Intelligent Diagnosis of Cardiac Disease Prediction using Machine Learning

Ravindhar NV, Anand, Hariharan Shanmugasundaram, Ragavendran, Godfrey Winster

Abstract*— Cardiac disease have become worldwide common public health issue, mainly due to lack of awareness of health, poor lifestyle and poor consumption. Practitioners may have different concerns when it comes to disease diagnosis, which result in different decisions and actions. On the other hand, even in the specific case of a typical disease the amount of information available is so massive that it can be difficult to make accurate and reliable decisions. With adequate patient and non-patient medical constraints, it is possible to accurately predict how likely it is that a person with heart disease and to obtain potential information from these systems. A mechanized framework for therapeutic analysis would also dramatically increase medical considerations and reduce costs. We developed a framework in this exploration that can understand the principles of predicting the risk profile of patients with the clinical data parameters. In this article, four machine learning algorithms and one neural network algorithm were used to compare performance measurements to cardiac diseases identification. We evaluated the algorithms with respect to accuracy, precision, recall and F1 settings to achieve the ability to predict cardiac attacks. The results show our method achieved 98 percent accuracy by neural network algorithm to predict cardiac diseases.

Index Terms*— Human Cardiac, Disease, Machine Learning, Prediction, Heart disease prediction System, CVD,Neural Network.

I. INTRODUCTION

Healthcare systems have dramatically increased human life expectancy by advancing the use of medications, healthcare services through the medical history management of patients. However, the health support systems are currently facing significant challenges with general healthcare such as minimal medical information, avoidable errors, a threat to information, inconsistencies in diagnosis and delayed transmission of medical data. On the other hand, even in a typical disease the amount of information available is so massive that it may be challenging to make consistent and accurate decisions. There exists many decision support systems for disease diagnosis process like Clinical decision support System (CDSS) and similar other systems like the Electronic Health Recorder (EHR). The subject of investigation aims for designing Disease Prediction Scheme (DPS) and several research communities around the world have considerable interest in this proposed research area. Computerized Medical information systems are always of great interest to researchers. It is possible to predict how likely a individual will suffer from illness with the medical parameters of a substantial number of patients and non-patients. For instance, of an patient was unable to recollect all the symptoms of the disease and medical diagnosis that exists earlier for the patient, which is a vital resource for clinician helps for efficient decision making process. The management of the health records of patients for more precise diagnosis of the disease at an early stage was used in support of the doctor in the decision-making process. This has also raised an important question as: "What means would we be able to make a viable decision from the Information into helpful data, which can help healthcare service practitioners?" The main objective of this study mainly relies on this. Data mining techniques in this scenario are suitable for the conversion of accumulated raw data to useful, influential data to predict the system. Many methods have been used recently in the mining of data to predict disease in patients [6]. DPSS with data mining approach can be more promising for prediction of accuracy level and to reduce the diagnosis time and cost of the disease [4]. This problem is motivated by the assessment by the World Health Organization (WHO). An estimated 23.6 million persons will be diagnosed as having heart disease by 2030 according to the World Health Organization. The expectation of coronary heart disease should be completed to minimize this risk. Coronary disease analysis is typically carried out with regard to signs, manifestation, and the patient's physical examination. This huge mass of raw information is the rule resource that can be used and inspected before the extraction of important information, which directs or reinforces decision making for costs by remedial society. In order to detect, analyze, and predict various diseases, the ML techniques are more effective. One important goal is the efficiency of disease identification and a reduction of mortality caused by heart disease and prevention. The organization of the paper goes in this order. The basic ideas related research study is initially presented in section I. Related research study is presented in section II. Section III briefs out the algorithm on the proposed idea. In section IV the analysis on the experiments and the evaluation metrics in section V. Section VI gives the conclusions and future work.

