ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2018.0253

HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

H. Benjamin Fredrick David and S. Antony Belcy
Department of Computer Science and Engineering, Manonmaniam Sundaranar University, India

Abstract

Data mining is a technique that is performed on large databases for extracting hidden patterns by using combinational strategy from statistical analysis, machine learning and database technology. Further, the medical data mining is an extremely important research field due to its importance in the development of various applications in flourishing healthcare domain. While summarizing the deaths occurring worldwide, the heart disease appears to be the leading cause. The identification of the possibility of heart disease in a person is complicated task for medical practitioners because it requires years of experience and intense medical tests to be conducted. In this work, three data mining classification algorithms like Random Forest, Decision Tree and Naïve Bayes are addressed and used to develop a prediction system in order to analyse and predict the possibility of heart disease. The main objective of this significant research work is to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person is carried out. Thus prevention of the loss of lives at an earlier stage is possible. The experimental setup has been made for the evaluation of the performance of algorithms with the help of heart disease benchmark dataset retrieved from UCI machine learning repository. It is found that Random Forest algorithm performs best with 81% precision when compared to other algorithms for heart disease prediction.

Keywords:

Data Mining, Classification, Prediction, Heart Disease

1. INTRODUCTION

In day to day life many factors that affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. In today’s world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living. The health of a human heart is based on the experiences in a person’s life and is completely dependent on professional and personal behaviours of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease. The term Heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behaviour, family history, smoking and hypertension.

The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the mentioned factors are taken into consideration when analysing and understanding the patients by the doctor through manual check-ups at regular intervals of time.

The symptoms of heart disease greatly depend upon which of the discomfort felt by an individual. Some symptoms are not usually identified by the common people. However, common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common to many types of heart disease is known as angina, or angina pectoris, and occurs when a part of the heart does not receive enough oxygen. Angina may be triggered by stressful events or physical exertion and normally lasts under 10 minutes. Heart attacks can also occur as a result of different types of heart disease. The signs of a heart attack are similar to angina except that they can occur during rest and tend to be more severe. The symptoms of a heart attack can sometimes resemble indigestion. Heartburn and a stomach ache can occur, as well as a heavy feeling in the chest. Other symptoms of a heart attack include pain that travels through the body, for example from the chest to the arms, neck, back, abdomen, or jaw, lightheadedness and dizzy sensations, profuse sweating, nausea and vomiting.

Heart failure is also an outcome of heart disease, and breathlessness can occur when the heart becomes too weak to circulate blood. Some heart conditions occur with no symptoms at all, especially in older adults and individuals with diabetes. The term 'congenital heart disease' covers a range of conditions, but the general symptoms include sweating, high levels of fatigue, fast heartbeat and breathing, breathlessness, chest pain. However, these symptoms might not develop until a person is older than 13 years. In these type of cases, the diagnosis becomes an intricate task requiring great experience and high skill. A risk of a heart attack or the possibility of the heart disease if identified early, can help the patients take precautions and take regulatory measures. Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis.

Data Mining is a task of extracting the vital decision making information from a collective of past records for future analysis or prediction. The information may be hidden and is not identifiable without the use of data mining. The classification is one data mining technique through which the future outcome or predictions can be made based on the historical data that is available. The medical data mining made a possible solution to integrate the classification techniques and provide computerised training on the dataset that further leads to exploring the hidden patterns in the medical data sets which is used for the prediction of the patient’s future state. Thus, by using medical data mining it is possible to provide insights on a patient’s history and is able to provide clinical support through the analysis. For clinical analysis of the patients, these patterns are very much essential. In simple

English, the medical data mining uses classification algorithms that is a vital part for identifying the possibility of heart attack before the occurrence. The classification algorithms can be trained and tested to make the predictions that determine the person's nature of being affected by heart disease.

In this research work, the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely Random Forest, Decision Tree and Naïve Bayes are used to make predictions. The analysis is done at several levels of cross validation and several percentage of percentage split evaluation methods respectively. The StatLog dataset from UCI machine learning repository is utilized for making heart disease predictions in this research work. The predictions are made using the classification model that is built from the classification algorithms when the heart disease dataset is used for training. This final model can be used for prediction of any types of heart diseases.

