Heart Disease Prediction System using Data Mining Techniques

[1] Dr.G.Srinaganya, [2] A. Kiruba

[1]Associate Professor, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract : Cardiovascular disease is one of the most prevalent causes of death around the world and has deemed as a vital illness in older and Middle ages. Coronary artery disease is a general cardiovascular disease involving high death rates. Angiography is more frequently than not, regarded as the best system for the examination of coronary artery disease; on the other hand, it is connected with significant side effects and high costs. Much investigation has been conveyed using data mining and machine learning to attempt alternative modalities. In this research work, a novel Thrice Filtered Information Energy based Particle Swarm optimization Feature Selection method for identifying the relevant features in the classification of heart disease. Diagnosing the existence of heart disease is really tedious process, as it entails deep knowledge and opulent experience. As a whole, the forecast of heart disease lies upon the conventional method of analysing medical report such as ECG (The Electrocardiogram), MRI (Magnetic Resonance Imaging), Blood Pressure, Stress tests by a Medicinal expert. Nowadays, a large volume of medical statistics is obtainable in medical industry and turns as a excessive source of forecasting valuable and concealed facts in almost all medical complications. Thus, these facts would really aid the doctors to create exact predictions. The innovative methods of Artificial Neural Network models have also been contributing themselves in yielding the main prediction accuracy over medical statistics. This work targets to predict the presence of heart disease utilizing back propagation MLP (Multilayer perceptron) of Artificial Neural Network with the help of MATLAB tool.

***IndexTerms* - Heart Disease, Data Mining, Feature Selection, Classification, MATLAB, Artificial Neural Network, Naïve Bayes.**

I. INTRODUCTION

According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases [1][2]. In 2008, five out of the top ten causes for mortality worldwide, other than injuries, were non-communicable diseases; this will go up to seven out of ten by the year 2030. By then, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs). Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non-communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths [3][4].

Cardiovascular disease includes coronary heart disease (CHD), cerebrovascular disease (stroke), Hypertensive heart disease, congenital heart disease, peripheral artery disease, rheumatic heart disease, inflammatory heart disease. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease [5]. Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques. This technique is routinely use in large number of industries like engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining [6][7]. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used. Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction [8][9]. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Several data mining techniques are used in the diagnosis of heart disease such as Naïve Bayes, Decision Tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies [10].

II. RELATED WORKS

Yekkala, Indu, and Sunanda Dixit [11] Heart disease or coronary artery disease (CAD) is one of the major causes of death all over the world. Comprehensive research using single data mining techniques have not resulted in an acceptable accuracy. Further research is being carried out on the effectiveness of hybridizing more than one technique for increasing accuracy in the diagnosis of heart disease. In this article, the authors worked on heart stalog dataset collected from the UCI repository, used the Random Forest algorithm and Feature Selection using rough sets to accurately predict the occurrence of heart disease.

Maji, Srabanti, and Srishti Arora [12] n this paper, hybridization technique is proposed in which decision tree and artificial neural network classifiers are hybridized for better performance of prediction of heart disease. This is done using WEKA. To validate the performance of the proposed algorithm, tenfold validation test is performed on the dataset of heart disease patients which is taken from UCI repository. The accuracy, sensitivity, and specificity of the individual classifier and hybrid technique are analyzed.

Mathan, K., et al [13] In this article, we proposed an altered calculation for classification with decision trees which furnishes precise outcomes when contrasted and others calculations. The proposed work is planned to show the data mining method in disease forecast frameworks in medicinal space by utilizing avaricious way to deal with select the best attributes. Our investigation demonstrates that among various prediction models neural networks and Gini index prediction models results with most noteworthy precision for heart attack prediction. A portion of the discretization strategies like voting technique are known to deliver more precise decision trees. To improve execution in coronary illness finding, this research work examines the outcomes in the wake of applying a scope of procedures to various sorts of decision trees and accuracy and sensitivity are attained by the execution of elective decision tree methods.

Maini, Ekta, Bondu Venkateswarlu, and Arbind Gupta [14] The biggest reason that accounts to the maximum number of death worldwide are cardiovascular disease. i.e. a huge section of people die due to Cardiovascular (CVDs) than from some other reason. According to WHO survey, nearly 80% of CVD deaths take place in underdeveloped or developing middle-income countries like India. Therefore, there is a great need to predict the disease at a premature phase to combat with this alarming situation. As tremendous quantity of data is generated by healthcare industry the data mining techniques can be efficiently explored to identify hidden patterns and interesting knowledge that may help in effective and efficient decision making. Purpose of this paper is to recommend development of a cloud based decision support system for the prediction and diagnosis of cardiovascular diseases using the methods of machine leaning. This cloud based solution will aid in making healthcare affordable in middle income groups.

Kalra, Ashima, Richa Tomar, and Udit Tomar [15] The motive of this study is to detect the risk of heart disease through a supervised learning network. The paper uses the back propagation approach of neural network. The data set here consists of 180 samples with four attributes adapted from UCI Machine Repository. We have set 70% data for training, 15% data for validation and 15% data for testing. The system gives a better accuracy as compared to the previous researches and with a good figure. It shows an accuracy of 91% for training part which is a good value for any data.

III. PROBLEM STATEMENT

The authors used three popular data mining algorithms (Naïve Bayes, RBF Network, J48) to develop the prediction models using a large dataset (683 breast cancer cases). The authors also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The genetic behaviors are analysed for prediction modeling. This existing project aims to demonstrate the working and the accuracy of a few machine learning models on a given set of data on heart disease and also, making a comparison between them to determine the best model suitable of a particular paradigm.

- Decreased classification accuracy.
- Increased error rates.
- The value of true positive, false positive and false negative are not more accurate.

IV. PROPOSED METHODOLOGY

Cardiovascular disease is one of the most prevalent causes of death around the world and has deemed as a vital illness in older and Middle ages. Coronary artery disease, in appropriate, is a general cardiovascular disease involving high death rates. Angiography is, more frequently than not, regarded as the best system for the examination of coronary artery disease; on the other hand, it was connected with significant side effects and high costs. Much investigation has, consequently, be conveyed using data mining and machine learning to attempt alternative modalities. In this paper, a novel Thrice Filtered Information Energy based Particle Swarm Feature Selection method for identifying the relevant features in the classification of heart disease.

