International Journal of Management, IT & Engineering
Vol. 9 Issue 4, April 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A REVIEW ON HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Adil Hussain Seh[*]

Dr. Pawan Kumar Chaurasia[**]

Abstract

Heart disease is one of the most fatal problems in the whole world, which cannot be seen with a naked eye and comes instantly when its limitations are reached. Therefore, it needs accurate diagnosis at accurate time. Health care industry produced huge amount of data every day related to patients and diseases. However this data is not used efficiently by the researchers and practitioners. Today healthcare industry is rich in data however poor in knowledge. There are various data mining and machine learning techniques and tools available to extract effective knowledge from databases and to use this knowledge for more accurate diagnosis and decision making. Increasing research on heart disease predicting systems, it become significant to summarize the completely incomplete research on it. The main objective of this research paper is to summarize the recent research with comparative results that has been done on heart disease prediction and also make analytical conclusions. From the study, it is observed Naive Bayes with Genetic algorithm; Decision Trees and Artificial Neural Networks techniques improve the accuracy of the heart disease prediction system in different scenarios. In this paper commonly used data mining and machine learning techniques and their complexities are summarized.

Keywords:

Data mining, Machine learning, Heart disease, Classification, Naive Bayes, Artificial Neural Networks, Decision Trees, Associative Rule.

[*] **Directorate of IT & SS, University of Kashmir, Srinagar (J&K) –India.**

[**] **Assistant Professor, Dept. of Information Technology, Babasaheb Bhimrao Ambedkar Central University, Lucknow, (UP) -India.**

1. Introduction

Heart disease describes a range of conditions that affect your heart. Heart disease term includes a number of diseases such as blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others. The term heart disease is sometimes used interchangeably with the term cardiovascular disease. Cardiovascular disease (CVD) generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack (Myocardial infarctions), chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease [3]. 17.9 million People die each year from CVDs, an estimated 31% of all deaths worldwide [4]. Nowadays healthcare sector produces large amount of information about patients, disease diagnosis etc. however this data is not used efficiently by the researchers and practitioners. Today a major challenge faced by Healthcare industry is quality of service (QoS). QoS implies diagnosing disease correctly & provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable [2]. There are various heart disease risk factors. Family history, Increasing age, Ethnicity and being male are some risk factors that cannot be controlled. But Smoking, Diabetes, High cholesterol, High blood pressure, not being physically active, being overweight or obese are those factors that can be controlled or prevented [5].

Data mining is the process of discovering unknown hidden patterns (knowledge) from large pre- existing data sets with the involvement of data mining and machine learning techniques, statistics, and database systems. The discovered knowledge can be used to build intelligent predictive decision systems in different fields like health care for accurate diagnosis at accurate time to provide affordable services and save precious lives. Machine learning provides computer programs the ability to learn from predetermined data and improve performance from experiences without human intervention and then apply what have learned to make an informed decision. At every successful decision machine learning program improves its performance. . Given below figure depicts the knowledge discovery from data (KDD) process.

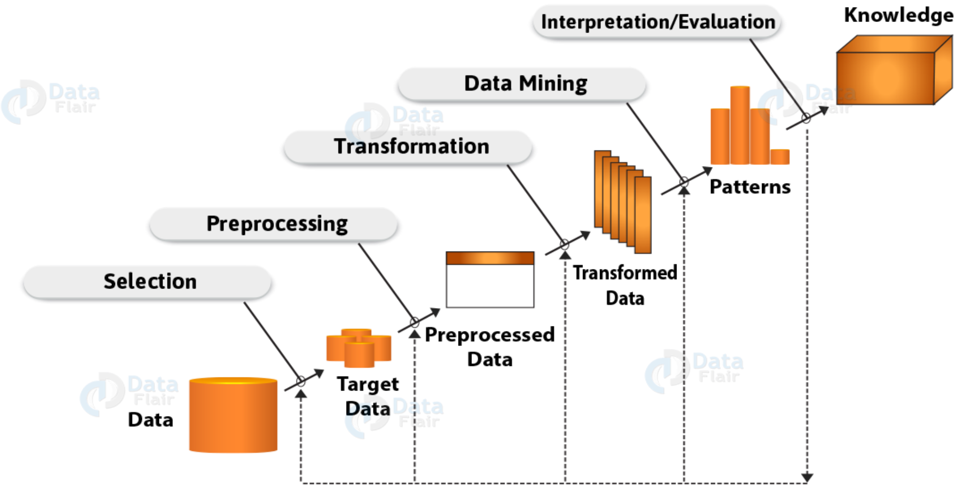

Figure.1 Steps in Knowledge Discovery Process [25].

