Research article **Available online www.ijsrr.org** **ISSN: 2279–0543**

International Journal of Scientific Research and Reviews

HCR-PSO Feature Selection for Heart Disease Prediction

Janani S.*[1] and Tamilselvi R.[2]

[1]Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, India. [2]Assistant Professor, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, India.

ABSTRACT

Data Mining is an important aspect of diagnosing and predicting diseases in automatic manner. It involves developing appropriate techniques and algorithms to analyze data sets in medical field. In present, heart disease has excessively increased and heart diseases are becoming the most fatal diseases in several countries. In this paper, heart patient datasets are investigate for building classification models to predict the heart disease. This paper implements feature extraction technique construction and comparative study for improving the accuracy of predicting the heart disease. By the use of HCR-PSO (Highly Co-Related PSO) feature selection technique; a subset from whole normalized heart patient datasets is acquired which have only significant attributes. The study emphasized on finding the effective heart disease prediction construction by using various machine learning algorithms that are KNN(K-Nearest Neighbor), Random forest, SVM(Support Vector Machine), Bayesian network and MLP(Multilayer Perceptron). The research work central point is on finding the efficient classification algorithm for the prediction of heart disease in the early stage based on the accuracy using validation metrics that are Mean Absolute Error(MAE), Relative Squared Error(RSE) and Root Mean Square Error(RMSE).

KEYWORDS: HCR-PSO, Feature extraction, Bayesian and heart disease prediction model .

*** Corresponding author**

S. Janani

Department of Computer Science,

Dr. SNS Rajalakshmi College of Arts and Science,

Coimbatore, India.

Tel: +919578995707, E-mail: janusarguru@gmail.com.

INTRODUCTION

Data mining is a process of finding previous applied unknown patterns and trends in databases. This pattern is further used to build predictive models. In this paper, the main objective is to study various data mining techniques/algorithms that are used in the forecasting of heart diseases by some data mining tool. Our life depends upon the efficient working of heart, so it is the most vital part of our body. There may be a huge reasons for heart diseases like high blood pressure, high cholesterol, unhealthy diet, smoking, irregular exercises and obesity. Data mining techniques are used in many fields especially in health care industry. Usually in health care industry, huge amount of data which are complex is generated and this data is about medicines, hospital resources, patients, medical devices, disease diagnosis etc. This complex data needs to be analyzed and processed for the extraction process which then helps in taking decisions and is also cost effective. According to world health organization, there were 17.5 million people losing their life due to cardio vascular diseases in 2012, represents 31 percent of global deaths. Due to the coronary heart disease several million (approximately 7.4) people were died and due to stroke around 6.7 million were died as per estimation. By the year 2030, around 23.6 million people will lose their life due to heart disease as stated. Thus, a best method to predict the heart diseases is for efficient heart disease prediction system. This system will find human interpretable patterns and will determine trends in patient records to enhance health care.

MATERIALS AND METHODS

Sample Data

In this paper, heart data set from UCI repository is taken into account. This data set has 583 samples which comprise 10 independent variables and 1 dependent variable.

Table 1: Heart dataset from UCI repository

S.No	Attributes Type	Gender Categorical
1	Age	Real number
2	Sex	String
3	Cp	Real number
4	Trestbps	Real number
5	Chol	Real number
6	Fbs	Real number
7	Restecg	Real number
8	Thalach	Integer
9	Exang	Integer
10	Oldpeak	Integer
11	Slope	Integer
12	Ca	Integer
13	Thal	Integer
14	Num	Binary

Normalization

Normalization is the technique that helps to reduce the redundancies thus the reliability will be increased. In this paper, Min-Max normalization approach is implemented. Min-max normalization is considered as one of the best normalization technique which linearly transforms the value from x to y= (x-**min**)/(**max-min**)(**min** and **max** represents the minimum value and maximum values in X, where X is the set of observed values of x). When x= min and y=0

$$\boxed{y=x-(\min(x)) / (\max(x)-\min(x))}$$

Feature Selection

Feature extraction is the most crucial concept for extracting important and most appropriate attributes in the data set. In this paper, HCR-PSO feature extraction technique is applied in Heart dataset to obtain some important attributes which are efficient in predicting the disease at earlier stage. It works in an iterative process. In each iteration, PSO tries to predict the optimal solution by changing the heart beat rate and attack disorder incident of own particle and of group towards two parameters, gbest and pbest location in consecutive iteration. Both the parameters pbest and gbest are determined by the particle swarm optimization in each iteration. gbest refers to the global best value which is acquired by any of the particles in population and pbest refers to the fitness value that is acquired so far. The Heart prediction computation time in HCR-PSO is much less than in GAs because all swam particles in HCR-PSO tend to meet to the best solution fast.

