Hybrid SVM-ANN Classifier is used for Heart Disease Prediction System

[1]S.Shylaja, [2]R. Muralidharan
[1]Research Scholar, [2]Academic Principal
[1]Rathinam College of Arts & Science, Coimbatore, India,
[2]Rathinam College of Arts & Science, Coimbatore, India

***Abstract* - Heart disease is one of the reason behind death of people universally, more people pass away from this disease compare to any other cause annually. To stay away from heart disease or find symptoms early, adults over 40 plus should have a complete cardiovascular checkup. Many experts developing intelligent decision support systems related to medical to get better ability of the detection of heart disease. In heart disease diagnosis and treatment, hybrid data mining techniques provide the reasonable accuracy level compare to other existing techniques. A hybrid classifier obtained by hybridization of Support Vector Machine and Artificial Neural Network classifier. In this proposed method a predictive analysis is carried out on UCI Heart disease dataset using SVM and ANN techniques. This SVM-ANN hybrid classifier performance much better than standard version of support vector machine and artificial neural networks. The obtained accuracy of this technique is 88.54%. This result shows that SVM-ANN is the best hybrid algorithm for diagnosis of heart disease**

***keywords* - Data Mining, Support Vector Machine, Artificial Neural Network, Hybrid SVM-ANN**

I. INTRODUCTION

Heart is one of the most important organ of individual human body. It pumps blood through blood vessels of the circulatory system. The circulatory system is very important because it transports blood and oxygen to the other organs of the body. Heart plays vital role in circulatory system. If the heart does not function properly then it will lead to serious health condition including death. In 2015, an expected 17.7 millions of people died from cardiovascular disease which is 31 % of worldwide deaths. An estimated 7.4 million peoples death out of total deaths were due to coronary heart sickness and similarly 6.9 million peoples have been died due to heart stroke [6]. Heart disease is the topmost reason for deaths of human all across the world and those experts who treat heart disease keep data records of patient to extract the valuable information. Therefore the data mining turns out to be a productive and useful tool. Data mining is obviously defined as "The monotonous and interactive way of near-term across logical, novel, beneficial and understandable knowledge (Patterns, models and guidelines etc) in huge databases. [7] [8]. Data mining does the survey of vast datasets to find out unrevealed and unknown patterns, connection between them and that are not easy to detect with statistical methods. Also statistical mining is increasing speedily and getting successful in various applications like organic compounds analysis, business forecasting, medical care and meteorology [9]. Data mining is the crucial step that comes within the uncovering of unrevealed data however beneficial details from significant databases. Researchers have suggested that the use of statistical mining in figuring out powerful remedies for patient can enhance practitioner overall performance. Researchers investigated and came to result that by applying specific data mining strategies in the analysis of coronary heart disease to identify which data mining technique can offer greater reliable accuracy. Many distinct data mining strategies were already used to help medical care departments to detect coronary heart disease. [10][11]. Data mining in medical care is a rising field of high importance for offering analysis and a deeper expertise of clinical statistics. A mixture of applications of data mining in medical care include health care centers analysis for finer medical policy – making and avoidance of health facility mistakes, early detection, prevention of sickness and preventable patient's deaths and financial savings. Data mining KDD process is given in Figure 1.