2. LITERATURE SURVEY

According to Ordonez [1] the heart disease can be predicted with some basic attributes taken from the patient and in their work have introduced a system that includes the characteristics of an individual human being based on totally 13 basic attributes like sex, blood pressure, cholesterol and others to predict the likelihood of a patient getting affected by heart disease. They have added two more attributes i.e. fat and smoking behaviour and extended the research dataset. The data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network are utilized to make predictions and the results are analysed on Heart disease database.

Yılmaz, [2] have proposed a method that uses least squares support vector machine (LS-SVM) utilizing a binary decision tree for classification of cardiotocogram to find out the patient condition.

Duff, et al. [3] have done a research work involving five hundred and thirty-three patients who had suffered from cardiac arrest and they were integrated in the analysis of heart disease probabilities. They performed classical statistical analysis and data mining analysis using mostly Bayesian networks.

Frawley, et al. [4] have performed a work on prediction of survival of Coronary heart disease (CHD) which is a challenging research problem for medical society. They also used 10-fold cross-validation methods to determine the impartial estimate of the three prediction models for performance comparison purposes.

Lee, et al. proposed a novel methodology to expand and study the multi-parametric feature along with linear and nonlinear features of Heart Rate Variability diagnosing cardiovascular disease. They have carried out various experiments on linear and non-linear features to estimate several classifiers, e.g., Bayesian classifiers, CMAR, C4.5 and SVM. Based on their experiments, SVM outperformed the other classifiers.

Noh, et al. suggested a classification method which is an associative classifier that is constructed based on the efficient FP-growth method. Because the volume of patterns can be diverse and huge, they offered a rule to measure the cohesion and in turn allow a tough choice of pruning patterns in the pattern-generating process.

Parthiban, et al. [7] have proposed a new work in which the heart disease is identified and predicted using the proposed Coactive Neuro-Fuzzy Inference System (CANFIS). Their model works based on the collective nature of neural network adaptive capabilities and based on the genetic algorithm along with fuzzy logic in order to diagnose the occurrence of the disease. The performance of the proposed CANFIS model was evaluated in terms of training performances and classification accuracies. Finally, their results show that the proposed CANFIS model has great prospective in predicting the heart disease.

Singh, et al. [8] have done a work using, one partition clustering algorithm (K-Means) and one hierarchical clustering algorithm (agglomerative). K-means algorithm has higher effectiveness and scalability and converges fast when production with large data sets. Hierarchical clustering constructs a hierarchy of clusters by either frequently merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones. Using WEKA data mining tool, they have calculated the performance of k-means and hierarchical clustering algorithm on the basis of accuracy and running time.

Guru, et al. [9] have proposed the computational model based on a multilayer perceptron with three layers is employed to enlarge a decision support system for the finding of five major heart diseases. The proposed decision support system is trained using a back propagation algorithm amplified with the momentum term, the adaptive learning rate and the forgetting mechanics.

Palaniappan, et al. [10] have carried out a research work and have built a model known as Intelligent Heart Disease Prediction System (IHDPS) by using several data mining techniques such as Decision Trees, Naïve Bayes and Neural Network.

Shantakumar, et al. [11] have done a research work in which the intelligent and effective heart attack prediction system is developed using Multi-Layer Perceptron with Back-Propagation. Accordingly, the frequency patterns of the heart disease are mined with the MAFIA algorithm based on the data extracted.

Yanwei, et.al [12] have built a classification method based on the origin of multi parametric features by assessing HRV (Heart Rate Variability) from ECG and the data is pre-processed and heart disease prediction model is built that classifies the heart disease of a patient.

3. MOTIVATION AND JUSTIFICATION

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using three classification algorithms namely Naïve Bayes, Decision Tree, and Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, the three algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better understanding and help them identify a solution to identify the best method for predicting the heart diseases.

4. DATASET DESCRIPTION

The database for this research work has been taken from the StatLog dataset in UCI repository. It includes 13 attributes. The heart disease dataset included in this research work consists of total 270 instances with no missing values. The dataset is typically used for various types of heart diseases such as typical angina, atypical angina, and non-anginal pain and asymptomatic. This research work is aimed at predicting the heart disease irrelevant of the disease types. The attribute is a numeric data type that represents the age of the patient and ranges from 29 to 65 years. The Cp is an attribute for determining the pain type, represented from the range1 to 4. The trestbpd is a resting blood pressure that lies between 92 and 100; the fbs is fasting blood sugar level that is either a 1 or 0 representing Boolean values true or false. The restecg is the resting electro cardio graphic result represented as three cases from 0 to 2. The thalach is the maximum heart rate achieved ranging from 82 to 185. The exang is the exercise induced angina that is a Boolean value. The disease is the target class of the dataset denoting the heart disease presence with a yes or a no. Similarly, all the attributes and their values are represented in Table.1.