Manuscript published on 30 September 2019.
***Correspondence Author(s)**

NV Ravindhar, received his B.E, M.E amd Ph.D degree specialized in Computer Science and Engineering. Currently he is pursuing his Ph.D degree from Saveetha University, Chennai (+91-9865855944, ravindhar@saveetha.ac.in)
Anand, Saveetha Engineering College and currently working in Saveeetha Engineering College, India (corresponding author to provide phone: +91-9444279907, e-mail: anand@svaeetha.ac.in).
Hariharan shanmugasundaram, is with Saveetha Engineering College and currently working as Professor in Saveeetha Engineering College, India (corresponding author to provide phone: +91-9003140372, e-mail: hariharans@saveetha.ac.in)
Ragavendran completed undergraduate B.E(CSE) programme in Saveetha Engineering College affiliated to Anna University, Chennai, India (e-mail: ragavendran1371997@gmail.com).
Godfrey Winster, is currently working as Professor in Computer Science and Engineering department at Saveetha Engineering College, Chennai, India. (godfreywinster@saveetha.ac.in)

Retrieval Number: J97650881019/19©BEIESP
DOI: 10.35940/ijitee.J9765.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

II. RELATED RESEAERHC STUDY

Over the past few years, a significant amount of research and work has been done to improve accurate model for Heart Disease prediction. The Rule-base approach to efficient prediction of cardiac disorders was presented by Purushottam et al [1] have dealt with several rules based constraints such as age, stereospaceous cholesterol, sex, chest pain, restoration of blood pressure, fast blood sugar, restoration electrocardiographic outputs, maximum cardiac rates, depression, peak exercises pitch, number of main vessels. Results using such work has produced the accuracy of 86.7% In the testing stage the accuracy achieved is 86.3% and in the training stage 87.3. The proposed CNN-MDRP (structured and unstructured hospital data, which produced 94.8% precision) based on the convolutional neural network[4]. Innovative clinical decision-making system to support diagnosis of cardiac disease, based on an analytical hierarchic process that collects information from 100 diagnosed patients [2].

Heart Disease Classification with forward heart network Algorithms using Clever Land Sets with common factors involving age, sex, chest pain, blood pressure, steroid cholesterol, blood sugar, electrocardiographic etc have lead to 98% accuracy [5]. The risk was estimated with the use of 1,174 participations, including 456 women and 718. The overall algorithm with several factors dealt in the star of this paragraph repeatedly to find some interesting conclusions. The use of Fuzzy set hypotheses, case-based reasoning and K-NN based on the hybrid reasoning contributes to improved forecast results [6]. Cardiovascular risk prediction model development and validation using NIPPON DATA have designed a tool for the prediction of cardiovascular risk using data from 2742 Japanese residents [7-8].

The diagnostic design models developed by research communities have been converted into point-based risk score sheets in which the data are divided in group based on age groups of 40 – 84 years. The midpoints for these categories were noted with an reference values with several categories. The marginal category reference values were determined and observations were recorded in this perspective. A reference group (0 points) was established for each variable as the lower risk class— e.g. the 40-49 age class, women with systolic blood pressure < 120 mmHg, and so on [15]. Palaniappan and Awang [12] have examined the comparison of different methods of data mining for the diagnosis of patients with heart disease. The comparison included naïve bays, decision tree and the network of neural networks. The results showed that in the diagnosis of heart disease patients the naïve bays can achieve the highest precision [12]. Clustering is one of the most popular clustering techniques, however the first selection of the center is an important problem which strongly affects its results. The study examined different methods of initial centric selection for K-means clustering technique is used for diagnosis of problems with range values, inlier, outlier, random attribute values and random row values. Back propagation algorithm wide range of improvements were proposed with weights for training in feed-forward neural network. Such scheme uses gradient descent technique for supervised learning in multilayer perceptron to increase the efficiency with various strategies [14].

The relationship between the input data and the desired output is more complex in nature. It also brings into mind that there is a higher number of processing elements between the input and output layer (i.e. at the hidden layer). The process is usually divided into several stages and has additional hidden layers as are required. The number of processing elements in the hidden layers and the amount of training data available sets an upper limit. Also the number of cases in the data sets to calculate this upper bound and divides the number by the total of input and output nodes in the network. For comparatively less noisy data, larger scaling factors are used. The training set has profound influence with too many artificial neuron and is too generalized and the network will no longer be used on new data sets [16].

There are different other ways to determined the hidden layers such as input layer and the output layer size and the number of hidden neurons [17]. The number of neurons that have been hidden must be 2/3 the input layer size plus the output layer size. The number of hidden layers should be less than twice as high as the number of input neurons [18,19]. Van Calster has proposed a prediction model assessment approach by estimating the difference between proportion of true positive and the proportion of false positive, multilie d by the risk threshold value [20].

Our study used Python and machine learning with support libraries such as sciki-learning, numpy, pandas or matplotlib. Many researchers have used 10-fold cross validations of overall data for medicinal data diagnosis and reported the results for the detection of disease with higher efficiency. We used both the idea of test split trains and cross validations for optimal selection of parameters.

