ORIGINAL RESEARCH

Submitted: 2.07.2022; Accepted: 24.07.2022; Published online: 20.11.2022

A Predictive Model for Cardiovascular Diseases Using Data Mining Techniques

Avneesh Kumar[1]*, Santosh Kumar Singh[2], Shruti Sinha[3]

[1]Department of Computer Applications, Galgotias University, Greater Noida, Uttar Pradesh, India

[2]School of Health & Allied Science, ARKA JAIN University, Jharkhand, India

[3]School of Allied Science, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India

*Corresponding author:
avneesh.kumar@galgotiasuniversity.edu.in

Abstract

In many countries, heart disease is the main cause of mortality. Heart diseases are often identified by doctors based on recent clinical trial results and their prior experience treating patients who present comparable symptoms. Patients with heart disease require early diagnosis, prompt treatment or constant monitoring. The purpose of this study is to look into the numerous data mining technologies that have recently been developed for predicting heart disease. According to observations, 15 feature neural networks outperform all other data mining methodologies. Another finding from the analysis is that decision trees using genetic algorithms or feature subset selection have good accuracy. The results illustrate that the same classifier can occasionally produce precise results that change depending on the data mining approach used. According to the findings, a neural network with 15 characteristics has so far achieved a maximum efficiency of 100%. Decision Tree, on the other hand, has also done well, with 98.65% accuracy and 15 features.

Keywords

Cardiovascular Disease, Cardiomyopathy, Data Mining, Decision Tree, Neural Network

Imprint

Avneesh Kumar, Santosh Kumar Singh, Shruti Sinha. A Predictive Model for Cardiovascular Diseases Using Data Mining Techniques. Cardiometry; Issue 24; November 2022; p. 367-372; DOI: 10.18137/cardiometry.2022.24.367372; Available from: http://www.cardiometry.net/issues/no24-november-2022/predictive-model-cardiovascular

1. INTRODUCTION

The heart is the physical part that exerts itself the most. Every day, day or night, the “normal heart beats 100,000 times to transport oxygen” or nutrients all over the body. Waste products such as carbon dioxide are delivered by blood, which is pushed by the heart to the lungs, where they can be expelled from the body [1], [2]. To sustain life, the heart must work properly. Heart disease, also known as “coronary artery disease” (CAD), is characterized by an “accumulation of calcium, cholesterol”, as well as other lipids in the blood arteries that supply the heart. When this substance hardens, a plaque develops, preventing blood from reaching the heart. Angina is a type of chest discomfort that occurs when a “coronary artery narrows” owing to plaque buildup or another ailment. Angina is commonly confused with a heart attack [3], [4].

1.1. History

Every year, large numbers of people are affected by coronary disease, which is the leading reason of death for both male and female in the worldwide. According to the “World Health Organization,” heart disease is the cause of twelve million deaths worldwide. One person dies of heart disease worldwide at regular intervals. “Medical diagnosis” is a crucial yet difficult process that must be completed quickly and accurately. A suitable “computer-based information” or decision support system must be used to help lower the cost of conducting clinical tests [5], [6]. The use of software algorithms to uncover consistency as well as patterns in big data sets is known as data mining. Furthermore, with the increase of data mining over the last 20 years, there is a significant chance that computers may create and classify new attributes or classes on their own. Medical workers are better positioned to identify persons who are at higher risk of developing heart disease while they are cognizant of the factors associated with the condition [7], [8]. Age, total cholesterol, “blood pressure”, family background of heart disease, diabetes, hypertension, inactivity, fasting blood sugar, and obesity, other characteristics have been statistically proven to be risk factors for heart disease.

1.2. Data Mining

Data mining is a methodology that uses methods from machine learning, statistics, as well as database systems to find patterns in massive amounts of data. In this essential step, data patterns are extracted using intelligent methods. Data mining may employ clustering, time series analysis, association, prediction, and classification.