Table.1. Attribute and Description of the dataset used for research

S. No.	Attribute Name	Type	Description	Range
1	Age	Numeric	Age in years	29-65
2	Sex	Nominal	Sex in number	Male = 0, Female = 1
3	Cp	Nominal	Chest pain type	typical angina = 1, atypical angina = 2, non-anginal pain = 3, asymptomatic = 4
4	trestbpd	Numeric	Resting blood pressure	92-200
5	serumCho	Numeric	serum cholesterol in mg/dl	126-564
6	fbs	Nominal	Fasting blood sugar level	Yes =1, No = 0
7	restecg	Nominal	Resting electrocardiographic results	Normal = 0, having ST-T wave abnormality=1, showing probable or definite left ventricular hypertrophy = 2
8	thalach	Numeric	Maximum heart rate achieved	82-185
9	exang	Nominal	Exercise induced angina	Yes = 1, No = 0
10	oldpeak	Numeric	ST depression induced by exercise	71-202
11	peakSlope	Numeric	the slope of the peak exercise ST segment	1-3
12	numVessels	Numeric	number of major vessels (0-3) coloured by fluoroscopy	0-3
13	thal	Nominal	The defect type of the heart	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Disease	Nominal	Identification of a heart attack.	Yes=2, No=1

5. METHODOLOGY

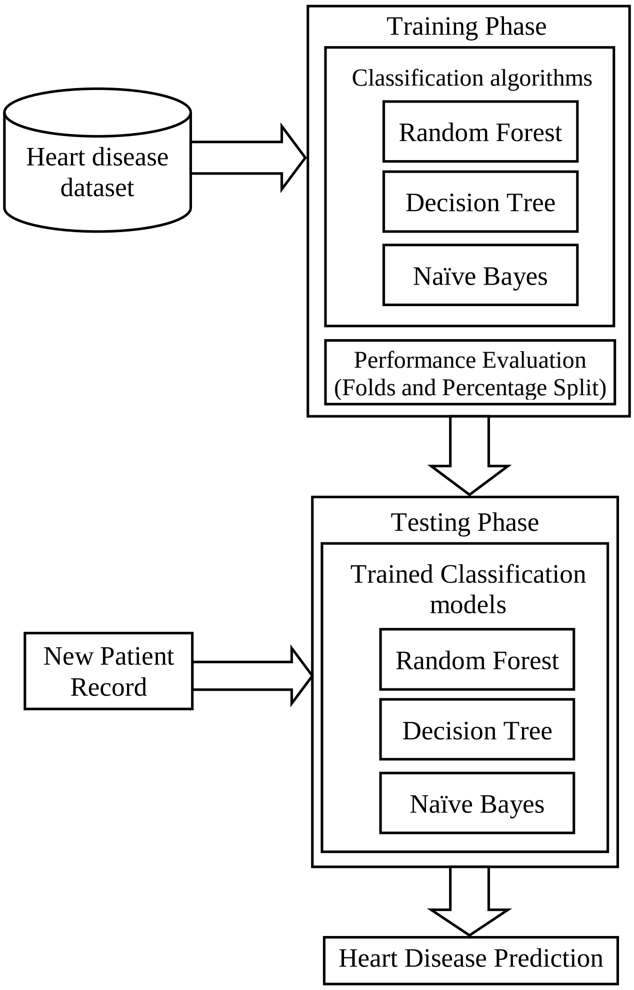

Fig.1. Methodology of the research work

The heart disease prediction can be performed by following the procedure which is similar to Fig.1 which specifies the research methodology for building a classification model required for the prediction of the heart diseases in patients. The model forms a fundamental procedure for carrying out the heart disease prediction using any machine learning techniques. In order to make predictions, a classifier needs to be trained with the records and then produce a classification model which is fed with a new unknown record and the prediction is made. The research methodology of this research includes the Performance Evaluation of the three classification algorithms i.e. Evaluation using cross validation and evaluation using percentage split. In the cross validation, the training and testing data is split up from the heart disease using several folds such as 10 folds and where each folds are recursively used for training and testing by replacement in the dataset for testing and training. It is discussed in detail in section 7.1. In the percentage split, the training and testing data is split up in percentage of data such as 80% and 20% where the 80% is used for training and 20% is used for testing. It is discussed in detail in section 7.2. In this work, the training phase includes training the three classification algorithms namely Naïve Bayes, Decision Tree, and Random Forest using the heart disease dataset and a classification model is built. All the three algorithms are described in the sections given below.

