Modeling of Optimized Data Processing Framework for Potential Knowledge Discovery And Recommendation Based on Healthcare Big Data

Karunamurthy A, M.Aramudhan

Abstract: ***The growth of data in healthcare application provides voluminous information about patients which are rich and meaningful insights using machine learning algorithms. In such cases, the volume and velocity of such high dimensional data requires new big data analytics framework where conventional machine learning tools cannot be directly applied. To overcome the issues like data uncertainty and misclassified data, we propose a better recommendation and optimization model which facilitates the healthcare systems. The proposed study is enhanced for data processing framework and better decision support systems. The proposed MR- FA devises the storage system of the big data which effectively generates the index for the received data. It works like partition based clustering algorithms which splits and then stores the given data to achieve the transactional database. Then, k-NN machine learning model is applied over transactional database to derive relevant knowledge using different k-values. It helps us to achieve suggest the knowledge for the users. The proposed framework is analyzed on four datasets collected from UCI machine repository. The performance metrics such as parallel processing time and reliability time are studied. It states that the computation of optimal fitness values helps the system to achieve the desired goal better parallel processing time with efficient reliability time.***

Index Terms: Healthcare systems; Uncertainty; Decision support systems; Recommendation systems; Machine learning models; Transactional database and the reliability

I. INTRODUCTION

The development made in big data analytics experienced by healthcare applications brings the opportunity to improve the accuracy of the data. The arrival of high dimensional data might bring the noise and bias to achieve better data processing framework [1]. In other cases, big data enables the effectiveness of advanced machine learning techniques. The advanced techniques engage optimal parameters than the conventional machine learning techniques. In some cases, the restricted data acquisition degrades the efficiency of the data processing systems which is an issue in big data systems. Henceforth, the machine learning techniques should be devised to improve the accuracy of the data. In data oriented systems, the patient's details have to be effectively store, retrieve and analyze the voluminous data [2]. Henceforth, the big data analytics helps the data analyzers to extract the insights for better decision making systems.

Big data analytics in complex healthcare environment is a trendy topic towards the principles of big data management and mining [3]. The objective of the big data analytics in healthcare is to support the research, availability and accessibility of the sensitive data. Clinical decision support system is a branch of science that supports clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care. The capability of the dealing with structured and unstructured data from variant sources demands big data analytics tools. A unique opportunity lies in the integration of traditional medical informatics with mobile health and social health, addressing both acute and chronic diseases in a way that we have never seen before. The big data is irresistible because of its volume and different data types. In the perspectives of data scientist, array of opportunities are available [4]. The discovery of association and percepting the data patterns assists the healthcare systems at lower expenditure.

Each big data constitute a value which hampers the progress of the data by imposing issues like heterogeneity, scale, timeliness and complexity [5]. Most of the data is not structured which challenges the domains like semantic content and search; transforming such content into a structured format for later analysis is a major challenge. The value of data enhances when it can be linked with other data, thus data integration is a major creator of value. Classification is the technique that resolves the challenges of the data processing framework, type of data analysis and data sources of the target variable on selected applications [6]. The rest of the paper is organized as follows: Section II presents the related work; Section III presents the proposed work; Section IV presents the experimental analysis and finally concludes in Section V.

II. LITERATURE SURVEY

This section explains about the related work of the big data analytics using machine learning techniques. The author in [7] discussed a data driven based stochastic model using machine learning techniques.

Revised Manuscript Received on March 10, 2019.

Karunamurthy A, Research Scholar, Department of Computer Science, Research and Development Centre, Bharathiar University, Coimbatore, India.

Dr. M. Aramudhan, Associate Professor & Head, Department of Information Technology, Perunthalaivar Kamarajar Institute of Engineering & Technology, Karaikal, India.

Most of their data models resolved the uncertainty presented in data of multi-class systems. It is noticed that computational tractability on large scale systems. In [8], studied a social network based patient data using evidence based knowledge systems. They discussed on patients and their relationships to predict the diseases. They presented and validated their approach using case-study of Brain Aneurysm with performance in terms of sensitivity and time-probability measures.

