IJCTA, 9(27), 2016, pp. 175-183
© International Science Press ISSN: 0974-5572

A Novel Hybrid Framework for the Prediction of Heart Disease

K. Sudhakar* and M. Manimekalai**

ABSTRACT

Numerous doctor's facility data frameworks are intended to bolster understanding charging, stock administration and era of straightforward measurements. A few hospital use choice emotionally supportive networks, yet they are to a great extent constrained. The significance of this research work is to dismember diverse systems in data mining to build up a Hybrid model for researching the seriousness of the Cardio Vascular Disease. The core ideas incorporates like PCA, Information gain, Naïve Bayes and Artificial Neural Network and Fuzzy Logic are joined together to propose another structure to anticipate the seriousness of the disease. With the end goal of the feature selection, Hybrid Feature Selection calculation is proposed. Naives Bayes is utilized to order the ideal set and Artificial Neural Network is prepared by Back Propagation Algorithm to arrange the ideal set. Fuzzy Logic is utilized to anticipate the coronary disease seriousness among the general population indications with coronary disease. In this paper, fuzzy expert system is utilized for creating knowledge based frameworks in solution. The proposed framework utilizes Mamdani interference strategy. The entire hybrid structure is then compared with existing work which utilizes Genetic Algorithm.

Keywords: Heart Disease, Artificial Neural Network, Naïve Bayes, Fuzzy Expert System, Information Gain. Mamdami Rules.

1. INTRODUCTION

Medicinal services industry today produces a lot of complex information about patients, hospital assets, sickness finding, electronic patient records, and therapeutic gadgets and so on. Bigger measures of information are a key asset to be handled and broke down for knowledge extraction that empowers support for cost-savings and decision making. Data mining applications in social insurance can be gathered as the assessment into general categories [1] [2].

The Healthcare business is for the most part "information rich", which is not achievable to handle physically. These a lot of information are essential in the field of Data Mining to extricate helpful data and produce connections amongst the properties. The doctors and specialists accessible are not in extent with the populace. Additionally, manifestations frequently are ignored. Coronary disease conclusion is an intricate undertaking which requires much experience and information. Coronary disease is a solitary biggest reason for death in created nations and one of the primary benefactors to ailment load in creating nations. In the social insurance industry the information digging is primarily utilized for predicting the diseases from the datasets.

Cardiovascular Disease CVD) has turned into the essential executioner worldwide and is relied upon to bring about more death later on. Prevention and prediction of CVD have along these lines get to be imperative social issues. Numerous gatherings have created expectation models for asymptomatic CVD by ordering its danger in view of built up danger variables (e.g., age, sex, and so on.).

2. RELATED WORKS

WHO, (2011) reported Cardiovascular Diseases (CVDs) are the main source of death all inclusive: a bigger number of individuals dies every year from CVDs than from some other cause. An expected 17.1 million

* Research Scholar, Department of Computer Science

** Director & Head, Department of Computer Application , Shrimati Indira Gandhi College, Trichy, Tamilnadu, India

individuals die from CVDs in 2004, speaking to 29% of every single worldwide death, of these deaths, an expected 7.2 million were because of coronary disease which is a standout amongst the most well-known sorts of coronary disease and 5.7 million were because of stroke. Low-and center salary nations are disproportionally influenced, 82% of CVD deaths happen in low-and center wage nations and happen similarly in men and ladies. By 2030, just about 23.6 million individuals will bite the dust from CVDs, for the most part from coronary disease and stroke. These are anticipated to remain the single driving reasons for death. The biggest rate increment will happen in the Eastern Mediterranean Region. The biggest increment in number of deaths will happen in the South-East Asia Region because of progress in way of life, work society and nourishment propensities. Subsequently, more careful and effective techniques for heart diseases and occasional examination are of high significance.

