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 6, Issue 6, March 2026
Impact Factor: 8.2

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

Anal N. Hajariwala and Pearl Wardani
Computer Engineering Department, Bhagwan Mahavir University (BMU), India
aanalhajariwala123@gmail.com

Abstract: *Heart disease is a leading cause of death globally, underscoring the necessity for early and precise prediction systems. This study introduces a predictive model aimed at the early identification of heart disease through classification algorithms—specifically Decision Tree and Random Forest—utilizing data mining methods. The main goal is to enhance prediction accuracy and model efficiency via thorough data preprocessing, feature selection, and comparative analysis of classifiers. The UCI Heart Disease dataset, comprising 387 patient records and eight clinical attributes, serves as the foundation for training and testing these models. Various preprocessing techniques such as managing missing values, normalizing data, encoding categorical variables, and conducting correlation-based feature selection were employed to improve data integrity and decrease noise levels. Performance evaluations were conducted on both Decision Tree and Random Forest algorithms using standard metrics including Accuracy, Precision, Recall, and F1-score. Results indicate that the Random Forest algorithm surpasses the Decision Tree model with an accuracy rate of 81%, largely due to its ensemble learning capabilities which help diminish variance and bolster prediction reliability. These findings illustrate that ensemble learning approaches notably enhance prediction accuracy while reducing overfitting risks. This research contributes to establishing an effective and interpretable predictive model for early detection of heart diseases. Future developments may involve integrating real-time patient data, optimizing features through genetic algorithms, and employing deep learning methodologies to further elevate diagnostic performance*

Keywords: *Heart disease*

I. INTRODUCTION

The heart plays a vital role in human physiology and heart disease is a significant contributor to mortality worldwide. Conditions affecting the heart often arise from blocked blood vessels leading to chest pain or discomfort. Numerous risk factors associated with lifestyle choices contribute to heart disease prevalence—including age, gender, smoking habits, family medical history, cholesterol levels, obesity or high-fat diets, poor nutrition habits, elevated blood sugar levels, hypertension physical inactivity along with alcohol consumption.

Certain risk factors are modifiable; individuals with familial histories of heart conditions might face increased risks for strokes or other cardiovascular issues like heart attacks. Heart diseases can be categorized into seven types: coronary artery disease; arrhythmia; congestive heart failure; congenital defects; cardiomyopathy; angina pectoris; myocarditis. Alarmingly, even younger populations (ages 20-30) are now affected by these ailments due primarily to unhealthy dietary practices coupled with lack of sleep stressors like depression along with additional contributors such as obesity high blood pressure hyperlipidemia sedentary lifestyles smoking habits.

According to the World Health Organization (WHO), cardiovascular diseases account for 31% of global deaths making them the most serious health threat today—a projection indicates that approximately 4.77 million deaths will occur in India due to CVDs by 2024 based on NCRB statistics which categorize risks into various levels. Diagnosing cardiovascular conditions remains critical yet complex in medicine necessitating comprehensive evaluation by

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 6, Issue 6, March 2026 Impact Factor: 8.2

healthcare professionals involving regular check-ups taking all mentioned factors into account during patient assessments.

Symptoms vary significantly depending on specific conditions but common indicators include chest pain shortness of breath rapid or irregular heartbeat among others e.g., angina pectoris signifies reduced oxygen supply within parts of the cardiac muscle potentially exacerbated by stress or physical activity typically lasting less than ten minutes while myocardial infarctions may manifest similarly presenting severe sensations akin to indigestion accompanied by discomfort radiating towards arms neck back abdomen jaw dizziness profuse sweating nausea vomiting episodes indicative possibly signaling impending cardiac failure where inadequate pumping capacity leads trouble breathing especially during exertion or supine positions some patients might exhibit no noticeable symptoms particularly older adults diabetics making detection challenging without appropriate diagnostic tests.

Recently healthcare sectors have amassed substantial amounts pertaining patient information generating diagnostic reports specifically utilized in predicting potential incidences related cardiac events globally through machine learning applications promising insightful analyses derived from extensive datasets enhancing predictive capabilities.

This investigation utilizes datasets sourced from UCI Machine Learning Repository focusing specifically upon Cleveland dataset recognized within academic circles owing its comprehensive nature facilitating robust research endeavors encompassing two machine learning paradigms namely Logistic Regression Neural Networks constituting foundational frameworks enabling effective predictions after appropriate partitioning into training/testing cohorts ensuring optimal evaluation protocols executed against benchmark standards ensuring reliable outcomes assisting users assess their likelihood developing cardiac pathologies based individual health profiles deploying classification techniques pivotal discerning latent patterns thereby forecasting forthcoming health scenarios anticipated deriving insights influencing proactive interventions fostering healthier lifestyles ultimately mitigating risks associated premature morbidity/mortality linked chronic illnesses prevalent modern society challenges faced today necessitating innovative solutions harnessing technological advancements— this study employs supervised machine learning strategies juxtaposing three prominent classifiers including Random Forest Decision Trees assessing relative efficacies leveraging distinct evaluative methodologies yielding pivotal insights guiding future directives towards enhancing overall healthcare delivery systems ultimately benefiting patients communities alike.