4.1 Symmetrical Uncertainty

The symmetrical uncertainty (SU) [15] between target concept and features are applied to obtain the best features for classification. The elements with higher SU values have the higher weight. SU measures the relationship among A, B variables based on the information theory. It was calculated as follows

$$SU\ (A,)\ B = 2\ \frac{I\ (A,B)}{H(B)A + H\ (B)}$$

Computing I(A, B) as the MI among A, B. H(..) as an entropy function for A, B features. The SU shows the normalized range value [0,1] as correction factor value is 2. If SU value is 1, then the information of one feature is predictable. If SU value is 0, then A, B are not associated[16][17].

4.2 Information Gain

Entropy is generally utilized in the information theory measure, which defines the purity of an absolute collection of examples. It is in the foundation of Gain Ratio, Information Gain and Similarity Uncertainty (SU) [15]. The entropy measure is considered a measure of the system's unpredictability. The entropy of Y is

$$H(Y) = \sum_{y \in Y} p(y) \log_2(p(y)) \qquad (3.1)$$

Where *p(y)* is the marginal probability density function for the random variable *Y*. If the observed values of *Y* in the training data set *S* have partitioned according to the values of a second feature *X*, and the entropy of *Y* for the partitions induced by *X* is less than the entropy of *Y* before partitioning, then there is a relationship between features *Y* and *X*. The entropy of *Y* after observing *X* is then:

$$H(Y|X) = \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x)\ \log_2\big(p(y|x)\big) \qquad (3.2)$$

where *p(y |x)* is the conditional probability of *y* given *x*.

Given the entropy is a criterion of impurity in a training set *S*, we can define a measure reflecting additional information about *Y* provided by *X* that represents the amount by which the entropy of *Y* decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y|X) = H(X) - \ H(X|Y) \qquad (3.3)$$

IG [9] is a symmetrical measure and it is given by equation (3.3). The information gained about *Y* after observing *X* is equal to the information gained about *X* after observing *Y*. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative [16][17].

4.3 Chi-Square

Feature Selection via chi-square χ^2test [16][17] is another, very commonly used method. Chi-squared method estimates the worth of a feature by calculating the value of the chi-squared statistic with respect to the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and it is tested by chi-squared formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{O_{ij}- E_{ij}}{E_{ij}}\right)^2 \qquad (3.6)$$

Where O_{ij} is the observed frequency, and E_{ij} is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of χ^2, the greater the evidence against the hypothesis H_0.

4.4 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is based on the social behavior associated with bird's flocking for optimization problem. A social behavior pattern of organisms that live and interact within large groups is the inspiration for PSO. The PSO is easier to lay into operation than Genetic Algorithm. It is for the motivation that PSO doesn't have mutation or crossover operators and movement of particles is affected by using velocity function. In PSO, every particle alters its own flying memory and its partner's flying inclusion keeping in mind the end goal to flying in the search space with velocity [16].

4.5 Artificial Neural Network

Artificial Neural Network (ANN) is an efficient computing system whose central theme is borrowed from the analogy of biological neural networks. ANNs are also named as "artificial neural systems," or "parallel distributed processing systems," or "connectionist systems." ANN acquires a large collection of units that are interconnected in some pattern to allow communication between the units. In order to form a feed-forward multi-layer in MLP, the collection of non-linear neurons is connected to one another. This technique is known to be very useful for prediction and classification issues. Cross-validation is used to determine the 'optimal' number of hidden layers and neurons which were relied on the experimental design of the HD classification framework. These units, also referred to as nodes or neurons, are simple processors which operate in parallel [18][19][20][21][22][23].

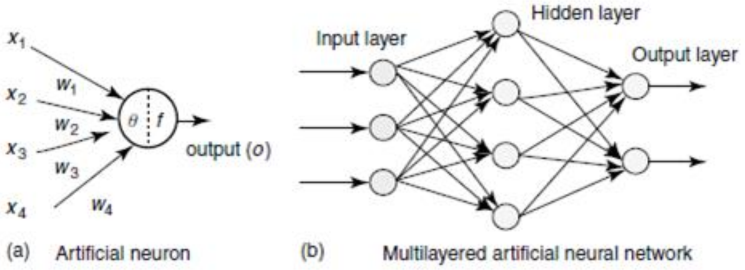

Figure 1: Architecture Diagram of Artificial Neural Network

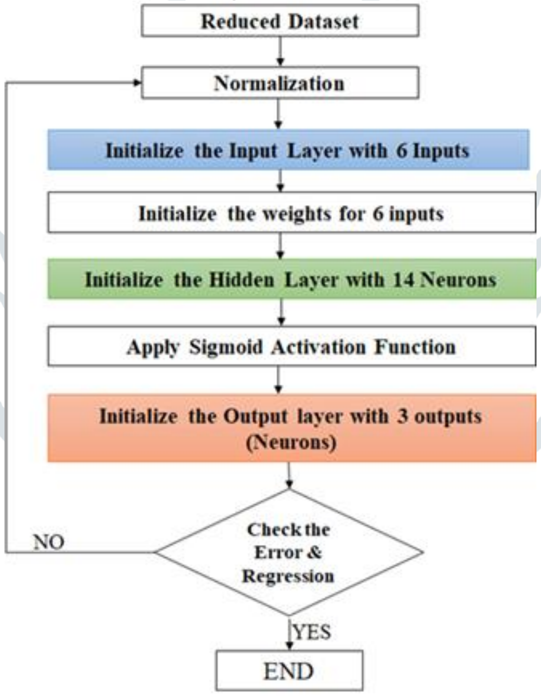
Reduced Dataset
Normalization
Initialize the Input Layer with 6 Inputs
Initialize the weights for 6 inputs
Initialize the Hidden Layer with 14 Neurons
Apply Sigmoid Activation Function
Initialize the Output layer with 3 outputs (Neurons)
Check the Error & Regression
NO
YES
END

Figure 2: Flowchart of proposed Artificial Neural Network Design

V. RESULTS AND DISCUSSION

5.1 Description of the Dataset

The dataset for this research is taken from the UCI repository. This database contains four benchmark datasets such as

- Cleveland data, data from the Cleveland Clinic Foundations.
- Hungarian data, data from the Hungarian Institute of Cardiology.
- Long-beach vs data, data from V.A. Medical Center.
- Switzerland data, data from the University Hospital, Zurich Switzerland.

