Global Scientific JOURNALS

GSJ: Volume 11, Issue 8, August 2023, Online: ISSN 2320-9186

www.globalscientificjournal.com

CITE UNIPORT

A Review Comparism on Algorithm-Based Techniques used for Heart Disease Prediction.

Anazodo William, Okengwu U. A. , Odia K. M.

CENTER FOR INFORMATION AND TELECOMMUNICATIONS ENGINEERING.

UNIVERSITY OF PORT HARCOURT

Abstract

A review of the Algorithm-based Techniques is very important in the prediction of heart disease given that heart disease is one of the world's most fatal problems, which cannot be seen with the naked eye and manifests itself instantly when its limits are reached, a review of algorithm-based techniques is crucial in the prediction of heart disease. Therefore, it requires a precise diagnosis at a precise moment. Every day, the health care sector generates enormous amounts of data about patients and diseases. However, researchers and practitioners do not effectively utilize this data. The healthcare industry is currently data-rich but knowledge-poor. To effectively extract knowledge from databases and use this knowledge for more precise diagnosis and decision-making, A variety of data mining and machine learning techniques and tools are available. As research on systems for predicting heart disease grows, it is important to review the research, which is still entirely inconclusive. This research paper's main goal is to summarize current studies that have been conducted on a proposed system for predicting heart disease, compare their findings, and draw analytical conclusions. The study shows that Artificial Neural Networks, Decision Trees, and Naive Bayes with Genetic Algorithms are all excellent techniques, but our proposed **Computational Value**

Algorithm increases the accuracy of the heart disease prediction system in various situations. This paper summarizes the complexity of the most popular data mining and machine learning techniques.

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

1. Introduction

A variety of conditions that affect your heart are referred to as heart disease. The term "heart disease" refers to a variety of illnesses, including congenital heart defects, blood vessel diseases like coronary artery disease, issues with heart rhythm (arrhythmias), and more. Heart disease and cardiovascular disease are sometimes used interchangeably. Cardiovascular disease (CVD) is a term used to describe ailments involving narrowed or blocked blood vessels that can cause myocardial infarctions, angina, and strokes. Heart diseases that affect the muscle, valves, or rhythm of your heart are also regarded as forms of heart disease. 17.9 million people worldwide die from CVDs each year, accounting for 31% of all fatalities. Today's healthcare industry generates a lot of data about patients, disease diagnoses, etc., but researchers and practitioners do not effectively use this data. The quality of service (QoS) issue currently faces the healthcare sector as a major challenge. Correct disease diagnosis and patient treatment are implied by quality of service. It is unacceptable when a poor diagnosis has disastrous results. There are numerous risk factors for heart disease. Some risk factors cannot be changed, including ethnicity, age, family history, and being a man. But factors like smoking, diabetes, high blood pressure, cholesterol, inactivity, and being overweight or obese can be prevented or controlled. Data mining is the process of identifying previously unidentified hidden patterns (knowledge) using data mining and machine learning techniques, statistics, and database systems. The knowledge that has been discovered can be used to create smart predictive decision-making systems in a variety of industries, including healthcare for accurate diagnosis at the right time to provide cost-effective services and save priceless lives. Without human intervention, machine learning enables computer programs to learn from predetermined data, enhance performance from experiences, and then use what they have learned to make wise decisions. The performance of a machine learning program gets better after each wise choice. The knowledge discovery from data

(KDD) process is illustrated in the figure below.