Exploring enormous databases for previously undiscovered patterns, correlations, and information that are challenging to find using conventional statistical techniques is called data. "Data mining" is the technique of extracting information from massive volumes of data [9], [10]. Uses for data mining would be utilized to improve health policy-making, illness prevention, hospital error prevention, early detection, or hospital avoidance. Utilizing patient clinical evidence, cardiovascular disease prediction systems can help doctors forecast heart diseases. As a result, it will be possible to more precisely estimate the likelihood that patients would be diagnosed with heart disease by applying data mining techniques to construct a computer - aided diagnostic system or doing data mining on various heart disease features [11], [12]. To increase the Decision Tree's accuracy in detecting patients with heart disease, this research introduces a novel model. It employs a different Decision Trees algorithm.

1.3. Heart Disease Types

Heart illness comes in a variety of forms that can affect various organ systems and manifest themselves in various ways:

1.3.1. Coronary Artery Disease

The most prevalent kind of cardiac illness is coronary artery disease (CAD). Plaque, which is comprised of components such as cholesterol and fat, lines the arteries that would deliver blood to the cardiac muscle in CAD. By narrowing the arteries as a result of plaque development (also known as atherosclerosis), the heart muscle receives less oxygen than is necessary for healthy function. Chest discomfort (angina) or even a heart attack could occur whenever the heart muscle is not getting enough oxygen.

1.3.2. Heart Failure

"Congestive heart failure" (CHF) occurs whenever the heart's ability to pump enough "oxygen-rich blood" to fulfill the needs of the body. This might be due to the heart's inability to beat with enough force or a lack of blood flow. Some folks experience both issues.

1.3.3. Heart valve disease

Whenever a portion of the heart's four valves malfunction, heart valve disease results. Heart valves aid in maintaining the forward flow of blood as it is pumped by the heart. It is challenging when the heart valves are ill (e.g., stenosis, mitral, and tricuspid prolapse).

1.3.4. Arrhythmia

Irregular and abnormal heartbeats are known as arrhythmias. This can be an irregular heartbeat, a rapid heartbeat (tachycardia), or even a sluggish heartbeat [13], [14]. Atrial fibrillation whenever the atria, as well as the upper heart chambers, contract abnormally, ventricular tachycardia contraction, extra beats that come from the ventricles, narrowing of the heart chambers, as well as Brady arrhythmias. Slow heart rhythm due to disease of both the conduction system of the heart. Some of the most common arrhythmias are.

As a result of cardiac muscle disease, the heart might expand or acquire thick heart walls (cardiomyopathy). The body's capacity to pump blood is diminished, which commonly results in heart failure. Figure 1 illustrates the many heart disease causes.

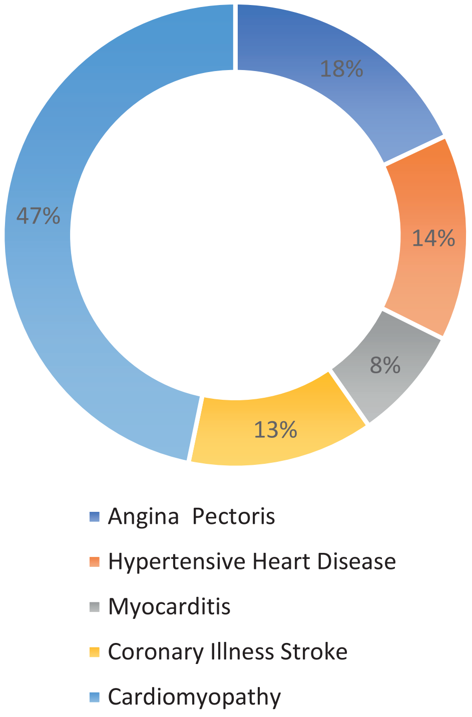

Figure 1: Illustrate the Various Reason behind the Heart Disease.