The author in [9] presented a group decision making and information quality in e-health systems. They discussed a framework using extended logic to represent the knowledge which further supports better decision making. It was experimented in virtual E care systems to sustain better healthcare services. Then, the study was extended to deep patient similarity learning model using Convolutional Neural Networks (CNN) [10] that captured the local information. Depends on similarity values, the patients data are clustered. It throws the location failure issue. The data driven adaptive nested oriented computation modelling framework to resolve data uncertainty was suggested [11]. They introduced Bayesian model which captured the nature of uncertainty data. A tailored column-and-constraint generation algorithm to solve the resulting problem occurs in design and operations of batch scheduling systems.

Discovery Engine [12] is a field used to find out the characteristics of the patients. They demonstrated the performance of DE in two clinical settings: diagnosis of breast cancer as well as a personalized recommendation for a specific chemotherapy regimen for breast cancer patients. By doing so, the authors have achieved an uncorrelated orthogonal transformation for breast cancer classification. Discrete event system was introduced by [13] that diagnosis breast cancer for a specific chemotherapy regimen for breast cancer patients. The author in [14] discussed the application of the discrete event system specification (DEVS) formalism within system of systems engineering (SoSE) to develop coordination models for transactions that involve multiple disparate activities of component systems and that need to be selectively sequenced to implement patient-centered coordinated care interventions. Then, an optimization model was introduced to perform metaheuristic algorithm. LOA is constructed based on simulation of the solitary and cooperative behaviors of lions such as prey capturing, mating, territorial marking, defense and the other behaviors. In order to evaluate performance of the introduced algorithm, they have tested it on a set of various standard benchmark functions. The results obtained by LOA [15] in most cases provide superior results in fast convergence and global optima achievement and in all cases are comparable with other meta-heuristics. Big NN was developed to build a highly extensible framework for text mining classification systems. It is mainly used for educational purposes to handle long length data. The consistencies of data have not studied by the traditional text analytics models. Then, a semantic based inference model [16] is studied using machine learning algorithms.

In [17], semantic inference on clinical documents using ML algorithms based on their requirements. Decision Tree is used to stream the data for multi-class classification. Finally, semantic based relationship is formed from heterogeneous information systems. Some semantic rules occupies higher memory rate. This resolved by suggesting deep analytics framework [18]. The patient details are elegantly stored and retrieved using deep analytics framework. Based on user's experience, three-level cognition model was deployed. Since the analytics systems treats physical and physiological data equally which throws high dimensional data. Context aware personalized healthcare [19] systems are investigated using artificial data generation process. Correlation of those data analyzed with different activities and symptoms.

III. PROPOSED WORK

This section explains about the working model of proposed scheme. The main theme of this study is to develop an efficient decision making systems for healthcare applications using advanced machine learning techniques.

A. Research Issues:

The study in data processing on medical datasets is still a challenging task because of the issues such as uncertainty and misclassified results. Since big data is a data centric, the changes should be easily adaptable and scalable. Temporal uncertainty is an emerging issue in the field of big data analytics. The analysis of new data is a hypothetical process due to irrelevant noise accumulation under big data framework. Thus, it leads to following research questions.

- Does it make sense to process all available data or focus only on the most recent data to gain the most relevant insight into the future?
- Where is the right trade-off, i.e. how far back into the history does it make sense to search?

Owing to this, the probability of misclassified results occurs that leads to improper decision making systems. So, making wiser decision is important. It is surveyed that most of the clinical databases composes of low temporal resolution information because of acquisition of time-series data. The development towards patient specific models for the available data is still under developmental stage. In particular, some chronic diseases are manifested with acute events that are unlikely to be predictable solely by sporadic measurements made within hospitals.

B. Research Process:

The proposed framework composes of following steps, namely,

(i) Data collection:

Data collection is an important part of the big data analytics. Different sets of medical data are collected from UCI machine repository, namely, Thyroid, Dermatology, Breast cancer and Heart diseases. The collected data is preprocessed by replacing the missing values with 0's or 1s (or) else just eliminating the null fields.

(ii) Data processing framework:

It is an important step that aims to store the data efficiently using firefly algorithm. Data uncertainty is the major issue of the data processing frameworks.

Handling of high dimensional and dynamical data often leads to uncertainty issue. To resolve this, we incorporate firefly algorithm into the Mapreduce system that helps to reduce the impact of recurring queries. We formulate the clustering task as an optimization problem to yield the best solution based on minimum distances between data points and the cluster centroids. The proposed MR-FA is a partitioning clustering algorithm that decides the cluster value by its light intensities. The cluster value is updated based on swarm brightness values. In MR- FA, each firefly F_i contains information such as L_i, FV, BLLI, BLFV, BGLI, and BGFV which is used for clustering process.

