AN Enhanced Model for Diabetes Prediction using Improved Firefly Feature Selection and Hybrid Random Forest Algorithm

B. Senthil Kumar, R. Gunavathi

Abstract:Diabetes is a chronic disease that causes numerous amount of death each year. Untreated diabetes disturbs the proper functionality of other organs in human body. Hence early detection is a significant process to have a healthy life style. Usually the performance of the classification is affected due to the existence of high dimensionality in medical data.In this study a system model is proposed on Pima dataset to enhance the classification accuracy by eliminating the irrelevant features. Therefore it is important to choose a suitable feature selection approach that provides the better accuracy in disease prediction compared to prior study.Hencenovel techniquesImprovedFirefly(IFF)and hybrid Random forest algorithmis proposed for feature selection and classification. The present study provides a better result with 96.3% accuracy.The efficiency of the present studyis compared with the prior classification approaches.

Keywords: Diabetics prediction, Pima dataset, Feature selection, Firefly optimization, Accuracy.

I. INTRODUCTION

Among the several health issues,diabetes causelarge number of death.Diabetes [1] is one kind of chronic disease occurs due to the imbalance glucose [2]. Compared to the previous century the diabetes disturbs billions of humans[3]. At present 425 million humansaffected worldwide [4].In 2000, less than one million people died whereas 1.6 million people died due to diabetes. According this information diabetes is the severe disease that is the seventh top cause of death [5]. The total diabetes patients in 1980 are108 million but it is increasing year by year in 2014, 422 million people are affected by this disease [6]. Nowadays people with below 18 years are affected by diabetes.

If the situation continues several human will affected and cause more death in upcoming years. The diabetes cause death in final stage before that it will leads to several disorders. Therefore it is important detect this disease in an early stage that allows the physicians to treat the patient with proper diagnosis.In many cases the physicians provides the erroneoustreatment without the proper experience of diabetes [8]. The contribution of the proposed paper is to develop a novel feature selection approach to reduce the unwanted features and provides the better classification accuracy that aids the patients to early predict the diabetics and prevents the death ratio due to diabetics. The upcoming sections organization is given: section II analyzes the feature selection and classification approaches for diabetesdataset developed by various researchers and their performance.

Revised Manuscript Received on October 15, 2019.

B. Senthil Kumar , Assistant Professor, Department of Computer Science, Sree Narayana Guru College.

Dr. R. Gunavathi, Associate Professor & Head, Department of MCA, Sree SaraswathiThyagaraja College.

Further feature selection methods applied to diabetic dataset with their accuracy is also surveyed. The section III provides the prior approaches in details then the section IV section presents the proposed feature selection and classification algorithm with their pseudocode.

Section V describes the experimental setup and the performance evaluation of the proposed approach.

II. RELATED WORK

This section provides the previous study on feature selection and classification on diabetes prediction. There are various techniques have been introduced by researcher. Only few researches on pima dataset is discussed below.

TalhaMahboobAlam et al (2019)present a diabetesprediction model with Principal Component Analysis (PCA) and Artificial neural network (ANN) which is used for attributes selection and classification respectively. In addition the authors found the association of diabetes using Apriori approach. 75.7% accuracy is achieved through ANN classification. HarleenKaur and Vinita Kumari (2018) applies the machine learning algorithms such as Linear Kernel SVM, Radial Basis KernelSVM,k-NN, Artificial neural network (ANN) and multifactor dimensionality reduction (MDR) on Pima dataset to develop a system to classify the diabetic and non-diabetic patients with the accuracy of 89%, 84%, 88%, 86% and 83% respectively. DeeptiSisodia and Dilip Singh Sisodia (2018) presents the diabetes classification model by utilizes the naïve bayes machine learning algorithm on Pima dataset and obtain the accuracy of 76.30%. Shut et al (2017) applied eight kinds of methods to extract the texture feature. Experiment is performed with various parameters to identify the most appropriate extractor with optimal parameters. For each extractor, the same dataset, classification methods such as Knn and SVM with validation method are used. Applying SVM this study obtainsbetter accuracy. YoichiHayashiand ShonosukeYukita (2016) introduced a diabetes prediction model using rule based algorithm. To gain a better accuracy the author utilizes the Re-RX with J48 graft on Pima dataset. The present study got 83.83 percent accuracy. T. Santhanam and M.S Padmavathi (2015)introduced a novel approach in diabetes classification using K-Means, genetic algorithms and SVM formeaningless data removal, feature selection and classification respectively.This system attained an outstanding result when compared to enhanced k-means approach. Muhammad WaqarAslam et al (2013) presents a genetic programming (GP) based method for diabetes classificationwith suitablefeature selection model with a novel combinations of diabetes features. This study utilizes the K-NN and SVM classifier for diabetes patient classification.

