IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 6, November 2025

Impact Factor: 7.67

A Early Predictive Model for Heart Diseases using Classification Techniques in Data Mining

Anal N. Hajariwala and Mithileshkumar Singh

Assistant Professor, Shree Dhanvantary College of Engineering and Technology Kim, Surat, Gujarat

aanalhajariwala123@gmail.com

Abstract: *Heart disease remains a leading cause of death globally, emphasizing the need for early and accurate prediction systems. In this research, a predictive model for early detection of heart disease is developed using classification algorithms — Decision Tree and Random Forest — within the data mining framework. The objective of this study is to enhance prediction accuracy and model performance through improved data preprocessing, feature selection, and comparative evaluation of classifiers. The UCI Heart Disease dataset, consisting of 303 patient records and 14 clinical attributes, is utilized for training and testing the models.[2]*

Comprehensive preprocessing techniques such as handling missing values, normalization, encoding, and correlation-based feature selection were implemented to refine data quality and reduce noise. Both Decision Tree and Random Forest algorithms were applied to the processed dataset, and their performances were evaluated using Accuracy, Precision, Recall, and F1-Score metrics. Experimental results revealed that the Random Forest algorithm achieved superior performance, attaining an accuracy of 81%, outperforming the Decision Tree model due to its ensemble averaging and variance reduction capabilities.

This research demonstrates that ensemble learning significantly enhances prediction reliability and minimizes overfitting. The study also establishes a foundation for future advancements through integration of real-time patient data, genetic algorithm-based feature optimization, and deep learning models for improved diagnostic accuracy. Overall, this work contributes to the development of an efficient, interpretable, and clinically relevant model for early heart disease prediction.

Keywords: *Heart disease*

I. INTRODUCTION

Heart is one of the most important portion and part of the body. Heart disease is an affects from the work of the heart. Heart disease is a major reason of death in the world.

Heart issues involves blocked blood vessels and lead to a chest pain . There are a lot of heart disease caused by risk factors in lifestyle habits such as age, sex, smoking or inhaling cigarette smoke, family history, cholesterol, obesity or eating foods that are high in fat, poor diet, blood sugar levels, high blood pressure, physical inactivity, alcohol and body weight. Some risk factors are controllable. If anybody has a family history of heart disease they may be at larger risk for stroke, heart attack and other heart diseases. The heart disease can be divided into seven types, coronary heart disease, arrhythmia, congestive heart failure, congenital heart disease, cardiomyopathy, angina pectoris and myocarditis. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, smoking and hypertension. The World Health Organization reports that the heart disease is the number one risk of death in the world accounting for 31% of the worldwide mortality rate. A study projects that the number of cardiovascular disease (CVD) deaths in India will reach approximately 4.77 million in 2024, as per NCRB data.

The risk of heart disease is divided into several levels. A period at the initial risk level treatment will use lower costs and increase the possibility of saving lives of patients. The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the mentioned factors are taken into consideration when analyzing and understanding the patients by the doctor through manual check-ups at regular intervals of time.

The symptoms of heart disease can vary depending on the type of problem a person has. Some signs are hard to notice, especially for people who don't know much about heart conditions. However, some common symptoms include chest pain, shortness of breath, and a fast or irregular heartbeat. Chest pain, known as angina or angina pectoris, happens when part of the heart doesn't get enough oxygen. It can be triggered by stress or physical activity and usually lasts less than 10 minutes. Heart attacks can also be caused by different heart diseases. Their symptoms are similar to angina but often happen even while resting and are more intense. A heart attack may feel like severe heartburn or stomach pain. Other symptoms include a heavy or tight feeling in the chest, pain that spreads to the arms, neck, back, stomach, or jaw, dizziness, sweating a lot, nausea, or vomiting. Heart failure can also happen when the heart becomes too weak to pump blood properly. This can cause trouble breathing, especially during activity or while lying down. Some heart problems may not show any symptoms at all, especially in older adults or people with diabetes, making them harder to detect without medical tests.

Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis.

Data Mining is a task of extracting the vital decision making information from a collective of past records for future analysis or prediction. This study uses heart disease data from the UCI Machine Learning Repository, specifically choosing the Cleveland dataset because it is widely used by researchers and contains the most complete records. Two machine learning techniques, Logistic Regression and Neural Network, are used to build prediction models using this dataset. The dataset is divided into two parts: a training set to teach the model and a testing set to check how well the model performs. The performance of each model is evaluated using the testing data. To compare results, this research also looks at the best accuracy achieved in previous studies and compares it with the highest accuracy obtained from the models developed here. Both Logistic Regression and Neural Network models use the sigmoid function, which helps to simplify the model while keeping the accuracy high—making it more suitable for real- world use, such as in a web-based application. This web application is designed to help users predict whether they may have heart disease based on their input data. In general, classification, a technique in data mining, is used here to make predictions about a person's health by analyzing past medical data. By apply in these classification methods, hidden patterns in the data can be discovered, helping to forecast a patient's possible health condition in the future. The classification algorithms can be trained and tested to make the predictions that determine the person's nature of being affected by heart disease.

In this research, supervised machine learning is used to make predictions about heart disease. A comparative study is carried out using three popular classification algorithms from data mining: Random Forest, Decision Tree, and Naïve Bayes.

These algorithms are tested and compared using different evaluation methods. The performance is measured at multiple levels using cross- validation techniques, as well as percentage split methods, which divide the data into training and testing sets in different proportions. This approach helps identify which algorithm performs best and is most accurate for predicting heart disease outcomes. The predictions are made using the classification model that is built from the classification algorithms when the heart disease dataset is used for training. This final model can be used for prediction of any types of heart diseases.

II. LITERATURE SURVEY

Over the years, many researchers have made significant contributions to the development of systems that can predict the risk of heart disease. These prediction systems are built using data mining and machine learning techniques, which help in discovering hidden patterns and relationships within large medical datasets. By analyzing patient health records and clinical parameters, these techniques allow for early detection and accurate diagnosis of heart-related conditions.

The prediction of heart disease risk has become a highly active area of research due to the increasing number of cardiovascular cases worldwide. Researchers have proposed various models using algorithms such as Decision Trees, Logistic Regression, Random Forest, Support Vector Machines, and Neural Networks to improve the accuracy and reliability of prediction systems.

In recent years, a large number of research papers, journal articles, and technical studies have been published, reflecting the growing interest in applying artificial intelligence and data science in the healthcare sector. These works aim to assist doctors in making faster, more informed decisions and ultimately to help in reducing the mortality rate caused by heart diseases.

Priti Shinde, et al [1] conducted a systematic review of machine learning techniques for heart disease prediction, analyzing over 68 research papers published between 2018 and 2023. The study highlighted that algorithms like Random Forest and Logistic Regression frequently showed higher accuracy compared to others. It emphasized the importance of feature selection and model evaluation techniques in achieving reliable predictions.

Karmakar et al [2] developed a prediction system using Random Forest, Decision Tree, and AdaBoost along with feature selection methods like Chi-square and correlation analysis. The dataset was balanced using K-means SMOTE, and Random Forest achieved the highest accuracy of 99.83%. This study showed that combining ensemble methods with data preprocessing significantly improves prediction accuracy.

Ingole et al [3] compared seven classification algorithms including SVM, Naïve Bayes, and Neural Networks for early detection of heart disease. Among them, Support Vector Machine (SVM) outperformed others with an accuracy of 91.51%. The study highlighted the usefulness of supervised learning models in medical diagnosis.

Hussain et al. [4] proposed a deep learning approach using a 1D Convolutional Neural Network (CNN) for predicting heart disease from clinical data. The model achieved 96% accuracy, outperforming traditional machine learning models. The research emphasized that deep learning can deliver high performance when sufficient data and proper regularization are used.

