Survey on Data Mining Algorithm used for Heart Disease Prediction

[1]Poonam Faldu, [2]Sumitra Menaria, [3]Tejal Patel
[1]M.Tech-CE Student, [2]Professor, [3]Assistant Professor
[1-3]Dept. of Computer Science & Engineering
[1-3]Parul Institute of Engineering & Technology
Vadodara, India.

***Abstract*—** Data mining is one of the richest areas of research that is more popular in health organizations. Data mining plays an effective role for uncovering new trends in healthcare organization which is helpful for all the parties associated with this field. Heart disease is the leading cause of death in the world over the past 10 years. Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the irregular health condition that directly affects the heart and all its parts. The healthcare industry gathers enormous amount of heart disease data which are not "mined" to discover hidden information for effective decision making. Data mining techniques are useful for analyzing the data from many different dimensions and for identifying relationships. This paper explores the utility of various decision tree and neural network algorithms to classify and predict the disease.

***Index Terms*—** Heart Disease Prediction, Disease prediction Techniques, Medical diagnosis using Machine Learning, Data Mining.

I. INTRODUCTION

Medical Industry has large quantity of medical oriented data that is not refined. The datasets for the prediction has hidden patterns which is necessary for data analysis in the detection and prediction of heart disease. According to WHO it is reported that 12 million peoples are affected by heart disease [1]. If the blood circulation to the body is not proper the organs of the body like brain, heart, etc., stop working and death occurs. The common risk factors associated to heart disease are age, family history, smoking, alcohol intake, obesity, etc. The diagnosis of heart disease is based on patient' s BP, cholesterol level number of major vessels blocked, etc. It is also based on patient's Echocardiography (ECHO) and Electrocardiogram (ECG) test results and good experience of doctors. It is one of the difficult task to diagnose heart disease which needs prior knowledge along with good skills.

Data mining is a significant extraction of indirect and valuable historical data called as knowledge from the patient's information using complex algorithm [10]. In most of the sectors of medicine it proved its efficiency where the results obtained achieve higher accuracy and performance when compared with other methodologies. For the predictions, many new algorithms and techniques of data mining have been discovered by many researchers.

A. Existing Heart Disease Prediction Algorithms

The existing 11 algorithms based on the prediction of heart disease which has been used by various researchers are explained below.

Naïve Bayes (NB): A classification approach relied on Bayes hypothesis having a presumption of distinct factors [13]. The dataset is categorized in a distinct manner for predicting the heart disease occurrence. It is particularly used for very large data sets such as medical related data. To predict the heart disease through probability Naïve Bayes is used.

Decision Tree (DT): Decision tree is a type of supervised learning algorithm. It works for both categorical and continues input and output variables [13]. In the diagnosis of heart disease, decision tree will segregate the datasets based on all the values of the attributes and identify the attribute, which creates the best homogeneous sets of data. It also provides classified report for the heart disease.

K-Nearest Neighbor (KNN): For predictive problems such as classification and regression KNN is used [13]. KNN algorithm is mainly used to find the values of the factors of heart disease by using the k user defined value. By using the K values, it is possible to make boundaries for each class of heart disease related attributes.

Average K-Nearest Neighbor (AKNN): To make the KNN a faster algorithm AKNN is proposed by C. Kalaiselvi [4]. AKNN reduces the training sample size of heart disease dataset to n super samples. When the test samples of heart disease are given, the AKNN searches only the sample data and find the closest input thus making the KNN a faster algorithm.

Java Implementation of C4.5 algorithm (J48): J48(C4.5) is used to generate decision tree which is an extension of ID3 (Iterative Dichomester 3). For each heart disease dataset all the attributes are measured based on the type of chest pain. Based on the chest pain type the J48 algorithm grows an initial tree by using divide and conquers technique [7].

Reduces Error Pruning Tree (REP TREE): REP Tree is a speed decision tree algorithm. It constructs a regression tree using data gathering and removes it using reduced-error pruning.

