ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

An Efficient Machine Learning Based Model To Predict Heart Disease

Ravindra Chauhan, Anshika yadav, Sneha Aggarwal, Gungun Tyagi, Tania

R.D. Engineering College Ghaziabad

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150300120

Received: 03 April 2026; 04 April 2026; Published: 23 April 2026

ABSTRACT

Across the globe, cardiovascular diseases remain a leading contributor to death rates and accurate prediction are essential to modern health systems. Unhealthy lifestyles are one of the elements leading to the increasing occurrence of heart disease, stress and aging, and it has become essential to create a system capable of delivering precise and reliable results diagnosis. With the growing accessibility of vast healthcare data, machine learning technology is emerging as an important tool for helping clinical decision-making by identifying hidden patterns and relationships in complex data sets. In this study, we developed a machine learning-based system for predicting heart disease. The proposed system uses a structured set of data obtained from a publicly available UCI source, contain important medical parameters. To guarantee high data quality and raise the level of performance of models, multiple preprocessing techniques were implemented, including data cleaning, feature normalization, and handling of missing values, classification variable encoding and outlier detection. Different approaches were tested to identify the most effective model. The models were evaluated based on performance indicators such as recall, accuracy, and precision and ROC-AUC points. The study focuses on the performance of ensemble learning using Random Forest, while comparative analysis shows that KNN achieved slightly higher accuracy on the given dataset. K-Nearest Neighbors performed the best, achieving an accuracy of around 91.8% and superior classification capabilities indicated by ROC curves and overall evaluation metrics. Our proposed approach can be used as an effective decision-making tool for medical professionals to identify high-risk patients in time. Finally, this approach helps reduce mortality rates and can assist doctors in early detection and better decision-making.

Keywords: Heart Disease Prediction, Machine Learning (ML) Techniques, Random Forest Classifier, Python, Supervised Learning, Logistic Regression Algorithm (LR), Support Vector Machine Algorithm (SVM), K-Nearest Neighbors (KNN), Decision Tree, Data Preprocessing, Feature Selection, Jupyter Notebook.

INTRODUCTION

Cardiovascular diseases rank as the foremost cause of mortality globally, representing a significant health concern in both developed and developing nations. The heart plays crucial part in preserving the system's appropriate operation of the human body by transporting oxygen and nutrients through the circulatory system. Any disruption to this process may result in serious health concerns, including heart related disease and sudden heart failure. The rise in heart disease is largely due to modern lifestyle elements like a poor diet and insufficient exercise, smoking, stress and genetic predisposition, and is therefore a crucial field of study within medicine.

Early diagnosis of heart disease is crucial for understanding effective treatment and prevention. However, accurate prediction is often difficult because of multiple interconnected clinical factors and various physiological aspects. Existing diagnostic methods largely based on clinical experience and cannot always capture hidden patterns in large and complex medical data sets. This limitation highlights the need for intelligent and data-driven systems to help doctors and professionals to make more accurate and timely decisions. Medical diagnosis field has received a lot of attention lately, due to its capacity to analyze enormous amounts of data and identify complex patterns and relationships between features.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

Machine learning techniques like logistic regression, support vector machines (SVMs), K nearest neighbors (KNNs), decision trees and aggregation methods, are widely used to predict heart disease. These techniques allow automated prediction by observing from existing patient data and identifying patterns connected to the emergence of illnesses. Despite this, many previous studies face challenges such as limited prediction accuracy and inadequate treatment of characteristic interactions, and insufficient generalization in different datasets. Furthermore, some models are too modified or do not precisely adapt non-linear connections between medical attributes.

This study aims to overcome these constraints proposes a model for the forecast of heart disease using many machine learning algorithms, focusing mainly on a random forest classification. The model undergoes testing and training on a UCI database containing relevant clinical features. To enhance data quality and model performance, a variety of preprocessing methods are employed. Our study shows that the KNN model achieves best performance which is 91.8 % accuracy, indicating its effectiveness in heart disease prediction. Our aims are to provide a reliable and efficient decision support system to help medical professionals identify and assess risks early. Our approach uses AI&ML which contributes to improving prediction accuracy and supporting better health outcomes.

