A Survey of Machine Learning Techniques for Heart Disease Prediction

Jyoti Tiwari
Research Scholar
Dept. of CSE
Sagar Institute of Research & Technology, Bhopal
Jyotitiwari1598@gmail.com

Dr Ritu Shrivastava
Head & Dean
Dept. of CSE
Sagar Institute of Research & Technology, Bhopal
ritushrivastava08@gmail.com

Prof Rupali Chaure
Research Scholar
Dept. of CSE
Sagar Institute of Research & Technology, Bhopal
rupali.cse@sirtbhopal.ac.in

ABSTRACT: Heart disease is one of the worst disorders in the world. It can't be seen with the naked eye, and it happens right away when its limits are reached. So, it needs to be diagnosed correctly at the right time. The health care business generates a substantial volume of data daily concerning individuals and disorders. Researchers and practitioners don't use this data very well, though. The healthcare business has a lot of data, but not a lot of information. There are several tools and methods for data mining and machine learning that can help you get useful information from databases and use that information to make better decisions and diagnoses. As research on cardiac disease prediction systems grows, it is important to highlight the incomplete studies on the subject. The primary aim of this research study is to synthesize recent comparative studies on heart disease prediction and to draw analytical conclusions. The study indicates that Naive Bayes combined with Genetic Algorithms, Decision Trees, and Artificial Neural Networks enhance the accuracy of the heart disease prediction system across many scenarios. This document provides a summary of regularly employed data mining and machine learning techniques along with their difficulties.

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

I. INTRODUCTION

Heart disease is a term that covers a number of problems that can affect your heart. Heart disease is a broad phrase that includes a number of conditions, such as coronary artery disease, which affects blood vessels; arrhythmias, which are abnormalities with the heart's rhythm; and congenital heart defects, which are heart disorders you're born with. People sometimes use the words "heart

disease" and "cardiovascular disease" to mean the same thing. Cardiovascular disease (CVD) usually means problems with blood vessels that are too thin or obstructed, which can cause a heart attack (myocardial infarction), chest pain (angina), or stroke. Heart disease is also caused by other problems with the heart, such as problems with the muscle, valves, or rhythm [3]. CVDs kill about 17.9 million people each year, which is about 31% of all deaths worldwide [4]. The healthcare sector generates substantial information regarding patients, disease diagnoses, and related matters; yet, this data is not utilized effectively by researchers and practitioners. Quality of service (QoS) is a big problem for the healthcare industry right now. QoS means effectively detecting diseases and giving patients therapies that work. A bad diagnosis can have terrible effects that are not acceptable [2]. There are many things that can make you more likely to get heart disease. Some risk factors that can't be changed are family history, age, race, and being male. But smoking, diabetes, high cholesterol, high blood pressure, not getting enough exercise, being overweight or obese are all things that can be stopped or changed [5].

Data mining is the process of using data mining and machine learning techniques, statistics, and database systems to find new, hidden patterns (knowledge) in massive quantities of data that already exist. The knowledge that has been found can be used to make smart predictive decision systems in many areas, such as health care, to make sure that the right diagnosis is made at the right time to save lives and give services that are cheap. Machine learning lets computer programs learn from data that has already been set and get better at what they do without any help from people. Then, they can use what they have learned to make smart choices. Machine learning programs get better every time they make a good decision. The graphic below shows the process of knowledge discovery from data (KDD).

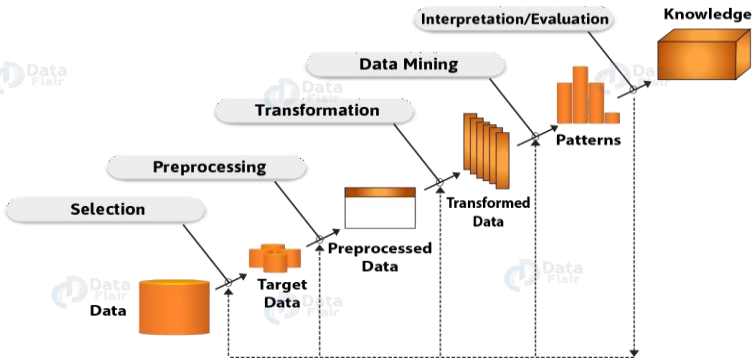


Figure.1 Steps in Knowledge Discovery Process [25].

II. 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:

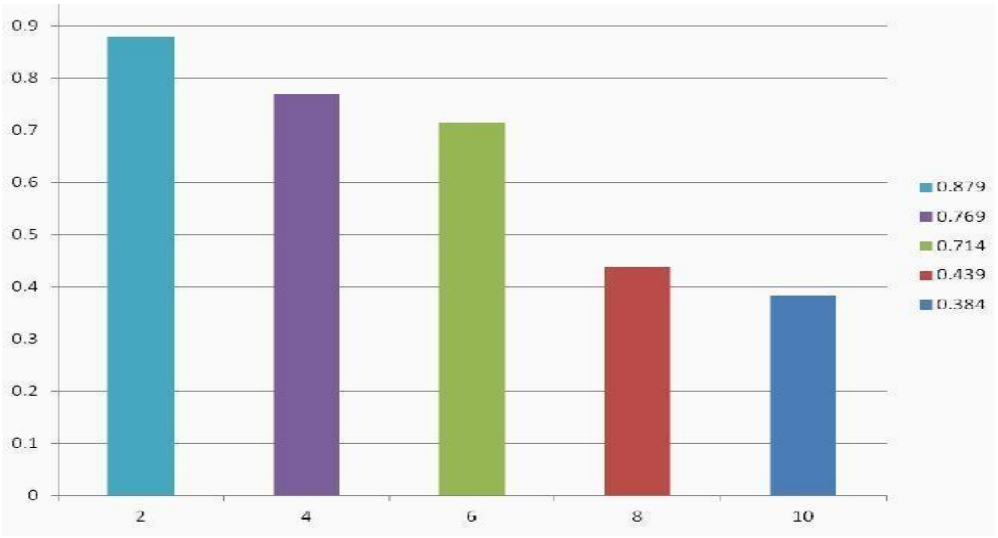
In 2010, A. Rajkumar and G. S. Reena applied machine learning algorithms such as Naive Bayes, KNN (K- nearest neighbors) and decision list for heart disease prediction. Tanagra tool is used to classify the data and the data evaluated using 10-fold cross validation and the results are compared in table 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 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

IV 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.
- 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.