2. Prior Knowledge

In every field of education we need prior knowledge to understand and analyze that field very well, prior knowledge become base for successful understanding and analyses of any study. So before we start to study the actual content of this paper we have to study and understand the basic concepts related to the paper that will help us to understand and comprehend the paper very well.

2.1) Classification: Classification is a supervised data mining and machine learning technique. It is a two step process, first step is called learning step where the model is constructed and trained by a predetermined dataset with class labels (training set) and second step is classification (testing) step where the model is used to predict class labels for given data (test data) to estimate the accuracy of classifier model[17].

2.2) Associative rule: Associative rule miming is a data mining technique which is used to find associative rules or patterns in data. In association rule mining, a pattern is discovered based on a relationship of a particular item to other items in the same transaction. It finds frequent item sets in data by using predefined support and confidence values. The association rule technique is used for heart disease diagnosis to discover the relationship of different attributes used for analysis and sort out the patient with all the risk factor which are required for prediction of disease [18].

2.3) Clustering: Clustering is basically an unsupervised machine learning technique. It is the task of dividing the dataset or population into a number of groups such that records or objects in the same groups are more similar to each other and dissimilar to the records or objects in other groups. Clustering helps to understand natural grouping or structure in a dataset and has no predefined classes [9] K-means algorithm is cluster based algorithm.

2.4) Decision Tree: Decision tree is a technique that is used as a decision support tool that uses a tree-like graph or model of decisions [19]. It takes as input a record or object described by a set of attributes and returns a "decision with predicted output value for the input”. The input attributes can be discrete or continuous. After performing a sequence of tests decision tree reaches its decision. Each non leaf node of a decision tree corresponds to a test for the relevant attribute value, and the branches from the node are labeled with the possible outcomes of the test. Each leaf node in the tree specifies the value (decision) to be returned if that leaf is reached [20]. J48, Random Forest (RF) and Logistic Tree Model (LTM) are Decision tree implementation algorithms.

2.5) Naive Bayes: A Naive Bayes classifier is a supervised machine learning technique based on Bayes theorem. In simple terms a Naive Bayes classifier assumes that the presence or absence of a particular attribute of a class is independent to the presence or absence of any other attribute of that class. It is often used to compute posterior probabilities of given observations and make decisions on higher probability [21].

2.6) Artificial Neural Networks: Artificial neural networks are machine learning algorithms having non linear data processing ability. Artificial neural network, sometimes just called a neural network, is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system [2]. Neural networks consist of input and output layers, as well as (in most cases) a number of hidden layers. They are excellent tools for finding complex patterns in data and improve performance continuously from past experiences.

*2.7) Genetic Algorithm**:*** Genetic algorithm is a method for solving optimization problems that is based on natural selection, the process that drives biological evolution. The genetic algorithm iteratively modifies a population of individual solutions. At each step individuals are selected randomly as parents from the current population by genetic algorithm to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution [22]. In genetic algorithm solutions are represents by chromosomes. Chromosomes are made up of genes, which are individual elements that represent the problem. The collection of all chromosomes is called population. The genetic algorithm uses three main types of rules (operators) at each step to create the next generation from the current population: a) Selection is used in selecting individuals for reproduction. b) Crossover is used to combine two parents to form children for the next generation. C) Mutation is used to alter the new solutions in the search for better solution. Mutation prevents the GA to be trapped in a local minimum [24].

2.8) Cross Validation: Cross-validation is a technique to evaluate predictive models by dividing the original dataset into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly divided into k equal size subsets. Of the k subsets, a single subset is taken as the validation data for testing the model, and the remaining k-1 subsets are used for training the model. The cross-validation process is then repeated k times (the folds), with each of the k subsets used exactly once as the validation data and average accuracy of k-folds is taken as final accuracy. In most experiments 10-fold cross validation technique is used. In 10-fold cross validation all the instances of the data set are used and are divided into 10 disjoint groups, where nine groups are used for training and the remaining one is used for testing. The algorithm runs for 10 times and average accuracy of all folds is calculated [23].