Light intensities (LI): Current light intensity value.

Attractiveness value (AV): Current attractiveness value

Fitness value (FV): Current fitness value for firefly at iteration t.

Best local light intensities (BLLI): Analyzed LI values for firefly F_i.

Best local fitness values (BLFV): Analyzed FV values for firefly F_i

Best global light intensities (BGLI): Analyzed LI values for firefly F_i.

Best global fitness values (BGFV): Analyzed FV values for firefly F_i

The main two operations needs to be adapted and implemented to apply the clustering task on large scale data are the fitness evaluation and firefly light intensities updating. The light intensity value is observed by the brightness value of every firefly. As we know that, the distance increases the brightness of firefly decreases. Henceforth, the fitness value is estimated by distance function.

$$\text{Fitness value (FV)} = \frac{\sum_{j=1}^{n_j} \text{Distance}(D_i, C_j)}{k} \quad (3.1)$$

Where,

N_j= No.of data belongs to j cluster.

D_i= i^{th} data

Distance (D_i, C_i) = Distance between the data D_i and the cluster value C_j.

Manhattan distance is employed to estimate the distance between the data points. It is given by:

$$Distance(D_i, C_i) = \sum_{v=1}^{D} |D_{iv} - C_{jv}| \quad (3.2)$$

Where,

D is the dimension of data D_i.

D_{iv} is the value of dimension v in data D_i

C_{jv} is the value of dimension v in class C_j

The proposed framework composes of three sub-phases, namely

- The first phase is a Mapreduce job to update the class values.
- The second phase is a Mapreduce job that updates the class values generated from previous phase.
- Atlast, the third phase merges the first two phases, in addition to, the global light intensities and global fitness is evaluated. Then, the process continues until finding of optimal solution.

a) First phase:

The Mapreduce job is employed to update the light intensities value. Mapreduce operator composes of two operations, namely, map that represents the data with its unique ID (Map key), whereas reduce function that sorts the output from map function. Let F_{id} represents the firefly with its ID. Map_{key} represents the F_{id} and Map_{value} represents the information such as FV, BLLI, BLFV, BGLI and BGFV. Then, the map function is given as:

$$Map_{func} \rightarrow \{Map_{kev}, Map_{value}\}$$

The cluster value is updated from the eqns. 1

$$x_i = x_i + \beta_0 e^{-\gamma r_{i,j}}(x_j - x_i) + \alpha E \quad (1)$$

Where, $\beta_0 e^{-\gamma r_{i,j}}(x_j - x_i)$ is the attractiveness value;

αE is the randomization parameters;

x_i is the current i^{th} data.

The other coefficients represent the configuration details of files. Then, the map function transfers the updated cluster value to the reduce function. In this work, MapReduce framework is denoted by its number of cluster node and the size of the swarm. The task of reduce function is to combine the output of the map function.

b) Second phase:

This phase again computes the fitness values for the updated fireflies. The above phase continues until global fitness value and global light intensity are found. The Mapreduce framework make use of caching files, thus, the map and reduces function combines together to generate' primary key'. Each firefly, the map function obtains the cluster value and estimates the distance between data and the current cluster value. Each value is updated and then a new value with new key is transferred to the Reduce function. Each center value of firefly is considered as fitness value, and aggregates the data under similar key. Again, the fitness value is obtained from the key emitted by Reduce function. This updated fitness value is processed for further distributed file system.

c) Third phase:

In this phase, the first two phases are combined together and the process continues until the generation of optimal solution. The cluster values obtained from second phase are collected and the global fitness value is estimated. The local light intensity value is compared with new firefly fitness value. If it's lesser, then new distance value is found. Then, new firefly with new information saved in distributed file system.