The instances of all the four databases are the same. It has 76 attributes, in which only the 14 attributes are suggested and used. The 14 attributes are listed in Table 1. Researchers in medical domain most commonly uses Cleveland dataset and Statlog dataset for testing purposes. This is because all the other dataset consists of a greater number of missing values than Cleveland dataset.

Table 1: Description of the Heart Disease Dataset

Sl.NO	Feature Name	Description	Range
1	*Age*	Patient‟s Age	29-77
2	*Sex*	1=male; 0=female	0-1
3	*Cp*	Value 1:typical angina Value 2: atypical anginal Value 3: non-anginal pain Value 4: asymptotic	1-4
4	*trestbps*	Resting blood pressure(in mm Hg)	94-200
5	*Chol*	Serum cholesterol in mg/dl	126-564
6	*Fbs*	(Fasting blood sugar .120mg/dl) (1=true; 0=false)	0-1
7	*Restecg*	electrocardiography results Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST Elevation or depression of>0.05mV) Value 2: showing probable or definite left	0-2
8	*Thalach*	Maximum heart rate achieved	71-202
9	*Exang*	Exercise induced angina(1=yes;0=no)	0-1
10	*OldPeak*	ST depression induced by exercise relative to rest	0-6.2
11	*Slope*	The slope of the peak exercise ST segment Value 1: up sloping Value 2: flat Value 3:down sloping	0-2
12	*Ca*	Number of major vessels (0-3) Colored by fluoroscopy	0-3
13	*Thal*	Normal, fixed defect, reversible defect	3-7

5.2 Result and Discussion on Feature Selection Method

The following tables represented the result obtained by existing feature selection and proposed feature selection method. The features with the rank 0 have removed to improve the classification accuracy. Table 2 depicts the result obtained by implementing the Symmetrical Uncertainty feature selection method on Heart Disease dataset. In the table 2, the features like thal, thalach, restecg, fbs have removed since its rank is 0.

Table 2: Result obtained by Symmetrical Uncertainty Feature Selection Method

Average Merit	Average Rank	Attribute name
0.193 +- 0.015	1 +- 0	13 thal
0.177 +- 0.004	2.2 +- 0.4	5 chol
0.162 +- 0.015	3.3 +- 0.9	3 cp
0.148 +- 0.013	4.2 +- 1.25	9 exang
0.149 +- 0.016	4.7 +- 0.9	12 ca
0.13 +- 0.005	6 +- 0	8 thalach
0.119 +- 0.007	6.7 +- 0.9	11 slope
0.108 +- 0.005	7.9 +- 0.3	10 oldpeak
0.077 +- 0.01	9.5 +- 0.81	2 Sex
0.072+-0.004	9.7+-0.46	1 age
0.068+-0.006	10.8 +- 0.4	4 trestbps
0.035 +- 0.005	12 +- 0	7 restecg
0.004 +- 0.001	13 +- 0	6 fbs

Table 3 gives the result obtained by using Information Gain feature selection method. The features chol, thalach, old peak, sex, restecg, fbs with rank 0 are removed to improve the prediction accuracy of the heart disease.

Table 3: Result obtained by Information Gain Feature Selection Method

Average Merit	Average Rank	Attribute name
0.697 +- 0.02	1 +- 0	5 chol
0.464 +- 0.019	2 +- 0	8 thalach
0.278 +- 0.014	3 +- 0	10 oldpeak
0.226 +- 0.02	4.9 +- 1.04	3 cp
0.218 +- 0.013	5 +- 0.77	1 age
0.216 +- 0.017	5.5 +- 1.12	13 thal
. 0.195 +- 0.022	7.1 +- 0.83	12 ca
0.19 +- 0.017	7.5 +- 0.67	4 trestbps
0.147 +- 0.014	9.3 +- 0.46	9 exang
0.141 +- 0.008	9.7 +- 0.46	11 slope
0.075 +- 0.01	11 +- 0	2 sex
0.036 +- 0.005	12 +- 0	7 restecg
0.003 +- 0.001	13 +- 0	6 fbs

Table 4 presents the result obtained from Chi-Square Feature selection method. The features with rank 0, i.e., chol, thalach, sex, restecg, fbs have removed.

Table 4: Result obtained by Chi-Square Feature Selection Method

Average Merit	Average Rank	Attribute name
459.807 +-84.083	1 +- 0	5 chol
284.386 +-55.823	2 +- 0	8 thalach
89.128 +-10.471	3.4 +- 0.49	1 age
86.108 +- 7.402	3.6 +- 0.49	10 oldpeak
66.103 +- 7.559	5.9 +- 0.94	4 trestbps
65.876 +- 5.532	6.4 +- 1.28	3 cp
65.249 +- 7.466	6.4 +- 0.92	12 ca
63.823 +- 4.639	7.3 +- 0.78	13 thal
44.131 +- 3.899	9.2 +- 0.4	9 exang
41.778 +- 2.366	9.8 +- 0.4	11 slope
22.094 +- 2.801	11 +- 0	2 sex
10.235 +- 1.501	12 +- 0	7 restecg
0.692 +- 0.278	13 +- 0	6 fbs

Table 5 gives the result obtained by proposed research method. This method also removes the features with rank 0. The features like sex, restecg, thalach, age, fbs, trestbps have removed for obtaining the optimal dataset.