5.1 CLASSIFICATION USING RANDOM FOREST

Random forests (RF) [13] are combination of tree predictors using decision tree such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. They are more robust with respect to noise. It is a supervised classification algorithm used for the prediction and it is considered as the superior due to its large number of trees in the forest giving improved accuracy than decision trees. Typically, the trees are trained independently and the predictions of the trees are combined through averaging. Random forest algorithm can use both for classification and the regression based on the problem domain. The algorithm for random forest is given below:

Step 1: Randomly select k features from entire m features, where $k << m$.

Step 2: Surrounded by the k features, calculate the node “d” using the best split point.

Step 3: Split the node into daughter nodes using the best split.

Step 4: Repeat 1 to 3 steps until l number of nodes has been reached.

Step 5: Construct forest by repeating steps 1 to 4 for n number times to create n number of trees.

Firstly, the k features are taken out of total m features. In the next stage, in each tree randomly select k features in order to find the root node by using the best split approach. The next stage involves calculating the daughter nodes using the same best split approach for the heart disease dataset. Similarly, the tree is formed from the root node and until all the leaf nodes are generated from the attributes. This randomly created tree forms the random forest that is used for making heart disease prediction in patients.

5.2 CLASSIFICATION USING DECISION TREE

Decision Tree (DT) [14] is a simple and easy to implement classifier. The bit through feature to access in depth patients’ profiles is only obtainable in Decision Trees. Decision tree builds classification or regression models in the structure of a tree making it simple to debug and handle. Decision trees can handle both categorical and numerical data. The algorithm works by finding the information gain of the attributes and taking out the attributes for splitting the branches in threes. The information gain for the tree is identified using the below given Eq.(1).

$$E(S)= -P(P)\log_2 P(P)-P(N)\log_2 P(N) \qquad (1)$$

The algorithm for the decision tree is given below:

Step 1: Identify the information gain for the attributes in the dataset.

Step 2: Sort the information gain for the heart disease datasets in descending order.

Step 3: After the identification of the information gain assign the best attribute of the dataset at the root of the tree.

Step 4: Then calculate the information gain using the same formula.

Step 5: Split the nodes based on the highest information gain value.

Step 6: Repeat the process until each attributes are set as leaf nodes in all the branches of the tree.

5.3 CLASSIFICATION USING NAÏVE BAYES

Naïve Bayes (NB) is a statistical classifier which assumes no enslavement between attributes. Naive Bayes [15] is based on Bayes rule and it assumes that attributes are independent of each other. The working principle of naïve Bayes classifier is as follows:

- *Training Step*: By assuming predictors to be conditionally independent given for a class, the method estimates the parameters of a probability distribution known as the prior probability from the training data.
- *Prediction Step*: For unknown test data, the method computes the posterior probability of the dataset which is belonging to each class. The method finally classifies the test data based upon the largest posterior probability from the set

6. PERFORMANCE METRICS

The metrics used for the research work is described in this section

6.1 PRECISION

Precision is the part of significant instances between the retrieved instances. The Eq.of precision is given in Eq.(2)

$$Precision = TP/(TP+FP) \qquad (2)$$

6.2 RECALL

Recall is the small part of appropriate instances that have been retrieved over the total quantity of relevant instances. The Eq.of recall is given in Eq.(3).

$$\text{Recall} = TP/(TP + FN) \qquad (3)$$

6.3 F-MEASURE

The f-score (or f-measure) is considered based on the two times the precision times recall divided by the sum of precision and recall. The equation of F-Measure is given in Eq.(4).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

6.4 ROC AREA

Roc Curves are commonly used to show in a graphical way the connection/ trade off involving clinical sensitivity and specificity for every potential cut off for a test or an arrangement of tests.