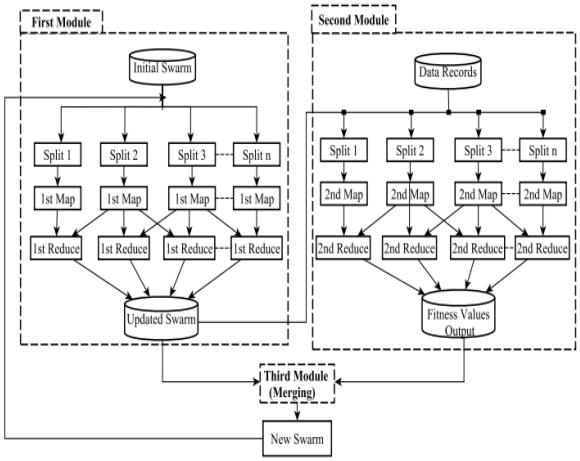

Fig.1. Working of firefly optimization in Mapreduce framework

Thus, the deployment of firefly optimization into MapReduce framework helps to achieve the organized data storage process.

i) Applying Machine learning based recommendation system:

This step helps to make wiser decision via machine learning process. Most of the researchers have studied about how to efficiently store the healthcare data. In this step, we have employed machine learning based recommendation systems that efficiently searches and retrieves the relevant data. This framework is proposed to provide integration and analysis of distributed data that further help patients to take better treatments and healthcare experts to take effective decisions. A recommendation based k-nn classification algorithm is designed and explain as follows:

a) Let P= {P_1, P_2…P_n} be the no.of patients; C = {C1, C2, C3, C4} be the set of classes; D= {D_1, D_2…D_m} be the set of organized transactional data (MapReduce).
b) Split the databases into training dataset and testing dataset.
c) The training dataset composes of common four diseases, namely, Breast Cancer, Dermatology, Heart Disease and Thyroid disease.
d) Each disease possesses its own features i.e predictive variables. In first phase, the obtained healthcare data is efficiently stored in MapReduce framework. The updated information with its new Unique ID is considered as transactional database.
e) Estimating the distance between the testing data and the trained data points using eqn. (3.2).
f) Finding k (distance) values for each trained sub- class.

ii) Decision Support System:

Based on k-values from previous step, the knowledge discovered for healthcare environment using multi-criteria based recommendation systems. The criteria dictate the different k-value estimation for the obtained distances. By doing so, the technicians, non –technicians and experts in healthcare environment can easily make decisions. As we know that, small ratio indicates a high degree of similarity and a large ratio indicating a low degree of similarity. Thus, the similar interval ranges are grouped together and form a final class. If any new data enters the system, the interval value depicts the predicted class of the data.

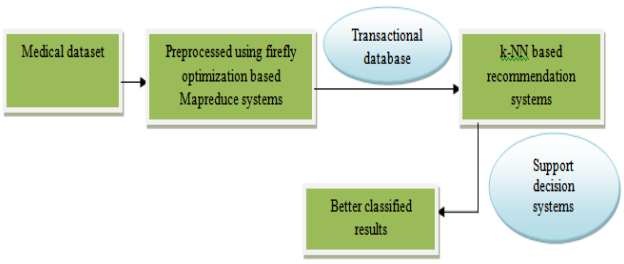

Fig.2. Proposed workflow

IV. EXPERIMENTAL RESULTS

This section presents the experimental analysis of our proposed model. Our proposed framework resolves the issues of uncertainty and the misclassified data. Since the healthcare environment is studied via collection of four datasets, namely, breast cancer, dermatology, heart disease and the thyroid disease. Each dataset composes of records is given in table 1.

Table1: Dataset details of healthcare systems

Sl. No	Database name	No. of instances	Attributes	Size (MB)
1	Breast cancer	286	10	3MB
2	Dermatology	366	11	5MB
3	Heart disease	303	24	5MB
4	Thyroid disease	7200	20	200MB

Let us assume that the process executes in several nodes of 2.304 TB of aggregated memory, 6GB of RAM, and 4 Intel cores (2.67 GHz). Let us assume that no. of fireflies be 100 with three parameter sets such as, Set 1 (α =0.1 and$\gamma = 0.01$); Set 2 (α =0.25 and$\gamma = 0.1$); Set 3 (α =0.5 and$\gamma = 1$)

Table 2: No. of iterations required to achieve the goal

No. of fireflies (N)	Average			Median			Minimum			Maximum			Expected no. of fitness function		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
20	1926	2769	3470	2659	3601	3563	2574	2801	3148	6440	6398	95325	110925	121560	121560
40	1709	2693	2561	2669	4520	3386	2709	2938	3709	6452	7303	189090	219930	219930	243120
60	1862	2801	2351	3654	4435	3316	2796	3883	3758	6398	8225	375360	375360	430560	479640
80	2759	3655	2340	3515	4438	2275	2826	2796	3649	7851	8312	28890	154500	208200	265920
100	2575	3541	6512	3479	3225	2017	2826	3148	3645	8001	7446	51270	132080	190680	377580

a) Parallel processing time:

Parallel processing time is the time taken for clustering the nodes when the size of the dataset increases. It is given as follows:

$$Parallel_{Time} = \frac{1}{1-P}$$

Where,

P is the number of parallel processors.