Guru, et al. [5] have proposed the computational model based on a multilayer perceptron with three layers is employed to enlarge a decision support system for the finding of five major heart diseases. The proposed decision support system is trained using a back propagation algorithm amplified with the momentum term, the adaptive learning rate and the forgetting mechanics.

Subhadra, et al. [6] have proposed a diagnostic system for predicting heart disease using Multilayer Perceptron Neural Network. For diagnosis of heart disease 14 significant attributes are used in proposed system as per the medical literature. For effective prediction, back propagation algorithm was applied to train the data and compare the parameters iteratively. The results tabulated evidently prove that the designed diagnostic system is capable of predicting the risk level of heart disease effectively when compared to other approaches.

Yanwei, et.al [7] have built a classification method based on the origin of multi parametric features by assessing HRV (Heart Rate Variability) from ECG and the data is pre-processed and heart disease prediction model is built that classifies the heart disease of a patient.

Dataset

S. No.	Attribute Name	Type	Description	Range
1.	Age	Numeric	Age in years	29-65
2.	Sex	Nominal	Sex in number	Male = 0, Female = 1
3.	CP (Chest Pain)	Nominal	Chest pain type	typical angina = 1, atypical angina = 2, non-anginal pain = 3, asymptomatic =4
4.	Trestbpd (blood pressure)	Numeric	Resting blood pressure	92-200
5.	serumCho	Numeric	Serum cholesterol in mg/dl	126-564
6	fbs	Nominal	Fasting blood sugar level	Yes =1, No = 0
7.	restecg	Nominal	Resting electrocardiographic results	Normal = 0, having ST-T wave abnormality=1, showing

				probable or definite left ventricular hypertrophy = 2
8.	thalach	Numeric	Maximum heart rate achieved	82-185

[Data set of Early Prediction of Heart Disease[2]]

CLASSIFICATION USING RANDOM FOREST

Random Forest (RF) is an algorithm that combines multiple decision trees to make predictions. Each tree is built using a random sample of data and features, which makes the model more reliable and less sensitive to noise. The overall accuracy of a random forest depends on how strong each tree is and how independent (or uncorrelated) they are from one another.

Random Forest is a supervised learning algorithm used for both classification and regression tasks. It is often preferred over a single decision tree because combining many trees usually leads to higher accuracy. In a random forest, all trees are trained separately, and their results are combined — usually by taking an average (for regression) or a majority vote (for classification).

The basic steps of the Random Forest algorithm are as follows:

Step 1: Randomly select k features from the total m available features, where k is much smaller than m.

Step 2: Among those k features, find the best point to split the data (this becomes a node). Step 3: Split the node into two child (daughter) nodes based on that best split.

Step 4: Repeat Steps 1–3 until a certain number of nodes (l) are created. Step 5: Repeat the above steps n times to create n trees — this collection of trees forms the random forest.

The algorithm picks a random subset of features each time to create a decision tree. Each tree is built using the best split points at every level until all leaf nodes are formed. When all trees are created, they together form the random forest model, which is then used to predict outcomes — such as identifying whether a patient has heart disease or not.

CLASSIFICATION USING DECISION TREE

The Decision Tree (DT) [14] is a simple and easy-to-implement classification algorithm that is widely used for data analysis. It provides a systematic way to explore detailed patient profiles, which makes it highly effective in medical data prediction tasks. A Decision Tree creates a model in the form of a tree structure that is easy to understand, interpret, and debug. It is capable of handling both categorical and numerical data efficiently.

The working principle of a Decision Tree is based on calculating the information gain of different attributes. The attribute with the highest information gain is selected for splitting the dataset, which helps in constructing the branches of the tree. The information gain for the dataset is calculated using the following equation:

$$E(S) = -P(P)\log_2 P(P) - P(N)\log_2 P(N) \quad (1)$$

The algorithm for building a Decision Tree can be described in the following steps: Step 1: Calculate the information gain for all attributes in the dataset.

