Prognosis of Heart Disease using Data Mining Techniques: A Comprehensive Survey

Ankita Naik
Computer Engineering Department
Goa College of Engineering
Farmagudi-Goa, India

Nitesh Naik
Computer Engineering Department
Goa College of Engineering
Farmagudi-Goa, India

ABSTRACT
Prediction and diagnosis of heart disease has become a formidable factor faced by medical practitioners and hospitals both in India and also worldwide. The early and timely diagnosis of heart disease plays a very crucial role in halting its advancement and reducing related medical costs. Taking into account the ever-increasing rise in heart disease-induced mortality, different techniques have been adopted to treat it. The idea intends to develop a heart disease prediction model, which will implement ensemble techniques, can help the doctors in detecting the heart disease status based on the patient's clinical data. This paper provides a quick and facile analysis and understanding of available prediction models using data mining from 2011 to 2017. The comparison shows the accuracy level of each model given by different researchers.

General Terms
Data Mining , Machine Learning , Artificial Intelligence.

Keywords
Prediction, heart disease, classification, ensemble, diagnosis

1. INTRODUCTION
Data mining is a novel field for finding out the patterns present in hidden information from huge raw data sets. In the modern world, cardiovascular diseases are the most commonly occurring diseases and in every year more than 13 million deaths occur worldwide due to heart problems. Cardiovascular diseases also cause maximum deaths in India and its diagnosis is a very difficult process. The rising high-performance computing has profited various disciplines in finding realistic solutions to their problems. Our health care is no special case to this. Data mining tools have been devised for effective study of medical data, so that it is able to aid medical practitioners to make a better diagnosis for treatment purposes.

In cardiovascular disease research, data mining technique have accomplished a crucial role. Different interpretations and explanations between the healthy persons and the diseased persons can help one in the research of cardiovascular (heart) related disease classification to find the hidden medical data. The Health care provides various services which are used to: (1) improve quality and efficiency; (2) occupy patients and families; improve care coordination, and society and public health; and (3) Maintain seclusion and safety of patient health information [11].

The most dominant health issue is heart failure which occurs often in elderly patients because of diet, anti-steroidal non-inflammatory drugs and can lead to even death. One of the frequently occurring diseases is cardiovascular diseases. Thus it is necessary to anticipate such diseases through suitable symptoms. There are different variety of algorithms which are present for the prognosis of heart diseases which are Decision Trees, Naive Bayes, Bayesian Net etc. Unfortunately all doctors do not acquire proficiency in every domain of specialization and moreover there is paucity of resource persons at certain places. Therefore, an automated medical diagnosis system would probably be more favorable for bringing the efficient and accurate result [11].

2. LITERATURE SURVEY
Different types of studies have been utilized in prognosis of heart disease. Numerous data mining techniques are utilized for diagnosis and have achieved different accuracy levels for different methods.

Mai Shouman, Tim Turner, Rob Stocker et al. (2011) have used a model that increased the decision tree accuracy in recognizing patients having heart disease. Multiple classifiers voting technique with different types of Discretization methods and several types of decision trees were combined by the model. Different types of decision tree impurity measures such as information gain, gini index and gain ratio are integrated with the various Discretization methods such as equal width, equal frequency, chi merge and entropy. The steps in research process includes methods like data discretization, data partitioning, Decision Tree type selection, and the application of reduced error pruning to produce a pruned Decision Tree that will reduce the overfitting problem. The data discretization is spilt into supervised and unsupervised methods. The data partitioning method has testing with and without voting. From the results it is decided that even though many of the researchers are utilizing the binary discretization with Gain Ratio Decision Tree for diagnosis of heart disease, using multi-interval equal frequency discretization with nine voting Gain Ratio Decision Tree gave good results in the determination of heart disease patients. The supervised discretization methods did not unveil any improvement in the Decision Tree accuracy either with or without voting. Applying voting indicated increase in the accuracy levels of different kinds of Decision Trees. The performance of the decision tree was also computed [2].