2. LITERATURE REVIEW

M. Anbarasi *et al.* studied about enhanced heart disease prediction. The initiative attempts to more reliably predict the existence of heart disease by using fewer features. Initially, thirteen parameters were used in order to anticipate heart disease. To limit the amount of tests a patient must do, our study use a genetic algorithm to identify the qualities most effective in the diagnosis of cardiac disorders. Within the same model design stage, the outstanding job of Naive Bayes was present both before and after attribute reduction. When compared to the other two approaches, "clustering-based classification" performs poorly. Furthermore, the findings show that after feature subset selection and very time-consuming model development, the "Decision Tree data mining technique" surpasses the other two data mining approaches [15].

T.Nagamani, *et al.* studied about "heart disease prediction" utilized Mapreduce and data mining. There are several research that have been done to develop models using Data Mining alone or in combination with computational methods like Naive Bayes (NB), Decision Tree (DT), "Machine learning, or unsupervised" classification algorithms such as KNN or "Support Vector Machine" (SVM). The suggested method takes a substantial number of medical incidents as input. When compared to traditional recurrent fuzzy neural networks, the experiment results show that the suggested technique could predict occurrences with an accuracy level of 98%. Furthermore, our Mapreduce technique beat earlier methods with prediction accuracy claims ranging from 95.00 to 98.00%. These findings imply that the "Mapreduce approach" may be capable of properly predict HD risks in a clinical setting [16].

Nidhi Bhatla and Kiran Jyoti studied about The goal of this effort is to lower the number of qualities utilized in prediction of heart disease, which could inevitably decrease the amount of tests required of a patient. A doctor uses a range of data or test sources to establish a diagnosis, although not all tests are required to diagnose a heart disease. Our efforts also aim to improve the efficacy of the offered approach. The findings demonstrated that Decision Tree and fuzzy logic-based Naive Bayes outperformed other data mining strategies [17].

Shamsher Bahadur Patel *et al.* studied about Heart disease patients' diagnoses using classification mining methods. Numerous applications, like as those in e-commerce, the healthcare industry, research, or engineering, depend on data mining. The healthcare industry uses data mining mostly for illness prediction. These 14 traits can be used to anticipate heart illness. The genetic algorithm, however, lowers fourteen qualities to six attributes. "Three classifiers Naive Bayes", "Decision Tree" as well as "Classification by Clustering", are used to prediction the analysis of "heart disease" after lowering the number of features [18].

3. METHODOLOGY

3.1. Design

To accurately diagnose heart disease, medical analysts and clinicians can benefit from examining the various data mining approaches presented in this research. The main method used in our research is the examination of recent articles, journals, or studies in the domains of "engineering or computer science", data mining, as well as heart disease. Figure 2 shows the model of data mining approach.

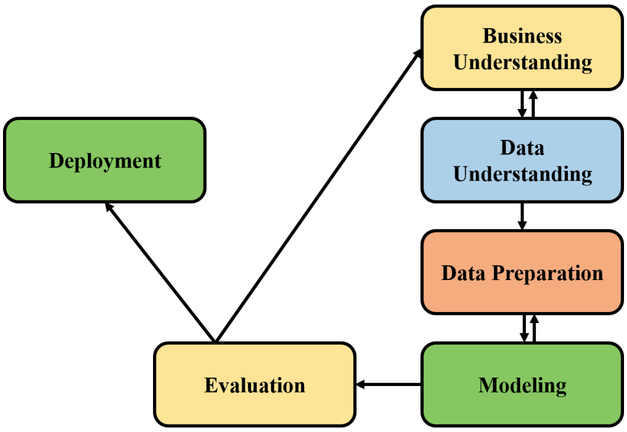

Figure 2: Model for the Data Mining Process.

3.2. Sample and Instrument:

Algorithms for Data Mining or "Supervised Machine Learning" were described in this paper for the classification of data using Decision List or "KNN supervised machine learning". The "Tanagra software" is utilized to categorize the data, or "10 folds cross-validation" is used to evaluate the data.