Table 5: Result obtained by proposed Thrice Filtered Information Energy based Particle Swarm Feature Selection (research) method

Average Merit	Average Rank	Attribute name
0.025 +- 0.007	**9.5 +- 1.36**	**3 cp**
0.014 +- 0.009	**11.5 +- 1.43**	**12 ca**
0.126 +- 0.017	**3.3 +- 0.46**	**13 thal**
0.107 +- 0.014	4.1 +- 0	2 sex
0.101 +- 0.012	**5.1 +- 1.04**	**11 slope**
0.082 +- 0.011	**5.8 +- 0.6**	**9 exang**
0.068 +- 0.011	6.7 +- 0	7 restecg
0.025 +- 0.007	8.8 +- 0	8 thalach
0.19 +- 0.022	**1.1 +- 0.3**	10 oldpeak
0.025 +- 0.008	9.6 +-0	1 age
0.015 +- 0.006	11.2 +- 0	6 fbs
0.171 +- 0.016	1.9 +- 0	4 trestbps
0.01 +- 0.003	**12.4 +- 0.66**	**5 chol**

Table 6 represents the number of features that are selected by Symmetrical Uncertainty, Information Gain, Chi-Square and Proposed Thrice Filtered Information Energy based Particle Swarm Feature Selection methods. From the table 6, the proposed research method gives the smaller number of features than the existing feature selection methods.

Table 6: Number of Features selected by SU, IG, CS and proposed research method

Feature Index	***Number of Features selected by existing feature selection methods and proposed research method***			
	SU	**IG**	**CS**	**research**
1	chol	cp	age	cp
2	cp	age	old peak	ca
3	exang	thal	trestbps	thal
4	ca	ca	cp	slope
5	slope	trestbps	ca	exang
6	old peak	exang	thal	chol
7	sex	slope	exang	
8	age		slope	
9	trestbps			

Figure 4 depicts the graphical representation of the number of features obtained by existing feature selection methods like SU, IG, CS and proposed research method. From the figure 4, it is clear that the proposed research method gives less number of features than the existing methods like SU, IG, and CS.

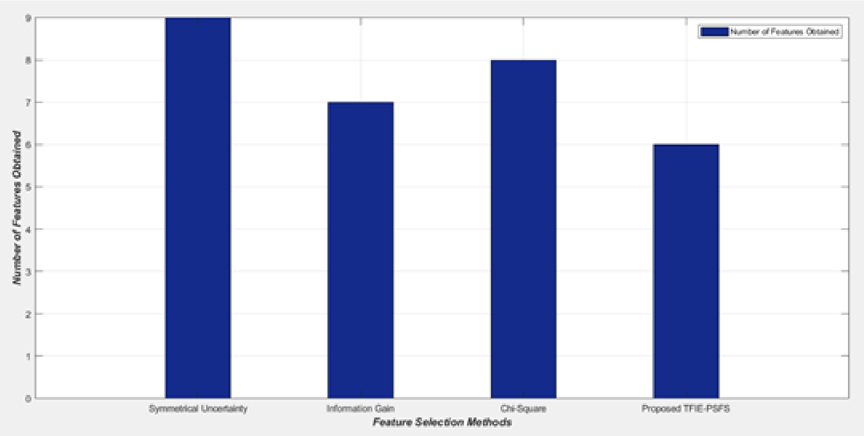

Figure 4: Number of features obtained by SU, IG, CS and Proposed research method

5.3 Result and Discussion on Classification Method

Figure 5 depicts the graphical representation of the performance analysis on the Classification accuracy (in %) of the original dataset and feature selection methods like Symmetrical Uncertainty, Information Gain, and Chi-Square analysis by using SVM, NB and ANN Classification. From the figure 5, it is clear that the classification accuracy has improved for proposed research method when using ANN classification technique.

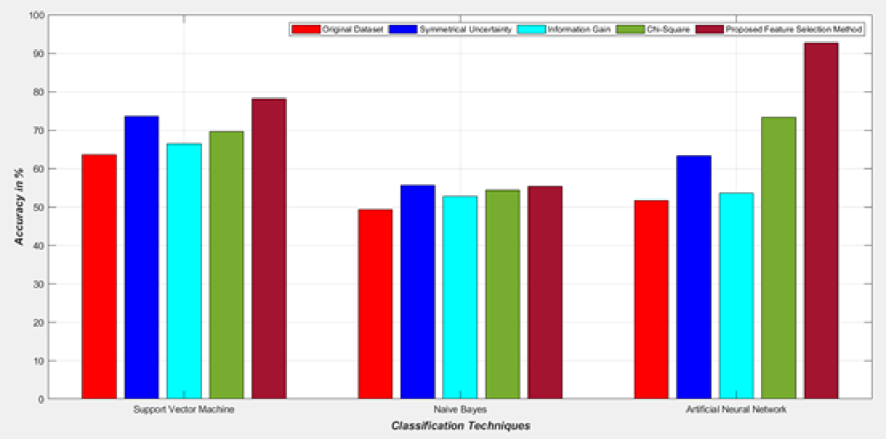

Figure 5: Classification Accuracy using SVM, NB and ANN

Figure 6 depicts the graphical representation of the Mean Absolute Error for the original dataset, existing feature selection techniques like Symmetrical Uncertainty, Information Gain, Chi-Square and proposed research method using ANN, NB and SVM classification techniques. From the figure 6, it is clear that the proposed research gives least error rate when using SVM as the classifier when it is compared with the proposed research method using ANN.

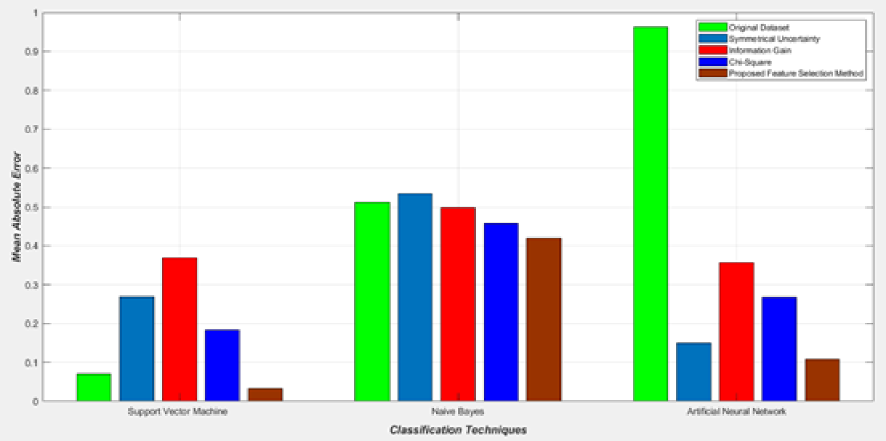

Figure 6: Mean Absolute Error using SVM, NB and ANN Classification

Figure 7 depicts the graphical representation of the Root Mean Squared Error for the original dataset, existing feature selection techniques like Symmetrical Uncertainty, Information Gain, Chi-Square and proposed research method using ANN, NB and SVM classification techniques. From the figure 7, it is clear that the proposed research gives least RMS error rate when using SVM as the classifier when it is compared with the proposed research method using ANN.