LITERATURE REVIEW

In recent years, various studies have been performed to improve early forecasting cardiovascular disease using machine learning techniques. Researchers have studied data-driven approaches for analyzing clinical data sets and identifying patterns associated with cardiovascular diseases. These methods are intended to assist medical professionals in making accurate and timely decisions and ultimately reduce mortality rates. Machine learning algorithms have demonstrated considerable promise in deriving valuable insights from intricate medical data that traditional methods cannot capture. Previous research focused on classical classification algorithms like Logistic regression and decision trees for predicting heart disease. These techniques' simplicity and clarity serve as a foundation for forecasting modeling.

Logistic regression is widely used in binary classification problems, and decision trees represent effective in handling nonlinear relationships between features. However, when applied to large and complex datasets, these models frequently experience issues with limitations such as lower accuracy and sensitivity to data change. To achieve overcome these challenges, advanced machine learning algorithms, such as support vector machines (SVMs) and K-nearest neighbors (KNNs), have been introduced. SVM is well-known for establishing optimal and managing high-dimensional data boundaries for medical classification tasks. Even so, the prediction performance is improved by these techniques, they may require careful parameter adjustments and require computational costs for large data sets.

In recent times, combined learning techniques such as random forest have received widespread attention due to, they have improved prediction accuracy and reduced over-adaptation. Random forests integrate numerous decision trees to produce more robust and reliable predictions. Various research has shown that ensemble models have the advantage of outperforming individual classifications by recording complex character interactions and reducing variances. In addition, techniques for choosing features are utilized in many research to find out the most important clinical traits, making models even better. Despite progress in this region, there exist still some limitations. Many studies indicate problems such as imbalanced data sets, insufficient generalization between different populations, and the lack of complex model interpretation.

Furthermore, some models do not effectively use all available clinical features, which can affect predictive accuracy. These difficulties underscore the need for greater efficient and scalable approaches to providing precise and trustworthy predictions. Considering this situation, the current study centers on creating a heart disease predicting system employing a range of machine learning algorithms, with a focus on KNN classification. This study aims to achieve higher prediction accuracy and better performance by applying appropriate pre-processing techniques and evaluating different models. The results contribute to ongoing research into intelligent health systems and show the efficacy of group learning in medical diagnosis.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

METHODOLOGY

This section explains the complete implementation of the forecasting of heart disease using AI&ML techniques. The process includes dataset collection, preprocessing, model training, and evaluation.

Dataset Description

In this research, the dataset used from a publicly available UCI heart disease dataset. It contains 303 records and 14 attributes, including 13 input features and 1 target variable.

Target:0 → No heart disease

1 → heart disease present

Important features such as age, gender, chest pain type (cp), cholesterol (chol), resting blood pressure (trestbps), and maximum heart rate (thalach) into training and testing sets using an 80:20 ratio to evaluate model performance.

Data Preprocessing

To improvise the data and model performance, below given steps were applied:

- Data Cleaning: Missing and inconsistent values were handled to ensure dataset reliability.
- Encoding: Categorical variables include cp, thal, and slope were changed into numerical form through label encoding.
- Normalization/Scaling: Feature values measured using Min-Max normalization defined as

$$X' = \frac{X - Xmax}{Xmax - Xmin}$$

to maintain uniformity across all features.

- Outlier Handling: Extreme values were identified and treated to minimize their impact on model performance. These steps of preprocessing considerably improved the precision and stability of the systems.

Model Implementation

Many machine learning techniques were deployed to predict heart disease, including Logistic Regression algorithm, Support Vector Machine (SVM) algorithm, K-Nearest Neighbors algorithm (KNN), Decision Tree, and Random Forest algorithm. Each model was supervised using the pre-processed dataset and estimated on test data.