III. DATASET AND MODEL DESCRIPTION

The In this section III, we describe about the implementation and the dataset obtained for the prediction of Cardiac Diseases.

Data Processing And Machine Learning Classification:

The UCI Machine Learning Repository is a database, domai n theory and data generator that is utilized for empirical anal yzes of machine learning algorithms in the machine learning community.

Table 1: Baseline Characteristics of Study Subject

Attribute	***Description***	***Values***
Age	*Age in Years*	*Continuous*
Gender	*Male or Female*	*0=male, 1=female*
Cp	*Chest Pain Variety*	*4 = typical type* *3 = typical type angina* *2 = no angina pain* *1 = asymptotic*
Thest bps	*Sleeping Blood Force*	*Continuous*
Chloe	*Serum Cholesterol*	*Continuous*
Rest Ecg	*Resting ECG*	*5 = normal* *4 = having ST_T wave abnormal* *3 = left ventricular hypertrophy*
Fbs	*Fasting Blood Sugar*	*1 >= 120 mg/dl* *0 <= 120 mg/dl*

Thalach	*Greatest heart speed reached*	*Continuous*
Ex ang	*Exercise induced angina*	*1 = no* *2 = yes*
Old peak	*ST despair induced by apply virtual to relax*	*Continuous*
Slope	*Slope of the height effect ST section*	*3 = unsloping* *2 = flat* *1 = downsloping*
Ca	*Number of key vessels painted by fluoroscopy*	*1-4 value*
Thal	*Desert Category*	*2 = normal* *4 = fixed* *5 = reversible defect*

IV. ALGORITHM FOR DISEASE DETECTION

Machine learning techniques have the ability to train and test the data using the raw data. For various application scenarios and types of data, a variety of machine learning algorithms have been developed and improved. We select five algorithms to detect disease or not in this paper. The algorithms selected are classified. The selected algorithms are widely used both in many fields. We would like to investigate how well these algorithms perform with respect to our datasets.

- ✓ Logistic Regression is a classification algorithm used as predictive analysis which determines the presence of Cardiac Diseases.
- ✓ Naïve Bayes technique is used as a classifier algorithm which classifies the diseases based on the parameters in medical diagnosis.
- ✓ Fuzzy Knn algorithm is mainly used as a classifier in supervised learning.
- ✓ K Means clustering algorithm is unsupervised learning used to solve clustering problems.
- ✓ Neural Network Back propagation used to solve complex problems.

Neural Network Back propagation perceptron is used for training and testing purpose. Every iteration, 50% of the data is used for training and testing purpose from the provided dataset in each iteration. The follow-up procedure after dataset splitting is equal for the testing and training of back propagation perceptron.

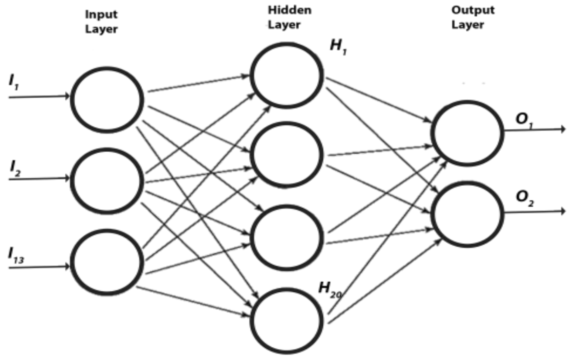

Fig 1. Neural Network back propagation perceptron

The neural network has 3 layers – an input layer, a hidden layer and an output layer. It uses forward propagation to compute, the activation (output value) of the kth output unit and θ represents the weights. The input layer comprises 13 inputs and two for the back propagation perceptron output layer. The Hidden layers are defined by the various aspects such as the Frontline Solvers and Rule of Thumb Methods.

$$N_{max} = \frac{N_s}{\alpha * (N_o + N_i)}$$

N_i - Number of input neurons.
N_o – Number of output neurons
N_s - Number of samples in training data set.
α - An arbitrary scaling factor usually 2-10.