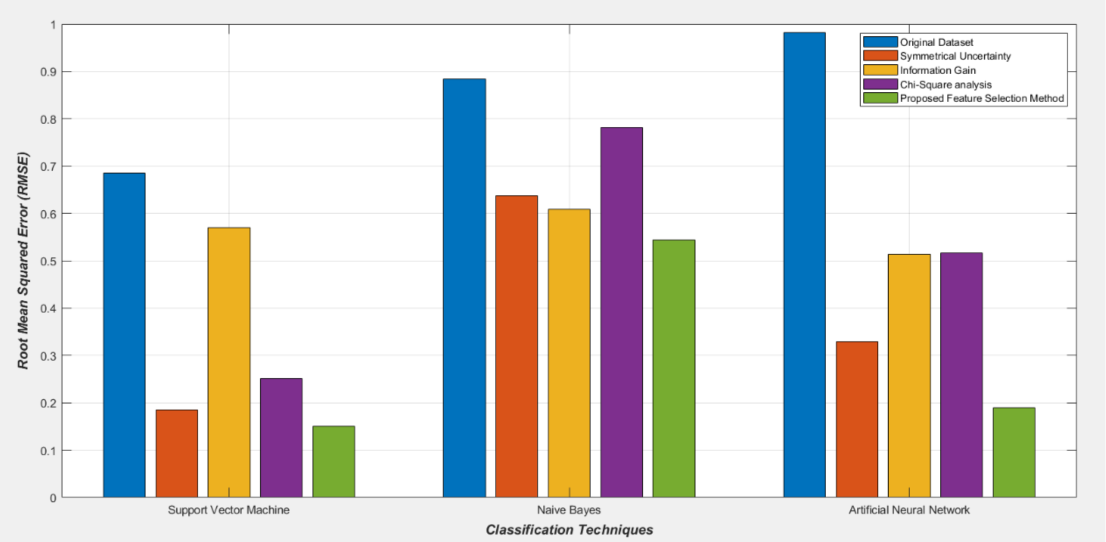

Figure 7: Root Mean Squared Error using SVM, NB and ANN Classification

Figure 8 depicts the graphical representation of the Relative Absolute Error (RAE) for the original dataset, existing feature selection techniques like Symmetrical Uncertainty, Information Gain, Chi-Square and proposed research method using ANN, NB and SVM classification techniques. From the figure 8, it is clear that the proposed research gives least RAE error rate when using ANN as the classifier when it is compared with the proposed research method using NB and SVM.

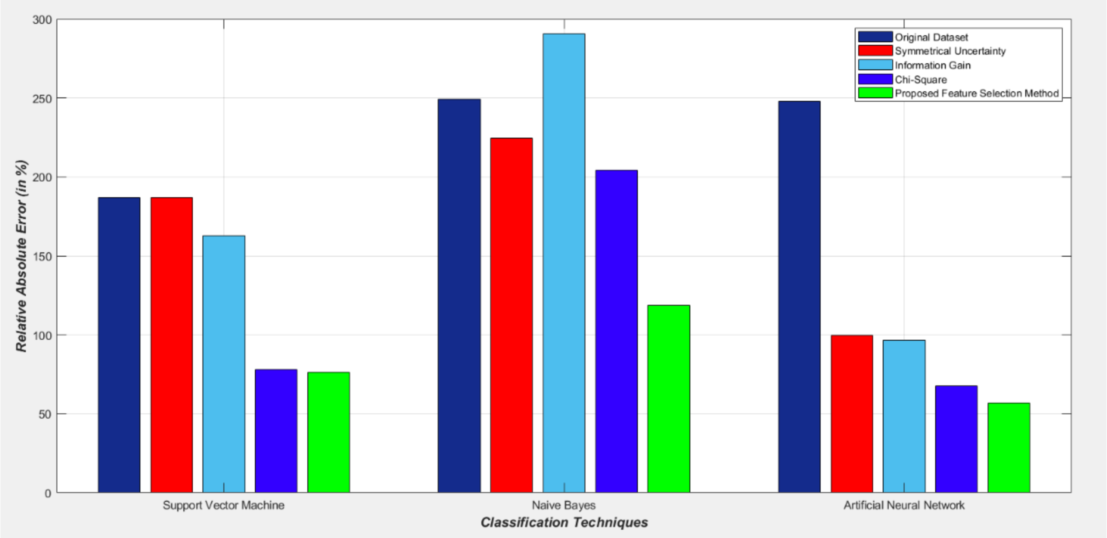

Figure 8: Relative Absolute Error using SVM, NB and ANN Classification

Figure 9 depicts the graphical representation of the Root Relative Squared Error (RRSE) for the original dataset, existing feature selection techniques like Symmetrical Uncertainty, Information Gain, Chi-Square and proposed research method using ANN, NB and SVM classification techniques. From the figure 9, it is clear that the proposed research gives least RRSE error rate when using ANN, NB as the classifier when it is compared with the proposed research method using SVM and NB.