3. Literature Survey

Till date different studies have been done on heart disease prediction. Various data mining and machine learning algorithms have been implemented and proposed on the datasets of heart patients and different results have been achieved for different techniques. But, still today we are facing a lot of problem faced by the heart disease. Some of the recent research papers are as follows:

In 2010, A. Rajkumar and G. S. Reena applied machine learning algorithms such as Naive Bayes, KNN (K- nearest neighbors) and decision list for heart disease prediction. Tanagra tool is used to classify the data and the data evaluated using 10-fold cross validation and the results are compared in table 4. The data set consists of 3000 instances with 14 different attributes. The dataset is divided into two parts, 70% of the data are used for training and

30% are used for testing. The results of comparison are based on 10-fold cross validation. Comparison is made among these classification algorithms out of which the Naive Bayes algorithm is considered as the better performance algorithm. Because it takes less time to build model and also gives best accuracy as compared to KNN and Decision Lists [7].

Table 1. Comparative Results

Classification Techniques	**Accuracy**	**Timing Taken**
Naive Bayes	52.33%	609ms
Decision List	52%	719ms
KNN	45.67%	1000ms

In 2011, G.Subbalakshmi, K. Ramesh and M. Chinna Rao developed a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modeling technique, namely, Naive Bayes. Using heart disease attributes such as chest pain, age, sex, cholesterol, blood pressure and blood sugar can predict the likelihood of patients getting a heart disease. It is implemented as web based questionnaire application. Historical data set of heart patients from Cleveland database of UCI repository was used to train and test the Decision Support System (DSS). The reasons to prefer Naive Bayes machine learning algorithm for predicting heart disease are as follows: when data is high, when the attributes are independent of each other and when we want to achieve high accuracy as compared to other models. When the dimensionality of the inputs is high in that case Naive Bayes classifier technique is particularly suited. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods [13].

In 2011, M. A. Jabbar, Priti Chandra and B.L.Deekshatulu in this study develop a prediction system by implement associative rule mining using a new approach that combines the concept of sequence numbers and clustering for heart attract prediction. By using this approach first dataset of heart disease patients has been converted into binary format then apply proposed method on binary transitional data. Data set of heart disease patients has been taken from Cleveland database of UCI repository with 14 essential attributes. The algorithm is well known as Cluster Based Association Rule Mining Based on Sequence Number (CBARBSN). Support is a basic parameter in associative rule mining. To become element of a frequent item set an item should satisfy support threshold. In this research transactional data table is divided into clusters based on skipping fragments (disjoint sub sets of actual transitional table) then Sequence Number and Sequence ID of each item has been calculated. On the basis of Sequence ID frequent item sets has been discovered in different clusters and common frequent item set has taken as Global Item set. It has been observed from the experiment that Age>45 and Blood pressure>120 and Max Heart rate>100 and old Peak>0 and Thal>3 =>Heart attack (Common frequent item set found in both clusters in this experiment). In our proposed algorithm execution time to mine association rules is less (i.e., 0.879 ms when support=3) and as support increases execution time changes drastically as compared to previously developed system. In Fig.3 Execution time is shown horizontally and Support vertically [14].

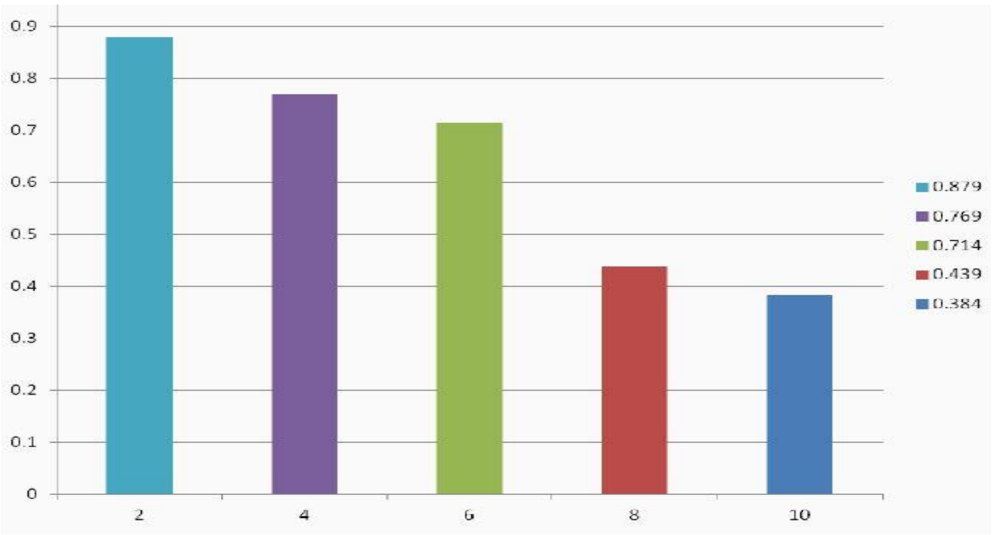

Figure.2 Our Proposed Algorithm CBARBSN [14]