Iterative Dichomester 3 (ID3): ID3 is the precursor to the C4.5 algorithms which is used to generate decision tree structure for the classified heart disease data. Different categories of the Heart disease data are investigated and every category are tested for heart disease prediction [5]. In each class the standard values are given to attributes such as Resting Blood Pressure, Serum Cholesterol, etc. In the tree structure each sub node represents the trained dataset of each class.

Support Vector Machine: SVM is a supervised machine learning algorithm which can be used for both classification and regression. It classifies the heart disease dataset by finding a hyperplane which separates the dataset into two classes. It performs better in terms of specificity and adaptability.

Artificial Neural Network (ANN): Artificial Neural Networks are the biologic outstanding act executed on the system to perform distinct functions such as classification, clustering, etc. The prediction of heart disease consists of the input layer, hidden layer, outputs and training classes which are based on the predicted class. It provides minimized error for the prediction of heart disease.

Fuzzy Analytical Hierarchical Process (Fuzzy AHP): Fuzzy AHP provides a hierarchical structure which shows the ultimate aim of the prediction of heart disease and its replacements are drafted. The local weight of each attribute and global weight of each attribute used in determining heart disease are calculated by using Fuzzy AHP. [2]

Feed Forward Neural Network (FFNN): The Feed Forward Neural Network is the first and simple type of artificial neural network except that the connections between the units do not form a cycle so that the information moves in a single direction. The given inputs are multiplied by weights, summed and combined by sigmoid activation to predict the output for the heart disease.

II. HEART DISEASE PREDICTION FACTORS

From the study it is found that several attributes have been used for better prediction. The following are the factors discussed by the various researchers. It is classified into Common Factors and Medical Oriented Factors.

A. Common Factors

Common Factors are the factors in which the values cannot be changed and obtained without any medical test.

Age in Years: Heart disease may occur at any phase of life. Specifically, in case of men heart attack might occur after the age of 45 and for women it occurs after the age of 55. The people in 20's and 30's also suffer from heart attack. (Values: Age in years (AGE)-Young (YNG) <33, Medium (MED) 34–40, Old (OLD) 41–52, Very Old (VOLD) >52)
[2]

Gender: It describes the gender (value 0 or 1) 1: Male; 0:
Female [4]

CPT-Chest Pain Type: Angina is a pain in the chest, caused when oxygen rich blood is not properly supplied to the heart muscle. Non-Angina chest pain is likely a chest pain with duration over 30 minutes or less than 5 seconds. (values: Typical Angina (TA) 1, Atypical (ATA) 2, Non-Angina pain (NAP) 3, Asymptomatic (ASY) 4) [2]

Ex-Ang-Exercise induced angina: Exercise induced angina occurs when the heart muscle does not get the oxygen it needs to function properly when you exercise. (values: True (YES) 1, False (NO) 0) [2]

Smoking: Smoking increases the risk of developing cardiovascular disease and coronary heart disease. Carbon monoxide in tobacco and nicotine in cigarettes makes the heart pump faster and raises blood pressure. (values: True (YES) 1, False (NO) 0) [7]

Alcohol: Long term intake of alcohol over a long period of time can lead to heart disease. It indirectly influences the risk of increase in blood pressure. It also weakens heart muscles so that heart can't pump blood efficiently. (values: True (YES) 1 False (NO) 0) [7]

B. Medical Oriented Factors

Medical Oriented factors are the factors in which the values can be changed and obtained from the medical test results.

FBS-Fasting Blood Sugar: A carbohydrate metabolism test is used to measure the blood sugar level. It is conducted after fasting. Where normal is 70 -108 mg/dl; High blood glucose level is 109 and above. (value 0 or 1) [4]

RECG-Resting electrocardiographic (ECG): The Resting electrocardiographic test measures the heart's electrical activity. Normally, the electric impulses cross the heart which contracts approximately 60 - 80 times per minute at rest. The left ventricular hypertrophy can be diagnosed when there is a large mass of myocardium for electrical activation to pass through (values: normal-0, ST-T wave abnormality- 1, definite left ventricular hypertrophy-2) [4]