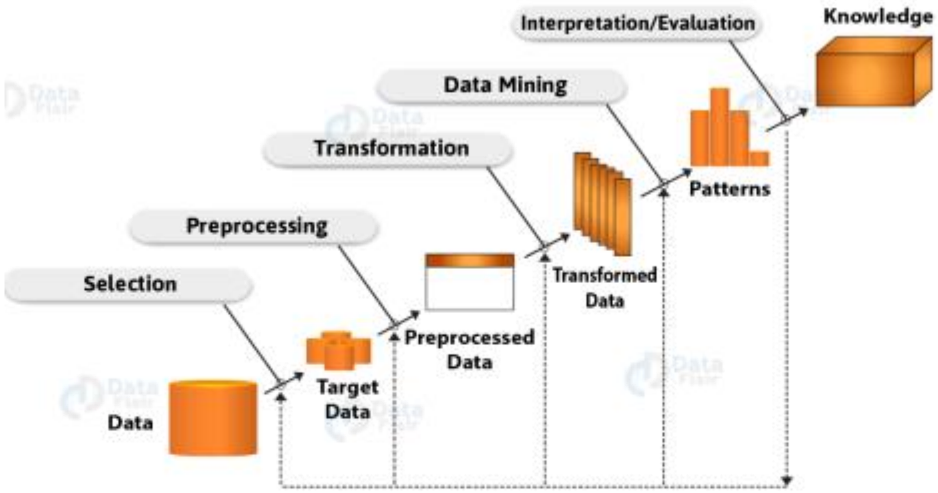


Figure.1 Steps in Knowledge Discovery Process

Prior Information

Prior knowledge serves as the foundation for successful understanding and analyses of any study and is required for success in every educational field. Therefore, before we begin to study the actual content of this paper, we must first study and comprehend the fundamental ideas that are related to it. This will enable us to fully understand and comprehend the paper.

Classification is referred to as supervised data mining and machine learning technique. It involves two steps: the first is learning, where the model is built and trained using a predetermined dataset with class labels (training set), and the second is classification (testing), where the model is used to predict class labels for provided data (test data) in order to gauge the accuracy of the classifier model.

Finding associative rules or patterns in data is done using the data mining technique known as associative rule miming. In association rule mining, a pattern is found based on how one item is related to other items that were purchased at the same time. Utilizing predefined support and confidence values, it identifies frequent item sets within the data. The association rule technique is employed in the diagnosis of heart disease to ascertain the relationships between various analytical attributes and to group patients according to all the risk factors necessary for disease prediction.

The task of grouping the dataset or population into a number of groups so that records or objects in the same groups are more similar to one another and dissimilar to the records or objects in other groups is known as clustering. Clustering is essentially an unsupervised machine learning technique. Clustering has no predefined classes but aids in understanding natural grouping or structure in a dataset. The cluster-based K-means algorithm is an algorithm.

Decision Tree: A decision tree is a technique that uses a tree-like graph or model of decisions as a decision support tool.

It receives as input a record or object that is described by a number of attributes and outputs a "decision with predicted output value for the input". Both discrete and continuous input attributes are possible. A decision is reached by the decision tree after performing a series of tests. The branches from each non-leaf node of a decision tree are labeled with the potential results of the test, and each node corresponds to a test for the relevant attribute value. The value (decision) that will be returned if a leaf node in the tree is reached is specified by each of its leaf nodes. Decision tree implementation algorithms include J48, Random Forest (RF), and Logistic Tree Model (LTM).

Naïve Bayes: A supervised machine learning method built on the Bayes theorem is known as a Naive Bayes classifier. A Naive Bayes classifier, to put it simply, believes that the presence or absence of one attribute of a class is unrelated to the presence or absence of any other attribute of that class. It is frequently used to determine higher-probability decisions by computing posterior probabilities for specific observations.

Artificial Neural Networks: Machine learning algorithms with nonlinear data processing capabilities include artificial neural networks. A mathematical or computational model based on a biological neural network is called an artificial neural network, or simply a neural network. It mimics the biological neural system, in other words. Input, output, and typically a number of hidden layers are all parts of a neural network. They are great tools for spotting intricate patterns in data, and they keep getting better as they learn from past mistakes.

Genetic Algorithm: Natural selection, the process that propels biological evolution, is the foundation of the genetic algorithm, a technique for resolving optimization issues. The population of unique solutions is modified iteratively by the genetic algorithm. Genetic algorithms are used at each stage to choose parents at random from the current population who will bear the progeny for the following generation. The population "evolves" over generations toward a perfect outcome. Chromosomes are used in genetic algorithms to represent solutions. Genes, which are discrete components that stand in for the issue, are what make up chromosomes. The term "population" refers to the totality of all chromosomes. To produce the next generation from the current population, the genetic algorithm employs three primary types of rules (operators): a) Individuals are chosen for reproduction using selection. b) Crossover is a technique used to create children for the following generation by combining two parents. C) In the search for a better solution, mutation is used to change the new solutions. The GA cannot be trapped in a local minimum due to mutation.