Chaitrali S. et al. (2012) used 13 attributes like sex, blood pressure, and cholesterol for prophecy of heart disease. Two more attributes called smoking and obesity was added. DM classification methods utilized were Neural Net, Decision Tree and Naïve Bayes. Neurat Network obtained an accuracy of 100%, decision tree got 99.62% accuracy and Naïve bayes got an accuracy of 90.74% respectively. Confusion matrix was found for three classification approaches for dataset of 13 attribute and 15 attribute data set. The accuracy with 15 attribute was 100% for neural network [18].

Shamsher Bahadur Patel et al. (2013) reduced 14 attributes to six by making use of the Genetic Algorithm. Then Classifiers Naïve Bayes, classification by clustering & Decision Tree were used to predict the analysis of Heart Disease. Genetic

search was useful on the 14 attributes and no. of attributes was reduced to 6. The deduced dataset was used on three classification models. For scheming the model four computing estimation measures used were namely, True positives refers to positive tuple, True Negative refers to negative tuple, False Negative and False positive. Implementation was done using the weka tool. Accuracy obtained by Decision Tree, Naïve Bayes, Classification using clustering were 99.2%, 96.5% and 88.3% respectively [15].

I.S.Jenzi et al. (2013) designed a consistent classifier model using data mining technique, Association rules and classification techniques like decision tree, Naïve Bayes and Neural Network. The data set that was used by them contained 14 attributes, where class attribute is well thought-out at the end of all attributes. The GUI was created in Microsoft .NET platform, with interconnections completed by using IKVM interface with some Java runtime libraries. The ROC curves was a tool used to symbolize the accuracy. The results received was represented on ROC which has the area under ROC of data mining was 0.807 which is found to be better than Naïve Bayes [16].

Hlaudi Daniel Masethe et al. (2014) presented a scheme using data mining algorithms namely J48, NB, REPTREE CART and Bayes Net for predicting heart attacks. The data set was collected from hospital and doctors who were practitioners in South Africa. 11 attributes that were considered by them are as follows: patient Identification Number, Gender, age, chest pain, cardiogram, BP , rating of heartbeat, cholesterol, tobacco 259 consumption, hot drinks intake & diabetes level. The tool called Waikote Environment for knowledge Analysis (WEKA) was used for prediction of heart disease. WEKA tool was remarkable in identifying, analyzing and predicting patterns. Accuracy obtained were 99.0741, 99.222, 98.148, 99.0741 for J48, REPTREE, Naïve Bayes, Bayes Net and simple CART algorithms respectively i.e. Bayesian Net algorithm outperformed the NB algorithm [7].

Lokanath Sarangi et al. (2015) created an efficient system by using Genetic Algorithm optimizer method. The optimized weights were given as an input to the specified network. The accuracy obtained was 90% the hybrid method of Genetic Algorithm and Neural networks was used [17].

Purushottama , Prof. (Dr.) Kanak Saxena, Richa Sharma et al. (2016) have designed an efficient Heart Disease Prediction System using data mining. The system was trained and tested using 10 fold method and an accuracy of 86.3 % in testing phase and 87.3 % in training phase was obtained. The proposed study used covering rules model for classification. The rules produced as output by the proposed system were ranked as Original Rules, Pruned Rules, and Rules without duplicates, Classified Rules, Sorted Rules and Polish. The top-down decision tree was constructed. A test for the actual node was selected in each step, which best segregates the given examples by classes. A hill climbing algorithm was then performed in order to find the best subset of rules. After the rules were generated performance was evaluated. The performance of this efficient heart based prediction system was compared with algorithms like svm, neural networks, part, multi layer perceptron etc. It was concluded that the system gave an accuracy of 86.7% which was highest in comparison with other algorithms [3].

Theresa Princy, J. Thomas et al. (2016) have used the KNN and ID3 algorithm for prediction. They have developed a system that contains two modules Initial module consist of classifier module and second module consist of prediction module. In Classifier module data was trained through KNN algorithm and classified. All the input parameters were observed and based on the attribute age the data were classified using KNN algorithm. This classified data was provided to test data. In Prediction module data was tested and predicted through ID3 algorithm. All the classes were observed and each class was verified to find the risk level of the heart disease. These risk values were obtained from the test data compared with the model generated for the systolic input factor. The accuracies of different number of attributes were found and plotted. The accuracy of the prediction was increased by adding additional attributes such as previous heart disease rate and smoking. The accuracy level of the prediction was found to be 40.3% when the basic attributes such as blood pressure, cholesterol, Pulse rate, age and gender were used. When two additional attributes such as smoking and previous heart disease were added the accuracy level of the prediction was increased up to 80.6%. They found that by adding number of prominent attributes increased the accuracy of the prediction system [5].