It suggests several data mining techniques from the fields of machine learning, databases, statistical learning, or exploratory data analysis. It offers a simple user interface and lets users evaluate both genuine and fake data. Additionally, this application offered users an architecture that made it simple for them to integrate their data mining techniques and assess how well they performed. There are several data sources, immediate access to databases and information warehouses, data purification, and interactive usage.

3.3. Data Collection

Experiments are done using the training data set, which consists of 2500 examples with 12 unique attributes. The dataset is divided in half based on the features, with 72% used for testing and the remaining 28% used for study. The accuracy and computing time of each approach are used to assess and compare their efficacy. Table 1, demonstrates the different algorithm performance, in the below table KNN show the 45. 69% accuracy and decision list 52% accuracy.

Table 1

Analysis of Various Algorithms' Performance.

S. no.	Used Algorithm	Accuracy	Time
1	KNN	45.69%	1000ms
2	Decision List	52.00	720ms

3.3.1. Neural Network and Data Mining

"Artificial neural networks" (ANN), often known as "neural networks" informally, are mathematical or computer models based on biological brain networks. To put it another way, it replicates the organic nervous system. In this study, a method for predicting heart disease was created utilizing 15 criteria. 13 factors were previously utilized for prediction, but this research included 2 new factors obesity or smoking for an accurate diagnosis of heart disease. Weka 3.6.7, a data mining program, is utilized in the experiment. The Replace Missing Values filter from 3.6.7 was first used to find missing values in the dataset or replace them with the proper values. Additionally, a database of cardiac diseases has been examined using different data mining approaches. For each classifier, a confusion matrix is found. The results of this research are shown in Table 2, which demonstrates that neural networks have performed better than other data mining techniques.

Table 2

Illustrate the Contrasting Different Data Mining Methods.

Classification of techniques	Accuracy
Decision Trees	98.65%
Neural Networks	100.00%

3.3.2. Genetic Algorithm and Fuzzy Logic

This study's method is an enhanced version of the model that combines fuzzy intelligent machines with evolutionary computation for effective classification as well as feature selection. It is incredibly acceptable to use fuzzy set theory or fuzzy logic to create knowledge-based procedures in healthcare for sickness diagnosis.

Using the fuzzy tools in Matlab, investigations are run. The Mamdani model of fuzzy inference system is employed in this. Based on their understanding of this field, the specialists who created the fuzzy rules. Researchers make use of a dataset from either the UCI machine-learning library, but we also discover that just 6 characteristics are both helpful and crucial for predicting heart disease in people. The suggested model's input is the collection of all the selected variables, and its objective is to provide a number between 0 and 1, which indicates whether or not patients have heart disease.

The input data is initially gathered as a crisp set and fuzzy, and then converted into a fuzzy set utilizing fuzzy linguistic factors, fuzzy linguistic words, or similarity measures as part of the fuzzy logic process. The conclusion is then reached using a set of principles, and the diffusion process is then completed. This system creates clear rules based on the gathered support sets, which are displayed in Table 3.

Table 3

Illustrate the values of the Support Set's Characteristics.

Variables	Supports Sets	
	Heart Patients	Non Heart Patients
Old peak	2.05-6.3	< 2.05
Exang (Exercise induces angina)	Yes	No
Ca	1, 2, and 3	zero
Chest Pain Type	4	1,2, and 3
Rbps	135 to 152	143 to 155
Thalach	72 to 135	166

4. RESULT AND DISCUSSION

4.1. Data Analysis

4.1.1. Using genetic Algorithms and Data Mining

The goal of this study was to minimize the number of criteria utilized in the diagnosis of heart disease. Originally, 13 characteristics were employed in this prediction, however this study had been using a genetic algorithm as well as feature subset selection to minimize the number of variables to six.