III. MATERIALS AND METHODS

A. Dataset

The commonly used PIDD dataset which is available in UCI repository is utilized in the proposed system because several machine learning researchers used these data on diabetic's prediction. Therefore this study performed the experiment on this database and compared the result to evaluate the efficiency.This dataset contains768 female patients' health data with eight attributes.

B. Firefly Algorithm

Firefly is an efficient optimization approach proposed by Yang [16] based on the behavior of the fireflies.It has an ideal feature that is admirable blinking lights. The fireflies flashing patterns which are produced by a bioluminescence procedure enjoy a special place for each of 2000current living fireflies' species. Two main purposes of these flashing are to attract thepotential prey and to mate partners.
In this study FA is utilized for feature selection problem. The basic firefly method is extended using the weight option to improve the efficiency of the feature selection process. The natural behavior of the fireflies is analyzed and applied this procedure in our Cleveland dataset. Generally intensity value based on the light attraction and the fitness function is computed to choose the best features. But in this study weight is applied on the intensity value to improve the result.

C. Random forest

The unsupervised machine learning Random forestalgorithm wasa combination of classification and regressionapproaches and developed through bagging technique [17]. This technique chooses the items from the source datasetwith replacement for every tree. RF utilizes the subspace method that chooses the fixed features for decision tree. The main reason for using the RF is because of these two important strategies. In the present study RF is used for the classification problems.

IV. PROPOSED METHOD

The PIMA dataset is utilized in our proposed approach for diabetic's prediction. The architecture and the steps involved in the proposed feature selection and a classification technique of our study is given in figure 1. The firefly algorithm isexpanded using the weighted intensity value for feature selection. On the other hand the Random Forest is enhanced using the back propagation approach to remove the unwanted tress to improve the final decision. The performance of the present study is compared with top classification algorithms: support Vector Machine (SVM) [18],k-Nearest Neighbors (KNN)[12], Naïve bayes(NB) [11], Artificial neural network (ANN)[19].

