillar coal mines. Journal of Systems and Software, 85(3):571-81.

[5] Ni Q, Pan Q, Du H, Cao C, Zhai Y. [2015] A novel cluster head selection algorithm based on fuzzy clustering and particle swarm optimization. IEEE/ACM transactions on computational biology and bioinformatics, 14(1):76-84.

[6] Potthuri S, Shankar T, Rajesh A. [2018] Lifetime improvement in wireless sensor networks using hybrid differential evolution and simulated annealing (DESA). Ain Shams Engineering Journal, 9(4):655-63.

[7] Sengottuvelan P, Prasath N. [2017] BAFSA: Breeding artificial fish swarm algorithm for optimal cluster head selection in wireless sensor networks. Wireless Personal Communications, 94(4):1979-91.

[8] Mann PS, Singh S. [2017] Artificial bee colony met heuristic for energy-efficient clustering and routing in wireless sensor networks. Soft Computing, 21(22):6699-712.

[9] Kaur S, Mahajan R. [2018] Hybrid meta-heuristic optimization based energy efficient protocol for wireless sensor networks. Egyptian Informatics Journal, 19(3):145-50.

[10] Rao PS, Banka H. [2017] Novel chemical reaction optimization based unequal clustering and routing algorithms for wireless sensor networks. Wireless Networks, 23(3):759-78.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

ARTICLE

FEATURES OF THE INTRODUCTION AND USE OF BIG DATA, DATA SCIENCE TECHNOLOGIES IN SINCERITY MARKETING

Viktoriya I. Tinyakova[1]*, Yaroslav B. Lavrinenko[2], Svetlana V. Bryukhovetskaya[3], Tatyana V. Karyagina[4], Maria A. Kirpicheva[3]

[1]*Institute of Industry Management, State University of Management, 99 Ryazansky Ave., Moscow, 109542, RUSSIA*

[2]*Faculty of Economics, Voronezh State Technical University, 84, 20-letiya Oktyabrya str. Voronezh, 394006, RUSSIA*

[3]*Department of Marketing, Financial University under the Government of Russian Federation, 49, Leningradsky Prospekt, Moscow, 125993, RUSSIA*

[4]*Faculty of Information Technology, Russian State Social University, 4, Wilhelm Pieck str., building 1, Moscow, 129226, RUSSIA*

ABSTRACT

This article discusses the problem of the introduction, features and complexity, as well as the feasibility of using big data technologies in sincerity marketing. The purpose of this work is the lack of coverage of the problems of introducing Big Data, Data Science technologies in company marketing. The analysis of the tasks that Data Science can solve on the basis of Big Data is carried out. The advantages for using these technologies in sincerity marketing are highlighted. The authors pay special attention to the reasons for the development of Big Data, Data Science technologies in Russia and current restrictions. The necessary actions before the introduction of the technology are identified and described. The authors focus on the implementation of these technologies, as well as on the use of metrics for sincerity marketing, which underlie the technologies. The difficulties that companies face when introducing Big Data, Data Science technologies into the company's work are considered. Examples of services for working with these technologies are given. In practice, an example of the use of the technologies in the work of a network of shoe stores, bags and accessories is given. ROPO analysis is performed. The value of the work is emphasized by its practical focus on marketers and the scientific community.

KEY WORDS
sincerity marketing, use of Big Data, Data Science, implementation, recommendations

Received: 17 Jul 2020
Accepted: 8 Oct 2020
Published: 17 Oct 2020

***Corresponding Author**
Email:
tviktoria@yandex.ru
Tel.: +7 925 047 98 97

INTRODUCTION

Big data increases profitability and reduces expenses of sincerity marketing. According to BCG research, companies leading in the implementation and use of big data tools have 2.5 times more profitability, while marketing costs are reduced by 1.4 times. Oddly enough, only 2 percent of the companies in the study mentioned above use the latest technologies in digital marketing. At the same time, real efficiency shows an increase in income by 20 percent and a decrease in expenses by 30 percent [11, 18]. Data Science is based on three Big Data properties: volume, regularity of updating and variety of forms. The Data Science and Big Data infrastructure consists of the following groups [Fig. 1]. Every year the level of service and the quality of marketing is growing. Consumers are getting used to a personalized approach. People like being offered the product that they are likely to enjoy. Personalization is not the future, and is not even a trend; it is already the present, taken for granted. One of the most important tasks of sincerity marketing is to attract the attention of consumers. The development of Data Science, Data Driven and Big Data technologies helps to understand what the customer needs.

Analytics tools facilitate the company's appeal to the target audience, which increases the likelihood of successful completion of transactions, that is, it increases conversion, and, as a result, profit. The analytics and application of information lead to a better understanding of consumers and taking informed decisions, rather than random actions [3, 5]. The use of approaches based on the processing of large volumes of data present new opportunities. First of all, this is the understanding of the company's work in specific figures and data; secondly, the study of competitors and the possibilities of competition; thirdly, the understanding of their consumers both current and potential.

If data is effectively collected, grouped, pre-analyzed and further analyzed, then it becomes valuable information. At the same time, there arises the problem of decoding and correct interpretation of the collected information. The development of Data Science now allows for more efficient information processing increasing the effectiveness of strategic and tactical sincerity marketing, as well as of the understanding of current consumers. Technologies and tools of Data Science help in the work of the marketing department and marketers in general. They solve the following problems [1, 17]:

Clustering of information; approximation of experimental information based on a descriptive or prognostic model; automation of various processes in decision making; assessment of the degree of objects potential; factor analysis of individual data; and construction and training of neural networks.

The advantages of using Big Data and Data Science in sincerity marketing are - Increasing the speed of planning advertising campaigns; optimization of budget funds; data online; increasing the degree of

loyalty; consumer segmentation; data visualization [9]; predictive analysis [2, 14]; decrease in the level of outflow coefficient of consumers [7, 10]; prediction of sales [16]; predicting customer reactions to various messages; and definition and understanding of the popularity of individual products and services.

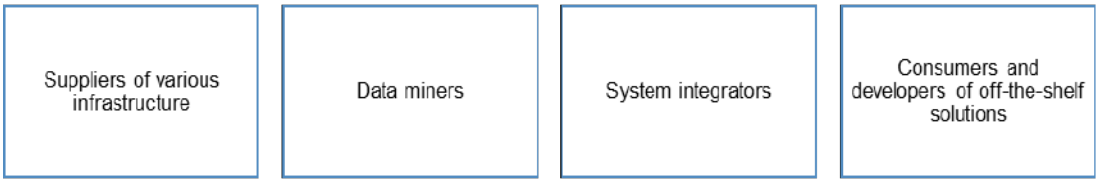

Fig.1: Data Science and Big Data Infrastructure

Big Data and Data Science are actively developing in Russia and abroad. The reasons for the rapid development and its limitations in the Russian Federation will be considered below in [Table 1].