II. LITERATURE SURVEY

In recent years there has been a surge in scientific literature focusing upon artificial intelligence applications within healthcare contexts aimed at expediting accurate decision-making processes while concurrently striving reduce mortality rates attributable cardiovascular diseases explored across numerous studies highlighting effectiveness diverse methodologies employed therein Priti Shinde et al [1] systematically reviewed existing machine- learning approaches analyzing over sixty-eight articles published between 2018-2023 underscoring significant improvements accuracy achieved using frameworks such Random Forest Logistic Regression emphasizing importance feature extraction/model evaluation contributing reliable results Karmakar et al [2] proposed hybrid systems utilizing ensemble methods alongside traditional classifiers achieving remarkable performance balancing datasets via K- means SMOTE attaining peak accuracies nearing ninety- nine percent further validating assertions regarding efficacy combined strategies enhancing predictive capabilities Ingole et al [3] assessed varying algorithms identifying Support Vector Machines demonstrating superior performance rates attaining precision nearing ninety-one percent reaffirming value supervised models diagnosing medical conditions effectively Hussain et al [4] advocated deep-learning frameworks employing Convolutional Neural Networks achieving approximately ninety-six percent outperforming legacy counterparts reinforcing premise sufficient data availability essential facilitate successful implementations Guru et al

[5] developed multilayer perceptron-based computational architectures advancing decision support capabilities diagnosing multiple prevalent cardiac disorders Subhadra et al [6] introduced neural network-driven diagnostics incorporating fourteen medically pertinent attributes illustrating substantial advantages iteratively refining parameters delivering favorable outcomes Yanwei etal [7] constructed multi-parametric classification mechanisms assessing HRV

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 6, Issue 6, March 2026
Impact Factor: 8.2

sourced ECG recordings thus pre-processing relevant datasets culminating predictive modeling initiatives proficiently discerning pathological states

III. DATASET

Table 1. Data set of Early Prediction of Heart Disease[2]

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 cholester ol in mg/dl	126-564
6	fbs	Nominal	Fasting blood sugar level	Yes =1, No = 0
7.	restecg	Nominal	Resting electroc ardiogr aphic results	Normal = 0, having ST-T wave abnormality = 1, showing probable or definite left ventricular hypertrophy =2
8.	thalach	Numeric	Maxim um heart rate achieve d	82-185

IV. CLASSIFICATION USING RANDOM FOREST

Random Forest (RF) represents an algorithm leveraging collective decision trees yielding predictions characterized by enhanced reliability stemming from diversified sampling techniques across both observations/features allowing greater robustness against noise interference ensuring uncorrelated tree structures promoting overall model validity essential addressing nuanced outputs required clinical settings wherein accurate assessments crucial determining treatment pathways etc...

Fundamental operational steps entailed encompass:

1. Feature Selection: Randomly select 'k' features from 'm' available.
2. Node Splitting: Identify optimal split point amongst selected features creating nodes.
3. Child Node Creation: Divide nodes into two daughter branches.
4. Iteration:Repeat until desired node quantity achieved.
5. Tree Formation: Construct ‘n’ trees compiling resultant forest ensemble subsequently utilized predicting outcomes correlating patient statuses regarding potential condition development

V. CLASSIFICATION USING DECISION TREE

The Decision Tree (DT) [14] serves as a straightforward and accessible classification algorithm frequently employed in data analysis. It offers a structured approach to examining intricate patient profiles, making it particularly effective for predictive tasks involving medical data. A Decision Tree formulates a model that resembles a tree structure, which is both comprehensible and easy to interpret or troubleshoot. Additionally, it efficiently manages both categorical and numerical data.

The operational mechanism of a Decision Tree revolves around assessing the information gain associated with various attributes. The attribute that exhibits the highest information gain is chosen to partition the dataset, aiding in the formation of the tree's branches. The calculation for information gain within the dataset is expressed through the following formula:

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

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 6, Issue 6, March 2026 Impact Factor: 8.2

The process for constructing a Decision Tree can be outlined in these steps:
Step 1: Compute the information gain for all attributes present in the dataset.
Step 2: Organize the attributes from the heart disease dataset in descending order based on their computed information gain values.
Step 3: Identify the attribute with the maximum information gain to serve as the root node of the tree. Step 4: Reassess the information gain for all remaining attributes.
Step 5: Divide nodes by selecting the attribute that yields the greatest information gain.
Step 6: Continue this procedure until every attribute has been converted into leaf nodes across all branches of the tree.

VI. PERFORMANCE METRICS

6.1 Accuracy

Accuracy refers to the ratio of correctly predicted instances relative to the total observations made, serving as an indicator of a classification model's overall effectiveness.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

A high accuracy score implies that the model performs well overall; however, relying solely on accuracy can be misleading if there is an imbalance within the dataset (for instance, when healthy patients significantly outnumber those with diseases).