Cross-validation: This is a technique for assessing predictive models that involves splitting the original dataset into training and test sets. The training set is used to train the model, while the test set is used to assess it. The original sample is randomly divided into k subsets of equal size for k-fold cross-validation. The remaining k-1 subsets are used for training the model, and one subset of the total k subsets is used as validation data for testing the model. The cross-validation procedure is then repeated k times (the folds), using the validation data from each of the k subsets exactly once, and the average accuracy over the k-folds is used as the final accuracy. The 10-fold cross validation technique is used in the majority of experiments. All of the data set's instances are used in 10-fold cross validation, which divides them into 10 disjoint groups, nine of which are used for training and the tenth for testing. The algorithm is executed ten times, and the average fold accuracy is calculated.

2. Literature Survey

Numerous studies on the prognosis of heart disease have been conducted to date. On datasets of heart patients, numerous data mining and machine learning algorithms have been proposed and put into practice. Different techniques have produced different results. But even now, heart disease continues to cause a lot of problems. The following are a few recent research articles:

In order to predict heart disease, A. Rajkumar and G. S. Reena used machine learning algorithms like Naive Bayes, KNN (K- nearest neighbors), and decision lists in 2010. The data are categorized using the Tanagra tool, evaluated using 10-fold cross validation, and the outcomes are contrasted. 3000 instances with 14 different attributes make up the data set. The dataset is split into two sections, with 30% of the data used for testing and 70% for training. The comparison's findings are supported by 10-fold cross validation. These classification algorithms are compared, and the Naive Bayes algorithm is determined to perform the best overall. Because it is more accurate than KNN and Decision Lists and requires less time to build a model.

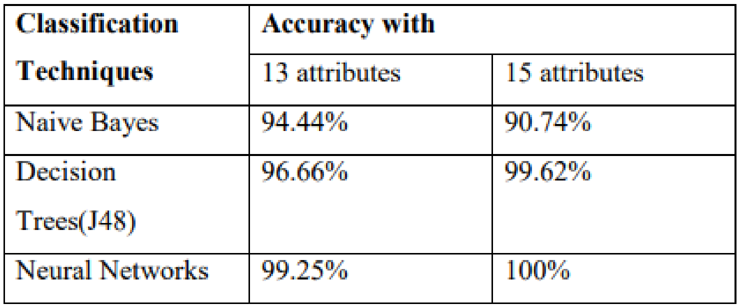
Table 1. Comparative Results

Classification Techniques	**Accuracy**	**Timing Taken**
Naïve Bayes	52.33%	609 ms
Decision List	52%	719ms
KNN	45.67%	629ms

Using the Naive Bayes data mining modeling technique, G. Subbalakshmi, K. Ramesh, and M. Chinna Rao created the Decision Support in Heart Disease Prediction System (DSHDPS) in 2011. Heart disease risk factors like chest pain, age, sex, cholesterol, blood pressure, and blood sugar can be used to predict a patient's likelihood of developing a heart condition. It is implemented as an online survey application. The Cleveland database of the UCI repository's historical data set of heart patients was used to train and test the Decision Support System (DSS). When data is abundant, when attributes are unrelated to one another, and when we want to outperform other models in terms of accuracy, we should choose the Naive Bayes machine learning algorithm for predicting heart disease. The Naive Bayes classifier technique is best suited in situations where the input dimensions are high. Naive Bayes can frequently outperform more complex classification techniques, despite its simplicity [13]. A new method that combines the idea of sequence numbers and clustering for heart attract prediction was developed by M. A. Jabbar, Priti Chandra, and B. L. Deekshatulu in 2011. They call it associative rule mining. This method involves converting a dataset of heart disease patients into binary format first, then applying the suggested method to binary transitional data. A patient data set with 14 key attributes for heart disease patients was taken from the Cleveland database of the UCI repository. Cluster Based Association Rule Mining

Based on Sequence Number (CBARBSN) is the popular name for the algorithm. A fundamental variable in associative rule mining is support. A piece of merchandise must meet the support threshold in order to be included in a frequent item set. In this study, the transactional data table is clustered according to the skipping of fragments (disjoint subsets of the actual transitional table), and then the Sequence Number and Sequence ID of each item have been determined. According to Sequence ID, frequent item sets have been found in various clusters, and the common frequent item set has been designated as the global item set. Age>45, Blood pressure>120, Max Heart Rate>100, Old Peak>0, and Thal>3 have been observed to be indicators of a heart attack (common frequent item set found in both clusters in this experiment). In contrast to a previously developed system, our proposed algorithm's execution time for mining association rules is lower (i.e., 0.879 ms when support=3) and drastically changes as support increases. Figure 3 shows Support vertically and Execution time horizontally.