Various studies have been done that have concentrated on determination of coronary disease. They have connected diverse information digging strategies for conclusion and accomplished distinctive probabilities for various techniques. An Intelligent Heart Disease Prediction System (IHDPS) is produced by utilizing data mining methods Naive Bayes, Neural Network, and Decision Trees was proposed by SellappanPalaniappan et al .[3]. Every strategy has its own particular quality to get suitable results. To build up the multi-parametric feature with direct and nonlinear attributes of HRV (Heart Rate Variability) a novel procedure was proposed by HeonGyu Lee et al. [4]. The expectation of Heart ailment, Blood Pressure and Sugar with the guide of neural systems was proposed by Niti Guru et al. [5]. The dataset contains records with 13 attributes in every record. The issue of recognizing compelled association rules for coronary disease expectation was considered via Carlos Ordonez [6]. The resultant dataset contains records of patients having coronary disease. Three imperatives were acquainted with decline the quantity of examples [7]. Franck Le Duff et al. [8] constructs a decision tree with database of patient for a medicinal issue. Latha Parthiban et al. [9] anticipated a methodology on premise of coactive neuro-fuzzy inference

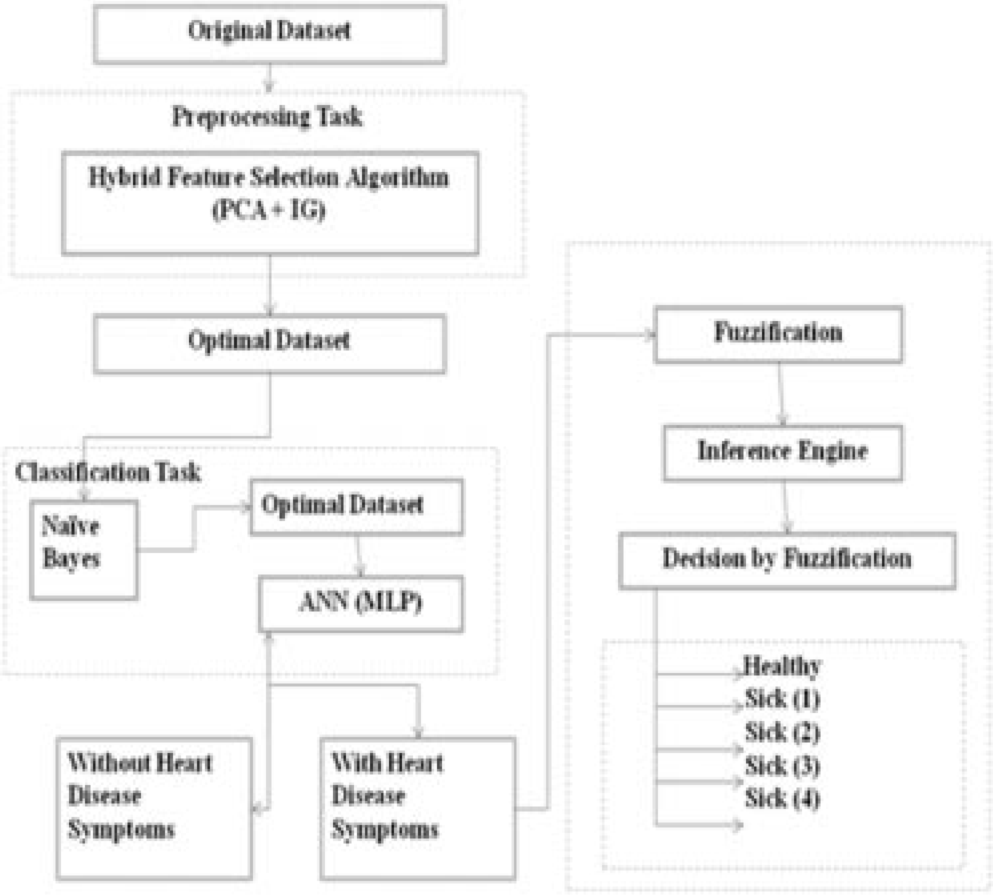

Figure 1: The Proposed Hybrid Framework for the Risk Severity Prediction of Heart Disease

framework (CANFIS) for prediction of coronary disease. The CANFIS model uses neural system capacities with the fuzzy logic and genetic algorithm. Kiyong Noh et al. [10] utilizes a clustering technique for the extraction of multi parametric attributes by surveying HRV (Heart Rate Variability) from ECG, information pre-preparing and coronary disease design.