Table 1: Causes and limiters of the development of Bid Data and Data Science in Russia

Causes	Limiters
Demand for competitiveness enhancing tools	Need for a high level of data protection
Increasing demand for the services of providers and integrators from Russia	Lack of adequate qualified staff
Placement of servers in Russia	Russian companies accumulate insufficient data
Improving media content processing possibilities	High cost of technology implementation
Techno parks development	The complexity of introducing new technologies into applied systems

METHODS

To implement an approach based on the collection, analysis and use of big data, intra-company changes are needed. Changes should be aimed at creating the foundation for the efficient operation of big data. The foundation is based on:

- willingness to invest in working with big data (time and costs);
- willingness to see and listen to the information received, making decisions based on the prepared figures;
- readiness to understand data. Analytical thinking and understanding of the data provided are necessary;
- willingness to trust data and the need to make decisions based on them. It is important to build trust between analysts and decision makers.

Before implementing Data Science and collecting big data, it is necessary to evaluate current and strategic goals and objectives, analyze the possibilities of application and prepare the ground for implementation. Of course, bonuses from the introduction of technologies are received by companies that carry out hundreds of transactions per day, have tens of thousands of consumers and thousands of positions in the product range. Consideration of future implementation can be carried out in three directions. The first direction is the database. It is necessary to evaluate its structure, interconnections and unity. If the company has several independent or weakly interconnected databases, then one needs to evaluate the possibility of setting up an effective and stable communication for the exchange of information. In addition, it is worth pre-evaluating the quality of the available data and its depth, as well as relevance. The second direction for evaluation is analytics. Who has the responsibility for conducting marketing analytics in the company, what tasks take the most time? What reports are generated at specific hours, weeks, and months? One should assess the current use of collected analytics for solving particular problems and for decision making and the level of development of consumer segmentation and product range. The third direction for evaluation before implementation is communication. Use of customer information by the company to target and apply personalization; whether profiles contain information about customers, the history of viewing advertising messages and responses to them. The level of data security and the load on them is also evaluated.

To start implementing the Data Science and Big Data approaches in sincerity marketing, the following steps should be taken:

- Adapting audience data aggregation.
- Building consumer segmentation, strategies, KPI.
- Combining campaigning.
- Establishing feedback with the used CRM systems.
- Implementing data analytics as part of ongoing marketing tactics

IIOAB JOURNAL
ISSN: 0976-3104

The implementation of technologies for big data processing and decision-making on their base is grounded on information about customer service, user satisfaction, visitor's behavior on a website or in applications. Before you start working with big data in sincerity marketing and making decisions based on big data, you need an action plan, an example is presented below:

i. Definition of data sources, verification of their relevance and accuracy, as well as the purity of the information received.
ii. Forming a team based on specialists and analysts. To begin with - a marketer and data scientist. It requires employees who are willing to share knowledge and work on a common task. And the task of the team at this stage is to understand how to collect, analyze and use the information received in the future. The basic goal is the coordination of hypotheses, ideas and a budget for the implementation of work with big data. It is important that the final data is clear to managers and ordinary specialists, the right questions give the right answers that marketing can apply in its work.
iii. Collection of all sources and information on one of the platforms. Data should be collected from as many sources as possible: information on products and services, advertising platforms, CRM and ERP systems. Here can be both online and offline data. Notes of the sales service, call center information, CRM and e-mail database. In addition, user behavior on the network and on the site [4, 19].
iv. Formation of infrastructure for storing the received information. It must be systematized in the required form. It is advisable to provide all information, except confidential, in the domain for access of employees of the organization.
v. Visualization of the data obtained on the basis of BI-platforms and dashboards. Here we segment the audience, calculate market shares, share of the target audience, history of purchases and sales. Among the consumers of the company, we choose the really best customers according to the Pareto principle. Having identified the best customers, their characteristics, one can start looking for other customers whose profile matches or resembles the best customers. Creation of detailed customer profiles [6, 8].
vi. Conducting experiments, evaluating various options of actions, the results obtained, interpreting data and testing hypotheses. Forecast modeling, impact on consumer behavior, anticipation. Creating a media plan for each of the segments.
vii. Optimization. The collected, analyzed information must be constantly cleaned, structured, and verified. Testing and pilot use of big data technologies in decision making. Analysis and review of the results obtained. Comparison of reality with set goals.
viii. Formation in the organization of a real culture of decision-making based on information and collected data. The launch of full-scale work with big data, the use of information. The work should be based on the principles of continuity and consistency of work, comprehensive and comparable data.

For sincerity marketing and Big Data, Data Science technologies, the use of effective metrics is very important. Effective metrics should be comparative. Metrics indicators should be comparable across different time ranges, advertising projects, or target groups. In addition, the metric should be understood by most team members. If they cannot adequately explain why the analysis of this or that metric is necessary, then achieving goals on it will be difficult.

Metrics should be expressed in relative indicators. This is important because relative indicators are convenient for making decisions. In addition, relative indicators are easy to compare with each other. It is easy to find out long-term effects or short-term outbreaks, to determine current and future trends. At each stage of work, tasks and goals, there should be no more than 3-5 metrics. Below are the possible metrics for online commerce and SaaS services in the [Table 2].

Table 2: Sincerity marketing metrics while using Big Data and Data Science

	Online Commerce Metrics	SaaS Services Metrics
1	CPC (Cost per click): The cost that the company pays, attracting one customer. It demonstrates the effectiveness of a particular channel, return on investment.	Regular monthly revenue: It gives an understanding of how much the company receives on average each month or during a specified time range.
2	Conversion rate: Various conversions: from leaving applications to purchases. The conversion contains valuable information for predicting future profits.	Outflow of customers: The volume of customers who were regular and quit. Such an indicator helps in forecasting losses and an adequate perception of the current situation.
3	Abandoned baskets: The volume of visitors to the online store who quit making their purchases at a certain stage. Using the indicator, one can evaluate and find the so-called bottlenecks in the sales funnel.	Lifetime customer value: Income from the user for the entire time of using the service. Here we can calculate how much to invest in advertising in order to attract a customer.
4	Average income per account / customer: The indicator shows how much income each customer brings.	Customer retention rate: Providing data on what percentage of consumers the company translates into regular customers

IIOAB JOURNAL
ISSN: 0976-3104

When implementing and using big data, it is worth remembering about the principle of 1-10-100. It is used in various fields and can be interpreted in different ways. Any project, also if it is the introduction of the use of big data technologies and decision-making based on them, consists of three stages: planning, implementation and using. What does this principle say? It states that, having discovered an error at the first stage (planning or data analysis), the cost of fixing it is 1 conditional ruble. At the second stage (when implementation or work with data is already underway), fixing the error costs 10 conditional rubles. At the third stage (when the product works, the big data technology is introduced), the correction of the same error will cost 100 conditional rubles.