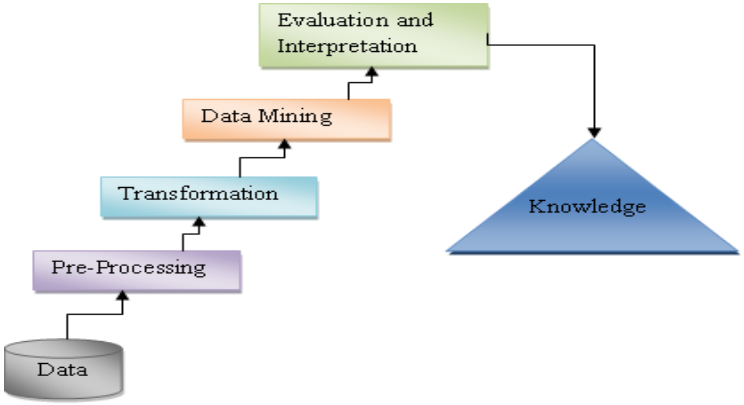

Figure 1: Knowledge Discovery Steps

The whole procedure is extraction of probably useful information from huge databases. It gives the choice for selection, cleaning and enhancing the data. Also data mining is the main step for understanding the procedure of knowledge [12]. A number of symptoms are connected with heart disease that makes it hard to diagnose it quicker and better. Focusing in cardiovascular disease patients databases could be compared to real life program, Doctor understands to assign the excess weight to each designated to the attribute having high effect on disease prediction. So that it become acceptable to make use of knowledge in order to have connection with various experts which are gathered to assist the process of medical decision when single and hybrid data mining techniques were compared to diagnose the heart disease they showed accuracies, where the hybrid one proved that it provides better accuracy as compared to single techniques [7].

In this proposed system is hybrid approach for heart disease prediction using Support Vector Machine (SVM) and Artificial Neural Networks (ANN) on UCI heart disease Dataset . The main intention is to obtain high accuracy rate of prediction. This paper is designed as Section I is a brief information about healthcare problem especially in heart disease with data mining use cases are added introduction section, section II is all about existing work. Section III specifies that proposed system architecture and algorithmic introduction. Then next section specifies about the steps required for SVM and ANN algorithm. In section V Implementation details steps are summarized along with dataset in detail. Section VI is about result and conclusion. This paper proposed a new model to boost the accuracy in recognizing the pattern of heart patients. It uses the different algorithm of Classification such as SVM and ANN combined.

II. RELATED WORK

Heart disease has easily identified over the last decade and has become the leading cause of the death for people in maximum countries around the world [1]. Coronary heart disease is defined as the problem of the heart that happens because of its irregular blood circulation. It leads to the fatty deposits build upon an inner layer of the blood vessels that provide the heart muscles with blood, resulting in contraction [2]. All heart oriented defects are called as heart disease which may leads to death. This hazardous disease causes many people in different countries including India. A large number of people died due to cardiovascular disease that is yearly increasing [3]. Heart disease is the leading cause of death all over the world in the past ten years. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of the heart disease. Using traditional data mining technique in the diagnosis of the heart disease has been expansively investigated showing acceptable levels of accuracy. In recent times, many researchers have inquire into the effect of hybridizing more than one technique showing enhanced results in the diagnosis of heart disease. The data mining is the important stage of knowledge discovery in database (KDD) which is extraction of implicit, unique and potentially useful information from data. The difference between mining of data and discovering knowledge is that the latter is the utilization of different intelligent algorithms to extract pattern from data where as information discovery is the complete process that is involved in discovery knowledge in data [4]. This hybrid classifier has been performed in the coronary heart disease (CHD) risk assessment problem and the experimental results are explained and analyzed. In this paper by using GSO algorithm performance is combined with the evaluation metrics such as accuracy, sensitivity and specificity are evaluated [5]. The Methodology of hybridizing two data mining techniques like Artificial Neural Network (ANN) and genetic algorithm (GA) which was implemented to achieve high accuracy with least error [7]. The very big disadvantages of GA are unsigned mutations. These mutations in genetic algorithm featured like including a randomly generated wide variety to a parameter of an individual of the procedure [14] which is the cause for completely slow convergence of genetic algorithm. Heart disease prediction using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have been implemented with numerous data mining techniques for prediction of cardiovascular disease such as Decision tree, Naive Bayes, Neural Network, Kernal Densisty , automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies in multiple databases of patients from around the world [14]-[18]. The proposed algorithm into two parts i.e. first part deals with evaluating attributes using genetic search and second part helps in structuring classifier and also to measure its accuracy. Comparison of accuracy of datasets with and without GA is done in this paper. When these two are combined accuracy showed a hike by 5 %. The accuracy obtained using k – Nearest neighbor and genetic algorithm is very less [19]. Also these algorithms take much more time for optimization. In this paper they used hybrid technique called neural network ensemble i.e. combination of neural network and ensemble based methods. Even they have used Cleveland dataset, the accuracy obtained using their proposed system is 89.01 %. They made use of SAS enterprise miner 5.2 to construct neural network ensembles based methodology. Even though they increase the number of neural network node in ensemble model but no improvement in performance was obtained [20].