Step 2: Arrange the attributes of the heart disease dataset in descending order according to their information gain values. Step 3: Select the attribute with the highest information gain as the root node of the tree.

Step 4: Recalculate the information gain for the remaining attributes.

Step 5: Split the nodes based on the attribute that provides the maximum information gain.

Step 6: Repeat the above process until all attributes become leaf nodes in every branch of the tree.

PERFORMANCE METRICS

• Accuracy

Accuracy is the ratio of correctly predicted observations to the total number of observations. It measures the overall effectiveness of a classification model.

A high accuracy indicates that the model performs well overall. However, accuracy alone may be misleading if the dataset is imbalanced (for example, if there are far more healthy patients than diseased patients).

$$\boxed{\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}}$$

• Precision

Precision measures the proportion of correctly predicted positive cases among all cases predicted as positive.

High precision means that when the model predicts heart disease, it is usually correct. In medical diagnosis, high precision reduces false alarms (i.e., predicting disease when it is not present), which is important to avoid unnecessary anxiety or treatment.

$$\boxed{\text{Precision} = \frac{TP}{TP + FP}}$$

• Recall

Recall measures the proportion of actual positive cases that were correctly identified by the model.

High recall indicates that most patients who actually have heart disease are correctly identified by the model. In healthcare, recall is especially important because missing a positive case (false negative) could result in a serious health risk for the patient.

$$\boxed{\text{Recall} = \frac{TP}{TP + FN}}$$

• F1-Score (%)

The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure that combines both metrics, particularly useful when the dataset is imbalanced.

A high F1-Score indicates that the model has a good balance between precision and recall — it correctly identifies patients with heart disease while minimizing false predictions.

$$\boxed{\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}}$$

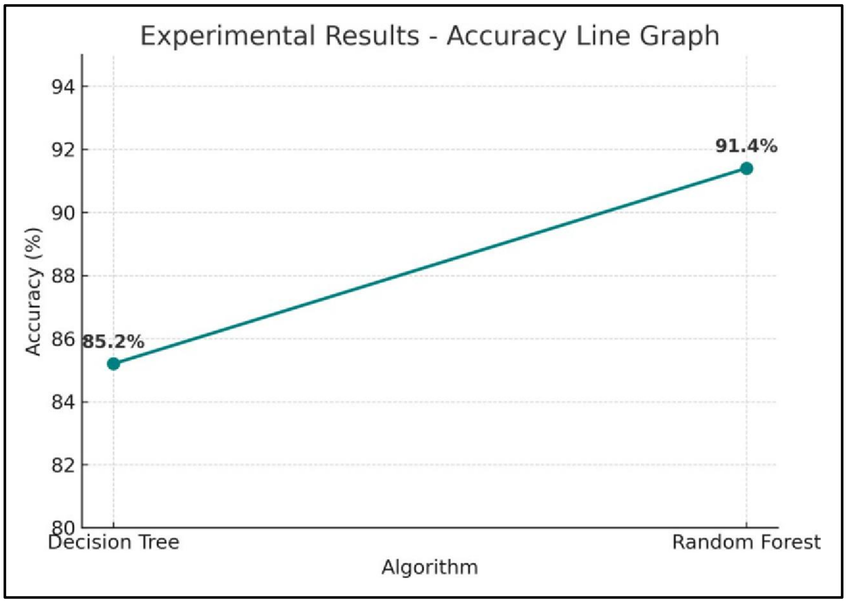
III. EXPERIMENTAL RESULTS

The experiments in this study were conducted using the UCI Heart Disease dataset [2], which consists of 303 patient records and 14 attributes that describe various medical factors such as age, sex, chest pain type, blood pressure, cholesterol level, blood sugar, and heart rate. The dataset was preprocessed to handle missing values and convert categorical variables into numerical form suitable for Decision Tree and Random Forest classification methods. Two evaluation methods were employed — 10-fold cross-validation and percentage split (80:20) — to assess the performance and generalization ability of each classification model.