The implementation of Big Data, Data Science and Data Driven technologies is accompanied by a number of difficulties in sincerity marketing, which are presented in [Fig. 2].

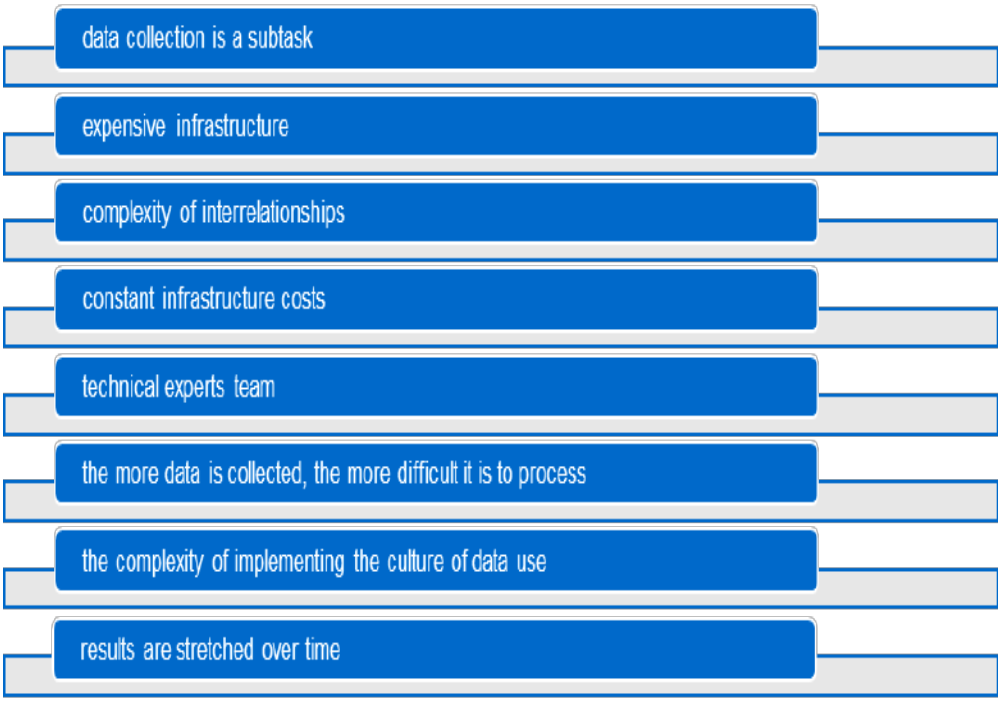

Fig. 2: Difficulties in implementing Big Data and Data Science technologies in sincerity marketing

- data collection is a primary but not the only task. The constant collection of information becomes insufficient, the data must be cleaned, processed, structured, analyzed and interpreted, based on which hypotheses are formed, tested, and the obtained data must be analyzed again;
- expensive infrastructure. The establishment of various services for analytics, resource accounting, CRM and ERP systems, as well as call tracking and tools for end-to-end analytics. Adequate visualization of the collected data, metrics and reporting clear to employees and management of the organization;
- the complexity of the interrelationships. Many elements must be configured to work effectively and consistently with each other. Services are updated, new ones appear, new tasks are set, etc.;
- constant expenses for the maintenance of infrastructure. The staff of specialists, rental and purchase of equipment / services may be underestimated at the implementation stage;
- a team of technical specialists. To begin with, a team may consist of one specialist and analyst. However, third-party service specialists will be involved in the work for the initial setup of interaction;
- the more data is collected, the more difficult it is to process. The more indicators, sites are included in the data collection, the more time and money is spent on processing them. It is more difficult to separate important indicators and their impact from the “garbage” ones. Hypothesis testing costs increase;
- the complexity of introducing the culture of data use. Managers and ordinary employees should be prepared to work with information, hypotheses, make decisions timely. The company must develop along with the development of big data technologies;
- the results are stretched over time. To get the first results from implementation, a long time must pass till there come the results from past decisions made on the basis of data.

The implementation of technologies for working with big data occurs using various tools to solve various tasks. Examples of tools are presented in [Table 3].

Table 3: Tools for working with big data

Tool type	Description	Examples
Web analytics	Collection and storage of information about website visitors and their behavior	Google Analytics, Yandex.Metrics, Google Tag Manager
Big Data Technologies	Provide the collection, storage, structuring of large amounts of data	SAP 4 HANA, Hadoop, Spark, Yandex Data Factory, IBM Pure Data and Watson, Microsoft Azure
End-to-end analytics	Conducts the assessment of the effectiveness and profitability of advertising channels	Roistat, CoMagic Mixpanel, Rick, Alytics
CRM	Implements sales forecasting and performance evaluation of the work of marketing department	Salesapcrm, Bitrix24, MangoCRM, Wrike, Megaplan
Data visualization	Presentation of information in the form of visualization and a set of dashboards	Google Data Studio, Qlik Easel.ly, Tableau, Power BI, Piktochart, Infogr.am
Unit Economics	Tips for identifying points of organization growth	1C, SAP, Oracle, Microsoft

RESULTS AND DISCUSSION

Here is an example of the implementation and use of big data technologies and analytics tools based on the work of one of the network stores of shoes, bags and accessories in the framework of sincerity marketing. The network has more than 90 stores, as well as a functioning online store. The online marketplace was quite convenient, but a network of physical stores with branded goods outperformed the online store. This was facilitated by the excellent personalized service of the stores chain.

However, marketers of the network of shoe stores noticed that many consumers before going to an offline store browse the site, evaluate the availability of models at retail outlets, immediately selecting models in various price categories. As a result, they come already prepared for a certain set of things.

The company's specialists had several hypotheses that could be tested on the basis of Big Data technology using Data Science and analytics tools:

- Find out the degree of influence of online marketing on the level of sales in stores.
- Evaluate and adjust network budgets for marketing on the Internet.
- Conduct a revision of the effectiveness of marketing channels and improve the current marketing strategy.