Ilayaraja M, Meyyappan T et al. (2016) have developed a method to generate frequent itemsets based on the user's clinical data (symptoms). The findings helped them to forecast the risk extent of patients affected. Frequent itemsets were produced based on the selected symptoms and minimum support value. The acquired frequent itemsets helped the doctors to make diagnostic conclusions and helped them to know the possibility of risks in patients at an early stage. The method can be applied to any medical dataset to predict the probability of risks with risk level of the sufferers based on selected factors. The study showed that the developed method could find the probability level of patients efficiently from frequent item sets. Besides this they have compared the performance of this method with methods like apriori, semi-apriori and association rule mining algorithm based on pattern generation[4].

Randa El Bialy , Mostafa A. Salama ,Omar Karam et al. (2016) developed an ensemble model for diagnosing the heart disease using techniques like bagging, boosting and stacking. The study involved use of different methods that adopt different strategies to combine the predictions made by numerous classifiers that are diverse are used. Two datasets namely collective heart disease dataset (CAD) and heart sound signals dataset for heart valve disease. Then, outlier detection and noise removal is done for both the datasets. The researchers have used six classifiers such as Naïve Bayes, Bayesian Network, MLP, and SMO. They have combined the classifiers using ensemble techniques and finally based on accuracies the best combination of classifiers are selected. The best classifier accuracy for the HVD dataset when applying a single classifier is Naïve Bayesian and the second best classifier is sequential Minimum Optimization for support vector machine as for the CAD dataset is Naïve Bayesian and second best classifier is Bayesian Net [1].

Ritika Chadha , Shubhankar Mayank et al. (2016) developed a prediction system by using ANN, Decision Tree and Naive Bayes methods. They implemented it by using C# and also used the Python platform. According to their research study, the prediction rate or accuracy for each of the data mining technique was computed. Based on the observations/technical experiments, it was found that Artificial Neural Networks gave highest accuracy followed by Decision Tree and Naive Bayes respectively. The accuracies of each of the technique are as follows:ANN achieved an accuracy of 90%, Decision tree got accuracy of 88.02% and accuracy of 85.86% was

obtained by the Naïve Bayes algorithm [9].

S. Radhimeenakshi et al. (2016) used SVM and ANN techniques for classification and prediction of disease risks. Two datasets namely Staglog and Cleveland datasets were used. The datasets were divided into training, validation and test sets. SVM and ANN were used to arrange the datasets into two classes. Performance evaluation was done in terms of accuracy, precision and sensitivity which is applied on both datasets and performance is evaluated. The outputs from the SVM classification are compared to the outputs of ANN classification which are extremely good. For Staglog dataset SVM model gave an accuracy of 84.7% and ANN gave an accuracy of 81.8% respectively. The precision equivalence of SVM and ANN obtained in the classification of Staglog output is 85.6% and 83.3% respectively. The sensitivity of SVM and ANN obtained was 84.12% and 68.7% [8].

Aigerim Altayeva , żSuleimenov Zharas ,Young Im Cho et al. (2016) , have designed a Heart Disease Prediction System (HDPS) using Naive Bayes and K means clustering algorithms which are one of the highly used clustering methods. The initial choice of the centroid has effect on the final result. The efficacy of unsupervised learning techniques, which are k-means clustering to improvise the learning technique, which is Naive Bayes, is displayed. Evaluation of the combination of K-means clustering with Naive Bayes in the detection of diseased patients was carried out. Several mechanisms involved in selection of initial centroid of the K-means clustering algorithm such as range, inliers, outlier, random attribute values, and random row methods were evaluated by them for the recognition of cardiovascular patients. The outcomes have shown that the combination of the K-means clustering with Naïve Bayes with variation in initial centroid selection, the Naive Bayesian improves accuracy in identification of the patient [10].