3. PROPOSED HYBRID FRAMEWORK FOR PREDICTION OF HEART DISEASE

In the pre-processing step, the proposed framework combines the concepts of PCA and Information Gain feature selection techniques. This hybrid technique is named as “Hybrid Feature Reduction Method”. The optimal dataset is obtained by developing the algorithm. Further the optimal dataset is reduced by using Naïve Bayes Classifier and Artificial Neural Network is used to to classify the patient without heart disease symptoms and with symptoms. The Neural Network is trained by the learning algorithm called Multi-Layered Percepton with Back Propagation algorithm. And the severity of the heart disease is predicted by using Fuzzy Logic. 63 Rules are generated by this Fuzzy Inference Engine. The Figure 1 depicts the Proposed framework for the heart disease prediction system, which includes the Hybrid Feature Selection algorithm, Naïve Bayes +ANN and Fuzzy Logic.

4. DATASET DESCRIPTION

The data set for this research is taken from the UCI repository [11]. This database contains four bench mark data sets such as Switzerland data, Long-beach-va data, Hungarian data and Cleveland data. This database contains 76 attributes, from this only 14 attributes are used after neglecting redundant and irrelevant attributes. Table 1 represents the list of 14 attributes. In particular, the Cleveland database is the only one that has been used by researchers to this date. This is because all the other data set contains more number of missing values than Cleveland data set [12].

5. IMPLEMENTATION RESULT AND DISCUSSIONS

In this section, the proposed method is compared with existing work to justify the results obtained. The attributes reduced by applying the proposed Hybrid Feature Selection method and methods like PCA and IG. And it is listed in Table 2.

Table 1
Dataset Attributes Information

No	*Name*	*Description*
1	Age	Age in Years
2	Sex	1 = male, 0 = female
3	cp	chest pain type(1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4	Restsbp	Resting blood sugar(in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl(1 = true, 0 = false)
7	restecg	Resting electrocardiographic results(0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy)
8	maxch	Maximum heart rate
9	Exang	Exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1=upsloping, 2=flat, 3= downsloping)
12	Ca	Number of major vessels colored by fluoroscopy
13	Thal	3 = normal, 6 = fixed defect, 7 = reversible defect
14	num	Class (1- With Symptoms and 0 without symptoms)

Table 2
List of features obtained from Hybrid Feature Selection Algorithm

PCA	*IG*	*Hybrid Algorithm*
Cholestrol	Chest Pain	Chest Pain
Maximum HR	Rest ECG	Exer Ind
ST by exercise	Exer Ind	Major Vessels colored
Thal	Slope Peak by ST	Thal
Chest Pain	Major Vessels Colored	
Major Vessels Colored	Thal	
Rest SBP		
Exer Ind		

Table 3
Attribute Reduction using Hybrid Feature Selection Algorithm and Genetic algorithm

Genetic Algorithm	*Hybrid Algorithm*
Chest Pain	Chest Pain
Rest ECG	Exer Ind
Fasting Blood Sugar	Major Vessels colored
Slope Peak by ST	Thal
Major Vessels Colored	

Table 4
Comparison of MSE, Regression and %E Values of Training Instances with Proposed method and Genetic algorithm

	Training Instances (213 Attributes)	
Error Type	*Proposed Hybrid Feature Selection Algorithm*	*Genetic Algorithm*
Mean Square Error (MSE)	1.27221e-1	2.35449e-1
Regression (R)	7.05969e-1	3.00049e-1
%E	21.11161e-0	22.05752e-0

Table 5
Comparison of MSE, Regression and %E Values of Validation Instances with Proposed method and Genetic algorithm

	Validation Instances (45 Attributes)	
Error Type	*Proposed Hybrid Feature Selection Algorithm*	*Genetic Algorithm*
Mean Square Error (MSE)	1.18958e-1	2.32248e-1
Regression (R)	7.14526e-1	2.66219e-1
%E	2.22222e-0	20.00000e-0

Table 6
Comparison of MSE, Regression and %E Values of Testing Instances with Proposed method and Genetic algorithm