6.5 PRC AREA

The Precision-recall curves are not impacted by the count of patients without disease and with low test results. It is extremely suggested to use precision-recall curves as a supplement to the regularly used ROC curves to obtain the full picture when evaluating and comparing.

7. EXPERIMENTAL RESULTS

The analysis and identification of the best classification algorithm in this research work is done and the results are provided here. For the validation of the results, several range of experiments are carried out using Cross validation and Percentage split methods which are described in the sections given below.

7.1 CLASSIFICATION USING CROSS VALIDATION

In *k*-fold cross-validation, the innovative sample is randomly partitioned into *k* subsamples. Then *k* subsamples, a single subsample is retained as the validation data designed for testing the representation, and the remaining *k*-1 subsamples are used as training data. This kind of situation is referred to as use training set generally referring to as entirely utilizing the dataset for training and testing.

For instance, in 10-fold cross validation, the original sample is randomly partitioned into 10 subsamples. From the 10 subsamples, a single subsample is retained as the validation data i.e. testing data which is used for testing the model and the remaining 9 subsamples is treated as training data that is used for training the classification algorithm. The cross validation process is then repeated in the same manner for 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. By shuffling and swapping the folds of data, the 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The experimental results of the classification of heart disease done using many folds of cross validation are given in Table.2. Here, (*A*) refers to the algorithms and (*F*) refers the number of folds in cross validation.

Based on the Table.2, the folds play a vital role in the improvement of the metric values for the precision, recall and f-measure. Although the performance has been improved based on implementation of several folds, the random forest algorithm outperforms the Decision Tree and Naïve Bayes algorithms.

7.2 CLASSIFICATION USING PERCENTAGE SPLIT

In percentage split, the data is split for both the training and testing. It actually splits the data and separates *x*% of the data for wisdom and the rest of it for testing. It is useful when your algorithm is time-consuming.

Table.2. Classification of heart disease using Cross Validation

(A)	(PS)	Metrics							
		TP Rate	FP Rate	Precision	Recall	F-measure	MCC	ROC Area	PRC Area
NB	60-40	0.56	0.56	0.313	0.56	0.401	0	0.5	0.507
	70-30	0.571	0.571	0.327	0.571	0.416	0	0.5	0.51
	75-25	0.577	0.577	0.333	0.577	0.422	0	0.5	0.512
	80-20	0.595	0.595	0.354	0.595	0.444	0	0.5	0.518
DT	60-40	0.714	0.311	0.714	0.714	0.71	0.415	0.846	0.815
	70-30	0.698	0.328	0.696	0.698	0.696	0.377	0.832	0.822
	75-25	0.712	0.321	0.709	0.712	0.708	0.401	0.811	0.8
	80-20	0.667	0.377	0.662	0.667	0.663	0.296	0.816	0.815
RF	60-40	0.753	0.283	0.756	0.75	0.774	0.492	0.733	0.682
	70-30	0.746	0.283	0.745	0.746	0.743	0.476	0.778	0.74
	75-25	0.731	0.282	0.731	0.731	0.731	0.448	0.768	0.773
	80-20	0.714	0.288	0.721	0.714	0.716	0.42	0.742	0.711

Table.3. Classification of heart disease using Percentage Split

(A)	**(F)**	**Metrics**							
		TP Rate	**FP Rate**	**Precision**	**Recall**	**F-measure**	**MCC**	**ROC Area**	**PRC Area**
NB	2	0.56	0.56	0.313	0.56	0.402	0	0.498	0.506
	5	0.56	0.56	0.313	0.56	0.402	0	0.488	0.501
	8	0.56	0.56	0.313	0.56	0.402	0	0.481	0.497
	10	0.56	0.56	0.313	0.56	0.402	0	0.482	0.498
DT	2	0.775	0.24	0.775	0.775	0.774	0.541	0.787	0.747
	5	0.785	0.241	0.788	0.785	0.781	0.562	0.821	0.797
	8	0.775	0.23	0.775	0.775	0.775	0.544	0.802	0.751
	10	0.77	0.241	0.77	0.77	0.77	0.532	0.819	0.774
RF	2	0.789	0.224	0.789	0.789	0.789	0.571	0.86	0.842
	5	0.804	0.201	0.804	0.804	0.804	0.602	0.864	0.847
	8	0.799	0.207	0.799	0.799	0.799	0.592	0.861	0.841
	10	0.809	0.192	0.81	0.809	0.809	0.614	0.864	0.848

For instance, 60%-40% percentage split the classification results will be evaluated on a test set that is a part of the original data. The percentage split is 60%, which means that 60% of the data go for training and 40% for testing. The classification model is built based on this and the experiment is conducted. The experimental results of the classification of heart disease done using many splits of percentage splitting are given in Table.3. Here, (*A*) refers to the algorithms and (*PS*) refers the range of percentage to split i.e. for e.g. 60-40 refers to splitting of the dataset as 60% and 40%. The first is to be utilized for training the classifier and the latter is to be used for testing the classifier.