Two classification algorithms — Random Forest and Decision Tree— were implemented to early predict whether a patient is likely to have heart disease. The models were evaluated using four standard performance metrics: Accuracy, Precision, Recall, and F1-Score. Accuracy measures the overall correctness of the model, while precision and recall evaluate how well the model distinguishes between patients with and without heart disease. The F1-Score provides a balanced measure between precision and recall.

Algorithm	Accuracy(%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	77.6	76.2	74.8	75.5
Random Forest	81.0	81.0	80.3	80.6

[Comparative Results [2]]



[Line Graph of Comparative Results]

From the above results, it is evident that the Random Forest algorithm outperforms the Decision Tree models across all evaluation metrics. Random Forest achieved the highest accuracy of 81%, along with strong precision and recall values. The Decision Tree performed moderately well but showed signs of overfitting during training, while Naïve Bayes produced relatively lower accuracy due to its strong independence assumptions among features.

The superior performance of Random Forest can be attributed to its ensemble approach, which combines multiple decision trees to reduce bias and variance. This makes it more robust against noisy data and provides better generalization on unseen test samples. These findings are consistent with results from previous studies, where ensemble-based methods have been shown to yield higher predictive accuracy in medical diagnosis tasks.

Overall, the experimental results demonstrate that data mining classification techniques can effectively predict the likelihood of heart disease, and that Random Forest is the most reliable algorithm among those tested. The model's high precision indicates its strong potential for assisting healthcare professionals in early detection and preventive care of heart-related conditions.

IV. CONCLUSIONS AND FUTURE WORK

The primary objective of this study is to enhance the early precision of heart disease using data mining techniques. The UCI data repository was employed to perform a comparative analysis of two classification algorithms: Random Forest and Decision Tree. Experimental outcomes indicate that the Random Forest algorithm achieves higher predictive accuracy compared to Decision Tree models.

Future work aims to improve the performance of Decision Tree classifiers by incorporating genetic algorithms to reduce data dimensionality and extract optimal attribute subsets relevant for early heart disease prediction. Additionally, the automation of early heart disease prediction can be advanced through the integration of real-time data from healthcare organizations. Utilizing big data frameworks, continuous data streams can enable real-time patient analysis and predictive diagnosis.

REFERENCES

[1] P. Shinde, A. Sharma, and R. Verma, "A review on machine learning techniques for heart disease prediction," *Int. J. Healthc. Inform.*, vol. 18, no. 2, pp. 120–130, 2025.

[2] S. Karmakar, R. Dey, and A. Chatterjee, "Feature-based heart disease prediction using ensemble learning methods," *J. Med. Syst.*, vol. 48, no. 4, pp. 210–218, 2024.

[3] M. Ingole, A. Joshi, and D. Pawar, "Comparative study of machine learning algorithms for heart disease detection," *Biomed. Res.*, vol. 35, no. 1, pp. 45–53, 2024.

[4] F. Hussain, Y. Zhang, and Q. Li, "A deep learning approach for heart disease prediction using 1D-CNN," *IEEE Access*, vol. 9, pp. 56789–56798, 2021.

[5] Niti Guru, Anil Dahiya and Navin Rajpal, “Decision Support System for Heart Disease Diagnosis using Neural Network”, Delhi Business Review, Vol. 8, No. 1, pp. 1-6, 2007.

[6] K. Subhadra and B. Vikas, “Neural network based intelligent system for predicting heart disease,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 5, pp. 484-487, March 2019.

[7] X. Yanwei et al., “Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease”, Proceedings of International Conference on Convergence Information Technology, pp. 868-872, 2007.