In 2012, Chaitrali S. Dangare and Sulabha S. Apte implement data mining and machine learning classification algorithms namely Decision Trees (J48), Naive Bayes, and Neural Networks on Heart disease datasets to build Intelligent Heart Disease Prediction System. In this research two datasets were used. The Cleveland Heart Disease dataset consists of 303 records & Statlog Heart Disease dataset consists of 270 records. With commonly used 13 attributes two more attributes, i.e. obesity and smoking were included in the dataset for efficient diagnosis of heart disease. Comparative results were examined on both 13 attribute dataset and 15 attribute dataset separately. Total 573 records were divided into two data sets one is used for training consists of 303 records and another for testing consists of 270 records. Weka 3.6.6 data mining and machine learning tool is used for experiment. Missing values in dataset were identified and replaced with most appropriate values using Replace Missing Values (RMV) filter from Weka 3.6.6. Give below table summarizes the comparative results of our research. From the results it has been observed that neural network provides accurate result as compare to decision trees & Naive Bayes [2].

Table 2. Comparative Results

Classification Techniques	**Accuracy with**	
	13 attributes	15 attributes
Naive Bayes	94.44%	90.74%
Decision Trees(J48)	96.66%	99.62%
Neural Networks	99.25%	100%

In 2013, A. Taneja, applied data mining and machine learning algorithms namely Decision Tree (J48 algorithm), Naive Bayes and Artificial Neural Networks (ANN) for heart disease prediction. A dataset of 7339 instance with 15 attributes has been taken from PGI Chandigarh. WEKA 3.6.4 tool was used for the experiment. For model training and testing 10-Fold Cross Validation techniques is used randomly. Best First Search method was used to select the best attributes from the already available 15 attributes and among them only 8 attributes has been selected. Each experiments was done on two different scenarios, first one containing all 15 attributes and the second case only 8 selected attributes. From all these experiments comparative results has been obtained and from these comparative results it has been found that J48 pruned in selected attributes case has performed best in accuracy with 95.56% and Naive Bayes with all attributes case gives less accuracy 91.96% but takes least time to build a model in the whole experiment [10].

Table 3. Comparative Results

Machine Learning Algorithms	**Accuracy with**		**Time to build Model (in sec.)**	
	15 Attributes	8 Attributes	15 Attributes	8 Attributes
J48 UnPruned	94.29%	95.52%	0.98 sec	0.36 sec
J48 Pruned	95.41%	95.96%	NM*	NM*
Naive Bayes	91.96%	92.42%	NM*	NM*
ANN	93.83%	94.85%	158.94 sec	93.83 sec

NM* (not mentioned in this research paper clearly)

In 2014, B.Venkatalakshmi and M.V. Shivsankar design and develop a prediction system for heart diseases diagnosis. In this proposed work, 13 attribute structured clinical dataset of only 294 records from UCI Machine Learning Repository has been used as a data source. WEKA tool is used for algorithm implementation. In table 6, Machine learning algorithms namely

Decision Tree and Naive Bayes are implemented and comparative results has been obtained. From the results it has been observed that Naive Bayes technique performs best in accuracy. In this research work implementation of Genetic Algorithm using MATLAB tool for attribute optimization to improve the accuracy and time complexity of system is also discussed for future work [9].

Table 4. Comparative Results

Classification Algorithm	**Classified Instances**		**Accuracy**	**Time Taken To Build Model (in sec)**
	Correctly	**Incorrectly**		
Decision Tree	247	47	84.013%	0.02
Naive Bayes	250	44	85.034%	0.11

In 2014, H. D. Masethe and M. A. Masethe applied data mining algorithms such as J48, Naive Bayes, REPTREE, CART, and Bayes Net in this research for predicting heart attacks. The patient dataset is collected from medical practitioners in South Africa. Only 11 attributes namely Patient Id, Gender, Cardiogram, Age, Chest Pain, Blood Pressure Level, Heart Rate, Cholesterol, Smoking, Alcohol consumption and Blood Sugar Level from the database are considered for the predictions required for the heart disease. WEKA data mining tool is used for experiment. The algorithms were applied on the data set using 10-fold cross validation technique in order to calculate average accuracy of all folds for each classification technique to predict a class. From the results it has been seen J48, REPTREE and SIMPLE CART algorithm performs best in this data set, while Bayes Net algorithm out-performed the Naïve Bayes algorithm [6].