	Testing Instances (45 Attributes)	
Error Type	*Proposed Hybrid Feature Selection Algorithm*	*Genetic Algorithm*
Mean Square Error (MSE)	1.67671e-1	1.89697e-1
Regression (R)	5.86872e-1	4.68699e-1
%E	11.11111e-0	28.88888e-0

Table 7
Performance Evaluation Comparison of Proposed method and Genetic algorithm

Performance Metrics	*Genetic Algorithm*	*Proposed Hybrid Feature Selection Algorithm*
Accuracy	72.98	81.95
RAE	56.56	42.01
RMSE	0.51	0.34
MAE	0.15	0.2
RRSE	74.39	101.23
Precision	0.81	0.74
Recall	0.81	0.74
True Positive Rate	0.81	0.74
False Positive Rate	0.19	0.26
Kappa Statistic	0.64	0.45
F-Measure	0.81	0.74
ROC	0.81	0.74

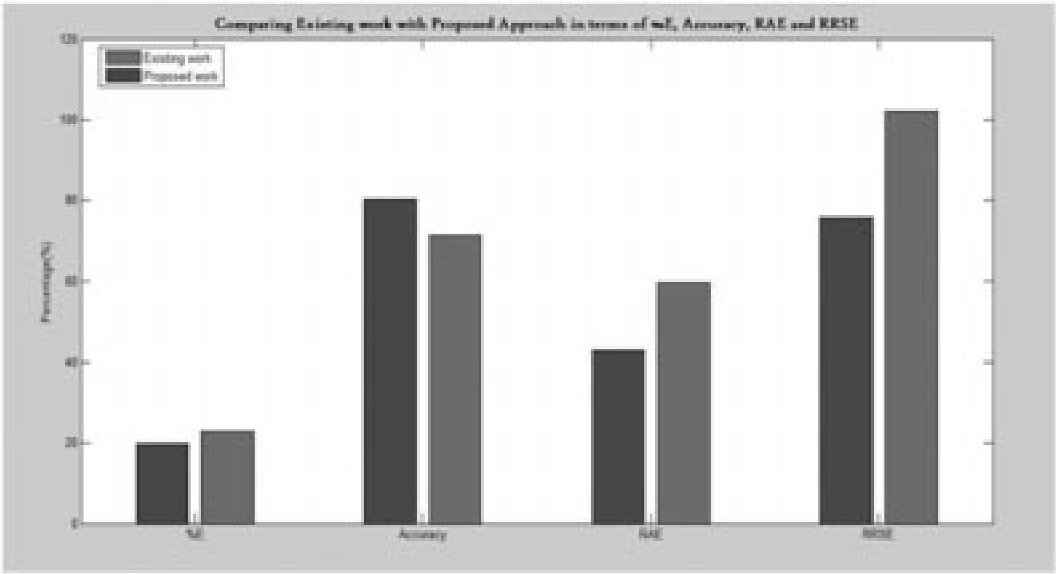

Figure 2: Performance Comparison of proposed Hybrid Feature Selection with Existing Genetic Algorithm

The MSE Values, Regression Values and portion of test which are misclassified (%E) values, got through Simulation are recorded in Table 4, table 5, table 6 and Table 7. The %E worth is zero implies that it has no misclassification and 100 means it has most extreme misclassification. While measuring the MSE, the lower qualities are better. The Regression (R) value measures the relationship between the target and the output. The value of regression is 1 means close relationship. Lower qualities ought to be evaded while measuring R. Alternate measurements like TPR (True Positive Rate), FPR (False Positive Rate), Recall, Precision, ROC value, F-Measure, RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) Values are additionally measured and the achieved results are organized.

The Figure 2 demonstrates the examination of the execution of Existing GA and the proposed Hybrid Feature Selection Algorithm. It is unmistakable from the measurements that are taken for estimation that the proposed Hybrid Feature Selection Algorithm performs well in lessening the attributes without impacting the knowledge gained from it.

Figures 3 and 4 show the regression plot for the Genetic Algorithm and Proposed Hybrid Feature Selection algorithm. It can be seen that the R worth is 0.3 in Genetic Algorithm where as it is 0.7 in the proposed Hybrid Feature Selection algorithm. The worth 0.7 demonstrates the close relationship between the target and output though 0.3 shows there is no plausibility of relationship between the target and output. Furthermore the fitness value lies nearer in the hybrid Feature Selection calculation then the Genetic Algorithm.