CLASSIFICATION ALGORITHM

Random Forest (RF)

Random forest is the supervised classification algorithms which construct the forest and similar to the typical forest, the algorithm generates more trees. If the number of trees generated by the Classifier is more, then the outcome will be accurate. Three basic techniques are used in Random Forest Classification algorithm; they are (1) Forest - RI (2) Forest - RC and (3) Combination of Forest - RI and Forest – RC.

Algorithm combines tree finders and each tree based on the values of a variable that evaluated individually with the normal distribution for all trees that exists in the forest. Random Forest Classifier is considered as a great tool for producing predictions but it is not well suited.

Support Vector Machine (SVM)

SVM is applied in classification technique by taking the sample heart dataset. Each data in the heart dataset placed like a point/tip in multi-dimensional space. Here, space n represents the number of attributes/features in a heart dataset and the attribute's value represents the particular coordinate value. SVM classification works on the principle that a hyper-plane is constructed to half the dataset into two categories to classify the data. Similarly, when applying SVM on heart dataset the hyper-plane is determined and constructed and divides the heart data set into two classes. The co-ordinates in each reflection is said to be the Support Vectors. SVM is a margin that separates two classes.

MLP (Multilayer Perceptron)

An MLP is a leading artificial neural network algorithm which maps input data to an appropriate output data, here input data is heart datasets. It has large number of nodes in a graph with multiple layers and each and every layer is connected to the adjacent layer. Except the input nodes, all layers in the nodes of the graph are called as neurons and also this MLP classification technique uses a supervised learning technique and is applied in Heart dataset so the network is generated with multiple nodes.

Bayesian Networks

NBC learns the data from heart dataset and typically uses Bayes rule to manipulate the probability of the class label "L" to the specific sample of attributes A1, A2, A3….An thus the best inductive probability is forecasted. The underlying aim of this NBC is to detect the distinct class variable's value by using a features' vector. NBC algorithm is easy in its computation and it is used widely in heart dataset because it yields good accuracy than other popular algorithms. It is also being efficient in learning dataset linearly with the help of some models to integrate the predictions. But if the number of attributes is increased, then classifier's accuracy may be decreased if the features are not equally distributed.

RESULTS AND DISCUSSIONS

In this paper, heart data set from UCI repository is taken into account. This data set has 583 samples which comprise 10 independent variables and 1 dependent variable. From this dataset, only fourteen attributes are chosen because of its importance in predicting the heart related disease.

Performances Metrics Analysis

The following table 2 and figure 1 describe a Mean absolute error analysis for HCR-PSO feature extraction model. In this table Mean Absolute Error details are shown,

Table 2: Mean Absolute Error

Classification Algorithm	Mean Absolute Error
MLP	0.402
SVM	0.398
Random Forest	0.392
Bayesian	0.389

Mean Absolute Error (MAE): MAE usually gives the average size of the error without taking their direction into account. MAE measures the average magnitude of the errors. It is evaluated by predicting the average over the test sample of the absolute differences between predicting and actual observation where all individual differences have equal weight.

Formula :MAE = 1/n $\sum |y_j - y\hat{}_j|$

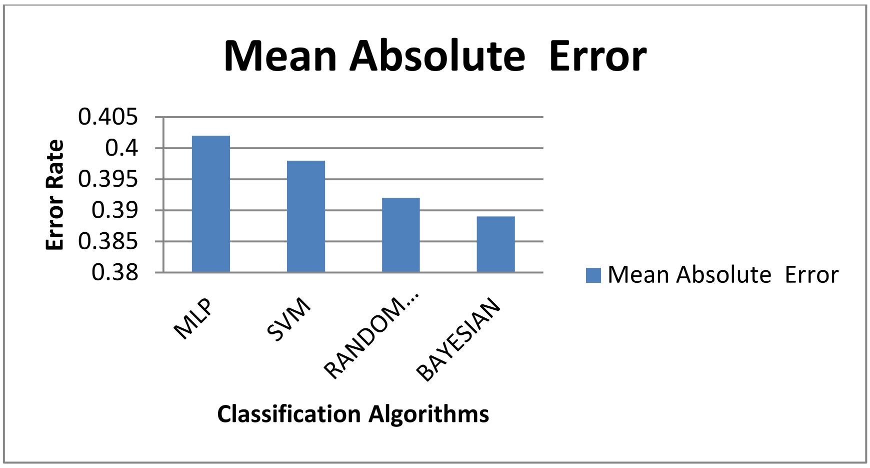

Figure 1: Mean Absolute Error Rate

The following table:3 contains a RMSE analysis for HCR-PSO feature extraction model. In this table RSME analysis details are shown,

Table: 3 Root Mean Square Error

Classification Algorithm	Root Mean Square Error
MLP	0.417
SVM	0.409
Random Forest	0.412
Bayesian	0.403

Root Mean Square Error (RMSE) is finding the standard deviation of the remaining values after evaluating all other values (prediction errors). RMSE is a measure of spread out these residuals is Heart dataset and it is mainly used in various fields like climatology, forecasting, and regression analysis to verify experimental results.