The Intelligent Heart Disease Prediction System was developed in 2012 by Chaitrali S. Dangare and Sulabha S. Apte using datasets on heart disease and the data mining and machine learning classification algorithms Decision Trees (J48), Naive Bayes, and Neural Networks. In this study, there were two datasets used. Both the Cleveland Heart Disease dataset and the Statlog Heart Disease dataset have 303 records each. In addition to the 13 commonly used attributes, smoking and obesity were added to the dataset for effective heart disease diagnosis. Both the 13 attribute dataset and the 15 attribute dataset were separately examined for comparative results. A total of 573 records were split into two data sets: 303 records were used for training, and 270 records were used for testing. The experiment makes use of the data mining and machine learning tool Weka 3.6.6. The Replace Missing Values (RMV) filter from Weka 3.6.6 was used to find missing values in the dataset and replace them with the most suitable values. The table below provides a comparison of the findings from our study. According to the findings, neural networks produce more accurate results than decision trees and naive bayes models.

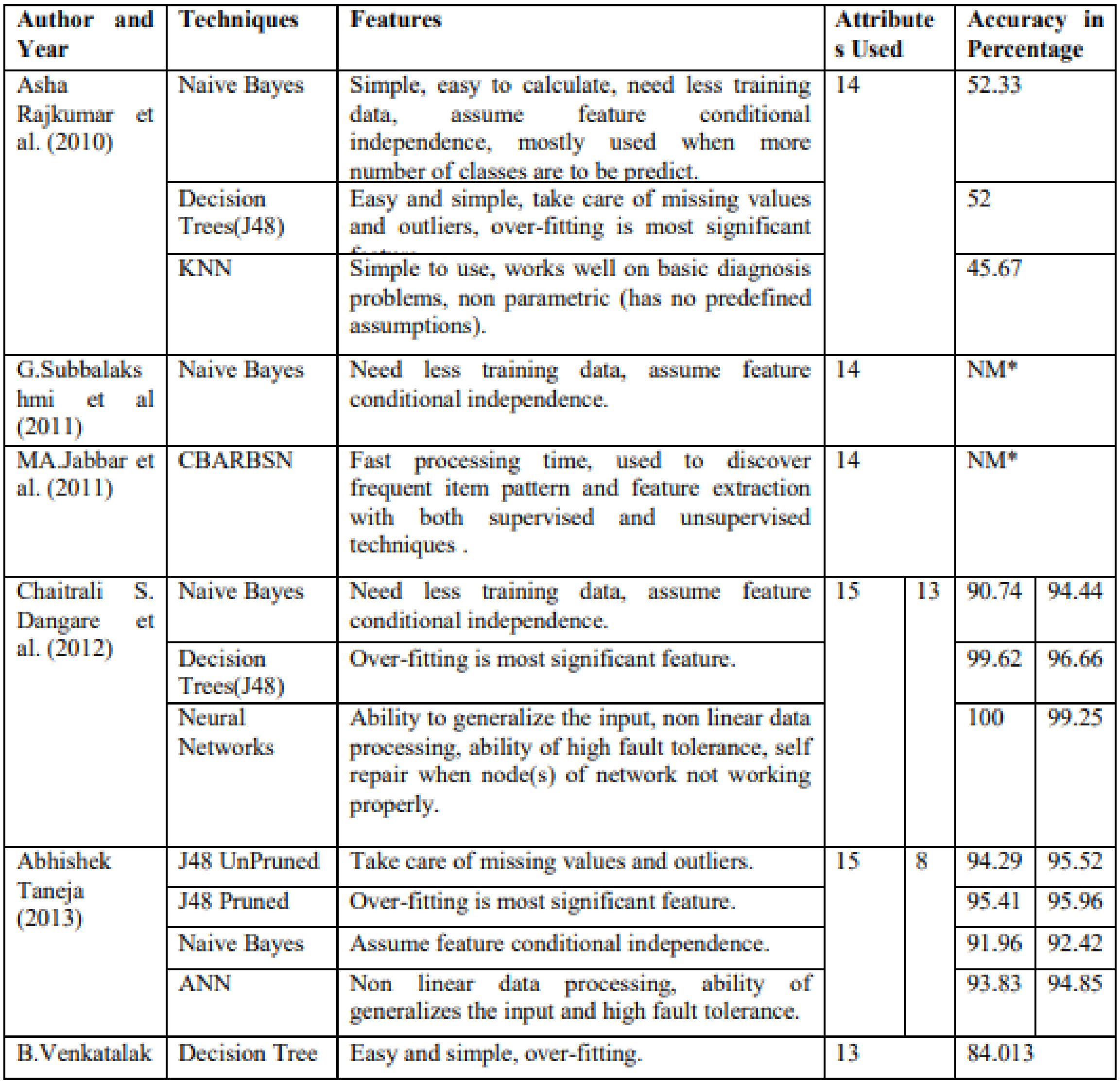


Classification Techniques	Accuracy with	
	13 attributes	15 attributes
Naive Bayes	94.44%	90.74%
Decision Trees(J48)	96.66%	99.62%
Neural Networks	99.25%	100%

Table 2. Comparative Results

N. Shirwalkar and T. Tak conducted an analytical study in 2018 and compared different data mining and machine learning techniques used in heart disease prediction to determine the most accurate method. The proposed heart disease prediction uses improved K-means