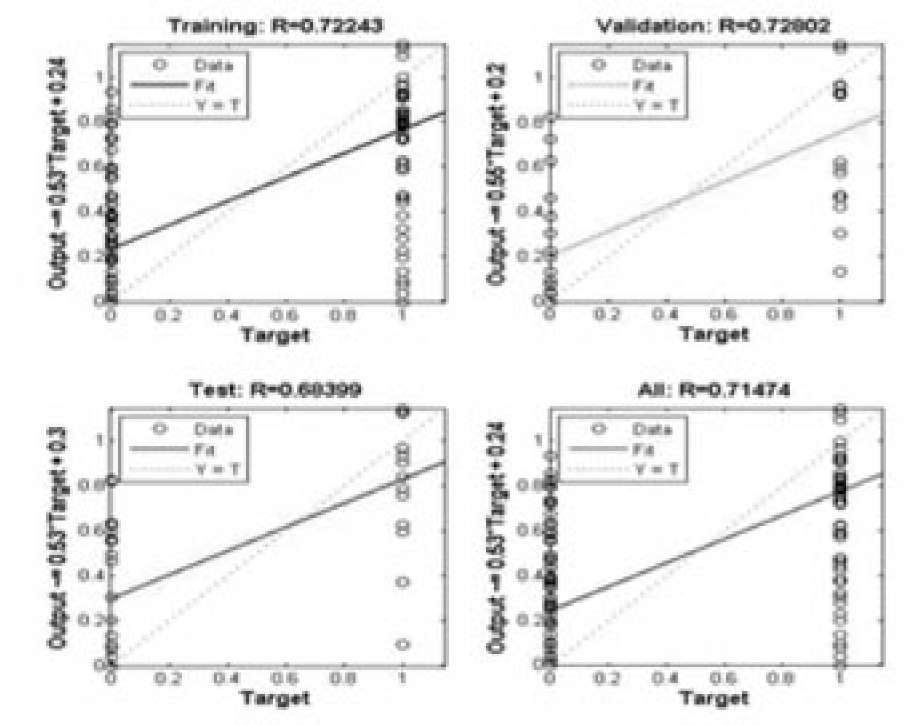

Figure 3: Regression Plot with Proposed Hybrid Feature Selection Algorithm

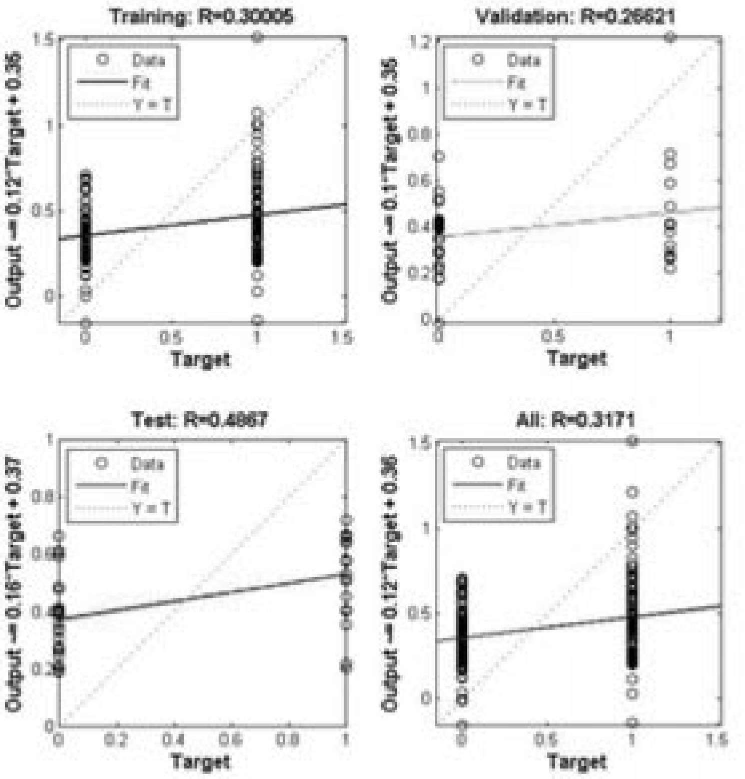

Figure 4: Regression Plot for Genetic Algorithm

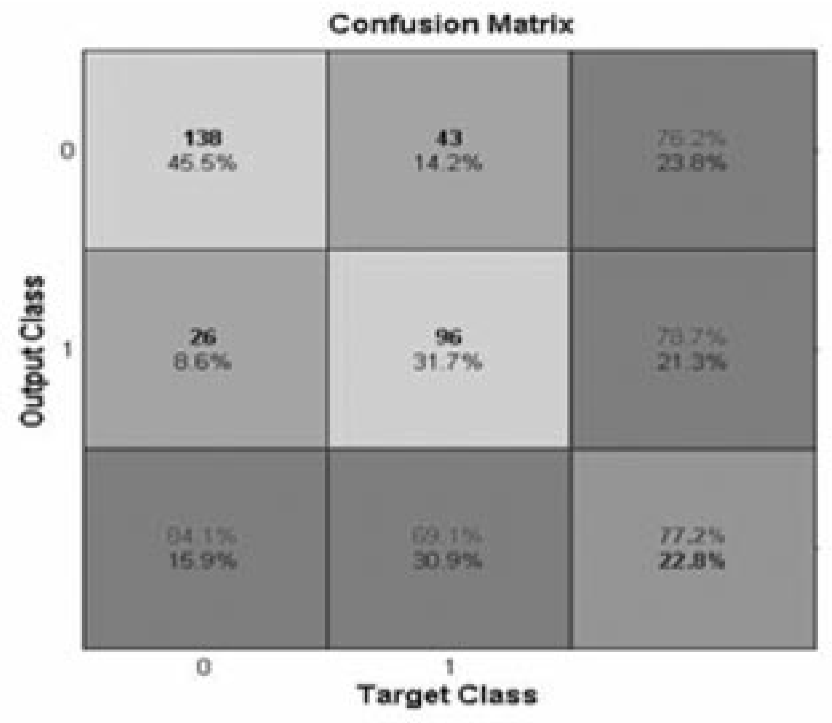

Figure 5a: Confusion Matrix for Genetic Algorithm

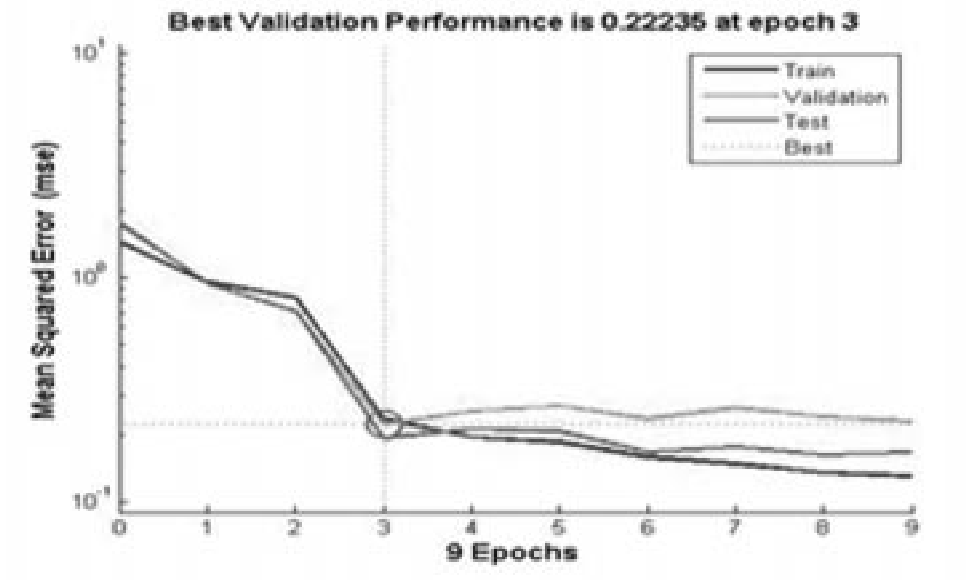

Figure 5b: Performance Analysis of Genetic Algorithm using MSE

Figure 5a and 5b illustrate the confusion matrix and execution plot of Genetic Algorithm and Hybrid Feature Selection calculation separately. It likewise bolsters the utilization of proposed Hybrid Feature Selection in the Feature Reduction. In the wake of ordering the cases, both the arrangements of qualities are assessed utilizing the Fuzzy Inference System. The exactness is 95.25 while utilizing the proposed Hybrid Feature Selection Algorithm though it is just 90.23 while utilizing the Genetic Algorithm. Figure 6a and 6b delineates the accuracy level acquired while looking at the existing and proposed calculation. Figure 7 demonstrates the training error obtained acquired while utilizing the calculations.

For predicting the Cardio Vascular Disease severity, it is concluded that the proposed Hybrid framework is the better diagnostic tool from the above all results. Since, the number of attributes is reduced from 13 to 4, the time factor is also reduced which is a major supporting factor this framework.