To test hypotheses, the goal was set to conduct a ROPO analysis, which includes researching information about online and offline acquisitions, as well as the paths that led to the sale.

The main problem was the need to combine data from online advertising and offline sales. The chain of stores previously generated a large amount of information on consumers who purchased or ordered in various ways:

- selected and bought goods in the store;
- browsed the products on the site and then came and purchased them in the store;
- selected and purchased on the site with delivery to a particular store or directly to their home address.

The stores chain generated information in various services, which is quite logical, given online and offline operating modes:

- information on the online store was collected in Google Analytics;
- information on orders, their execution, as well as all data on offline buyers were stored inside the CRM system.

Prior to this, the company collected data from Google Analytics in Google Big Query based on OWOX services. Now the goal was to implement ROPO analysis. The following tasks were set:

- collect information from advertising campaigns, actions of site visitors, offline purchases and executed orders within the same structure;
- link online orders with offline sessions;
- generate visual reports based on structured information to assess the contribution of various online sources.

The information transfer scheme is presented in [Fig. 3].

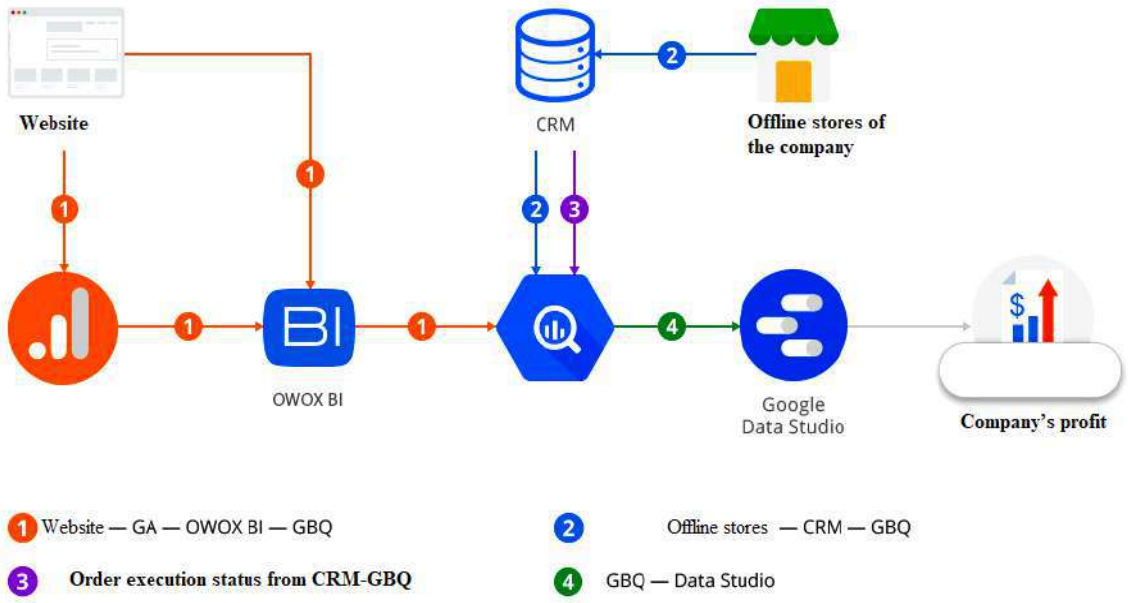

Fig. 3: Information transfer scheme

Now let us consider in detail the necessary steps.

Stage 1. Data collection in Google BigQuery.

Each registered visitor of the online store received his personal ID (user_id), after the transaction was executed, he was also assigned another identifier (transaction_id). All information was sent to Google BigQuery. Every day, data on online and offline orders from CRM was transferred to Google BigQuery and connected based on the two keys presented above.

Stage 2. Combining the information obtained.

Information on online orders was supplemented by information on the execution of each order (purchased or not purchased). Here transaction_id was used. Next, user sessions are integrated with offline purchases data based on user_id. For most users who bought in a physical store, there was a history of sessions on the site. The unification scheme is presented in the following [Fig. 4].

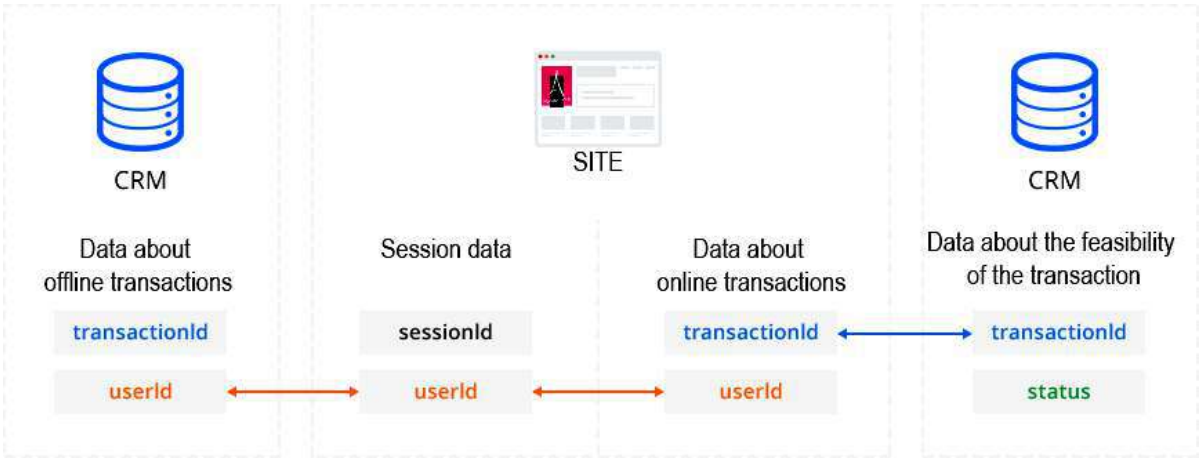

Fig. 4: The scheme of combining information

According to the results of the combination, the analysts had at their disposal the following information:

- does the transaction relate to online, offline or ROPO?
- what led the consumer to the purchase or what was the last source?
- the number of days passed between the last session in the online store and the purchase? Here we will explain in more detail. For consumers who have made purchases online, the indicator will always be "0", because the last visit to the online store is always a purchase. For offline

IIOAB JOURNAL
ISSN: 0976-3104

purchases, this will also be zero, because the consumer does not have sessions offline. All other buyers can be attributed to the target sample of ROPO;

- location of the last user session.

Stage 3. Visual presentation of the formed analytics.