SL: Slope (Peak exercise- ST): The Isoelectric section of the ECG which represents the S-wave, T-wave and the ST segment (values: Up-sloping (UPS) 1 Flat (FLT) 2 Down- sloping (DWS) 3) [2]

Col-Ves: It represents the count of the dominant vessels coloured by fluoroscopy (values: Fluoroscopy-0 (FL-0) 0 Fluoroscopy-1 (FL-1) 1 Fluoroscopy-2 (FL-2) 2
Fluoroscopy-3 (FL-3) 3) [2]

Thal: Thallium Scan taken when the heart's ability to pump is impaired. In a normal flow, blood is supplied to the entire heart muscle. In reversible defect the blood flows less to the heart muscle which takes place due to insufficient blood supply from a specific coronary artery. In fixed defect, the blood flows less to the heart muscle which takes place as a result of permanently damaged muscle (Values: Normal (NOR) 3 Fixed defect (FDE) 6 Reversible defect(RDE) 7)
[2]

Serum Cholesterol(mg/dl): The Serum Cholesterol(SC) contains triglycerides, high density lipoprotein and low density lipoprotein. The normal level of Serum Cholesterol is 200 mg/dl. [4]

Thalach: It represents the maximal heart rate achieved. A valuation of a person's highest age-related value can be obtained from deducting the age of the person from 220. (values: Low (LOW) <112 Medium (MED) 112–152 High (HIGH) >152) [2]

Old Peak: ST depression which is caused due to exercise associated to rest. Coronary insufficiencies are the major cause for ST depression. While an exercise stress test requires an ST depression of at-least 2mm to significantly indicate reversible inchaemia. (values: Low (LOW) <1.5 Risk (RSK) 1.5–2.55 Terrible (TER) >2.55) [2]

RBP-Resting Blood Pressure: It is the pressure as the heart relaxes. The normal blood pressure is defined as 120mm systolic and 80mm diastolic. (values: Low (LOW) <128 Medium (MED) 128–142 High (HIGH) 143–154 Very High (VHIGH) >154) [2]

Num: The anticipated factors for the investigation of heart disease gives the angiographic disease condition. (value 0: Absent; value 1: Present) [11]

Obesity (Diet): It is calculated using BMI -for- age percentile. Being obese can increases the chances of Heart disease (CHD) and heart attack. Blood Pressure, Cholesterol and Sugar level indirectly influences obesity. (values of BMI: below 18.5 (less weight-skinny) – 0, 18.5 to 24.9 (Healthy-Physically fit) – 1, 18.5 and above(overweight- obese) - 2) [10]

III. HEART DISEASE PREDICTION – A SURVEY

A survey is carried out on different data mining techniques and the different parameters used for prediction of heart disease are discussed. The accuracy obtained with these models are also mentioned.

A. John Peter T et al., (2012) used classification technique for the heart disease prediction. Dataset of heart disease which is in the ARFF format uses 14 attributes which has large quantity of intrinsic linear combination of variables. The limitations of medical scoring systems are handled and the data is classified by using classification models that assign data in a collectio3n of target classes. The classifiers were implemented on the reduced data. Accuracy of the classifiers was 83.70% for Naïve Bayes, 76.66% for decision tree, 75.18% for K-Nearest Neighbour and 78.148% for Neural Network. [11]

B. Challenges: The dimensionality of the data is reduced using attribute selection method, since it consumes more time for classification. This decrease in the number of attributes does not give the correct prediction.

C. Chaitrali S et al., (2012) used classification methods such as Neural Network, Decision Tree and Naive Bayes for classifying 13 common attributes like age, gender, blood pressure, cholesterol, etc. for the prediction of heart disease. Two more attributes called smoking and obesity were also added. Confusion matrix was obtained for 3 classification methods using 13 attribute datasets and 15 attribute datasets. Accuracy obtained by these techniques were 100%, 99.62% and 90.74% respectively. [10]

Challenges: Prediction is carried out using 15 factors. Still indirectly influencing factors such as alcohol are not considered for the prediction. Since more attributes are used in the prediction, text mining could have been implemented for reducing the time for classification.