The above equation is the proposed approach by Frontline Solvers [16]. It indicated that the Hidden layers is defined by the ratio of the Total number of data in the training dataset N_s divided by the sum of the inputs N_i and outputs N_o, multiplied by an arbitrary scaling factor α. There is also diversity approach such as rule-of-thumb [17-19] methods for determining the correct number of neurons to use in the hidden layers, such as the following:

- The number of hidden neurons should be between the size of the input layer and the size of the output layer.
- The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.
- The number of hidden neurons should be less than twice the size of the input layer

The neurons are all similar in MLP. Each has multiple input links and output links. The values obtained from the preceding layer are summarized with some weights if the individual neuron and the bias term. The amount is processed by activation function. In order words, given the outputs of X_j of the layer n, the output Y_j of the layer n+ 1 are computed as:

$$u_i = \left(\sum_j w_{i,k}{}^{n+1} * x_j\right) + w_{i,bias}{}^{n+1}$$

$$y_i = (u_i)$$

Different Activation functions may be used. Machine Learning Implements three standard functions such as Identity function, Sigmoid and Gaussian function.

Sigmoid Function : $\delta = \frac{1}{1-e^{-z}}$

In Machine Learning, all the neurons have the same activation functions, with the same free parameters, that are specified by user and are not altered by the training algorithms. Training algorithm calculates weights. The algorithm uses a training set of multiple input vectors and adapts the weights to the desired response from the network to the provided input vectors iteratively. The larger the size of the network, the more flexibility the network is. Arbitrarily small could be the mistake with the training set. But the learned network is also able to "learn" the noise in the training set, so that the error in the test set usually increases after the network size is limited. Moreover, larger networks are much more trained than smaller ones, so that only essential features are reasonable to prepare the data and train a smaller network.

This is can be optimized using the cost function. In Back Propagation, cost function to measure how close the predicted values are to their corresponding real values has follows.

$$Cost(\theta)=\frac{1}{m}+\sum_{i=1}^{m}\sum_{k=1}^{k}\left[-y_k^{(i)}ln\left(\left(h_\theta\left(x^{(i)}\right)\right)_k-\left(1-y_k^{(i)}\right)ln\left(h_\theta\left(x^{(i)}\right)\right)_k\right]+\frac{\lambda}{2m}[\sum_{i=1}(\theta_i)^2]$$

Fig 2. Initial Count Of Colours

where $\left(h_\theta\left(x^{(i)}\right)\right)$ is the activation function. The activation function can be calculated using the formula given below

$$h_\theta\left(x^{(i)}\right)=\sigma\left(\sum_i w^i I^{i-1}+b^i\right)$$

where,
w^i – *Weight matrix* for each layer i.
I^i – *Input matrix* for each layer i
b^i – *Bias vector*
The gradient helps in optimizing the weights in order to minimize the value of the cost function.

V. EVALUATION METEICS

We compare the performance of diseased samples taken by Logistic Regression, Naïve Bayes, K–means clusters, KNN and back propagation algorithms. This section examines the methods chosen to assess performance, stability and scalability.

A. PERFORMANCE

In any setup of binary classification, all the test data points are classified as one of two classes, namely the positive and the negative classes. The corresponding labels (the "correct" or "true" label) are also included in the test set. Under such a framework, we define the following terms -

1. True positives (TP) - These are the data points which actually belong to the positive class and are classified as such
2. True negatives (TN) - The are the points which belong to the negative class and are classified as such
3. False positives (FP) - These are the points which actually belong to the negative class, but have been incorrectly classified as positive
4. False negatives (FN) - These are the points which actually belong to the positive class, but have been incorrectly classified as negative

The true positives and true negatives constitute the set of correctly classified data whereas false positives and false negatives are the set of points which were incorrectly classified.
All the above metrics are expressed in terms of absolute numbers.
To have comparability across datasets of different sizes, we express them in the form of ratios :

1. True positive rate (TPR) - This is the fraction of all positive class data points which were classified correctly
2. True negative rate (TNR) - This is the fraction of all negative data points which were correctly classified
3. False positive rate (FPR) - Since false positives are essentially negative data points in our test set which got incorrectly classified, we define the FPR as the proportion of all negative data points which got incorrectly classified.
4. False negative rate (FNR) - Similar to (3), FNR is the proportion of positive data points in our test set which got incorrectly classified.