and naive Bayes algorithms. From the UCI repository's Cleveland database, 303 records of patients with heart disease were extracted. Discreteization is the process of transforming the original dataset table from one form to another. Clusters are created from the dataset table using an improved version of the k-means algorithm. The model is trained using the Naive Bayes algorithm to forecast patients with heart disease. On the basis of the probability ratio produced by the Naive Bayes algorithm, we can predict four stages of heart disease using this model, namely Normal, Stage 1, Stage 2, and Stage 3. Navdeep Singh and Sonika Jindal created the Hybrid Genetic Model in 2018. Naive Bayes Model predicts heart disease with high accuracy using two supervised machine learning algorithms, namely Genetic Algorithm and Naive Bayes. A dataset of 303 records with the 14 required attributes has been taken from the online Cleveland database of the UCI repository and used in this propped model. Results were obtained in terms of precision (98%) recall (97.14%), and accuracy (97.14%), three different performance parameters. According to the findings, our proposed hybrid model outperforms existing models with the highest level of accuracy.

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	

Table 3. Summary of Literature Review

3. Our Proposed Algorithm Methodology

Computational Value Algorithm: Now in this Research Taking the analogy of Disease Prediction, rather than using some Data Mining Algorithm already in Existence a New algorithm is developed to handle the Prediction Accuracy Most Effectively, So this New Algorithm Model is Integrated into the Software with the Formula