Genetic algorithms take into consideration natural evolution theory. Genetic discovery's beginning population has random rule selection but no initial features. The concept of survival of the fittest in the present population was utilized to encourage population expansion in order to meet any offspring of these regulations. To produce children, the optimization algorithms of cross-over or mutation were applied. The generation process went on indefinitely until population P formed with all the rules satisfying the fitness requirements. With a cross-over chance of 0.6 and a mutation probability of 0.034, the generations proceeded to the twentieth generation, with an initial population of 20 occurrences. Through a DNA search, six of the thirteen characteristics were discovered.

In addition to the genetic algorithm, the CFS Evaluator is also employed. Weka 3.6.0 tool is used to conduct the observations. 912 records with 15 characteristics made up the first data set. For simplicity, all qualities were made categorical, so contradictions were removed. Several classifiers are employed on the dataset to correspond to the 7 attributes after the 13 variables were reduced to 7, to predict heart disease. Table 4 displays a performance study of various classifiers. The table demonstrates that Decision Tree fared better than other methods with the best accuracy as well as the lowest average absolute error.

Table 4

Illustrate the Comparison for Two Classifiers.

Data Mining Techniques	Model Construction time	Accuracy	Mean Absolute Error
Classification Via Clustering	1.07s	88.5%	0.118
Decision Tree	0.08s	99.5 %	0.00017

4.1.2. Techniques for Data Mining and "Intelligent Heart Disease Prediction Systems" (IHDPS)

In this study, a prototype "Intelligent Heart Disease Prediction System" (IHDPS) was created by combining Decision Trees, or Neural Networks with data mining approaches. The NET framework is used to construct the web-based, scalable, user-friendly, dependable, and extendable IHDPS system. IHDPS can discover and retrieve hidden heart disease information from a "historical heart disease database". It can respond to intricate questions about the diagnosis of heart disease, assisting analysts or practitioners in the healthcare industry to make wise clinical judgments that "conventional decision-support" systems cannot. Offering efficient therapies also aids in lowering treatment costs. Additionally, it shows the outcomes in both tabular and visual formats, it's based on 15 characteristics, this IHDPS.

There were 910 entries in the Cleveland Heart Disease database. The records were used to construct two datasets: a training dataset (455.00 records) as well as a testing dataset (455.00 records) (456 records). The research found that "Decision Trees or Neural Networks" had the highest proportion of correct predictions for heart disease patients (86.53%). Nevertheless, decision trees scored (89.00%) when comparing to the other two models for predicting patients without heart disease, as seen in Table 5.

Table 5

Illustrate the Intelligent Heart Diseases Prediction Systems (IHDPS) Performance Evaluation.

Data Mining Technology	Accuracy
ANN	86.53%
Decision Trees	89.00 %

For ease of understanding, results from each data mining technique have been presented separately in distinct tables. To predict cardiac disease, several classifiers are employed in conjunction with various data mining approaches. The findings illustrate that the same categorization might produce precise results that differ on occasion for different data mining methodologies.

According to the investigation, the 15-featured neural network has an accurate results of 100% so far. The "Decision Tree", on the other hand, with 15 criteria, obtained 98.65% accuracy but also performed well. Additionally, the decision tree combined with the genetic algorithm and 7 features has been shown to have 99.5% efficiency.

5. CONCLUSION

This study compares two taxonomic functional data mining algorithms for the predicting of cardiovascular disease utilizing fewer variables. They are decision trees as well as neural networks, respectively. Our research decreases the amount of tests that a patient is required to undergo by utilizing a genetic algorithm to find the features that are most beneficial in "diagnosing cardiovascular disease". 15 characteristics are reduced to 7 traits via genetic search. Our study intends to provide an examination of several “data mining techniques” that might be applied in automated systems for “heart disease prediction”. This study describes several methods and data mining classifiers recently created for rapid and precise heart disease identification.

References

1. J. Premsmith and H. Ketmaneechairat, “A predictive model for heart disease detection using data mining techniques,” J. Adv. Inf. Technol., vol. 12, no. 1, pp. 14–20, 2021, doi: 10.12720/jait.12.1.14-20.