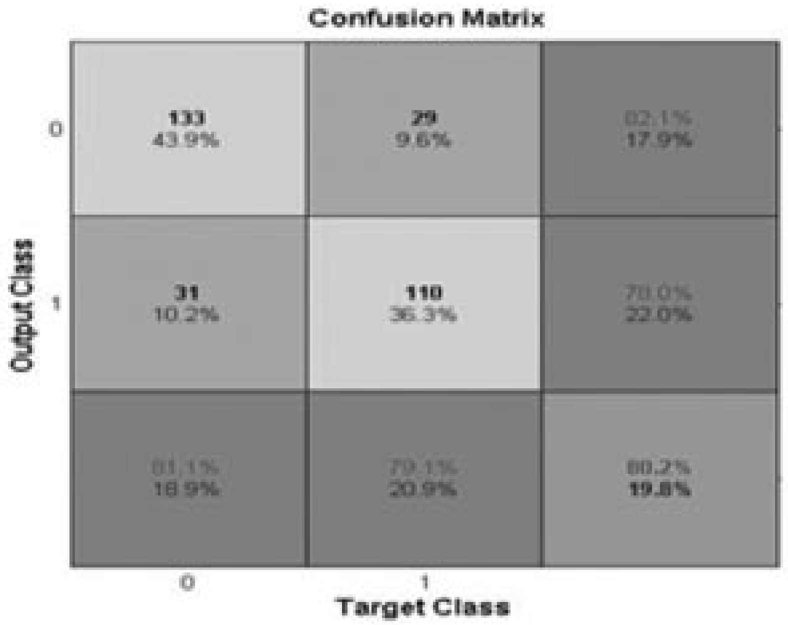

Figure 6a: Confusion Matrix of Proposed Hybrid Feature Selection Algorithm

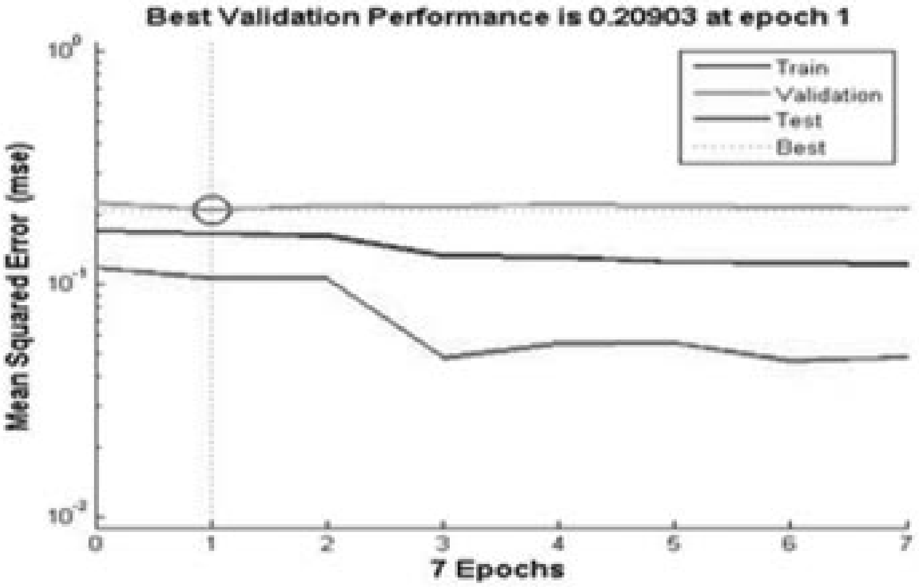

Figure 6b: Performance Analysis of Proposed Hybrid Feature Selection using MSE

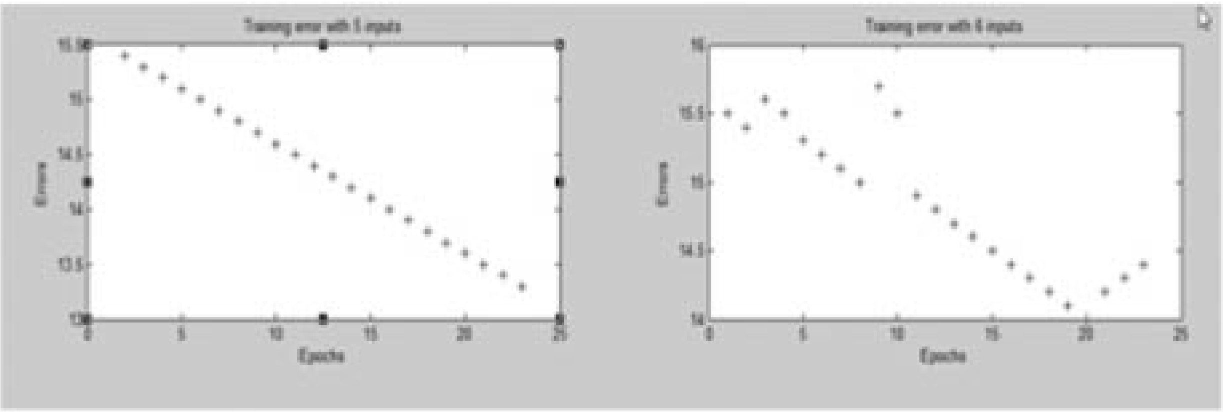

Figure 7: Training Error Comparison

6. CONCLUSION

The proposed Algorithm can be likewise be recommended as a superior Feature Reduction calculation in every one of the fields, which is additionally taken as future work and to be tried in the real time environment. Since this work does not contain information from the continuous, the future direction is gathering the ongoing information set and testing the same progressively to demonstrate the accuracy of the structure in seriousness prediction. The practical execution of the proposed calculation in hospitals is one of the future directions in the work.

REFERENCES

[1] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", *Journal of Healthcare Information Management*, Vol. 19, No. 2.

[2] PrasannaDesikan, Kuo-Wei Hsu, JaideepSrivastava, "Data Mining For Healthcare Management", *2011SIAM International Conference on Data Mining*, April, 2011.

[3] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", *IJCSNS International Journal of Computer Science and Network Security*, Vol. 8 No. 8, August 2008.

[4] HeonGyu Lee, Ki Yong Noh, KeunHoRyu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV" *LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining*, pp. 56-66, May 2007.

[5] Niti Guru, Anil Dahiya, NavinRajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", *Delhi Business Review*, Vol. 8, No. 1 (January - June 2007).

[6] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," *Seminar Presentation at University* of Tokyo, 2004.

[7] ShantakumarB.Patil, Y.S.Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network". ISSN 1450-216X Vol.31 No.4 (2009), pp. 642-656.

[8] Franck Le Duff, CristianMunteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", *Studies in health technology and informatics*, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[9] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", *International Journal of Biological, Biomedical and Medical Sciences*, Vol. 3, No. 3, 2008.

[10] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and KeunHoRyu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", *Springer*, Vol: 345, pp: 721- 727, 2006.

[11] UCI machine learning repository: http://archive.ics.uci.edu/ml/ datasets/ Heart Disease: Last visited 30th December, 2014.

[12] K.Rajeswari, "Prediction of Risk Score for Heart Disease in India using Machine Intelligence",*IPCSIT*, Vol 4, 2011.