The above definition also leads to the following result. Consider the set of all positive class data points in our test set. Each of them will get classified either as "positive", or as "negative". The positive points which get classified as "positive" end up in the set of "true positives". The positive points which get incorrectly classified as "negative" end up in the pile of "false negative" points.
The performance measurements shall be used for accuracy calculation and efficiency tests on True Positive, False Negative, True Negative and False Positive algorithms. The accuracy rate is the ratio of the true positive disease rate correctly identified over the total sample quantity as shown in the equation (1). The accuracy rate is calculated by the correctly identified disease ratio, as shown in the equation (2) by the total total amount of PT and FP. The recall sample shall be computed with the accuracy with the difference between the ratio of TP and the total sum of TP and FN as indicated in equation (3).

$$Accuracy=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}\dots\dots\dots(1)$$

$$Precision\ (P)=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}\dots\dots\dots(2)$$

$$Recall\ (R)=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}\dots\dots\dots(3)$$

$$F1\ Score=\frac{2*\mathrm{PR}}{(\mathrm{P}+\mathrm{R})}\dots\dots\dots(4)$$

Five different types of machine learning algorithm (Logistic regression), Naivé Bayes, Fuzzy K nearest neighboring, K–clustering and back propagation) are used in the experimental analysis of cardiac conditions. The technique of 10-fold cross-validation is used. Table 2 provides the best performance measurements for the diseased and not as 1 and 0. Cross-validation has been done to achieve the accuracy of each algorithm implementation. The neural network shows maximum precision among all five algorithms (Fig 2).

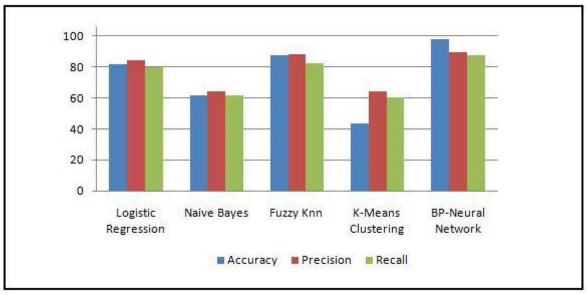

Fig 2 Comparison of Various Machine Learning Algorithms

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Table 2: Tabulated Accuracy, Precision and Recall

Algorithm	*Accuracy*	*Precision*	*Recall*
Logistic Regression	*81.86*	*84.32*	*80.006*
Naive Bayes	*61.46*	*64.21*	*61.5*
Fuzzy K-NN	*87.33*	*87.9*	*82.6*
K-Means Clustering	*43.24*	*64.01*	*60.5*
BP-Neural Network	*98.2*	*89.65*	*87.64*

VI. CONCLUSION AND FUTURE WORK

In the present life style, cardiovascular disease is a major cause of dreariness and death. This paper proposes a model to recognize cardiac diseases using a machine learning approach based on regularized logistic regression, Naïve Bayes, KNN, K-Means Clustering, Back Propagation (deep learning). The assessment compares several approaches to machine learning. The results obtained confirm that the algorithm of neural networks competes strongly against these algorithms in the detection of diseases according to the data set presented. In our model, the accuracy of our neural back propagation network is 98.2 percent, while the other algorithms are less exact than the other algorithms used.