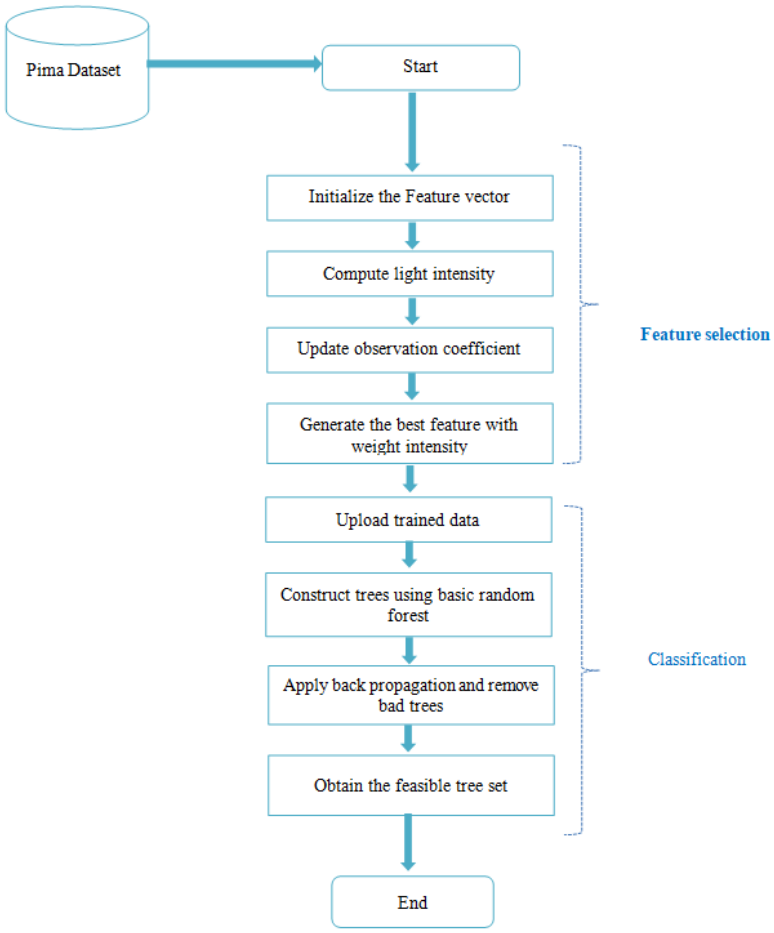

Fig 1. Proposed architecture

A. Intensity Weighted Firefly Optimization

Feature selection is a significant method to select the required feature from dataset. Fire fly is a kind of a flash light which tries to communicate with the other members of their nature. As the intensity of light vanishes with respect to distant locations, its accuracy can be defined at local horizons for finding the best solution for any function. In our research, the fire flies are the extracted features from peak estimations. Each fire fly is assigned by light intensity and Out of all the extracted features, the distinct features are selected as the best one. This is best explained by the contours in which random regions are created based on the nature of features extracted and the particles of similar species are attracted towards the centre of the regions of the contours.

The random regions are created based on the feature categories and the particles of similar nature will follow their own regions. Out of all the particles, some are in the centre of the regions and these are defined as the best features for better disease classification. Hence, fire fly optimization will serve the purpose of feature reduction technique by considering similar natured particles and neglecting the others.

Let us consider F_v as the feature vector or feature matrix. On selecting a training feature, T_f, Define m, n *and* $o$with some random values (here 0.2, l.0 and l.0 are considered respectively).

Let $X = X_i\ (i = 1,2,3 \ldots\ldots\ldots\ldots z)$(1)

Where 'n' is the number of particles and 'X' is the population of fire flies.

$Define\ I = rand(F_v)$(2)

Where I is the light intensity. Updating the observation coefficient as

$$m_i = \sqrt{(p_i - p_j)^2 + (q_i - q_j)^2} \qquad (3)$$

Where i=I,2, 3, ... z, j=I,2, 3, ... m. Final updates are expressed as

$p_n = p_n(i) \times (1 - n) + p_n(j) \times n + m(\text{rand} - 0.5)$ (4)

$q_n = q_n(i) \times (1 - n) + q_n(j) \times n + m(\text{rand} - 0.5)$ (5)

When the light intensity gets updated after some iteration, the final values are indicated as

$fn_t = I(x, y)$ Exact fitness value

$bv = \min(fn_t)$ Exact best fitness value

$T_f = F_v(bv)$ Selected best feature

B. Improved Random Forest

A single decision tree can learn one type of logic rules during the training process. While for certain trees, the logic rules may not investigate the relationship between the model inputs and outputs effectively because of random subspace algorithm and bagging. So the trees reduce the performance ofthe random forest. Therefore it is important to found those trees and has to delete them [20,21].

The present study concentrates on the classification issues. The basic RF is enhanced by applying the back propagation (BP) neural network approach. BPmeasures the efficiency of the inputs on the outputs by network weights [22]. Hence in our work every output treeestablishes the input vector and produces the final predicted result of the BP network,

The significance of the tree is calculated once the training process is completed. Then the feasible tree set is extracted with the best performance as a result by applying the forward search methodon the tree significance. The steps involved in enhancing the basic random forest are given below:

Step 1

Compute the RF to train the data

Step 2

Perform back propagation on each tree of step 1.

Step 3

Find the significance of the tree using

$$\begin{cases} II_{a,b} = \dfrac{C(p_a, q_b) d_{a,b}}{V(p_a) V(q_b)}, & a = 1, \ldots, n; b = 1, \ldots, m \\ II_a = \dfrac{C(P_a, o) d_a}{V(P_a) V(o)}, & a = 1, \ldots, n \end{cases} \qquad (6)$$

where $q$denotes the input value, o representsdenotes the output value, $d_{a,b}$denotes the weight amonga^{th}unknown value and the b^{th}known value, vi denotes the weight among the a^{th}unknown valueand the result value, hidenotes the input of the a^{th}unknownvalue, Hidenotesthe output of the a^{th}unknown value, $n$denotes the neurons counts in the known layer, $m$represents the neurons countsin the unknownlayer, Ii represents the impact index of the b^{th} input node on the a^{th}unknown value, and Iirepresents the influence index of the a^{th}unknown valueon the output node. $C(\,)$and$V(\,)$represents the covariance and variance respectively. The significance of a tree is computed by the below equation.

$$II_b = \left| \sum_{a=1}^{n} II_{a,b} II_a \right|, a = 1, \ldots, n; b = 1, \ldots, m \qquad (7)$$

Step 4

After receiving the significance of every tree the system applies the forward search method to find the feasible tree set. This process starts from initial tree with high significance and includes the next trees based on the significance level. The final tree set is selected with the highest overall performance. These feasible tree set is result of the improved random forest approach. In case of regression mean value is computed for these tree set to obtain the final result.As already discussed, this study considers the RF for diabetes classification problem. Hence the final feasible tree set will provide result of diabetes possibilities.