III. PROPOSED METHOD

In this section we mentioned about the system architecture. Fig 2 represents the outline of system architecture. The nucleus modules of the proposed system consist of

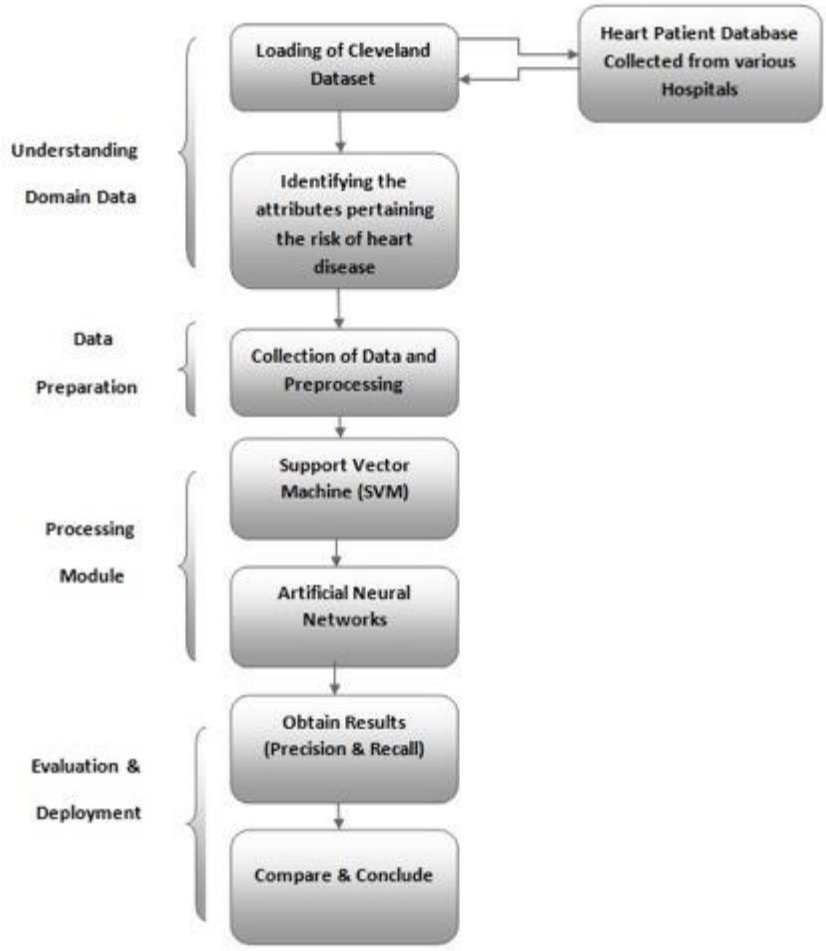

Figure 2: Heart Disease Prediction system architecture

Understanding Domain Dataset

Cleveland Dataset is provided as input to system, detail about the dataset is discussed in section V and feature of loading excel dataset or manual entries into the system are also mentioned.

Data Preparation

When data is manipulated in a way that it can be suitable for future examination is called data preparation.

Module

It talks about the algorithmic approach applied over the system in order to high accuracy result. As an algorithmic approach we use SVM Classifier and ANN Classifier in data mining techniques.

Evaluation and Deployment

Final analyzing modules states information related to generated output. We obtain a confusion matrix; our system compares and concludes about measurable resultant like sensitivity, specificity, accuracy, true positive rate and false positive rate.

A detailed flow of system architecture of heart disease prediction using data mining approach is shown in figure 3.