Table 5. Comparative results

Evaluation Criteria	**Classification Techniques**				
	J48	**REPTREE**	**NAVE BAYES**	**BEYES NET**	**SIMPLE CART**
Timing to build model (in sec)	0	0	0	0.02	0.1
Predictive Accuracy	99.0741	99.0741	97.222	98.1481	99.0741

In 2015, Jaymin Patel, Prof.Tejal Upadhyay and Dr. Samir Patel in this study implement decision support system using three data mining and machine learning algorithms viz. J48, Logistic Model Tree, and Random Forest algorithm to develop a system for accurate heart disease prediction. In this experiment WEKA 3.6.10 tool is used for implementation. A data set of 303 records of heart patients has been taken from Cleveland database of UCI repository to train and test the system. To evaluate the system 10-fold cross validation technique is used for model training and testing. Algorithms are analyzed generally on the basis of three parameters viz. sensitivity (The sensitivity is proportion of positive instances that are correctly classified as positive), specificity (The specificity is the proportion of negative instances that are correctly classified as negative), and the accuracy (The accuracy is the proportion of instances that are correctly classified). From the comparative results it has been observed J48 algorithm achieved higher sensitivity and accuracy while LMT achieved higher specificity. So overall it is concluded that J48 (with Reduced Error Pruning) has got the best overall performance [11].

Table 6. Comparative Results

	Decision Tree(J48)	**Logistic Model Tree (LMT)**	**Random Forest (RF)**
Train Error	0.1423221	0.1656716	0
Test Error	0.1666667	0.237931	0.2
Accuracy	56.76%	55.77%	NM*

In 2016**,** *K.Gomath* and Shanmugapriyaa applied three machine learning algorithms viz.

Naïve Bayes, J48, and Artificial Neural Network (ANN) to achieve best accuracy in heart disease prediction for male patients. A dataset of 210 records with 8 attributes has been used in this experiment. In order to carry out experiments and implementations WEKA was used as the data mining tool. From the experiments comparative results has been drawn in table 8 and from the comparative result has been found that Naïve Bayes performed best as compared to J48 and ANN to predict heart disease with an accuracy of 79.9043% and takes less time 0.01 seconds to build a model [12].

Table 7. Comparative Results

Classification Techniques	**Accuracy**	**Timing Taken**
Naive Bayes	79.9043%	0.01 sec.
Decision List (J48)	77.0335%	0.01 Sec.
ANN	76.555 %	1.55 Sec.

In 2017, Zeinab Arabasadi et al., proposed a hybrid diagnosis model for coronary artery disease using machine learning algorithm namely Artificial neural network (ANN) and genetic algorithm. In this research Z-Alizadeh Sani dataset is used consists of 303 patient records with 54 attributes (only 22 essential attributes were used in experiment), among them 216 patients suffered from coronary artery disease (CAD). First weights to artificial neural network were identified by genetic algorithm then ANN model was trained by using training data. In this experiment ANN with one input and output layer also consists of one hidden layer having five neurons employ feed forward approach. 10-fold cross validation technique is used for system evaluation in this experiment. From the results we observe that our proposed model performed high in accuracy as compared to existing simple ANN model. We also test our model in other four world famous heart disease data sets with comparative results. Our proposed model also provides high accuracy as compared to existing ANN model.

Table 8. Comparative Results

Data sets (with No. of Attributes)	**Proposed Model (Genetic ANN) Accuracy**	**Existing Model (ANN) Accuracy**
Z-Alizadeh Sani dataset	93.85 %	84.62 %
Hungarian dataset (14)	87.1	82.9
Cleveland dataset (14)	89.4	84.8
long-beach-va dataset	78.0	74.0
Switzerland dataset	76.4	71.5

In 2018, S. Kodati and R Vivekanandam applied classification techniques namely Random

Forest, Decision Table, Naive Bayes and J48 algorithm. Waikato Environment for Knowledge Analysis (WEKA) data mining tool is used for experiment. Dataset of 303 records with 14 attributes used in this research is taken from UCI repository. In this research it has been found from comparative results in table 5 that Decision Table performs best in accuracy with 84.81% followed by Naive Bayes and Random Forest. J48 algorithm gives less accuracy as compared to other algorithms in this research [8].