For a visual presentation of the available information, the Google Data Studio service was used. An informative Dashboard was formed, which was equipped with dynamic charts, the capabilities of in-depth analytics and forecasting of the advertising budget. Dashboard is presented in [Fig. 5].

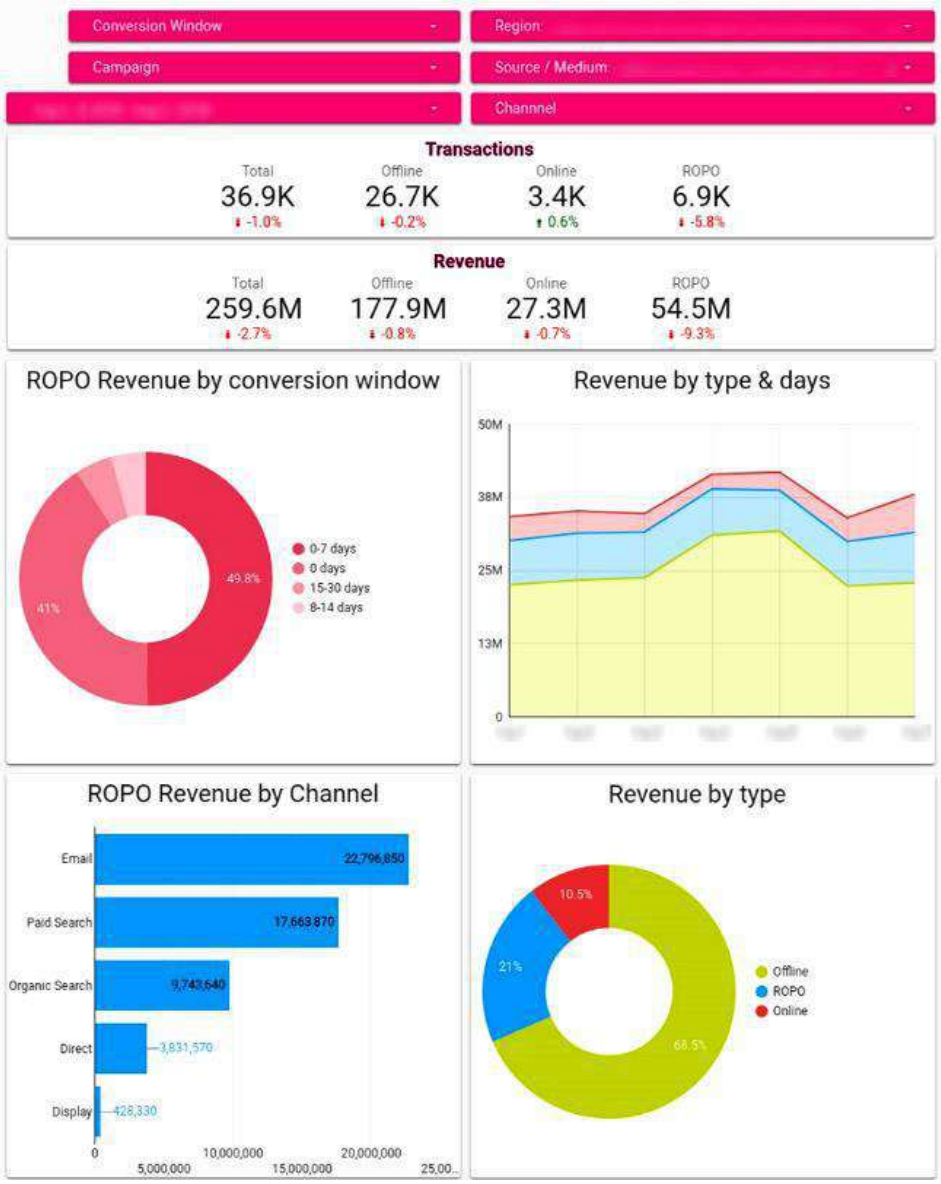

Fig. 5: Dashboard for data management

Let us explain its contents. For example, the pie chart on the right side provides us with information that ROPO purchases account for about 20 percent of the revenue share. The diagram on the left explains that almost 50 percent of all ROPO customers visited the website in a range of up to 7 days. In addition, the created template can be customized for different cities, regions of presence, sources, channels presented, campaigns, individual target segments.

The table at the bottom of the Dashboard determines what amount of additional income from ROPO purchases can be obtained as part of forecasting marketing activities and development of a marketing strategy. In addition, we see the amount of profit from specific sources, channels and campaigns conducted.

As a result of using Big Data, Data Science technologies and analytics tools, the store chain made the following conclusions:

- the organization realized that online advertising contributed to 20 percent of online revenue. In fact, every fifth visitor to a physical store has already visited an online store;
- ROPO analysis showed that the organization underestimated offline companies as a source of revenue;
- the organization has changed its marketing strategy of sincerity based on the conducted research and increased advertising on the Internet.

In the research of Chintagunta et al., they say that marketing is becoming more and more a quantitative science. Previously, marketing focused on institutional knowledge and empirical rules. Now Big Data technologies are at the forefront. However, in their study, they did not take into account the implementation of the Data Driven approach in all areas of the company [12]. At the same time, in his article, Nadler says that it is important to take into account the principles of behavioral economics in marketing that uses data [13]. However, a timely question arises: "Can marketers and marketers handle the flow of data?" This question is given in the work by Sheth and Kellstadt. The presence of a huge amount of information, even processed, does not always affect the increase in profitability or the growth of competitiveness [15]

In our article, we offer a different perspective on the use of Big Data, Data Science and Data Driven technologies. With their help, we can know our client more, express our genuine concern, impersonality, impartiality, and inner freedom. Technology helps marketers and companies strive to produce value for consumers, customers, and society generally. Big Data technologies, Data Science and Data Driven are not just new tools or mechanisms. They are a different paradigm of marketing, where the old philosophy of marketing is giving a way to a new paradigm, the marketing sincerity paradigm. Sincerity marketing is based on three principles: trust, facts and truth. It is worth remembering that data is interpreted by analysts, decisions are made by managers. Here, more than ever, the following values of sincerity marketing should be used: transparency, fairness, responsibility and honesty. Note that these are not just principles and values for marketing. Their use, demonstration in their work and promotion draw new customers and keep current ones. In the modern world there is a lot of data, but little sincerity. Sincerity marketing helps companies stand out using Big Data, Data Science and Data Driven technologies.