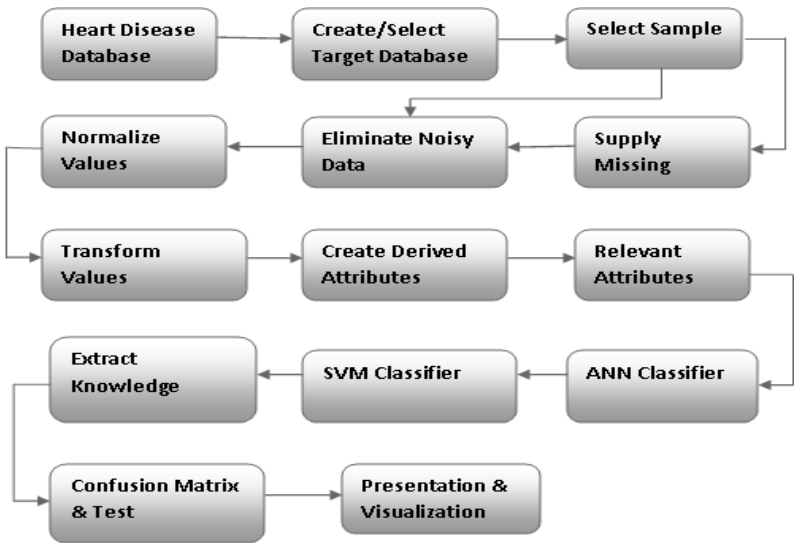

Figure 3: Heart disease prediction system detailed architecture

Data set

UCI Machine Learning repository Cleveland's heart disease database contains 76 attributes, but all published experiments refer to using a subset of 14 out of which we have reduced it to 11. Clinical database have collected large quantities of significant information about patient and their medical conditions. Records set with minimum significant medial attributes was obtained from the Cleveland heart disease database with the help of the dataset the pattern significant to the cardiac predictions are extracted. The records for each set were selected randomly to avoid bias.

Table 1: Data set Information

Data set characteristics	Multivariate
Attribute characteristics	Categorical, Integer, Real
Associated Tasks	Classification
Number of Instances	303
Number of attributes	75
Missing Value	yes

Prediction of heart disease is done by considering following 11 attributes instead of 14 that have been traditionally used to explore and analyze the result

Table 2: Attribute description

Attribute	Description	Value
CP	Chest Pain Type	1 : typical angina 2 : atypical angina 3 : non- angina pain 4 : asymptomatic
Trestbps	Resting blood pressure	Continues value in mm hg
Chol	Serum cholesterol	Continues value in mm/dl
Fbs	fasting blood sugar	1 ≥ 120 mg/dl 0≤ 120 mg/dl
Restecg	Resting electro cardio graphics results	0 : Normal 1:having_ST_T Wave abnormal 2: Left Ventricular Hypertrophy
Thalach	Maximum heart rate achieved	Continuous value
Exang	Exercise induced angina	0 : No 1 : Yes
Slope	The slope of the peak exercise ST segment	1 : up sloping 2 : flat 3 : down sloping
Thal	Defect Type	3 : Normal 6 : Fixed 7 : Reversible defect
Sex	Male or Female	1: Male 0 : Female
Age	Age in years	Continuous

The above mentioned five algorithms with SVM-ANN hybrid classifier, the attribute set of the heart patient is analyzed using MATLAB 7.10. The efficiency of the algorithms is analyzed using the sensitivity, specificity, accuracy, true positive rate and false positive rate. The test has been conducted with different set of data ranging from 25, 50, 75, 100 patient data sets. In the data set 60% belongs to heart disease patients and 40% belongs to normal patients with slight indications of heart disease. The entire data set is divided into two sets one as training data set and another one as test data set. Initially the algorithm was trained for the training data set and then the test data set is given as input to predict the disease.

IV. ALGORITHM DESCRIPTION

RIPPER

RIPPER is expansion of Repeated Incremental Pruning to Produce Error Reduction. RIPPER is a rule based algorithm that builds the set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules. It is based on association rules with Reduced Error Pruning (REP), a very general and effective method found in decision tree algorithms. In Reduced Error Pruning (REP) for rules algorithms, the training data is separated into a growing set and a pruning set. First of all, an initial rule set is formed that is the growing set, using some heuristic method. At each phase of Simplification, the pruning operator preferred is the one that yield the greatest reduction of error on the pruning set. Generality ends when applying any pruning operator would increase error on the pruning set.