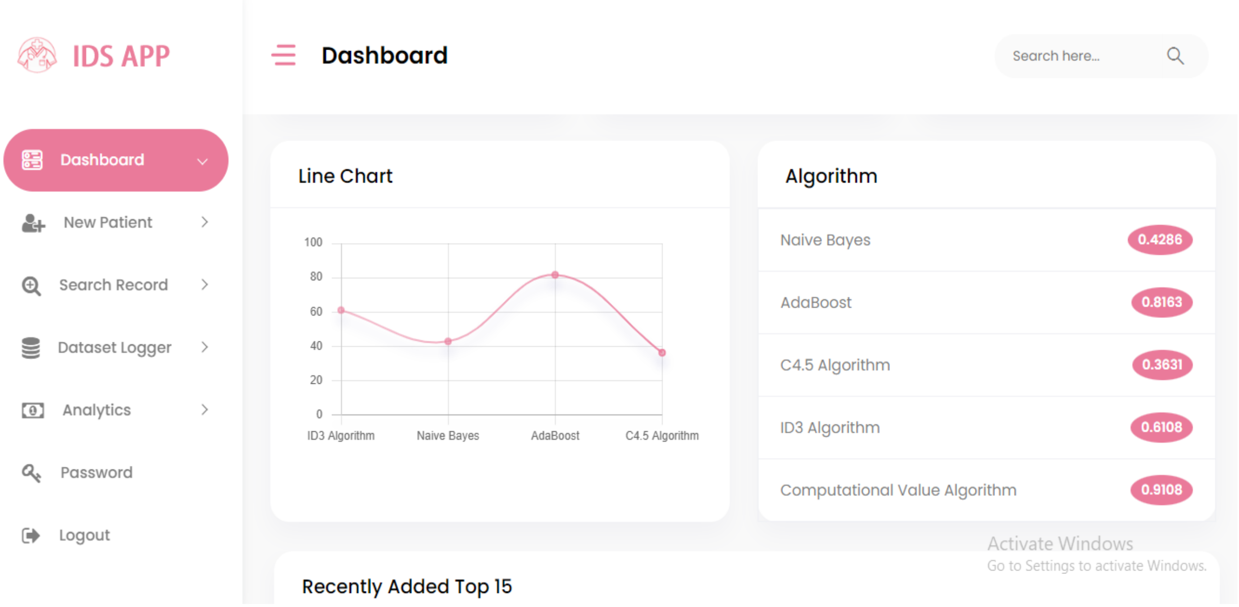
$$\mathbf{P(V)= (AnXn)+(Xn+1)}$$

Where the probability that we are interested in calculating P(V) is called the Computational value, (Xn) is the Data set Instance and (An) is the pulse average Reading. Below is a table showing a Comparative analysis of the Algorithms used in the proposed System.

Classification	Accuracy	Timing Taken
Naïve Bayes	85%	0.42 secs
AdaBoost	81%	0.81 secs
ID3 Algorithm	61%	0.61 secs
C4.5 Algorithm	36%	0.36 secs
Computational Value Algorithm	91%	0.31 secs

Table 4. Comparative analysis of prediction accuracy

The comparison findings are supported by 10-fold cross validation. These classification algorithms are compared, with our proposed computational value algorithm emerging as the superior performing algorithm. This compared to other algorithms in the system, it is more accurate and takes less time to build the model.



Algorithm	
Naive Bayes	0.4286
AdaBoost	0.8163
C4.5 Algorithm	0.3631
ID3 Algorithm	0.6108
Computational Value Algorithm	0.9108

Algorithm Analysis Interface

Results and Discussion

Now, based on an examination of numerous recent studies that used various data mining and machine learning techniques and algorithms to predict the development of heart disease. We discover that various data mining and machine learning techniques are applied to forecast heart disease with the aid of various experimental tools like WEKA, MATLAB, etc. In various experiments, different patient datasets with heart disease are used. The majority of experiments use data from the UCI repository's online Cleveland database. The dataset consists of 303 records with a total of 75 attributes, 14 of which are essential, and some of which have missing values. On various datasets, fewer experiments have been conducted. According to the study, neural networks with 15 attributes perform with 100% accuracy in one experiment while neural networks with 8 attributes perform with 76.55% accuracy. In the majority of experiments with various numbers of attributes, Naive Bayes also provides high accuracy above (90%). The accuracy of decision lists (J48) in a case increases to 99.62%, which is a very good performance. As a result, the number of attributes used and the implementation tool used determine the accuracy of the various techniques. We draw the following conclusions from this study, which should be taken into account in future research for high accuracy and more accurate heart disease diagnosis using intelligent prediction systems. Most experiments train prediction models using a small, uniform dataset. So, in order to train and test our prediction models, we must use real data from a large number of heart disease patients at reputable medical facilities in our nation. The accuracy of our prediction models on large datasets must then be evaluated. For a more accurate diagnosis and high accuracy, we must consult highly qualified cardiology experts to prioritize the attributes based on their impact on the patient's health and, if necessary, add more crucial attributes of heart disease. It is necessary to create more intricate hybrid models for precise prediction by combining various data mining and machine learning techniques, as well as text mining of the unstructured medical data that is readily available in large quantities in medical institutions. Additionally, the use of genetic algorithms for feature selection and optimization significantly improves the overall performance of intelligent prediction models. In this study, classification techniques received more attention than regression and association rules did. Therefore, we must take these factors into account in order to produce better comparative results in future research. The choice of research tools and procedures directly affects the accuracy of the research. Therefore, selecting the right experimental tool (WEKA, METLAB, etc.) for technique implementation is also a crucial consideration.

References

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

[2] 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.

[3] 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.

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

[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] 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.

[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.