REFERENCES

1. Kanak Saxenab, Richa Sharmac, Efficient Heart Disease Prediction System Vol :85, pp: 962 – 969, (2018)
2. Y.Jin,Advanced Fuzzy system design and applications, New york.(2003)
3. Ralph B. D’Agostino, Sr, PhD; Ramachandran S. Vasan, MD; Michael J. Pencina, PhD, Philip A. Wolf, MD; Mark Cobain, PhD,Joseph M. Massaro, PhD, William B. Kannel, “General Cardiovascular Risk Profile for Use in Primary Care”, Vol:117, pp: 743–753, (2008)
4. Chen M , Hao Y , Hwang K , Wang L , Wang L . “Disease prediction by machine learning over big data from healthcare communities”. IEEE Access Vol:88, pp: 69–79, (2017)
5. Manjusha B. Wadhonkar , Prof. P. A. Tijare and Prof. S. N. Sawalkar , Classification of Heart Disease Dataset using Multilayer Feed forward back propogation Algorithm, ISSN 2319 – 4847,Vol:4,(2013)
6. Malathi D, Logesh R., Subramaniyaswamy V , Vijayakumar , Arun Kumar Sangaiah, Hybrid Reasoning-based Privacy-Aware Disease Prediction Support System, Vol: 73 ,pp: 114–127,(2019)
7. Takanori Hondaa, Daigo Yoshidaa,b, Jun Hataa,b,c, Yoichiro Hirakawaa,b,c, Yuki Ishidaa, Mao Shibataa,b, Satoko Sakataa,c, Takanari Kitazonoc, Toshiharu Ninomiyaa,b, Development and validation of modified risk Prediction models forcardiovascular disease and its subtypes: The Hisayama Study, Vol:279, pp:38–44,(2018)
8. J. Nahar, “Computational intelligence for heart disease diagnosis: A medical knowledge driven approach,” Expert Systems with Applications, Vol:4 pp: 96-104, (2013).
9. H. Arima, K. Yonemoto, Y. Doi, T. Ninomiya, J. Hata, et al., Development and validation of a Cardiovascular risk prediction model for Japanese: the Hisayama Study, Hypertens. Res. Vol:32, pp:1119–1122, (2009)
10. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu et al., “International application of a new probability algorithm for the diagnosis of coronary artery disease,” American Journal of Cardiology, Vol. 6, pp. 304-310, (1989).
11. H.C.Yuan F.L.Xiong X.Y.Huai , A method for estimating the number of hidden neurons in feed-forward neural networks based on information entropy Vol: 40 Pp 57-64 (2003).
12. Palaniappan, S. and R. Awang, WebBased Heart Disease Decision Support System using Data Mining Classification Modeling Techniques. Proceedings of iiWAS, (2007).
13. Front Matter, Introduction to Neural Network 2nd Edition (1991).
14. Martin Riedmiller Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms Vol:16 , pp 265-278(1994)
15. L.M. Sullivan, J.M. Massaro, R.B. D'Agostino, Presentation of multivariate data for clinical use: the Framingham Study risk score functions, Stat. Med. Vol: 23 pp.1631–1660 (2004)
16. Frontile Solver Training on Artifical Neural Network Jan 17(2017).
17. K. Gnana Sheela and S. N. Deepa Review on Methods to Fix Number of Hidden Neurons in Neural Networks pp.11 , (2013)
18. Jeffrey Theodore Heaton Introduction to Neural Networks for Java Heaton Research, Inc, (2008)
19. Gaurang Panchal, Amit Ganatra , Y P Kosta and Devyani Panchal, Behaviour Analysis of Multilayer Perceptrons with Multiple Hidden Neurons and Hidden LayersVol. 3, (2011)
20. A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ 352 (2016).

AUTHORS PROFILE

NV Ravindhar received his B.E, M.E amd Ph.D degree specialized in Computer Science and Engineering. Currently he is pursuing his Ph.D degree from Saveetha University, Chennai. He is a member of various professional associations and has several years of experience. Currently he is working as Assistant Professor (OG) in Department of Computer Science and Engineering, Saveetha Engineering College, India. His research interests includes cloud comuting, comuter networks and image processing. He has published several research papers in conference and referred journals.

Dr.K.Anand received his B.E, M.E amd Ph.D degree specialized in Computer Science and Engineering. He is a member of various professional associatiosn and has 18 years of experience. Currently he is working as Professor in Department of Computer Science and Engineering, Saveetha Engineering College, India. His research interests include Information Retrieval, Web mining. He has to his credit several papers in referred journals.

Dr.S.Hariharan received his B.E degree specialized in Computer Science and Engineering from Madurai Kammaraj University, Madurai, India in 2002, M.E degree specialized in the field of Computer Science and Engineering from Anna University, Chennai, India in 2004 and Ph.D degree in the area of Information Retrieval from Anna University, Chennai, India in the year 2010. He is a member of IAENG, IACSIT, ISTE, CSTA and has 14 years of experience in teaching. Currently he is working as Professor in Department of Computer Science and Engineering, Saveetha Engineering College, India. His research interests include Information Retrieval, Data mining, Opinion Mining, Web mining. He has to his credit several papers in referred journals and conferences.He also serves as editorial board member and as program committee member for several international journals and conferences.

Ragavdendran completed his Bachelor of Engineering under the domain Computer Science and Engineering at Saveetha Engineering College affiliated to Anna University, Chennai, Tamilnadu, India. Her interest inlcudes Human Computer interaction, Data mining, Natural Language Processing and Information Retreival. He has developed models and innovative concepts using the programming Python and has won several academic prizes and credits and has also interest in developing the Android Applications that are helpful for the society.

S Godfrey Winster, Professor in Computer Science and Engineering department at Saveetha Engineering College, Chennai, India. He received his Ph.D in Information and Communication Engineering from Anna University. He has 16 years of experience in techiningand published several papers in reputed journals. His area of specialization includes data mining, big data and cloud computing.