Table 9. Comparative Results

Classification Algorithm	**Number Of Correctly Classified Instances**	**Accuracy**	**Time Taken To Build Model (in sec)**
Decision Table	229	84.81%	0.11
J48	207	76.66%	0.05
Naive Bayes	226	83.70%	0.02
Random Forest	221	81.85%	0.23

In 2018, N. Shirwalkar and T. Tak in this paper make an analytical study on various data mining and machine learning techniques used in heart disease prediction and compare them to find the best method for prediction. Naive Bayes and improved K-means algorithms are chosen for proposed heart disease prediction. Dataset of 303 records of heart disease patients is taken from Cleveland database of UCI repository. Original dataset table is transformed from one form to another known as discretization. Improved k-means algorithm is used to make clusters from the dataset table. Naive Bayes algorithm is used to train the model on training to predict heart disease patients. Using this model we can predict four levels of heart disease on the basis of probability ratio generated by Naive Bayes algorithm i.e., Normal, Stage 1, Stage 2, Stage 3 [15].

In 2018, Navdeep Singh and Sonika Jindal devise a model called Hybrid Genetic Naive Bayes Model using two supervised machine learning algorithms i.e., Genetic algorithm and Naive Bayes to predict heart disease with high accuracy results. In this propped model a dataset of 303 records with 14 necessary attributes has been taken from online Cleveland database of UCI repository. The results are obtained in the type of the various performance parameters as precision (precision=98 %), recall (recall=97.14%) and accuracy (accuracy=97.14%). From the results it has been found our proposed Hybrid Model performs with highest accuracy as compared to existing models [16].

Table 10. Comparison of Models

MODEL	ACCURACY
Weighted Fuzzy Rules	57.85%
Logistic Regression	77%
Naive Bayes	81.48%
Attribute Weighted Artificial Immune	82.59%
Artificial Immune System	84.5%
Modified Artificial Immune System	87.43%
Neural Networks ensemble	89.01%
ANN_Fuzzy_HP	91.10%
GA_Fuzzy_Naive (Proposed)	97.14%

Table 11. Summary of Literature Survey

Author and Year	**Techniques**	**Features**	**Attributes Used**		**Accuracy in Percentage**	
Asha Rajkumar et al. (2010)	Naive Bayes	Simple, easy to calculate, need less training data, assume feature conditional independence, mostly used when more number of classes are to be predict.	14		52.33	
	Decision Trees(J48)	Easy and simple, take care of missing values and outliers, over-fitting is most significant feature			52	
	KNN	Simple to use, works well on basic diagnosis problems, non parametric (has no predefined assumptions).			45.67	
G.Subbalakshmi et al (2011)	Naive Bayes	Need less training data, assume feature conditional independence.	14		NM*	
MA.Jabbar et al. (2011)	CBARBSN	Fast processing time, used to discover frequent item pattern and feature extraction with both supervised and unsupervised techniques .	14		NM*	
Chaitrali S. Dangare et al. (2012)	Naive Bayes	Need less training data, assume feature conditional independence.	15	13	90.74	94.44
	Decision Trees(J48)	Over-fitting is most significant feature.			99.62	96.66
	Neural Networks	Ability to generalize the input, non linear data processing, ability of high fault tolerance, self repair when node(s) of network not working properly.			100	99.25
Abhishek Taneja (2013)	J48 UnPruned	Take care of missing values and outliers.	15	8	94.29	95.52
	J48 Pruned	Over-fitting is most significant feature.			95.41	95.96
	Naive Bayes	Assume feature conditional independence.			91.96	92.42
	ANN	Non linear data processing, ability of generalizes the input and high fault tolerance.			93.83	94.85
B.Venkatalak	Decision Tree	Easy and simple, over-fitting.	13		84.013	