Table 3: Determination of parallel processing time.

No. of clusters	P= 0.50	P= 0.90	P= 0.95	P= 0.99
10	1.82	5.26	6.82	9.17
20	1.98	9.17	16.79	50.35
30	1.99	9.91	19.35	90.89
40	1.98	9.89	19.42	99.04
50	1.97	9.87	19.69	99.20

b) Reliability time:

Reliability time is the estimation of running time taken between the cluster nodes. It is certainly increased when the cluster sizes increases. It is given as follows:

$$Reliability_{Time} = \frac{T_2}{T_n}$$

Where,

T_2 is the running time between Node 1 and Node 2.
T_n is the running time among n nodes.

Table 4: Average reliability time

No. of dimensions (n)	Average computation time (s)	
	Proposed	Existing [20]
20	36.8	40.5
40	53.3	58.2
60	63.2	66.5
80	73.1	75.6
100	95.3	105.3

Table 5: k -value estimation for trained data

Dataset	No. of training instances	K value				
		K=1	K=2	K=3	K=4	K=5
Breast cancer	150	75.18	76.03	78	77	79.12
Dermatology	200	77.2	77.36	75	81.6	80.3
Heart disease	150	79.2	78.25	81.4	78.23	81.4
Thyroid disease	3000	78.3	81.1	81.7	81.96	78.36

The table 5 depicts the k-value estimation for training data. This k-value dictates the similarity finding of the upcoming data. Based on those different k-values, the tested data are classified.

V. CONCLUSION

Due to the enormous growth of information in healthcare systems, the need of big data analytics is highly challenging study. There is a growing issue such as data uncertainty and misclassified results which degrades the decision making systems. In this paper, we propose an enhanced big data optimization model and better decision support systems. Initially, we collect big data of four datasets from UCI machine repository, namely, breast cancer, dermatology, heart disease and the thyroid diseases. These data are collectively grouped together to provide better clinical decision support systems. The collected data are pre-processed and then arranged using Mapreduce framework. Here, we introduce firefly optimization model under Mapreduce framework which efficiently stores the data. Then, a recommendation based k-NN classification is deployed to classify the transactional data and recommends the right knowledge to the system users. Experimental analysis is carried out in different healthcare datasets to prove efficiency such as parallel processing and reliability time of the proposed systems. The results states that the use of firefly optimization outperforms better than prior work. We also suggest that the obtained k-values assist the test data with better accuracy systems.