The results described in the above Table.3 are achieved by splitting the dataset into percentage for training and testing part that are taken as randomly stratified data points. The metric results show that the values for the precision, recall and f-measure of the Random Forest algorithm are higher. Finally, the results show that the Random Forest is best suited for the prediction of heart disease than the Decision Tree and Naïve Bayes classification algorithms.

8. CONCLUSIONS AND FUTURE WORK

The overall objective of the work is to predict more exactly the occurrence of heart disease using data mining techniques. In this research work, the UCI data repository is used for performing the comparative analysis of three algorithms such as Random Forest, Decision trees and Naive Bayes. From the research work, it has been experimentally proven that Random Forest provides perfect results as compare to Decision tree and Naive Bayes.

The Future work of this research work can be made to produce an impact in the accuracy of the Decision Tree and Bayesian Classification for additional improvement after applying genetic algorithm in order to decrease the actual data for acquiring the optimal subset of attribute that is enough for heart disease prediction. The automation of heart disease prediction using actual real time data from health care organizations and agencies which can be built using big data. They can be fed as a streaming data and by using the data, investigation of the patients in real time can be prepared.

REFERENCES

[1] Carlos Ordonez, “Improving Heart Disease Prediction using Constrained Association Rules”, Technical Seminar Presentation, University of Tokyo, 2004.

[2] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, “Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method”, *Studies in Health Technology and Informatics*, Vol. 107, No. 2, pp. 1256-1259, 2004.

[3] W.J. Frawley and G. Piatetsky-Shapiro, “Knowledge Discovery in Databases: An Overview”, *AI Magazine*, Vol. 13, No. 3, pp. 57-70, 1996.

[4] Heon Gyu Lee, Ki Yong Noh and Keun Ho Ryu, “Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV”, *Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining*, pp. 56-66, 2007.

[5] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, “Associative Classification Approach for Diagnosing Cardiovascular Disease”, *Intelligent Computing in Signal Processing and Pattern Recognition*, Vol. 345, pp. 721-727, 2006.

[6] 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, pp. 1-8, 2008.

[7] Niti Guru, Anil Dahiya and Navin Rajpal, “Decision Support System for Heart Disease Diagnosis using Neural Network”, *Delhi Business Review*, Vol. 8, No. 1, pp. 1-6, 2007.

[8] Sellappan Palaniappan and Rafiah Awang, “Intelligent Heart Disease Prediction System using Data Mining Techniques”, *International Journal of Computer Science and Network Security*, Vol. 8, No. 8, pp. 1-6, 2008.

[9] Shantakumar B. Patil and Y.S. Kumaraswamy, “Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network”, *European Journal of Scientific Research*, Vol. 31, No. 4, pp. 642-656, 2009.

[10] X. Yanwei et al., “Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease”, *Proceedings of International Conference on Convergence Information Technology*, pp. 868-872, 2007.

[11] Ersen Yilmaz and Caglar Kilikcier, “Determination of Patient State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree”, Master Thesis, Department of Electrical Electronic Engineering, Uludag University, 2013.

[12] Nidhi Singh and Divakar Singh, “Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time”, Ph.D Dissertation, Department of Computer Science and Engineering, Barkatullah University Institute of Technology, 2012.

[13] A. Liaw and Matthew Wiener, “Classification and Regression by Random Forest”, *R News*, Vol. 2, No. 3, pp. 18-22, 2002.

[14] J. Ross Quinlan, “Induction of Decision Trees”, *Machine Learning*, Vol. 1, No. 1, pp. 81-106, 1986.

[15] K. Ming Leung, “Naive Bayesian Classifier”, Master Thesis, Department of Computer Science and Engineering, Polytechnic University, 2007.