shmi et al. (2014)	Naive Bayes	Assume feature conditional independence.		85.034
Hlaudi Daniel Masethe et al. (2014)	J48	Best for holding missing value, decision tree pruning, good for both discrete and continuous attribute values.	11	99.0741
	REPTREE	Reduced-error pruning, generate multiple trees and pick best one from them, deals with missing values.		99.0741
	NAVE BAYES	Assume feature conditional independence.		97.222
	BEYES NET	Handle incomplete data sets, provides effective methods for preventing data over-fitting.		98.1481
	SIMPLE CART	Used for both classification and regression, easily handle outliers, uncover complex interdependencies between sets of variables by using same variable multiple times in different parts of a tree.		99.0741
Jaymin Patel et al. (2015)	J48 Algorithm	Best for holding missing value, decision tree pruning.	14	56.76
	Logistic Model Tree	Handle nonlinear effects, robust, best for binary classification problems.		55.77
	Random Forest	Needs less data to train model, reduce over-fitting, individual decision trees can be trained in parallel.		NM*
K.Gomath et al. (2016)	Naive Bayes	Assume feature conditional independence.	8	79.9043
	Decision List (J48)	Over-fitting is most significant feature.		77.0335
	ANN	Ability to generalize the input and high fault tolerance.		76.555
Zeinab Arabasadi et al. (2017)	ANN with Genetic algorithm	Attribute selection with optimization, ability of generalizes the input and high fault tolerance	22	93.85
Nikita Shirwalkar et al. (2018)	Naive Bayes and improved K-means algorithm	K-means: Easy to implement, used when data is unlabeled, when variables are huge, produce tighter clusters.	14	NM*
Sarangam Kodati et al. (2018)	Decision Table	Easy to draw and easy to change, completeness by drawing every possible combination of condition values.		84.81
	J48	Best for holding missing value.		76.66
	Naive Bayes	Assume feature conditional independence.		83.70
	Random Forest	Individual decision trees can be trained in parallel, reduce over-fitting.		81.85
Navdeep Singh et al. (2018)	Genetic algorithm and Naive Bayes	GA: Works with populations and string coding of variables, narrowing search spaces, propagate adaptively.	14	97.14%

NM* (not mentioned in this research paper clearly). Where data set is taken from Cleveland database of UCI repository and number of attributes are not defined in research paper we assume 14 attributes.

4. Conclusion

From the study of various recent research papers written on heart disease prediction using various data mining and machine learning techniques and algorithms. We find that different techniques of data mining and machine learning are used to predict heart disease with the help of different experimental tools such as WEKA, MATLAB etc. Different datasets of heart disease patients are used in different experiments. In most experiments dataset used is taken from online Cleveland database of UCI repository. The dataset consists of 303 records with 14 essential attributes (total attributes 75) with some missing values also. Fewer experiments have been done on different datasets. From the study we also find that Neural Networks with 15 attributes provide 100% accuracy in one experiment whereas in another experiment gives 76.55% accuracy with 8 attributes. Naive Bayes also gives high accuracy above (90%) in most experiments with different number of attributes. Decision lists (J48) also performs very well in accuracy goes up to 99.62 % in a case. So, different techniques used indicate the different accuracies depend upon number of attributes taken and tool used for implementation. From this study we come up with following observations that should be taken in consideration in future research work for high accuracy and more accurate diagnosis of heart disease by using intelligent prediction systems.

- In most experiments Small and same dataset has been used to train prediction models. So, we have to take real data in a large quantity of heart disease patients from reputed medical institutes of our country and use that data to train and test our prediction models. Then we have to examine the accuracy of our prediction models on large datasets.
- We have to consult highly experienced experts of cardiology to prioritize the attributes according to their effect on patient’s health and also if necessary add more essential attributes of heart disease for more accurate diagnosis and high accuracy.
- There is need to develop more complex hybrid models for accurate prediction by integrating different techniques of data mining and machine learning and also include text mining of unstructured medical data available in large quantities in medical institutes. Also use of Genetic algorithm for optimization and feature selection make intelligent prediction models much better in overall performance.
- In this study we find more focus was given on classification techniques as compared

to regression and association rule. So, for better comparative results in future research we have to take these things in our consideration.

- Accuracy of research is directly proportional to the selection of research tools and procedures. So, Choice of appropriate experimental tool (WEKA, METLAB etc.) for implementation of techniques is also an important parameter.

References

[1] Nidhi Bhatla, and Kiran Jyoti, “An Analysis of Heart Disease Prediction using Different Data Mining Techniques”, *International Journal of Engineering Research & Technology (IJERT)*, Vol. 1, Oct.2012.

[2] Chaitrali S.Dangare, and Sulabha S.Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques*”, International Journal of Computer Applications (0975 – 888),* Vol. 47, No.10, June.2102.

[3] Heart disease webpage on MAYO CLINIC [Online]. Available: *https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118 , 2019*

[4] Cardiovascular disease webpage on WHO [Online]. Available: *https://www.who.int/cardiovascular_diseases/en/ , 2019.*

[5] Dr. T. Karthikeyan, and V.A.Kanimozhi, “Deep Learning Approach for Prediction of Heart Disease Using Data mining Classification Algorithm Deep Belief Network” , *International Journal of Advanced Research in Science, Engineering and Technology*, Vol. 4, Issue 1, January 2017.