D. Jenzi.I.S et al., (2013) used data mining technique such as Decision Tree, Naïve Bayes and Neural Network for building a reliable classifier model. There are 14 factors in the data set. It identifies related patterns for quality decision making. [9]

Challenges: ARFF format is used in User interface which is not efficient for classification. Automatic Conversion to ARFF is not Enabled. Additional analysis in the preference of the factors is not accomplished.

E. Shamsher Bahadur Patel et al., (2013) used Decision tree and Naive Bayes in the prediction of heart disease. Genetic Algorithm is used for reducing 14 factors to 6. This reduced dataset was applied to 3 classification models. WEKA tool was used for implementation. [8]

Challenges: The Missing values and Inconsistencies are not resolved. In this case the intensity of the disease is unpredictable, which could have been resolved by applying fuzzy learning model.

F. Hlaudi Daniel Masethe et al., (2014) used data mining algorithms like J48, NB and REP TREE for predicting heart attacks. The medical database was collected from the doctors in South Africa. The various attributes considered were Gender, age, CPT, ECG, RBP, Thalach, serum cholesterol, alcohol, obesity(diet) and smoking. WEKA - Waikato Environment for Knowledge Analysis tool was used for discovering, analysing and predicting patterns for heart disease. The accuracy obtained were 99.0741, 99.222, 98.148 for J48, REPTREE and NB respectively. [7]

Challenges: Some important attributes such as Col-Ves, Thal, Ex-Ang, etc., are not considered for prediction. Here various data mining algorithms are implemented and compared to find the best method for prediction. They came to a conclusion that algorithms such as J48 and REPTREE are efficient in the prediction of heart disease. This conclusion is derived by not considering some efficient Data mining algorithms such as Regression and Artificial Neural Network algorithms.

G. Venkatalakshmi B et al., (2014) aims to predict heart disease using predictive mining. 13 factors from UCI Repository is taken as the source data, to compare the performance of DT and NB where the accuracies such as 84.01% and 85.03% was obtained respectively.

Challenges: Further investigation in attribute selection is not carried out. Here automatic conversion to ARFF is not possible and Large dataset are not applied in the classifier.

H. Theresa Princy R et al., (2016) predicted heart disease using ID3 and KNN algorithm. The ID3 algorithm is used as a classifier, KNN algorithm organizes and pre-processes the incorrect values which are considered as the training set. The basic factors along with some additional factors such as smoking were included. The accuracy level increased to 80.6%. [5]

Challenges: Very few factors are considered for the prediction. Some common influencing factors such as CPT (Chest Pain Type), RECG (Resting electrocardiographic (ECG)) and indirectly influencing factors such as Alcohol and Obesity are not considered. Without using these important factors, the prediction could not be given precisely.

I. Kalaiselvi C (2016) proposed a new website called average K-nearest neighbour algorithm which is to improve classification accuracy and efficiency. The advantage in AKNN is that grouping the samples based on super classes reduces the number of samples used for training, thus making the KNN the faster algorithm. It uses 13 attributes and the accuracy obtained is 96.5% and with 12 attributes the accuracy obtained is 97%. This works well if the data are well segregated and does not work if the data are noisy. [4]

Challenges: The accuracy is simply increased by decreasing the number of attributes, which does not result in the better prediction of heart disease.

J. Radhimeenakshi S (2016) compared support vector machine as well as artificial neural network. Examination is completed among two strategies on the premises of accuracy and training time. The target of this paper is to break down the use of AI devices for order and expectation of heart disease. The dataset utilized are the Cleveland Heart Database and Stat Log Database. The data records are classified into two classes in support vector machine as well as artificial neural network which attained 84.7% and 81.8% accuracies respectively. [3]

Challenges: Multiple thresholds are not used in progressing. The speed of the performance in Artificial Neural Network is less thus resulting in less accuracy.