CONCLUSION

It should be noted that the use of Big Data, Data Science and Data Driven technologies is not without drawbacks. Therefore, we give recommendations on the use of Big Data, Data Science and Data Driven technologies in the work of the company. The first recommendation is the use of information for decision making. It is not necessary to analyze the data for too long, because the goal is to help organizations make effective decisions, evaluate current processes, internal and external information. The second recommendation is involving senior management in analytics. Their mission is not only to support the development of analytics and big data technologies. They should maintain continuous communication with the heads of analytical departments, and, if necessary, gain new skills in analytics. The third recommendation is to train the team to work with data. If the implementation takes place on a top-down basis, then this is most often not that effective. For ordinary employees, it is necessary to create motivation for the use of data and analytics. The fourth recommendation is to attract individual employees as leaders in the transformation of the entire company. They will subsequently motivate, promote the use of data analytics and big data in their work. The effectiveness of Big Data, Data Science technologies and decisions made on their basis is proved by the success of many companies. Leading companies in the application of big data technologies achieve better results while reducing costs by 30 percent and increasing revenues by 20 percent. Sincerity marketing in companies, as never before, reaches a different level of understanding of ongoing processes based on Big Data and Data Science technology, the possibilities of their intermediate control and personalization, which allows to reduce costs and increase profitability.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Purohit BGL, Yanenko MB, Yanenko ME. [2017] On the issue of formation and improvement of a digital platform for organizing and managing a company's marketing activities: problems and tasks, Problems of the modern economy. 2(62):127-132.

[2] Bochkova EV, Avdeeva EA, Scherbakov DS. [2016] Features of the application of Big Data information technology in the marketing activities of Russian B2C-sector companies, Scientific and Methodological Electronic Journal "Concept", 17. URL: http://ekoncept.ru/2016/76193.htm.

[3] Gorelova AA. [2017] Big data and directions of their use in marketing, Actual problems of the humanities and natural sciences, 4-2:11-16.

[4] Ilyin IV, Shaban AP. [2018] Features of the use of big data in marketing activities, Science Week SPbPU: materials of a scientific conference with international participation. SPB.: St.

Petersburg Polytechnic University of Peter the Great, 123-126.
[5] Kurochkina AA, Kulikova KA, Rumyantseva ER. [2019] Big Data as a New Competitive Advantage in Retail, Global Scientific Potential. 12:224-228.
[6] Lavrinenko YaB, Shitikov DV. [2019] Evaluation of the effectiveness of the implementation of information systems in the organization (for example, ERP-systems), Economics in the investment and construction complex and housing and communal services. 1:96-100.
[7] Simakina MA. [2018] Features of the use of big data technologies in marketing, Bulletin of science and practice, 4(6):255-260.
[8] Shaban AP, Ilyin IV. [2019] New opportunities for BI-systems in marketing activities, Technological perspective within the Eurasian space: new markets and points of economic growth, 370-373.
[9] Shevchenko MV, Kosyanova EA. [2016] BIG DATA: problem, prospects and relationship with the economic system, Belgorod Economic Bulletin, 4:31-35.
[10] Shchetinina IS, Zhuravleva OV. [2018] Problems of using big data technologies in small and medium-sized businesses, Sustainable development of Russia in the period of instability: external challenges and prospects: Materials of the XII full-time international scientific-practical conference. Yelets: Yelets State University. IA Bunina, 50-54.
[11] Chintagunta P, Hanssens DM, Hauser JR. [2016] Marketing and data science: Together the future is ours, GfK Marketing Intelligence Review. 8(2):18-23.
[12] Hossain TMT, et al. [2017] The impact of integration quality on customer equity in data driven omnichannel services marketing, Procedia Computer Science. 121:784-790.
[13] Nadler A, McGuigan L. [2018] An impulse to exploit: the behavioral turn in data-driven marketing. Critical Studies in Media Communication, 35(2):151-165.
[14] Peyne B, Chan A. [2017] Data-driven decision making in Marketing: A theoretical approach. URL.: http://www.diva-portal.org/smash/get/diva2:1081339/FULLTEXT01.pdf
[15] Sheth J, Kellstadt CH. [2020] Next frontiers of research in data driven marketing: Will techniques keep up with data tsunami?. Journal of Business Research. https://doi.org/10.1016/j.jbusres.2020.04.050
[16] Stolle CS, et al. [2018] Data-Driven Development of a Roadside Safety Marketing Campaign for Tree Removal – Phase I.
[17] Tinyakova V, Lavrinenko Y, Blinov A. [2018] Usage of various metrics for clustering key queries in contextual advertising, Economic and Social Development: Book of Proceedings. 907-917.
[18] Verhoef P, Kooge E, Walk N. [2016] Creating value with big data analytics: Making smarter marketing decisions. Routledge. DOI: 10.4324/9781315734750.

ISSN: 0976-3104
ISSUE: Computer and Allied Sciences

EXPERT OPINION

CUSTOMER AND PRODUCT PROFITABILITY IN LARGE ENTERPRISES USING SAP S/4 OR ANY MAJOR ERP SYSTEM

Sarat Chandra Tella*

SAP S/4 HANA Solution Architect, Archer Daniels Midland Co., 1260 Pacific Ave, Erlanger, KY 41018, USA

ABSTRACT

In major organizations, to achieve customer and product profitability lot of manual processes is required at month end depending upon the systems involved in the organizations. Customer and product profitability can be referred as Gross Margin or Margin Analysis by customer and product. Margin Analysis is one of the key capablities to support business decisions globally for large organizations and also required for best and least performed products in different regions of the world. Different business processes and teams are involved with in organization to achieve customer and product profitability which includes production, purchasing, sales and Finance. Different ERP systems like SAP, Oracle are used as transactional system to perform daily activities like creation of sales order or purchase order, customer/vendor invoice. This article covers functional and technical aspects how gross margin can be achieved in an organization and this is broadly classified as 1. Business processes, 2. Maintenance of Master data, 3. Transactions from different modules in SAP S/4 to COPA (controlling – profitability Analysis), and 4. Reporting.

KEY WORDS
SAP, S/4, S/4 HANA, COPA, SAP ECC, Gross Margin, Margin Analysis. Profitability Analysis.

INTRODUCTION

Customer and product profitability is very important step in month end process for strategic business units such as sales, customer service and finance with in organization. This is the critical step for any organization to make important decisions from business leader perspective and also to perform analysis whether particular product line and customer is profitable or not.

SAP S/4 HANA system uses in-memory database and used as transactional system in major organizations. Although there is no specific defined process for profitability analysis, it is an important factor for management reporting which is used for analytical purposes across the organization. COPA data can be viewed as data cube that takes information with in transactional system from different modules like sales, purchasing and production and presented in different ways for analytical purpose with in organization. The purpose of gross margin capability is to provide visibility and insight into sales and margin activity by customer and product. As part of this functionality lot of other characteristics on customer and product are available to perform analysis in a flexible manner suited for different purposes within the organization.