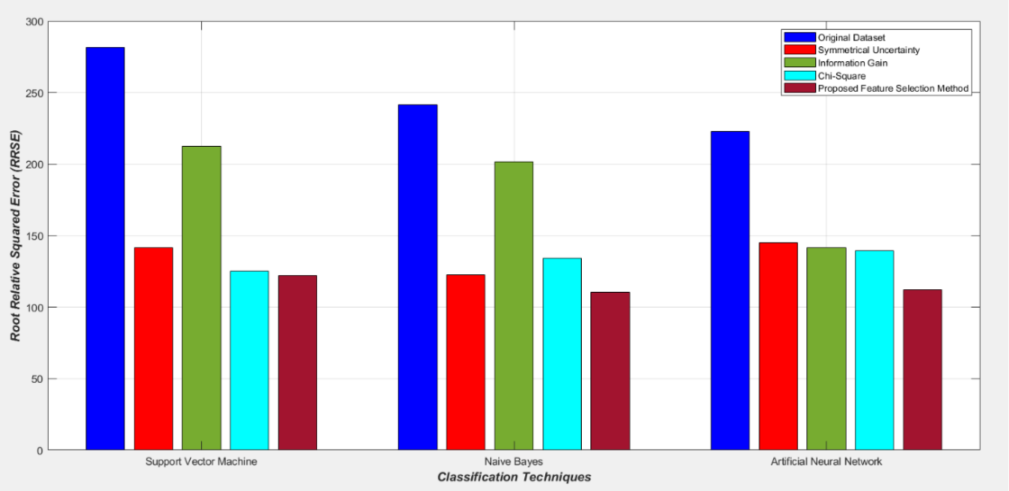

Figure 9: Root Relative Squared Error using SVM, NB and ANN Classification

Figure 10 represents the performance analysis on the True Positive Rate (TPR) of the original dataset, and feature selection methods like Symmetrical Uncertainty, Information Gain and Chi-Square analysis by using SVM, NB and ANN classification techniques. From the figure 10, it is clear that the proposed research method gives high TPR when using ANN as the classifier when it is compared with the proposed research method using NB and SVM.

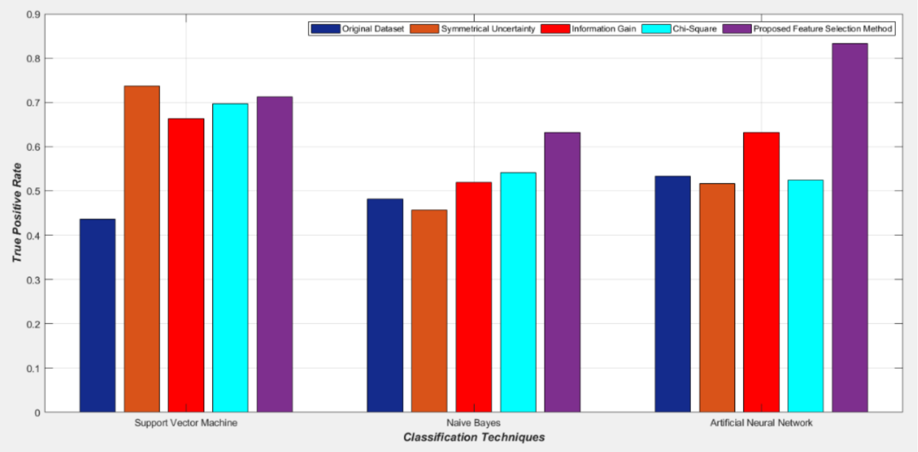

Figure 10: True Positive Rate using SVM, NB and ANN Classification

Figure 11 depicts the graphical representation of the performance analysis on the False Positive Rate (FPR) of the original dataset, and feature selection methods like SU, IG and proposed research method by using SVM, NB and ANN Classification. From the figure 11, it is clear that the proposed research method gives least FPR with ANN as the classification, than the other methods using ANN, NB and SVM.

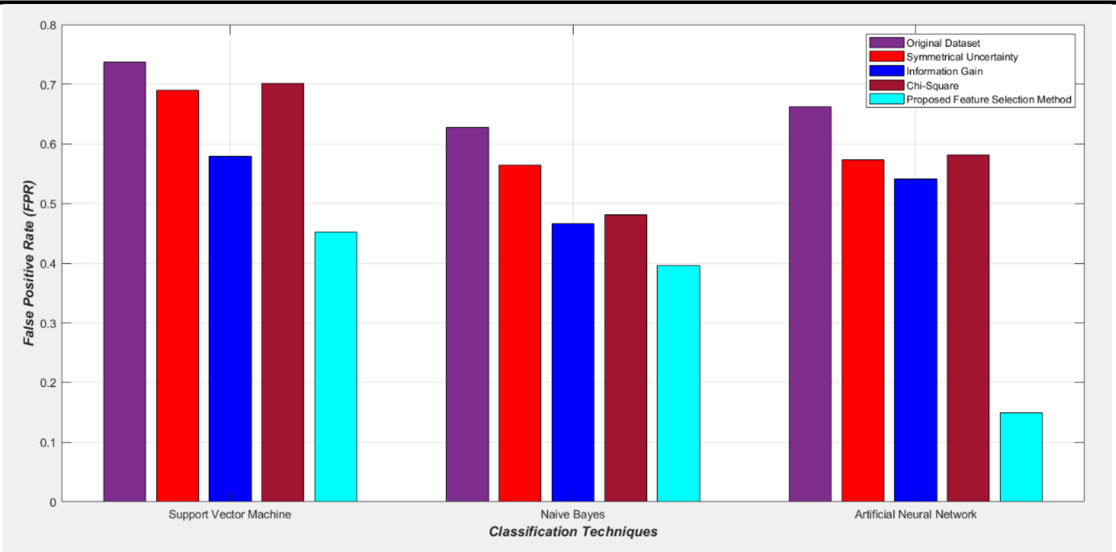

Figure 11: False Positive Rate using SVM, NB and ANN Classification

Figure 12 depicts the graphical representation of the performance analysis on the Precision of the original dataset, and feature selection methods like SU, IG, CS and proposed research method by using SVM, NB and ANN Classification. From the figure 12, it is clear that the proposed research method gives maximum Precision with ANN as the classification, than the other methods using ANN, NB and SVM.