RMSE = $\sqrt{}$ (f-o)^2 (f = predicted value, o =actual value).

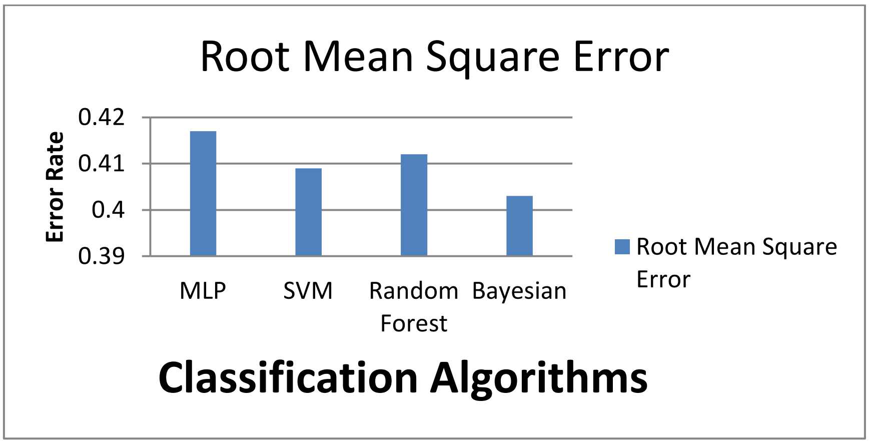

Figure 2: Root Mean Square Error

The following table 4 and Fig 3 illustrates a RSME validation metrics for HCR-PSO feature selection. In this table Relative Squared Error details are shown,

Table: 4 Relative Squared Error

Classification Algorithm	Relative Squared Error
MLP	65.87
SVM	65.58
Random Forest	65.73
Bayesian	65.33

Root relative squared error is done by predicting the total squared error and by splitting the predictor’s total squared error, it performs normalization.

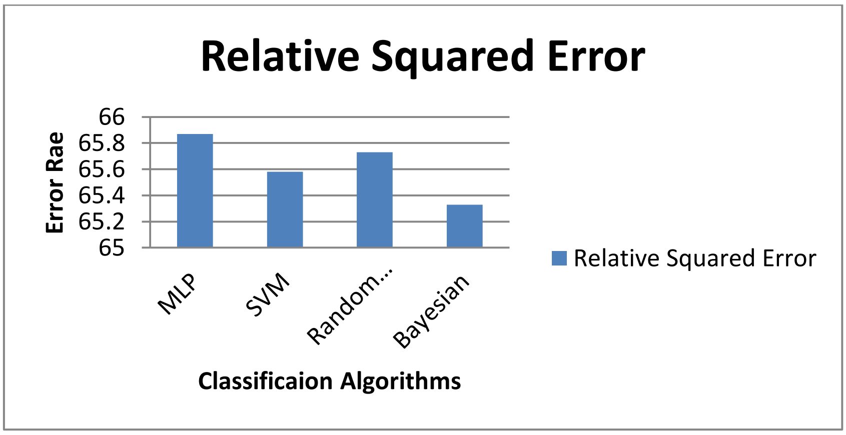

Figure 3: Relative Squared Error

The **RRSE** T_i of a single program j is typically computed by the equation. Here $P_{(ij)}$ is the value that is found by the individual program i for an instance case j (out of n sample cases); where T_j is the target value for an instance case j; and $\overline{T}$ is given by the formula:

$$\mathbf{RRSE = \ T_j = 1/n \sum_{j\text{-}1} T_j}$$

The following table 5 represents an accuracy value of various classification algorithms. In this table existing and proposed accuracy values are shown,

Table 5: Accuracy

Classification Algorithm	Existing Accuracy	Proposed Accuracy
MLP	68.26	77.54
SVM	71.35	73.44
Random Forest	70.32	80.22
Bayesian	67.23	90.33

The following Figure 4 describes a overall classification algorithm for accuracy values analysis. In this figure Greedy and HCR-PSO accuracy values are shown,