Decision Tree

A Decision tree is a decision support tool that uses a tree like graph or model of decision and their possible consequences together with chance event, resource cost, outcome and utility. This algorithm based on flow chart like tree structure in which each internal node represent a "test" on an attribute and each branch represent the outcome of the test, then each leaf node represents a class label (The decision will take after complete all the attribute). It is a popular classification algorithm and the paths from root to leaf represent classification rule. This algorithm includes Quinlan's ID3, C4.5, C5, and Breiman et al.'s CART. The main objective of partition algorithm is to find a variable-threshold pair that maximizes the homogeneity of the resulting two or more subgroups of sample. In most of the case generally used mathematical algorithm for split up process that Entropy based information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID).

Artificial Neural Networks

Artificial neural networks (ANNs) often just called a neural network, and it is a mathematical model by biologically inspired, a neural network consist of an interconnected group of artificial neurons, and it is a set of connected input and output network in which weight is associated with each connection. It consists of one input layer, one or more in between layer and one output layer. It processes information utilizes connectionist approach to computation. This algorithm is highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. One of eminent ANN structural design is called multi-layer perceptron (MLP). It has three layers namely input layer, output layer and hidden layer. It feed the input data to the input layer and take the output from the output layer. So it will increase the number of the hidden layer as much as we want, to make the model more complex. The MLP is known to be a powerful function approximate for prediction and classification problems. Specified the correct size and the structure, MLP is efficient of learning arbitrarily complex non linear functions to arbitrary accuracy levels. The MLP is basically the collection of nonlinear neurons (perceptron) organized and connected to each other in a feed forward multi-layer structure.

Support Vector Machine

The support vector machine (SVM) is peradventure one of the most important and talked about machine learning algorithm. The SVM is major technique for classification of jointly linear and non-linear data. It utilizes a non-linear mapping to transform the original training data into a higher dimension. Contained by this new dimension it searches for linear optimal separating hyper plane. With the suitable non linear mapping to an adequately high dimension, data from two classes can always be separated by a hyper plane. The SVM find this hyper plane using support vectors and margins. SVM perform classification tasks by maximizing the margin distinct both classes while minimizing the classification errors.

Naive Bayes

Naive Bayes is a classification algorithm for binary (Two class) and multi class problems. The Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. This technique is easiest to understand when describing a numerical or categorical input values. The main rule of this algorithm is used to create models with predictive capabilities. It grants new ways of exploring and understanding data.

SVM-ANN Hybrid

In this proposed work, the combination of SVM and ANN is used to classify the heart disease prediction system in earlier stage. The hybridization method is used firstly to classify normal and cardiovascular dataset using linear SVM kernel and then classify the heart disease data into different stages through ANN. In SVM there will be a hyper plane between the set of data points as the decision boundary. In this case, there are two classified data of normal patients and heart disease patients are used by this SVM-ANN hybrid algorithm.

V. RESULT & DISCUSSION

The table 3 shows the analysis of the six algorithms considered for the prediction of heart disease patient based on the accuracy, sensitivity and specificity. Based on the analysis, it clearly shows that the SVM-ANN Hybrid has higher accuracy compared to the other algorithms. In all the test runs with different size of data set, the SVM-ANN provides higher sensitivity and specificity compared to other algorithms. In figure fig.1 the performance of the algorithm over the prediction of heart disease is shown.

Table 3: Comparative Analysis of Data mining techniques

Algorithm	Accuracy	Sensitivity	Specificity
ANN	85.30 %	81.75 %	77.73 %
SVM	86.12 %	84.87 %	79.21 %
RIPPER	81.08 %	80.25 %	75.82 %
Decision Support	79.05 %	72.17 %	73.54 %
Naive Bayes	82.97 %	76.10 %	76.27 %
SVM-ANN Hybrid	88.54 %	91.47 %	82.11 %