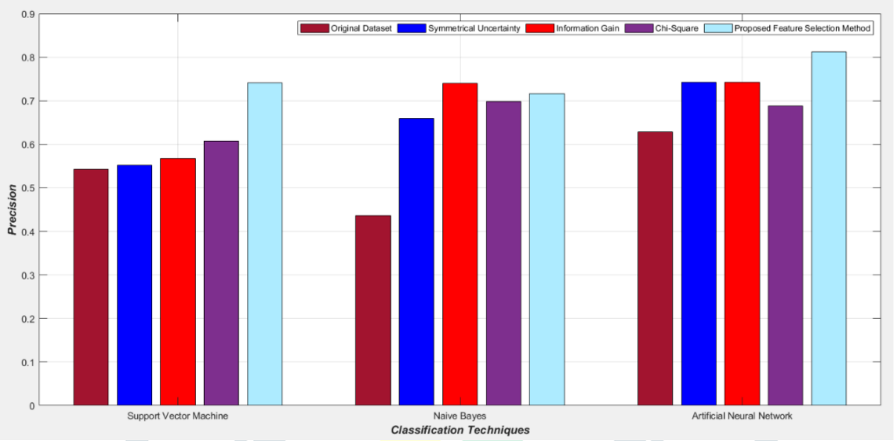

Figure 12: Precision using ANN, NB and SVM classification

Figure 13 depicts the graphical representation of the performance analysis on the True Negative Rate (Specificity) of the original dataset, and feature selection methods like SU, IG, CS and proposed research method by using SVM, NB and ANN Classification. From the figure 13, it is clear that the proposed research method gives maximum True Negative Rate (Specificity) with ANN as the classification, then the other methods using ANN, NB and SVM.

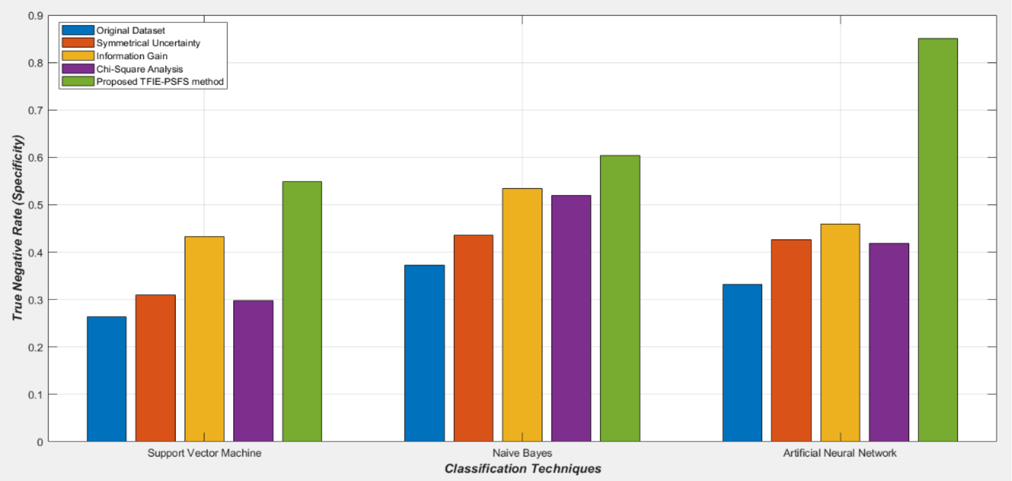

Figure 13: True Negative Rate (Specificity) using ANN, NB and SVM classification

Figure 14 depicts the graphical representation of the performance analysis on the False Negative Rate (Miss Rate) of the original dataset, and feature selection methods like SU, IG, CS and proposed research method by using SVM, NB and ANN Classification. From the figure 14, it is clear that the proposed research method gives maximum False Negative Rate (Miss Rate) with ANN as the classification, then the other methods using ANN, NB and SVM.

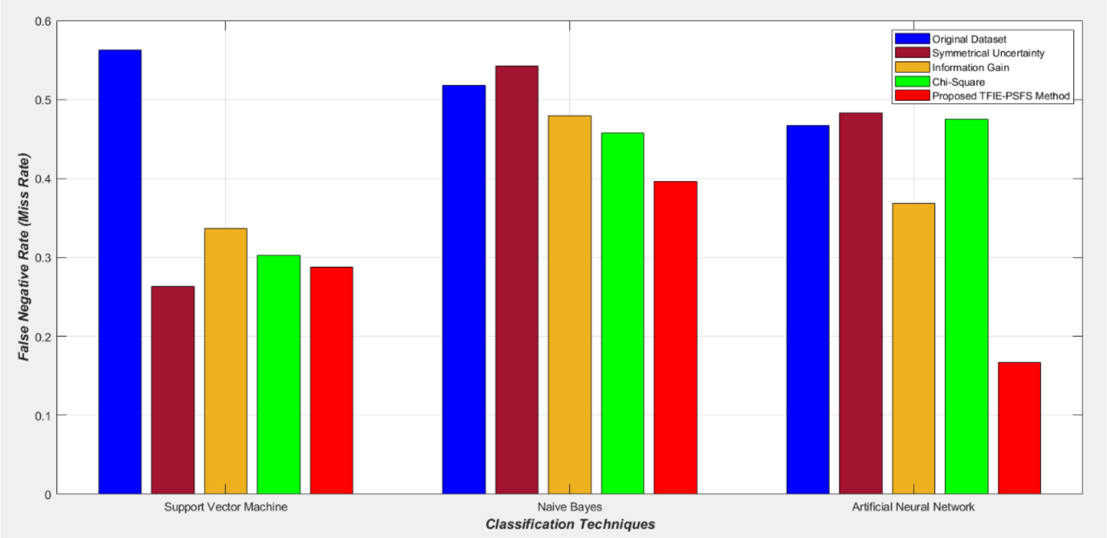

Figure 14: False Negative Rate (Miss Rate) using SVM, NB and ANN

Figure 15 depicts the graphical representation of the performance analysis on the F-Measure of the original dataset, and feature selection methods like SU, IG, CS and proposed research method by using SVM, NB and ANN Classification. From the figure 15, it is clear that the proposed research method gives maximum F-Measure with ANN as the classification, then the other methods using ANN, NB and SVM.