K. VivekanandanTet al., **(**2017) proposed the challenging tasks of selecting critical features from the enormous set of available features and diagnosing heart disease. DE (Modified Differential Evolution) algorithm is used to perform feature selection. Prediction of heart disease was carried out using Fussy AHP and Feed-Forward neural network. Using 9 attributes an accuracy of 98% was achieved. [2]

Challenges: More number of inputs are not listed and Error minimization is not carried out properly. It could have been carried out by using effective back -propagation model. Without proper testing they have mentioned that large data sets can also be adopted.

L. Sharmila S et al., (2017) proposed a method to improve Naïve Bayes performance. It takes only two values for the prediction. It classifies the data into two classes 0-Absent and 1-Present. The proposed algorithm is used for identification of the values ranging from 0 to 1. It uses 14 attributes available in UCI machine repository which contains 303 records. It takes only two values for the prediction with accuracy of 97%. [1]

Challenges: The existing Naïve Bayes algorithm has been modified into new algorithm which takes only two values for prediction. By considering just two values they have come to a conclusion that their proposed algorithm achieves higher accuracy when compared to other algorithms.

The following Table 1 describes various research work done by many researchers in the prediction of heart disease using different data mining techniques and corresponding accuracy level has been obtained. The number of attributes used by each

researcher has also been mentioned. It is found that the maximum number of attributes is used by Chaitrali S et al., [10] and maximum accuracy has also been obtained in that work. It is observed that most of them have implemented Naïve Bayes approach in prediction.

Table 1. Research Works Addressing the Prediction of Heart Disease

Contributors	Data Mining Algorithm	Accuracy	Number of Attributes used
T John Peter et al., 2012 [11]	NB, DT, k-NN, NN	83.70% , 76.66%, 75.18% , 78.148%	14
Chaitrali S et al., 2012[10]	NN,DT,NB	100%, 99.62%, 90.74%	15
Shamsher Bahadur et al., 2013[9]	DT, NB	99.2%, 96.5%	6
I.S. Jenzi., 2013[8]	NB	80.7%	14
Hlaudi Daniel et al.,2014 [7]	J48, REP TREE, NB	99.07%, 99.07%, 97.22%,	11
B. Venkatalakshmi et al., 2014[6]	NB,DT	85.03%, 84.0136%	13
Theresa Princy. et al., 2016[5]	KNN , ID3	40.3%, 80.62%	5,7
C.Kalaiselvi, 2016[4]	Average KNN	87%	12
S.Radhimeenakshi, 2016 [3]	SVM, ANN	84.7%, 81.8%	eaveland-14 Statlog-13
T.Vevikanandan et al., 2017[2]	Fuzzy AHP, FFNN	98%	9
Sharmila et al., 2017 [1]	NB	97%	14

The comparison between the different attributes used in each work has been listed out in the following Table 2. New attributes have been included by few researchers. It is observed that most of them have considered only common factors for prediction. By comparing different works containing different attributes, a consolidated 17 attributes have been obtained.

Table 2. Comparison between Different Attributes

Contributors	Factors / Attributes																
	Age	Gender	CPT	FBS	RECG	Ex-Ang	SL	Col-Ves	Thal	SC	Thalach	Old Peak	RBP	Num	Smoking	Alcohol	Obesity
T John Peter et al., 2012 [11]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Chaitrali S et al., 2012 [10]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
Shamsher Bahadur et al., 2013 [9]			✓			✓		✓			✓	✓	✓				
I.S Jenzi,2013[8]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Hlaudi Daniel et al., 2014 [7]	✓	✓	✓	✓	✓					✓	✓		✓		✓	✓	✓
B.Venkata-lakshmi et al., 2014 [6]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Theresa Princy. et al., 2016 [5]	✓	✓		✓						✓			✓	✓	✓		
C.Kalaiselvi 2016 [4]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
S.Radhimeenakshi 2016 [3]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
T.Vivekanandan et al., 2017 [2]	✓		✓			✓	✓	✓	✓		✓	✓	✓				
Sharmila, et al., 2017 [1]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			

IV. DISCUSSIONS

- From the above investigation it is found that the researchers have not considered many attributes for prediction. Many important factors such as alcohol and smoking could have been included.