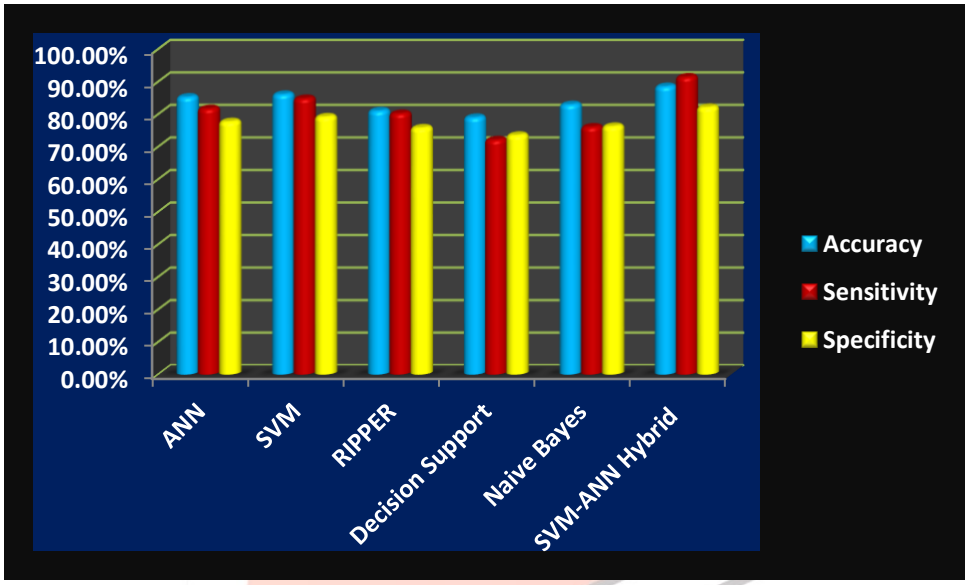

Figure 4: Comparison of prediction performance of the algorithms

From the above Figure 4 it is shown that the SVM-ANN Hybrid classifier gives the Highest Accuracy, Specificity and Sensitivity.

Table 4: True Positive Rate and False Positive Rate algorithms (100 Samples of Heart patients Database)

Algorithm	True Positive rate	False positive Rate
ANN	0.8735	0.1265
SVM	0.8848	0.1152
RIPPER	0.8548	0.1452
Decision Support	0.8274	0.1726
Naive Bayes	0.8199	0.1801
SVM-ANN Hybrid	0.8875	0.1125

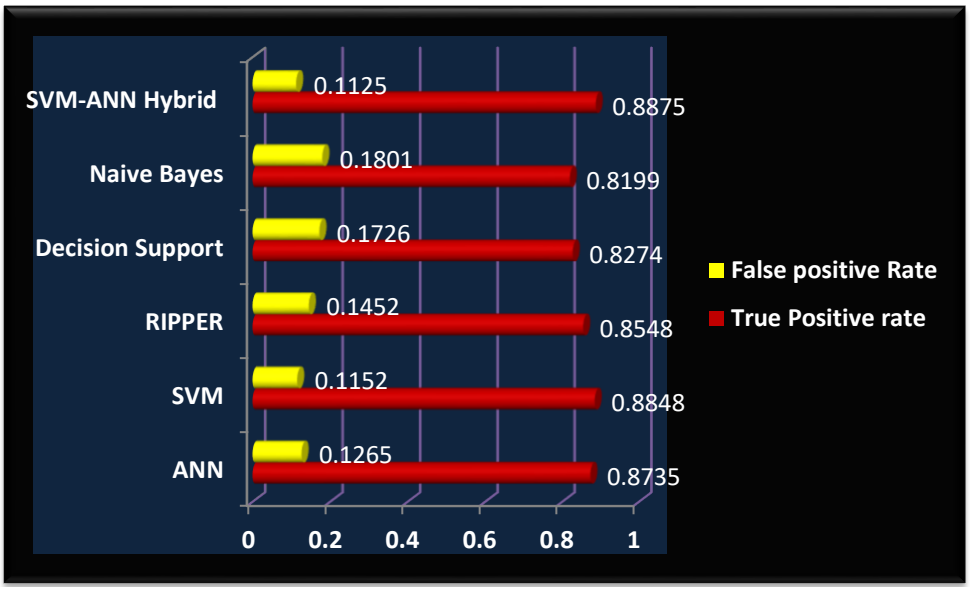

Figure 5: Prediction of true & false positive rate (100 Samples of Heart patients Database)

From the above Figure 5 it is shown that the SVM-ANN Hybrid classifier gives higher true positive rate compared to other algorithms.