V. RESULTS AND DISCUSSION

The proposed model has been evaluated the performance of the presented feature selection and classification algorithm.

Various classification approaches were applied on PIMA dataset and obtain the result with few variations due to their different working criteria.The outcome of this study is based on the performance metrics such as precision, recall, f-measures and Accuracy.The experimental was performed with the help of Netbeans IDE. Six significant attributes are selected from total attributes. Pima dataset contains 500 non-diabetic patients (class = 0) and 268 diabetic ones (class = 1) as shown inFigure 2**.**

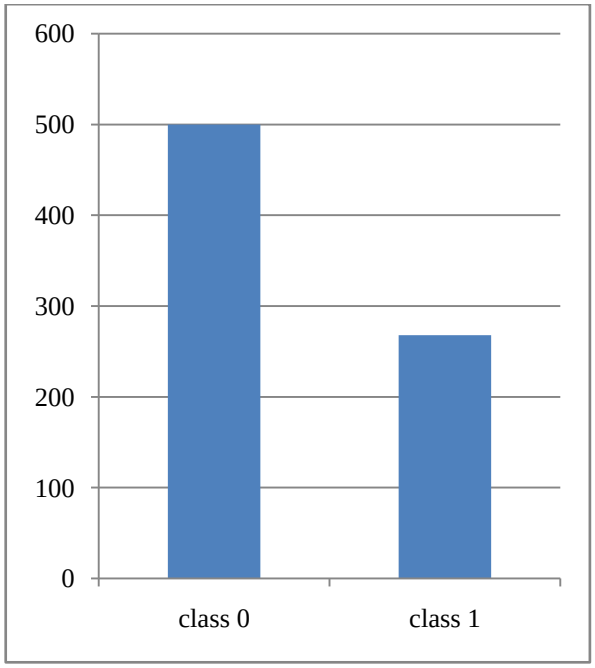

Figure 3. Outcome attribute in PIMA Dataset

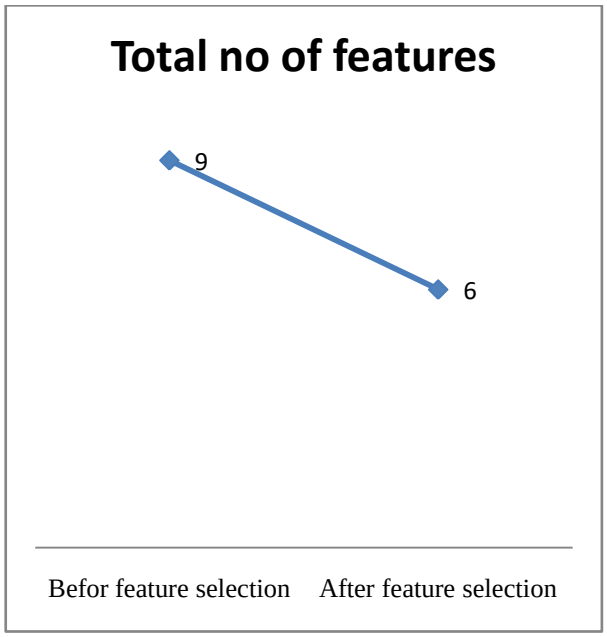

Figure 4. Features selection result comparison

Figure 3 gives the total count of features before and after feature selection process. From this plot it is clear that after feature selection two features are reduced to obtain the best attributes.The proposed system selects 6 features Glucose, Blood Pressure, Insulin, BMI, Diabetes Pedigree Function and age where as basic Firefly selects 8 features such as Pregnancies,Glucose, Blood Pressure, Skin Thickness,Insulin, BMI, Diabetes Pedigree Function and age. It is clear that the weighted firefly provides efficient result compared to basic fireflythatgives the better accuracy in diabetes prediction.