[6] Hlaudi Daniel Masethe, and Mosima Anna Masethe, “Prediction of Heart Disease Using Classification Algorithms”, *in Proceedings of the World Congress on Engineering and Computer Science 2014* Vol. II WCECS 2014, 22-24 Oct. 2014, San Francisco, USA.

[7] Asha Rajkumar, and Mrs. G. SophiaReena, “Diagnosis of Heart Disease Using Data Mining Algorithm”, *Global Journal of Computer Science and Technology*, Vol. 10, pp. 38-43, Sept. 2010.

[8] Sarangam Kodati, and Dr. R Vivekanandam, *“A Comparative Study on Open Source Data Mining Tool for Heart Disease” , International Journal of Innovations & Advancement in Computer Science*, Vol. 7, Issue 3, March-2018.

[9] B.Venkatalakshmi, and M.V Shivsankar, “Heart Disease Diagnosis Using Predictive Data mining”, *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 3, Special Issue 3, March-2014.

[10] Abhishek Taneja, “ Heart Disease Prediction System Using Data Mining Techniques”, *Oriental Journal Of Computer Science and Technology*, Vol. 6, pp. 457-466, Dec.2013.

[11] Jaymin Patel, Prof.Tejal Upadhyay, and Dr. Samir Patel, “ Heart Disease Prediction Using Machine Learning and Data Mining Technique”, *IJCSC*, Vol. 7, No. 1, pp.129-137, September-2015.

[12] K.Gomath, Dr. Shanmugapriyaa, “Heart Disease Prediction Using Data Mining Classification”, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Vol.4, Issue 2, February-2016.

[13] G.Subbalakshmi, K. Ramesh, and M.C. Rao, “Decision Support in Heart Disease Prediction System using Naive Bayes*”, Indian Journal of Computer Science and Engineering (IJCSE),* Vol. 2, No. 2, Apr-May 2011.

[14] MA.Jabbar, Dr. Priti Chandra, and B.L. Deekshatulu, “Cluster Based Association Rule Mining For Heart Attack Prediction”, *Journal of Theoretical and Applied Information Technology*, Vol. 32 No.2, October-2011.

[15] Nikita Shirwalkar, and Tushar Tak, *“ Human Heart Disease Prediction System Using Data Mining Techniques”, International Journal of Innovations & Advancement in Computer Science*, Vol. 7, Issue 3, Mar.2018.

[16] Navdeep Singh and Sonika Jindal, “ Heart Disease Prediction System using Hybrid Technique of Data Mining Algorithms*”, International Journal of Advance Research, Ideas and Innovations in Technology,* Vol.4, Issue 2, 2018.

[17] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining Concepts and Techniques*, 3rd ed., USA: Morgan Kaufmann Publishers, 2012.

[18] Beant Kaur, and Williamjeet Singh, *"*Review on Heart Disease Prediction System using Data Mining Techniques", *International Journal on Recent and Innovation Trends in Computing and Communication,* Vol.2 Issue 10, October-2014.

[19] Animesh Hazra, S.Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee*, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology,* Vol. 10, No.7, July-2017.

[20] Stuart J. Russell and Peter Norvig, *Artificial Intelligence A Modern Approach*, 2nd ed., New Jersey: Pearson Education Inc., 2003.

[21] Shadab A. Pattekari, and Asma Parveen, "Prediction System For Heart Disease Using Naive Bayes", *International Journal of Advanced Computer and Mathematical Sciences,* Vol.3, pp. 290-294, 2012.

[22] What is the Genetic Algorithm webpage on MATHWORKS [online]: Available *https://in.mathworks.com/help/gads/what-is-the-genetic-algorithm.html , 2019*

[23] M.A. Jabbar, B.L.Deekshatulu, and Priti Chandra, *" Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing*, Vol. 4, pp.174-184, 2016.

[24] M.Akhil jabbar, B.L Deekshatulu and Priti Chandra, *"Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA)* 2013.

[25] Google Images: Data mining and knowledge discovery [online]: Available https://www.google.com/search?q=data+mining+and+knowledge+discovery/

[26] Zeinab Arabasadi et al. , " Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm" , *Computer Methods and Programs in Biomedicine-ELSEVIER*, Vol. 141, pp.19- 26, April- 2017.