- Most of them have considered only common attributes. It is found that the attribute size is reduced to increase the accuracy. This reduction in the number of attributes does not predict heart disease accurately. All the factors influencing heart disease should be considered.
- Few researchers have used large data sets but they did not apply proper error handling mechanism for classification. The error handling mechanisms such as removal the row of data, usage global values for filling the empty values and taking factor mean for the values could have been used.
- Some researchers have justified that the algorithm which they have proposed is more efficient than other algorithms without proper investigation of various efficient algorithms.

V. CONCLUSION

In this survey, major influencing factors for determining the heart disease and various research works in predicting the heart disease has been identified and reported. It is observed that not all attributes are taken into consideration by every researcher. Few attributes are eliminated to provide more accuracy by few researchers. We have carried out a detailed discussion about the key challenges of various research works for heart disease prediction that are not yet addressed. In future, the researchers should include all the factors for determining the heart disease using an effective algorithm.

REFERENCES

[1] Sharmila S et al., "Analysis of Heart Disease Prediction using Data Mining Techniques", International Journal of Advanced Networking & Applications (IJANA), Volume: 08, Issue: 05, Pages: 93-95 (2017).

[2] Vivekanandan T et al., "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease", www.elsevier.com/locate/compbiomed, https://doi.org/10.1016/j.compbiomed, Pages: 125-136 (2017)

[3] Radhimeenakshi S, "Classification and Prediction of heart disease risk using Data Mining Techniques of Support Vector Machine and Artificial Neural Network", InternationalConference on Computing for Sustainable Global Development(INDIACom), IEEE, Pages: 3107-3111 (2016)

[4] Kalaiselvi C, "Diagnosing of heart diseases using Average K- Nearest Neighbor Algorithm of Data Mining", 2016 InternationalConference on Computing for Sustainable Global Development (INDIACom), IEEE, Pages: 3099-3103 (2016)

[5] Theresa Princy R, "Human Heart Disease Prediction System using Data Mining Techniques", International Conference on Circuit, Power and Computing Technologies [ICCPCT], IEEE (2016)

[6] Venkatalakshmi B et al., "Heart Disease Diagnosis using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology [IJIRSET], ISSN (Online): 2319-8753, Volume: 3, Issue: 3, Pages :1873- 1877 (2014)

[7] Hlaudi Daniel Masethe et al., "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science [WCECS], Volume: II, ISBN: 978-988-19253-7-4 ISSN: 2078-0958 (Print), ISSN: 2078-0966 (Online) (2014)

[8] Jenzi I.S et al., "A Reliable Classifier Model using Data Mining approach for Heart disease prediction", International Journal of Advanced Research in Computer Science and Software Engineering

[9] Shamsher Bahadur Patel et al., "Predict the Diagnosis of Heart Disease patients using Classification Mining Techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR- JAVS), e-ISSN: 2319-2380, p-ISSN: 2319-2372, Volume: 4, Issue: 2, Pages: 61-64 (2013)

[10] Chaitrali S. Dangare et al., "Improved Study of heart disease Prediction system using Data Mining Classification Techniques", International Journal of Computer Applications, Volume: 47, Pages: 44-48 (2012)

[11] John Peter Tet al., "An Empirical study on Prediction of Heart Disease using Classification Data Mining Techniques", International Conference On Advances in Engineering, Science and Management (ICAESM -2012), IEEE, ISBN: 978-81-909042-2-3, Pages: 514-518

[12] Andras janosi et al., "Heart Disease Data Set", URL: archive.ics.uci.edu/ml/datasets/heart+Disease, Retrieved on- 09.01.2018

[13] Sunil Ray, "Learn Naïve Bayes algorithm" and "Decision tree- Simplified", URL: www.analyticsvidhya.com, Retrieved on- 10.01.2018