Shan Xu, Zhen Zhang, Daoxian Wang, Junfeng Hu, Xiaohui Duan et al. (2017) used a risk prediction method based on CFS Subset Evaluation and Random Forest algorithm. In Feature Selection, CFS Subset Evaluation strategy and Best-First-Search method were combined to reduce dimensionality. In Classification, after many tests and experiments of different kinds of data mining methods, Random Forest proved to be best classifier than others, which is also a prior trial in CVD risk prediction field. Two data source were both tested to confirm accuracy as well as practicality. In CHDD test, the system had a notably greater accuracy of 91.6% than all the other methods. It achieved an accuracy of 97% in People's Hospital dataset test. This was superior than most of other classifiers except SVM which achieved an accuracy of 98.9%. However random forest only took half of time than SVM. Taking into consideration the risk prediction system showed good importance in accuracy and practical use for patient's treatment and doctor's diagnosis [6].

3. DATA DESCRIPTION

Standard heart disease dataset from UCI repository has been utilized for training and testing purpose in almost all the studies. The dataset has 76 attributes, but only 14 of them have been utilized in order to obtain more accurate results.

The dataset consists a total of 303 instances, of which 164 are healthy and 139 have heart disease. 297 rows do not contain any missing values and 6 rows have missing values that are shown by -9 and replaced by the attribute mean . No outliers and noise was found by the researchers as the data has already been processed by them. Table I displays the chosen heart disease dataset attributes [14].

TABLE I. Heart Disease Dataset Information

Name	Possible Values	Description
Age	Continuous	age in years
gender	1=male 0=female	gender
Cp	1:typical angina 2:atypical angina 3:non-anginal pain 4:asymptomatic	Chest pain type
trestbps	Continuous	Resting blood pressure(in mm Hg on admission to the hospital)
chol	Continuous	Serum cholesterol in mg/dl
fbs	1:true 0:false	Fasting blood sugar >120
restecg	0:normal 1: having ST-T wave abnormality 2: left ventricular hypertrophy	Resting electrocardiographic results
Thalach	>0	Maximum heart rate achieved
exang	1: yes 0:no	Exercise induced angina
oldpeak	continuous	ST depression induced by exercise relative to rest
slope	1:upsloping 2:flat 3:downsloping	The slope of peak exercise ST segment
ca	0-3	Number of major vessels colored by fluoroscopy
thal	1:Normal 2:Fixed defect 3:Reversable defect	Thalium Stress Test Results
AHD	1:Yes 0:no	Depicts if a person has heart disease or no

4. COMPARISON TABLE

The table below lists the different data mining algorithms and methods used by various researchers used in their studies along with the accuracies obtained by them for each technique.

TABLE II. Comparison Table

Year	Author	Data Mining Technique/ Algorithm	Accuracy
2011	Mai Shouman, Tim Turner, Rob Stocker	Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree	84.1%
2012	Chaitrali S	Neural Network, Decision Tree, Naïve Bayes	100%, 99.62%, 90.74%
2012	Mai showman	SVM with Bagging Algorithm	84.1%
2013	Shamsher Bahadur	Decision Tree, Naïve Bayes, Classification using Clustering	99.2%, 96.5%, 88.3%
2013	I.S. Jenzi	Naïve Bayes	80.7%
2014	B.Venkatalakshmi, M.V Shivsankar	Naïve Bayes (NB), Decision Tree(DT)	85.03% , 84.01%
2014	Hlaudi Daniel Masethe, Mosima Anna Masethe	J48, REPTREE, Naïve Bayes, Bayes Net CART	99.07%,99.07%, 97.22%,98.148%, 99.07%
2015	Ilayaraja M, Meyyappan T	Association Rule Mining Algorithms	85%
2015	Lokanath Sarangi	GA Technique	90%
2016	Theresa Princy. R ,J. Thomas	K-NN ,ID3	80.6%
2016	Ritika Chadha , Shubhankar Mayank	Artificial Neural Networks(ANN), Naive Bayes and Decision Tree	85.86% ,88.02%
2016	Jagdeep Singh, Amit Kamra, Harbhag Singh	IBk with Apriori Algorithm	99.19%
2016	Purushottama , Prof. (Dr.) Kanak Saxena, Richa Sharma	Rule based classifier	86.7%
2016	Randa El Bialy ,Mostafa A. Salama,Omar Karam	Bagging Boosting Stacking	90% ,90%,91%
2017	Luo Y, Li Z, Guo H, Cao H, SongC, Guo X	WSVM,Logit, WRF	94.76%,98.08%,92.98%
2017	Shan Xu ,Zhen Zhang, Daoxian Wang, Junfeng Hu, Xiaohui Duan	Random Forest,C4.5,SVM,Bayes,RBF, Adaboost	97.0%,87.7%,98.9%,54.1%,51.8%, 52.6%