VI CONCLUSION

The paper explores the hybrid of SVM-ANN as the finest binary classification system for predicting the heart disease. Most of the heart diseases are incurable by its nature and these diseases make dangerous complexities such as heart attacks and deaths. Our SVM-ANN Hybrid Classifier approach gives higher accuracy rate of about 88.54 % than earlier proposed method. The result of this research shows that this hybrid SVM-ANN is more precise than the single data mining algorithm. This hybrid classifier is suggesting that it is very effective for classification of that patient having or not having heart disease.

REFERENCE

[1] Nguyen CL, Phayung M, Herwig U. " A Highly accurate firefly based algorithm for heart disease prediction" J Exp Sys Appl 2015; 1-11
[2] Shamsher BP, Pramod KYS. " Predict the diagnosis of the heart disease patients using classification mining techniques. J Agr Veter Sci 2013; 4: 61-64.
[3] Srinivas RRG. Rough- fuzzy classifier. " A system to predict the heart disease by blending two different set theories. J Sci Eng 2014; 39: 2857-2868.
[4] Ajmal Mohamed , Mr. Balamurali " Data mining based heart attack prediction from the medical datasets of the patients " .International journal of pure and applied mathematics, Volume 119, No.17 2018, 1511-1516.
[5] Vahid K, Gholam AM, " A fuzzy evidential hybrid inference engine for coronary heart disease risk assessment. J Exp Sys Appl 2010; 37:85236-8542.
[6] Amita malav, Kalyani Kadam, " A hybrid approach for heart disease prediction using artificial neural networks and K means ", International journal of pure and applied mathematics, Volume 118, No.8,2018, 103-110.
[7] Fayyad U.M, Piatetsky-Shapiro G, Smyth P., " Knowledge discovey of data mining : Towards a unifying framework in KDD", 1996 august , Vol 96, pp.82-88.
[8] Kurgan L.A, Musilek P ," A survey of knowledge discovery and data mining process model," The knowledge engineering review 2006, 21(1), 1-24.
[9] Das R Turkoglu I, Sengur A 2009, " Effective diagnosis of heart disease through neural network ensembles", Expert system with applications 36(4), 7675-7680.
[10] Dewan A, Sharma M, " Prediction of heart disease using hybrid technique in data mining classification," In computing for sustainable global development (INDIACom), 2015 2nd international conference on (pp. 704-706). IEEE March 2015.
[11] Chandna D, "Diagnosis of heart disease using data mining algorithm", International journal of computer science and information technologies, 5(2), 1678-1680.
[12] Sumathi S, Sivanandam S.N " Introduction to data mining and its applications," 2006, Vol 29, Springer.
[13] Zhang L, Li J, Shi Y, Liu X, " Foundations of intelligent knowledge management", 2009, human systems management 28(4), 145-161.
[14] Ratnaparkhi D, Mahajan T, jadhav V, " Heart disease prediction system using data mining techniques ", International research journal of engineering and technology (IRJET),2015, 2(08), 2395-0056.
[15] Devi S.K, Krishnapriya S, Kalita D., " Prediction of heart disease using data mining techniques ", Indian journal of science and technology 2016, 9(39).
[16] Tanjeja A, " Heart disease prediction system using data mining techniques", 2013, Oriented journal of computer science and yechnology, 6(4), 457-466.

[17] Rajkumar A, reena G.S, " Diagnosis of heart disease using data mining algorithm", 2010, Global journal of computer science and technology , 10(10), 38-43.
[18] Soni J, Ansari U, Sharma D, Soni S, " Predictive data mining for medical diagnosis : An overview of heart disease prediction", International journal of computer applications, 2011, 17(8), 43-48.
[19]Chaltrali S. Dangare and Sulabha, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques",IJCA, Vol 47(10), pp 44-48, June 2012.
[20] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur, "Effective diagnosis of heart disease through neural networks ensembles",Elsevier, 2009.