2. A. Ishaq et al., “Improving the Prediction of Heart Failure Patients’ Survival Using SMOTE and Effective Data Mining Techniques,” IEEE Access, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

3. J. S. Phul and S. Vatta, “Diagnose And Predict Diabetic Heart Diseases Using Data Mining Classification Techniques,” Int. J. Adv. Res. Sci. Eng., vol. 5, no. 11, pp. 187–193, 2016.

4. P. Singh, S. Singh, and G. S. Pandi-Jain, “Effective heart disease prediction system using data mining techniques.,” Int. J. Nanomedicine, vol. 13, no. T-NANO 2014 Abstracts, pp. 121–124, Mar. 2018, doi: 10.2147/IJN.S124998.

5. S. Maji and S. Arora, “Decision Tree Algorithms for Prediction of Heart Disease,” in Lecture Notes in Networks and Systems, vol. 40, pp. 447–454, 2019, doi: 10.1007/978-981-13-0586-3_45.

6. D. Ananey-Obiri and E. Sarku, “Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms,” Int. J. Comput. Appl., vol. 176, no. 11, pp. 17–21, 2020, doi: 10.5120/ijca2020920034.

7. L. Yahaya, N. David Oye, and E. Joshua Garba, “A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques,” Am. J. Artif. Intell., vol. 4, no. 1, p. 20, 2020, doi: 10.11648/j.ajai.20200401.12.

8. V. Haribaabu, V. Sivakumar, Selvakumarasamy, and V. Dixit, “Prediction of heart disease risk using machine learning,” Int. J. Mech. Prod. Eng. Res. Dev., vol. 8, no. Special Issue 2, pp. 605–614, 2018, doi: 10.24247/ijmperdfeb201867.

9. F. F. Firdaus, H. A. Nugroho, and I. Soesanti, “A Review of Feature Selection and Classification Approaches for Heart Disease Prediction,” IJITEE (International J. Inf. Technol. Electr. Eng., vol. 4, no. 3, p. 75, 2021, doi: 10.22146/ijitee.59193.

10. “Prediction of Heart Disease using Artificial Neural Network,” VFAST Trans. Softw. Eng., pp. 102–112, 2018, doi: 10.21015/vtse.v13i3.511.

11. G. Parthiban, A. S.K.Srivatsa, and A. Rajesh, “Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method,” Int. J. Comput. Appl., vol. 24, no. 3, pp. 7–11, 2011, doi: 10.5120/2933-3887.

12. R. Dbritto, A. Srinivasaraghavan, and V. Joseph, “Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods,” Int. J. Appl. Inf. Syst., vol. 11, no. 2, pp. 22–25, 2016, doi: 10.5120/ijais2016451578.

13. B. Martins, D. Ferreira, C. Neto, A. Abelha, and J. Machado, “Data Mining for Cardiovascular Disease Prediction,” J. Med. Syst., vol. 45, article no. 6, 2021, doi: 10.1007/s10916-020-01682-8.

14. A. Jerline Amutha, R. Padmajavalli, and D. Prabhakar, “A novel approach for the prediction of treadmill test in cardiology using data mining algorithms implemented as a mobile application,” Indian Heart J., vol. 70, no. 4, pp. 511–518, 2018, doi: 10.1016/j.ihj.2018.01.011.

15. N. C. S. N. I. M Anbarasi, E Anupriya, “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,” Int. J. Eng. Sci. Technol., vol. 2, no. 10, pp. 5370–5376, 2010.

16. T. Nagamani, S. Logeswari, and B. Gomathy, “Heart disease prediction using data mining with mapreduce algorithm,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 3, pp. 137–140, 2019.

17. N. Bhatla and K. Jyoti, “A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic,” Int. J. Comput. Appl., vol. 54, no. 17, pp. 16–21, 2012, doi: 10.5120/8658-2498.

18. S. Bahadur, “Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques,” IOSR J. Agric. Vet. Sci., vol. 4, no. 2, pp. 60–64, 2013, doi: 10.9790/2380-0426164.