5. CONCLUSION

In this paper, a survey conducted, gives the idea of different models available and the various data mining techniques used from 2011 up to 2017 for prognosis of cardiac disease. The accuracy obtained with these models is also mentioned. Different data mining techniques and classifiers discussed in these studies are used for efficient and efficacious heart disease diagnosis. As per the analysis mode, it is seen that many authors use various technologies and different number of attributes for their study. Hence, different technologies have given different results depending on a number of attributes that are considered. In future, Fuzzy models can be applied to know the intensity of the cardiac disease.

6. REFERENCES

[1] Randa El Bialy, Mostafa A. Salama, Omar karam. 2016. An ensemble model for Heart disease data sets: a generalized model, ACM May 2016.

[2] Mai Shouman, Tim Turner, Rob Stocker ,"Using Decision Tree for Diagnosing Heart Disease Patients , ACM,2011.

[3] Purushottama , Prof. (Dr.) Kanak Saxena, Richa Sharma "Efficient Heart Disease Prediction System", Elsevier ,2016.

[4] Ilayaraja M, Meyyappan T , " Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets", Elsevier 2016.

[5] Theresa Princy ,J. Thomas "Human Heart Disease Prediction System using Data Mining Techniques" ,IEEE ,2016.

[6] Shan Xu ,Zhen Zhang, Daoxian Wang, Junfeng Hu, Xiaohui , "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework" ,IEEE 2017 .

[7] Hlaudi Daniel Masethe, Mosima Anna Masethe , "Prediction of Heart Disease using Classification Algorithms", Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[8] S Radhimeenakshi , "Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural networks" ,IEEE, 2016.

[9] Ritika Chadha ,Shubhankar Mayank , "Prediction of heart disease using data mining techniques" ,Springer , December2016.

[10] Aigerim Altayeva , żSuleimenov Zharas ,Young Im Cho "Medical Decision Making Diagnosis System Integrating k-means and Naïve Bayes algorithms" , IEEE October 2016.

[11] B.Venkatalakshmi, M.V Shivsankar, "Heart disease diagnosis using predictive data mining", ICIET, March 2014.

[12] Jagdeep Singh, Amit Kamra, Harbhag Singh , "Prediction of Heart Diseases Using Associative Classification" ,IEEE ,2016.

[13] Saba Bashir, Usman Qamar, M.Younus Javed, "An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis" ,IEEE 2014.

[14] https://archive.ics.uci.edu/ml/datasets/Heart+Disease?spm =5176.100239.blogcont54260.8.TRNGoO

[15] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D.P. Shukla, "Predict the Diagnosis of Heart Disease Patients using classification Mining Techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), 2013.

[16] I.S.Jenzi, P.Priyanka, Dr.P.Alli, "A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction", International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

[17] Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik, "An Intelligent Decision Support System for Cardiac Disease Detection", IJCTA, International Press 2015.

[18] Chaitrali S. Dangare Sulabha S Apte, "Improve study of Heart Disease prediction system using Data Mining Classification techniques", International journal of computer application, 2012.

[19] Luo Y, Li Z, Guo H, Cao H, SongC, Guo X, et al..,Predicting congenital heart defects: A comparison of three data mining methods.PLoS ONE 12(5):e0177811,2017.

[20] Mai Shouman, Tim Turner, Rob Stocker, " Using data mining techniques in heart disease diagnosis and treatment", IEEE Japan-Egypt Conference on Electronics, Communications and Computers, 2012.