REFERENCES

1. Aleman, D. M., A. Kumar, R. K. Ahuja, H. E. Romeijn and J. F. Dempsey (2008). "Neighborhood search approaches to beam orientation optimization in intensity modulated radiation therapy treatment planning." Journal of Global Optimization 42(4): 587-607.
2. Allaudeen, N., J. L. Schnipper, E. J. Orav, R. M. Wachter and A. R. Vidyarthi (2011). "Inability of providers to predict unplanned readmissions." Journal of general internal medicine 26(7): 771-776.
3. Allaudeen, N., A. Vidyarthi, J. Maselli and A. Auerbach (2011). "Redefining readmission risk factors for general medicine patients." Journal of Hospital Medicine 6(2): 54-60.
4. Araz, C., H. Selim and I. Ozkarahan (2007). "A fuzzy multi-objective covering-based vehicle location model for emergency services." Computers & Operations Research 34(3): 705-726.
5. Bard, J. F. and H. W. Purnomo (2005). "Preference scheduling for nurses using column generation." European Journal of Operational Research 164(2): 510-534.
6. Bednarz, G., D. Michalski, C. Houser, M. S. Huq, Y. Xiao, P. R. Anne and J. M. Galvin (2002). "The use of mixed-integer programming for inverse treatment planning with pre-defined field segments." Physics in Medicine and Biology 47(13): 2235.
7. Benbassat, J. and M. Taragin (2000). "Hospital readmissions as a measure of quality of health care: advantages and limitations." Archives of Internal Medicine 160(8): 1074-1081.
8. Berretta, R., A. Mendes and P. Moscato (2005). Integer programming models and algorithms for molecular classification of cancer from microarray data. Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38, Australian Computer Society, Inc.
9. Bertsimas, D. and R. Shioda (2007). "Classification and regression via integer optimization." Operations Research 55(2): 252-271.
10. Bester, M., I. Nieuwoudt and J. H. Van Vuuren (2007). "Finding good nurse duty schedules: a case study." Journal of Scheduling 10(6): 387-405.
11. Joynt, K. E., E. J. Orav and A. K. Jha (2011). "Thirty-day readmission rates for Medicare beneficiaries by race and site of care." Jama 305(7): 675-681.
12. Kansagara, D., H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman and S. Kripalani (2011). "Risk prediction models for hospital readmission: a systematic review." Jama 306(15): 1688-1698.
13. Lee, E. K., F. Yuan, D. A. Hirsh, M. D. Mallory and H. K. Simon (2012). A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department. AMIA Annual Symposium Proceedings, American Medical Informatics Association. Annu Symp Proc 2012; 2012: 495-504.
14. Lee, E. K., F. Pietz, B. Benecke, J. Mason and G. Burel (2013). "Advancing Public Health and Medical Preparedness with Operations Research." Interfaces 43(1): 79-98.
15. Lee, E. K., F. Yuan, A. Templeton, R. Yao, K. Kiel and J. C. Chu (2013). "Biological Planning for High-Dose-Rate Brachytherapy: Application to Cervical Cancer Treatment." Interfaces 43(5): 462-476.
16. Lee, E. K., F. Yuan, Y Cao, A. Templeton, R. Yao, K. Kiel and J. C. Chu (2016). "PET-image guided TCP-driven Treatment Planning Optimization for High-Dose-Rate Brachytherapy" International Journal of Radiation Oncology* Biology* Physics.
17. Pato, M. V. and M. Moz (2008). "Solving a bi-objective nurse rerostering problem by using a utopic Pareto genetic heuristic." Journal of Heuristics 14(4): 359-374.
18. Persson, M. and J. A. Persson (2009). "Health economic modeling to support surgery management at a Swedish hospital." Omega 37(4): 853-863.
19. Phoungphol, P., Y. Zhang, Y. Zhao and B. Srichandan (2012). Multiclass SVM with ramp loss for imbalanced data classification. Granular Computing (GrC), 2012 IEEE International Conference on, IEEE.
20. Sangeetha, B., Saranya, E., & Saranya, G., (2015) "A Novel Framework for Secure Sharing of Personal Health Records (PHR) in Cloud Computing". International Journal of Advanced Scientific Research & Development (IJASRD), 2 (2), pp. 08 - 14.
21. Sankari, E., Priya, S. R., Suriya, R., Prabhakaran, G., & Sowkarthika, T., (2016) "Enhancing Security through Data Hiding". International Journal of Advanced Scientific Research & Development (IJASRD), 3 (1), pp. 126 - 133.
22. Saranya, E., Praveenkumar, S., & Usha, N. S., (2014) "A Review Based Study of Classification of X-Ray Images Using Content Based Image Retrieval (CBIR)". International Journal of Advanced Scientific Research & Development (IJASRD), 1 (1), pp. 25 - 32.

23. Navneet and Nasib singh gill (2017). "A novel algorithm for big data classification based on Lion Optimization", Journal of theoretical and applied information technology. 95(7): 1525-1532.

AUTHORS PROFILE

Mr. A. Karunamurthy obtained the M.Tech., and MCA degree from Manonmaniam Sundaranar University at Tirunelveli and Pondicherry University at Puducherry respectively. Current he pursuing his Doctor of Philosophy in Computer Science in Bharathiar University, Coimbatore. He has more than five years of teaching and research experience. He had published number of research articles in referred journal and also has participated and presented the research and review articles in many national and international conferences and seminars.

Dr. M. Aramudhan, Ph.D., is a Head & Associate Professor in Information Technology at Perunthalaivar Kamarajar Institute of Engineering and Technology (PKIET), Karaikal. He received his Ph.D., from Anna University, Chennai. He is an academician, a Research Supervisor in Computer Science, with more than 20 years of accomplished experience in teaching. He has published over 49 articles in national and international referred journals and one text book for the college students. He has received Young Teacher Award from AICTE and also received best paper award 3 times.