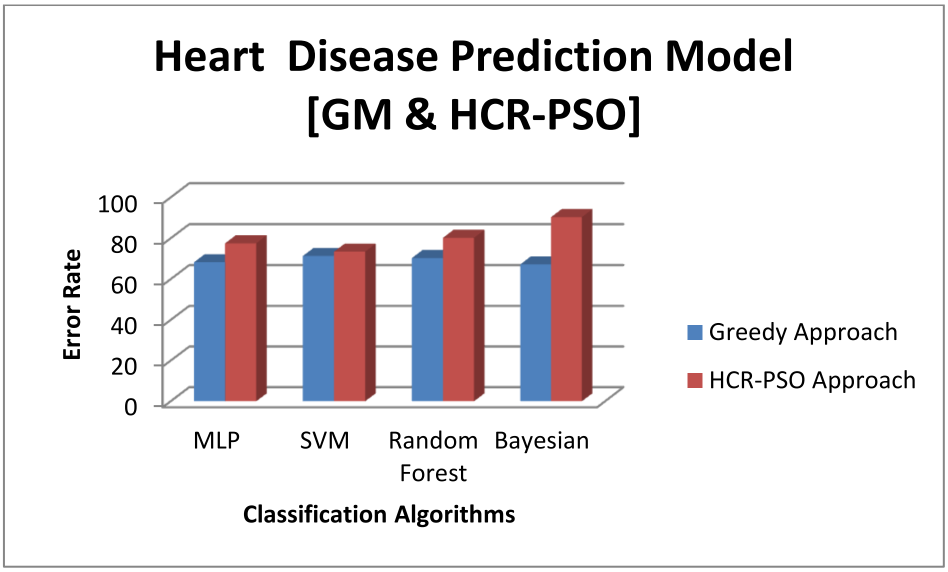

Figure 4: Accuracy

CONCLUSION

In this paper the HCR-PSO feature selection technique for Indian Heart Patient Dataset is proposed. HCR-PSO is the most widely used feature selection technique. It helps to improve Heart dataset classification performance. Additionally, HCR-PSO for feature selection improves Heart dataset classification performance and also reduces the number of attributes from the whole data set. There is a scope to further reduce search space for better Heart dataset classification accuracy if enhanced selection and mutation procedures are being used.

After the analysis, HCR-PSO feature selection algorithm is efficient for selecting optimal feature subset from the actual dataset. This feature extraction technique is applied in the heart data set to predict the heart related disease. Various classification algorithms are implemented for analyzing the data set such as SVM, Random Forest, MLP, and Bayesian Classification. While applying HCR-PSO feature extraction technique in those machine learning algorithms, various results will be obtained. There exist many criterions for computing the selected feature subset, in this paper various features like exang, oldpeak, slope, restecg, thalach are taken in order to assess the performance of various machine learning algorithms. It is also seen that Bayesian classification algorithm gives better result compared to other classification algorithms.

REFERENCES

1. Hai Wang, Medical Knowledge Acquisition through DM, IEEE (ISIME), 2008; 978-1-4244-2511-2/08.
2. Victor-Emil Neagoe. A N-Fuzzy Approach to Classification of ECG Signals for Ischemic Heart Disease Diagnosis , AMIA Annu Symp Proc. 2003; 494–498.
3. Franck Le Duff, Munteanb Cristian, Cuggiaa Marc, Philippe Mabob et al. Prediction of Survival Causes After Out of Hospital Cardiac Arrest using DM , Studies in health technology and informatics, 2004; 107: 1256-1259,.
4. Palaniappan Sellappan, Awang Rafiah et al. IHDPS Using DM Techniques (IJCSNS). August 2008; 8:108-115.
5. Latha Parthiban, R.Subramanian. Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, (IJMHS). 2007; 1(5): 278-281.
6. Vazirani Harsh, Rahul Kala, Anupam Shukla, et al. Using Modular NN technique for Heart Disease, IJCCT 2010; 1(2): 3, 4 for International Conference (ACCTA-2010) 3-5 August 2010; 88-93.
7. S.Vijiyarani. An Efficient Classification Tree Technique for Heart Disease Prediction (ICRTCT -2013) published in (IJCA) (0975 – 8887), 2013; 6-9.
8. R. Tamilselvi, S. Kalaiselvi. An Overview of Data Mining Techniques and Applications. IJSR.Vol.No.2. issue.2. ISSN: 2319-7064.2013;506-509.
9. Noh Kiyong, Ho-Sun Shon, Heon Gyu Lee, et al. Associative Classification Technique for Cardiovascular Disease Diagnosis. Springer 2006; 721- 727.
10. Neagoe Victor-Emil. A Neuro-Fuzzy Technique to Classification of ECG Signals for IHDD. AMIA Proceedings. 2003; 494–498.