PROCESS FRAMEWORK TO ACHIEVE PROTIFABILITY

Received: 15 Oct 2020
Accepted: 10 Nov 2020
Published: 16 Nov 2020

This paper explains an approach to determine gross margin by customer and product using SAP S/4 HANA system. Almost all ERP systems are not configured correctly to pull much of the information that would normally goes to profitability analysis module, resulting in material differences in Finance ledger and Hyperion systems which is used for consolidation. Therefore to understand the complexity article is divided in to different sections as per below:

1. Business processes

The goal is to use profitability analysis module to deliver management reporting that is communicated to top management in an organization. Although this process never reconciles completely with Hyperion consolidation system due to exchange rates and other details involved with full consolidation process that will not occur in any transactional system. Fundamental requirement from leadership is to have profitability by region and in order to achieve this profitability analysis module will have to use ship to party information included in every sale as way to split all sales by region.

In terms of what level of profitability would be shown in an organization, the intention is to provide profitability to actual contribution and gross margin level before Sales General and Administrative costs. Since profitability analysis is exclusively used for managerial reporting the common practice is to include information directly related to sales and production of goods. Also planning information can be loaded to profitability analysis module which is used to compare actuals vs plan to make effective managerial decisions. Profitability reports are generated to analyze contribution margins of market segments along with other COPA characteristics like sold-to country, source/production plant, ship-to party, customer, distribution channel and profit center.

***Corresponding Author**
Email:
sarat.tella@adm.com
Tel.: +1-217-450-7128

IIOAB JOURNAL
ISSN: 0976-3104

2. Maintenance of master data

In SAP S/4 system, defining and maintaining master data is a crucial task and in this section will explain what kind of master data is required in order to achieve customer and product profitability in an organization. First need to define organizational units like company code/legal entity, plant where production happens, sales organization, distribution channel and division where sales happen in an organization. In controlling module of SAP S/4 HANA system, controlling area is defined where company code is assigned to and in profitability analysis module operating concern is defined which represents a data cube with certain defined structure which contains predefined characteristics like customer, product and ship-to party [1], and value fields/GL accounts.

In SAP S/4 system there are two types of profitability analysis: Account based and costing based analysis. In Account based and costing based analysis, characteristics like customer, product, product hierarchy, sold-to party, ship-to party is created and included in operating concern so that we can analyze transactional information based on characteristics what we defined. Difference between account based and costing based analysis is that in account based General Ledger accounts are used where as in costing based analysis value fields are defined along with characteristics and assigned to operating concern. Characteristics include customer, product, segment, trading partner, payer, material account assignment group, customer account assignment group and Incoterms. Value fields which is required in costing based analysis include quantity and amount fields. Quantity fields include sale volume whereas amount fields include Gross Revenue, sales deduction and Net revenue fields.

Customer master/business partner, material master is defined in transactional system with certain defined characteristics in SAP S/4 HANA system. Business partner contains lot of information where in general view name of customer, address and in company code view contain reconciliation accounts, and in sales & distribution contains account assignment group which identify third party or intercompany, partner functions where Sold-To Party, Bill-To Party, Ship-To Party and payer is defined. Material master in transactional system which is an crucial thing to define and have lot of information like sales information, material type, material group, product hierarchy, foreign trade, accounting and costing view which contains standard cost.

3. Transactions from different modules in SAP S/4 HANA system to COPA (controlling – profitability analysis)

Mainly COPA is used as a data cube to gather information from different modules of SAP S/4 HANA system. In Sales and distribution module, product sold to customer starts with a process of creating a sales order where customer and material is being sold by customer service agent. Outbound delivery is created to ship product to customer location from warehouse. Finally invoice to customer and collect cash from customer. During billing all the information is passed to COPA via condition types defined in sales pricing procedure. It contains information like quantity sold to customer, gross revenue, sales deduction and other characteristics defined on customer and product are passed to COPA or derived by using derivations in COPA. And also there are certain characteristics that is derived during transactions or directly derived in the reporting layer based on values defined in customer and product.

From billing documents, net revenue can be calculated in COPA but costs need to be updated in each and every transaction to calculate gross margin by customer and product. Standard costs are calculated for every product at the beginning of the month in the organization. While posting a billing document standard costs is populated based on material and plant combination in COPA and used to calculate gross margin by customer and product. Standard cost comprised of variable and fixed costs and this will provide flexibility in COPA to calculate contribution margin and gross margin by customer and product. Fig 1 explains data flow from different modules such as sales, production, costing, overhead management and direct posting from finance to profitability analysis module.

4. Reporting

Transactional data from different modules are collected in profitability analysis and is passed to business warehouse system for reporting purposes. The main purpose of the report is to provide visibility and insight in to sales and margin activity. Actual costs are not available until month end processing is complete, so during the month standard costs and standard margins are available. The period end processing is finalized by the close of work day 4 and actual costs/periodic unit price is populated from material ledger. After actual costs are available, billing documents are revaluated with actual costs and post to COPA after month end close is completed. Statutory view reflects from legal entity perspective so inclusive of intercompany activity whereas management view eliminates intercompany transactions. The source data in BW system is replicated from SAP S/4 HANA system and extraction happens daily and during month end replication happens more frequently (usually every 4 works) to support month end reconciliation process

between finance and consolidation system. Table 1 depicts the information required for user selection criteria.