Table 1. Performance comparisonof classification model

Algorithm	Precision	Recall	F-measures	Accuracy
SVM	0.831	0.896	0.873	90.7
KNN	0.86	0.904	0.884	91.4
NB	0.81	0.86	0.835	89.1
ANN	0.84	0.896	0.875	91
RF	0.895	0.927	0.906	93.5
Hybrid RF	0.917	0.953	0.94	96.3

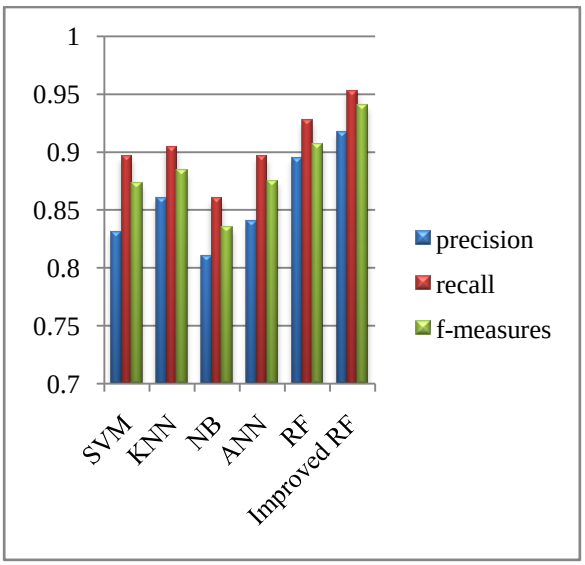

Figure 5. Precision, Recall, F-measure comparison in diabetes prediction

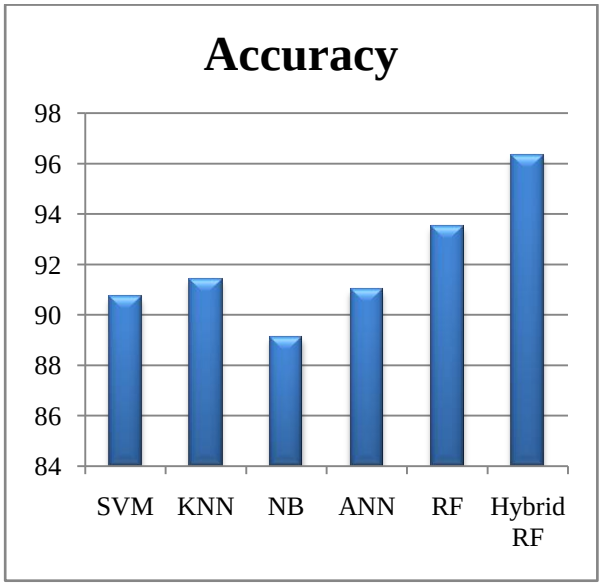

Figure 6.Accuracy comparison in diabetes prediction

Performance measures of the proposed approach have been compared. The SVM provided an accuracy of 90.7%, KNN gave 91.4%, NB gave 89.1% accuracy, ANN has given 91%, RF provides 93.5% and Hybrid RF has given 96.3%. ANN outperforms other methods, as shown in Fig. 5.

VI. CONCLUSION

This study introduced a feature selection approachby applying the weight to the basic firefly algorithm.The working principle and the pseudocode of the prior and proposed algorithm are discussed. The results showed that the hybrid Random forest algorithm obtain the better accuracy compared to other approaches such as SVM,NB,KNN,ANN and Random forest.An experimentwas conducted onpima-Indians-diabetes-databaseusing Improved Firefly (IFF) and the top six data mining techniques.In future,research can be conducted to test different combination of dataset to verify the performance of the proposed algorithm.

REFERENCES:

1. P. C. Thirumal and N. Nagarajan, “Utilization of data mining techniques for diagnosis of diabetes mellitus—A case study,” ARPN J. Eng. Appl. Sci., vol. 10, no. , pp. 8–13, 2015.
2. R. Valdez, P.W. Yoon, N. Qureshi, R.F. Green, M.J. Khoury, “Family history in public health practice: a genomic tool for disease prevention and health promotion,” Ann. Rev. public. health. no. 31, pp. 69-87, 2010.
3. International Diabetes Federation, “Idf diabetes atlas 2017,” 2017.
4. S.M. Grundy, “Obesity, metabolic syndrome, and cardiovascular disease,” J. Clin. Endocrinol. Meta., no. 89, pp. 2595-2600, 2004.
5. M. Mashayekhi, F. Prescod, B. Shah, L. Dong, K. Keshavjee, A. Guergachi, “Evaluating the performance of the Framingham Diabetes Risk Scoring Model in Canadian electronic medical records,” Can. J. diabet., no. 39, pp. 152-156, 2015.
6. Retrieved http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, Accessed 27th Jul 2018.
7. http://www.who.int/news-room/fact-sheets/detail/diabetes retrieved 27/07/2018.
8. M. Opper , O. Winther , Gaussian processes for classification: mean-field algo- rithms, Neural Comput. 12 (20 0 0) 2655–2684
9. Alam, T.M., Iqbal, M.A., Ali, Y., Wahab, A., Ijaz, S., Baig, T.I., Hussain, A., Malik, M.A., Raza, M.M., Ibrar, S. and Abbas, Z., 2019. A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, p.100204.
10. Kaur, H. and Kumari, V., 2018. Predictive modelling and analytics for diabetes using machine learning approach. *Applied Computing and Informatics*.
11. Sisodia, D. and Sisodia, D.S., 2018. Prediction of diabetes using classification algorithms. Procedia computer science, 132, pp.1578-1585.
12. Shu, T., Zhang, B. and Tang, Y.Y., 2017. An extensive analysis of various texture feature extractors to detect Diabetes Mellitus using facial specific regions. *Computers in biology and medicine*, *83*, pp.69-83.
13. Hayashi, Y. and Yukita, S., 2016. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. *Informatics in Medicine Unlocked*, *2*, pp.92-104.
14. Santhanam, T. and Padmavathi, M.S., 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 47, pp.76-83.
15. Aslam, M.W., Zhu, Z. and Nandi, A.K., 2013. Featuregeneration using genetic programming with comparative partner selection for diabetes classification. Expert Systems with Applications, 40(13), pp.5402-5412.
16. Yang, X. S. (2009). Firefly algorithms for multimodal optimization. In 5th symposium on stochastic algorithms, foundations and applications, SAGA 2009 (pp 169–178).
17. L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5−32, Kluwer Academic Publishers. Manufactured in The Netherlands.
18. Utkin, L.V., 2019. An imprecise extension of SVM-based machine learning models. *Neurocomputing*, *331*, pp.18-32.
19. Ali, J.B., Hamdi, T., Fnaiech, N., Di Costanzo, V., Fnaiech, F. and Ginoux, J.M., 2018. Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network. *Biocybernetics and Biomedical Engineering*, *38*(4), pp.828-840.
20. M.N. Adnan, M.Z. Islam, Optimizing the number of trees in a decision forest to F. Wang et al. *Measurement 125 (2018) 303–312* 311 discover a subforest with high ensemble accuracy using a genetic algorithm, Knowl.-Based Syst. 110 (2016) 86−97.
21. J. Abellán, Ensembles of decision trees based on imprecise probabilities and uncertainty measures, Inform. Fusion 14 (4) (2013) 423−430.
22. T.T. Yang, C. Cui, Y. Shen, Y. Lv, A novel denitration cost optimization system for power unit boilers, Appl. Therm. Eng. 96 (2016) 400−410.

AUTHORS PROFILE

B. Senthil Kumar,has completed his M.Phil. in Computer Science in Bharathiar University. He has 11 years of teaching experience and currently working as Assistant Professor, Department of CA & IT at Sree Narayana Guru College, Coimbatore. He has 8 years of research experience. He is now a Doctoral Student in Computer Science, Sree Saraswathi Thyagaraja College, Bharathiar University. His current field of research is Data Mining and Health Informatics. He guided 11 M.Phil scholars and published 22 papers in international journals.

Dr. R. Gunavathi, has completed her Ph.D. in Computer science in Mother Teresa Women’s University,Kodaikanal, and her research is on “Efficient Cluster head selection algorithms to improve the Quality ofservice in Mobile Ad hoc networks”. She has 20 years of teaching experience and currently working asAssociate Professor and Head, Department of MCA at SreeSaraswathiThyagaraja College, Pollachi.She has 15 years of research experience. Her current research interest is in mobile ad hoc networks,Vehicular Ad hoc Networks and big data analytics. She has published around 30 research articles in therefereed International journals with good impact factor and also presented 25 research papers in theNational and International level conferences. She has organized many National level Seminars,workshops and Conferences.