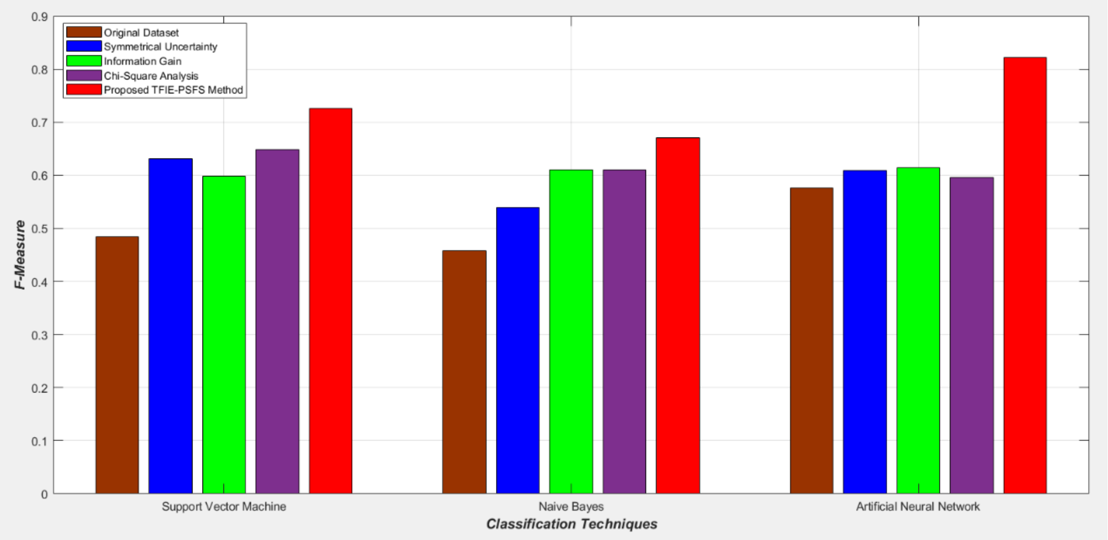

Figure 15: F-Measure using SVM, NB and ANN

VI. CONCLUSION

Data mining techniques clearly outperformed CVD risk analysis and it is suggested that the tool is accurate and can result whether to consult the physician or not. This research work has derived into three contributions. The research work is, a new features selection method has proposed in the pre-processing stage to reduce the number of features. SU method gives nine features, IG method gives seven features, CS method gives eight features and the proposed research method gives only six features among 13 features of heart disease data set.

In this research work, ANN has utilized the reduced dataset for the classification of heart disease into three categories No Class (not having symptoms for heart disease) YES Class (having symptoms for heart disease) Hesitant Class (Symptoms target that they may or may not have heart disease). The proposed ANN design for classification gives More accuracy, less error rates for metrics such as RAE, RRSE and False positive rate Increased the TPR and precision. In this method when compared with other classification methods most of the metrics are relatively good.

Since in this work, the bench mark dataset collected from the UCI repository is used to validate the proposed model. The future direction of this research work is to collect the real time dataset in Indian environment to further validate the proposed model. The proposed algorithms can be applied for other fields for the effective results. This proposed model can be implemented in hospitals and based on the feedbacks and results obtained, the proposed model can be improved further as a part of standardizing the tool.

REFERENCES

[1] Nahar, Jesmin, et al, "*Association rule mining to detect factors which contribute to heart disease in males and females*", Expert Systems with Applications, Vol. 40 Issue. 4, pp. 1086-1093, 2013.
[2] Vijiyarani, S., and S. Sudha, "*An efficient classification tree technique for heart disease prediction*", International Conference on Research Trends in Computer Technologies (ICRTCT-2013) Proceedings published in International Journal of Computer Applications (IJCA)(0975–8887). Vol. 201, 2013.
[3] Gayathri, P., and N. Jaisankar, "*Comprehensive study of heart disease diagnosis using data mining and soft computing techniques*", 2013.

[4] Shouman, Mai, Tim Turner, and Rob Stocker, "*Integrating clustering with different data mining techniques in the diagnosis of heart disease*", J. Comput. Sci. Eng, Vol. 20 Issue.1, 2013.

[5] Amato, Filippo, et al, "*Artificial neural networks in medical diagnosis*", pp. 47-58, 2013.

[6] Persi Pamela, I., and P. Gayathri, "*A fuzzy optimization technique for the prediction of coronary heart disease using decision tree*", 2013.

[7] Chaurasia, Vikas, and Saurabh Pal, "*Data mining approach to detect heart diseases*", 2014.

[8] Thenmozhi, K., and P. Deepika, "*Heart disease prediction using classification with different decision tree techniques*", International Journal of Engineering Research and General Science, Vol. 2, Issue. 6, pp. 6-11, 2014.

[9] Kim, Jae-Kwon, et al, "*Adaptive mining prediction model for content recommendation to coronary heart disease patients*", Cluster computing, Vol. 17, Issue. 3, pp. 881-891, 2014.

[10] Seera, Manjeevan, and Chee Peng Lim, "*A hybrid intelligent system for medical data classification*", Expert Systems with Applications, Vol. 41, Issue. 5, pp. 2239-2249, 2014.

[11] Yekkala, Indu, and Sunanda Dixit. "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection." *International Journal of Big Data and Analytics in Healthcare (IJBDAH)* 3.1 (2018): 1-12

[12] Maji, Srabanti, and Srishti Arora. "Decision Tree Algorithms for Prediction of Heart Disease." *Information and Communication Technology for Competitive Strategies*. Springer, Singapore, 2019. 447-454.

[13] Mathan, K., et al. "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease." *Design Automation for Embedded Systems* 22.3 (2018): 225-242.

[14] Maini, Ekta, Bondu Venkateswarlu, and Arbind Gupta. "Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System." *International Conference on Intelligent Data Communication Technologies and Internet of Things*. Springer, Cham, 2018.

[15] Kalra, Ashima, Richa Tomar, and Udit Tomar. "Heart Risk Prediction System Based on Supervised ANN." *Proceedings of International Conference on Recent Advancement on Computer and Communication*. Springer, Singapore, 2018.

[16] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems*, Preprint: 1-18.

[17] M. Durairaj, T S Poornappriya, "Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.", *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.

[18] M. Durairaj, T. S. Poornappriya, "Importance of MapReduce for Big Data Applications: A Survey", *Asian Journal of Computer Science and Technology*, Vol.7 No.1, 2018, pp. 112-118.

[19] M. Lalli, V.Palanisamy,(2016), "Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks", International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572

[20] M. Lalli, V.Palanisamy,(2017), "Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques", Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).

[21] M. Lalli, V.Palanisamy, (Sep 2014), "A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN", International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.

[22] M. Lalli, "Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET", *Journal of Applied Science and Computations*, Volume VI, Issue VI, JUNE/2019, 1795-1813.

[23] M. Lalli, "Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network", *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.