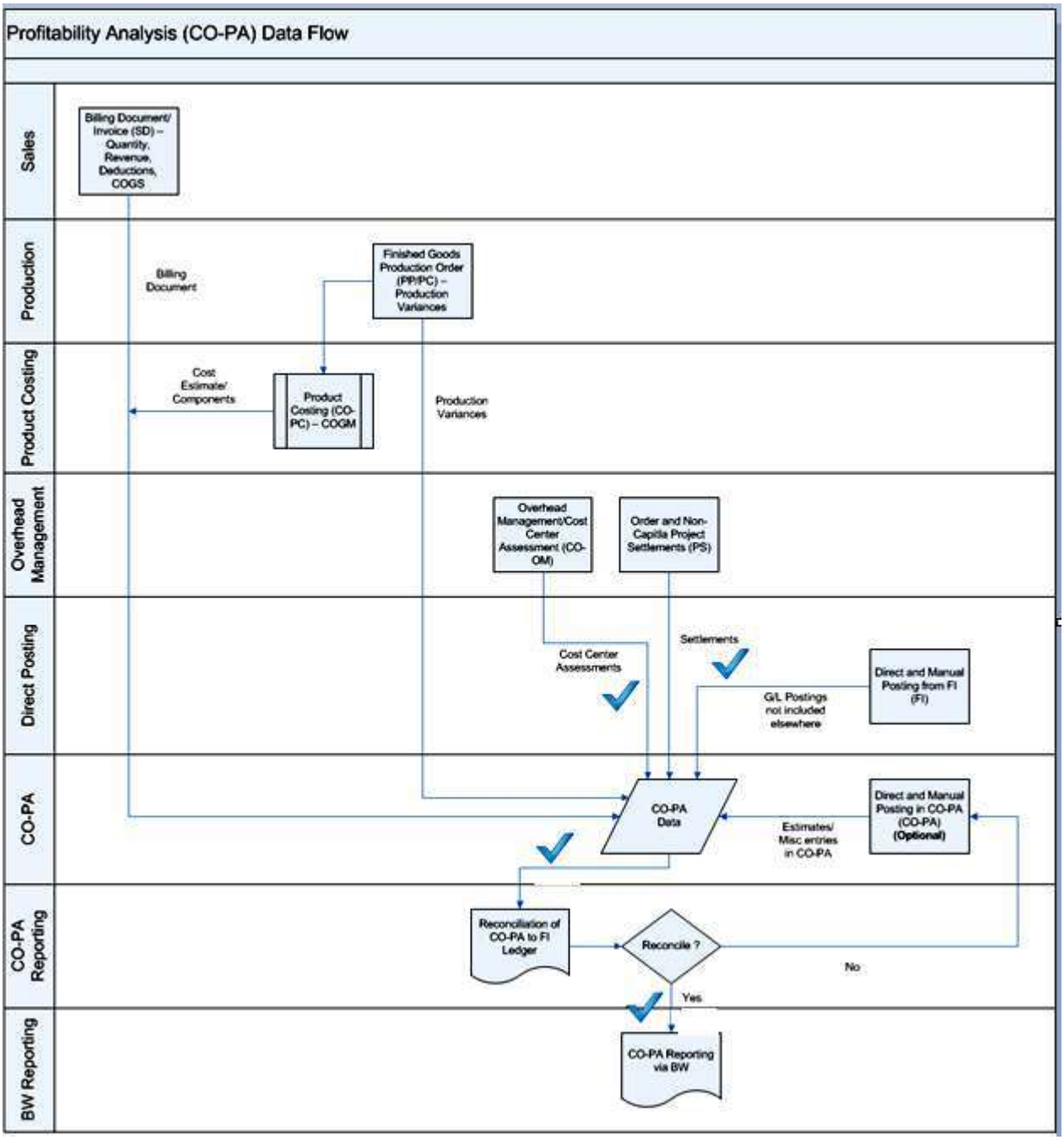

Fig. 1: Profitability data flow in an organization from different process teams

Table 1: User Selection criteria for customer and product profitability report

Field	Multiple value	Range	Optional	Mandatory
Customer	X		X	
Product	X			X
Location	X			
Sales org	X			
Time Period		X		X

Report free characteristics should be defined in the report. If an aggregated level of the product or customer hierarchy is selected, the margin and other values should be displayed at that level. There should be ability to display multiple, single values and ranges in the report. Table 2 displays free characteristics usually available in any of the gross margin reports for drill down reporting. [2].

Table 2 : Free Characteristics in customer and product profitability report

Free Characteristics
Product Attribute
Customer Segmentation
Customer Group
Industry code
Application Code which is on product level
Contract validity date
Product Tiers
Target Price

Report will be displayed in columns and based on time period options report should display actual gross margin based on customer, customer group, product or product family, region or sales organization. Table 3 displays output of the report with necessary information to get customer and product profitability in an organization.

Table 3 : report output for customer and product profitability report

Label	Description
Sales Volume	Volume sold to customer
Gross Product Revenue	Sales Price of product to customer
Intercompany Revenue	Sales price sold between intercompany
Sales Deductions and allowances	Sales deductions provided to customer
Freight costs	Freight costs
Net Revenue	Net sales price charged to customer.
Standard Variable costs	Variable cost of product
Standard Contribution Margin	Calculated from Net Revenue and standard variable costs
Standard Fixed costs	Variable cost of a product
Standard Gross Margin	Calculated from Net Revenue, standard Variable and Fixed costs
Actual Variable Costs	Actual variable cost of product after month end is completed
Actual contribution Margin	Calculated from Net revenue and actual variable costs
Actual Fixed costs	Actual fixed cost of product after month end is completed
Actual Gross Margin	Calculated from Net revenue, actual variable costs and Actual Fixed costs
SG&A Costs	Sales, general and Administrative costs not allocated to customer and product
Commissions	Other expenses or any sales commissions. This alignment may differ from organization to another.
Net Income	Calculated from Net Revenue, Actual Gross margin, SG&A costs and other commissions.

RESULTS

As per defined approach, organization able to get better insights to customer and product profitability and also helps leadership to make decisions faster based on output. This paper defined an approach for successful implementation of gross margin or customer and product profitability using SAP S/4 HANA system in large complex organization. Approach followed is account based profitability where we don't have perform reconciliation between finance and COPA with in controlling where as all functionalities available in costing based still possible with Account based analysis. Gross margin reports can be useful for finance, sales and production teams to analyze the information and take informed decisions whether customer is profitable or not, product line is profitable or not and also used for planning and forecasting of production of that particular product.

IIOAB JOURNAL
ISSN: 0976-3104

CONCLUSION

Using SAP S/4 HANA system, customer and product analysis make it easier and also provides drill down capabilities up to actual gross margin by customer and products based on the defined approach. With out capabilities there will be lot of manual process involved to come up actual gross margin by customer and product which is a herculean task in any organization.

CONFLICT OF INTEREST

There is no conflict of Interest.

ACKNOWLEDGEMENTS

I would like to express my special thanks to all project team in Tate and Lyle, it is truly an honor to work with all of you.

FINANCIAL DISCLOSURE

There are no financial conflicts of interest to disclose.

REFERENCES

[1] Schmalzing K. [2020] Profitability Analysis with SAP S/4 HANA, Rheinwerk Publishing, Quincy, MA, USA.
[2] Christensen J, Darlak K, Harrington R, Li Kong, Poles M, Savelli C [2017], SAP BW/4 HANA, Rheinwerk Publishing, Quincy, MA, USA.