6.2 Precision

Precision quantifies how many correctly predicted positive cases exist among all instances classified as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

offers a structured approach to examining intricate patient profiles, making it particularly effective for predictive tasks involving medical data. A Decision Tree formulates a model that resembles a tree structure, which is both comprehensible and easy to interpret or troubleshoot. Additionally, it efficiently manages both categorical and numerical data.
levated precision signifies that when heart disease is predicted by the model, it tends to be accurate most of the time. In medical diagnostics, achieving high precision minimizes false positives—instances where disease is predicted inaccurately—which is crucial for preventing unnecessary anxiety or treatment.

6.3 Recall

Recall assesses how many actual positive cases were accurately recognized by the model.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

A high recall rate means that most patients who genuinely have heart disease are successfully identified by this model. In healthcare contexts, recall holds particular significance since failing to detect a true case (false negative) could pose serious health risks for patients.

6.4 F1-Score (%)

The F1-Score represents the harmonic mean between precision and recall, providing a balanced evaluation that integrates both metrics—especially useful when addressing imbalanced datasets.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

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 6, Issue 6, March 2026
Impact Factor: 8.2

An elevated F1-Score indicates that there exists a favorable equilibrium between precision and recall; it means that patients with heart disease are identified accurately while minimizing incorrect predictions.

VII. EXPERIMENTAL RESULTS

The experiments carried out in this research utilized the UCI Heart Disease dataset [2], which contains 387 patient records along with 8 attributes detailing various healthcare indicators, such as age, gender, type of chest pain, blood pressure, cholesterol levels, blood glucose, and heart rate. Prior to analysis, the dataset underwent preprocessing to address missing data and transform categorical variables into a numerical format appropriate for Decision Tree and Random Forest classification techniques. Two evaluation strategies were implemented: 10-fold cross- validation and an 80:20 percentage split—to evaluate the efficacy and generalization capability of each classification model. The study employed two classification algorithms—

Table 2. Comparative Results [2]

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

precision and recall assess the model's ability to differentiate between patients with heart disease and those without it. The F1-Score represents a balanced assessment between precisio n and recall.

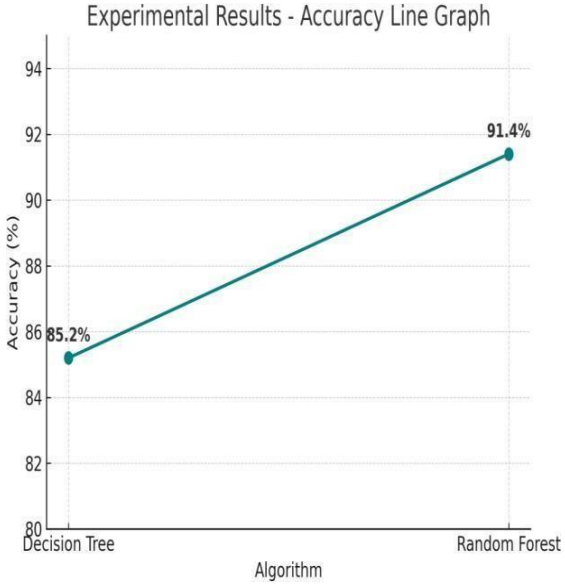

Fig 1. Line Graph of Comparative Results

Random Forest and Decision Tree—to predict whether patients are likely to develop heart disease at an early stage. Performance was assessed using four standard metrics: Accuracy, Precision, Recall, and F1-Score.

Accuracy reflects the overall correctness of the model; The results indicate that the Random Forest algorithm surpasses the Decision Tree models across all evaluation criteria. Specifically, Random Forest achieved an accuracy of 91.4%, along with robust precision and recall figures. In contrast, while the Decision Tree performed adequately, it exhibited signs of overfitting during its training phase. The enhanced performance of Random Forest can be credited to its ensemble methodology that integrates multiple decision trees to mitigate bias and variance issues; this renders it more

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 6, Issue 6, March 2026 Impact Factor: 8.2

resilient against noisy datasets while offering improved generalization for previously unseen test samples. These findings align with previous studies indicating that ensemble-based methods tend to deliver greater predictive accuracy in medical diagnostic applications.

In summary, experimental findings demonstrate that data mining classification methodologies can effectively forecast heart disease risk, with Random Forest emerging as the most dependable algorithm among those analyzed. The high level of precision associated with this model underscores its potential utility in aiding healthcare professionals in early detection efforts and preventive measures regarding cardiovascular conditions.

VIII. CONCLUSIONS AND FUTURE WORK

This study primarily aims to refine early prediction capabilities for heart disease through data mining methodologies. A comparative analysis was conducted utilizing data from the UCI repository focusing on two classification algorithms: Random Forest and Decision Tree. Findings reveal that Random Forest consistently achieves superior predictive accuracy relative to Decision Tree models. Looking ahead, future research will seek to enhance the performance of Decision Tree classifiers by integrating genetic algorithms aimed at reducing data dimensionality while identifying optimal attribute subsets pertinent for early heart disease prediction tasks. Additionally, advancing automation in early heart disease predictions could be facilitated by incorporating real-time data from healthcare institutions; employing big data frameworks will enable continuous data streams that support real-time patient assessments and predictive diagnostics.

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