- Logistic Regression: Predicts probability using the sigmoid function

$$P(Y = \frac{1}{X}) = \frac{1}{(1 + e^{\wedge} - (\beta 0 + \beta 1X1 + \ldots + \beta nXn))}$$

- Support Vector Machine (SVM): Separates data using a hyperplane defined as

$$\mathbf{w \cdot x + b = 0}.$$

- K-Nearest Neighbors (KNN): Classifies data based on similarity using Euclidean distance

$$d = \sqrt{[(x1 - y1)^{\wedge}2 + \ldots + (xn - yn)^{\wedge}2]}$$

- Decision Tree: Splits data based on feature conditions using measures such as Gini Index.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

• Random Forest: A collective approach that integrates various decision trees, where prediction is given as

$$\hat{y} = (\frac{1}{N}) \Sigma Ti(x)$$

Among these, Random Forest provided better performance because of its capacity to handle complex data patterns and reduce overfitting.

D. Performance Evaluation

The model's performance was calculated through standard classification metrics:

• $Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$

• $Precision = \frac{TP}{(TP + FP)}$

• $Recall = \frac{TP}{(TP + FN)}$

• ROC-AUC Score: Evaluates how well the model can differentiate between classes.

According to the evaluation's findings Logistic Regression achieved approximately 88.5% accuracy, SVM 90.2%, Random Forest 88.5%, and the highest accuracy was attained by KNN of approximately 91.8%, making it the model that works best for heart disease prediction.

RESULTS AND DISCUSSION

The performance of various machine learning models was assessed to identify the most efficient method for predicting heart disease. The study incorporates several models, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest. The evaluation was done using key performance metrics such as accuracy, precision, recall, and ROC-AUC score.

Accuracy Comparison

The accuracy of all implemented models is compared in Fig.2. The findings suggest that the K-nearest neighbor classifier perform better than other algorithms, achieving the highest accuracy of approximately **91.8%**.

In comparison, SVM, LR and Random Forest achieved moderate accuracy levels, while Decision Tree showed relatively lower performance.

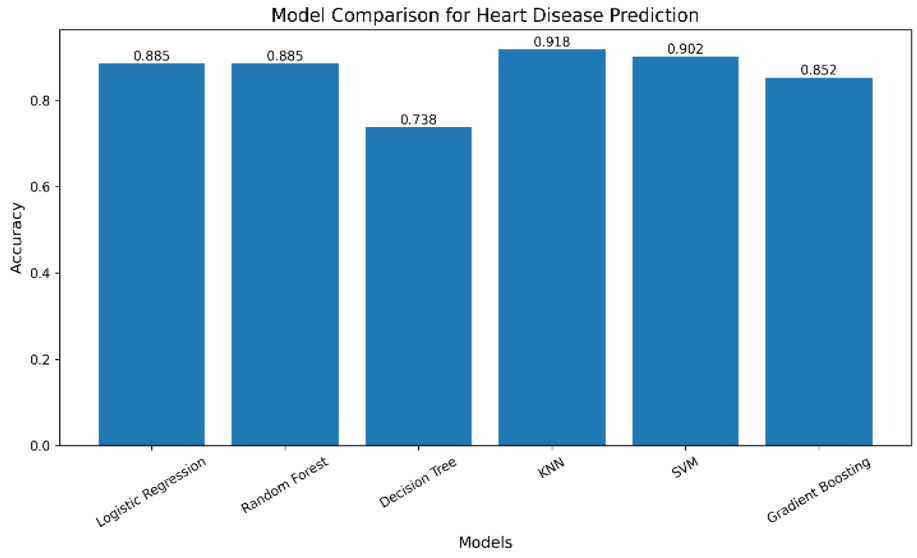

Fig.1: Comparison of accuracy among different machine learning systems, where KNN attained the highest accuracy.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

Dataset Distribution Analysis

The dataset distribution utilized in this research is illustrated in Fig. 3. The chart illustrates the count of patients categorized into two categories: heart disease and no heart disease. It is evident that 526 instances belong to the heart disease class, while 499 instances correlate with the no heart disease class. This indicates that the dataset is relatively balanced, approximately the same quantity of samples in each class. A balanced dataset is important for machine learning models, as it prevents bias toward a particular class and improves the reliability of predictions. The slight difference in class distribution does not significantly impact model performance.

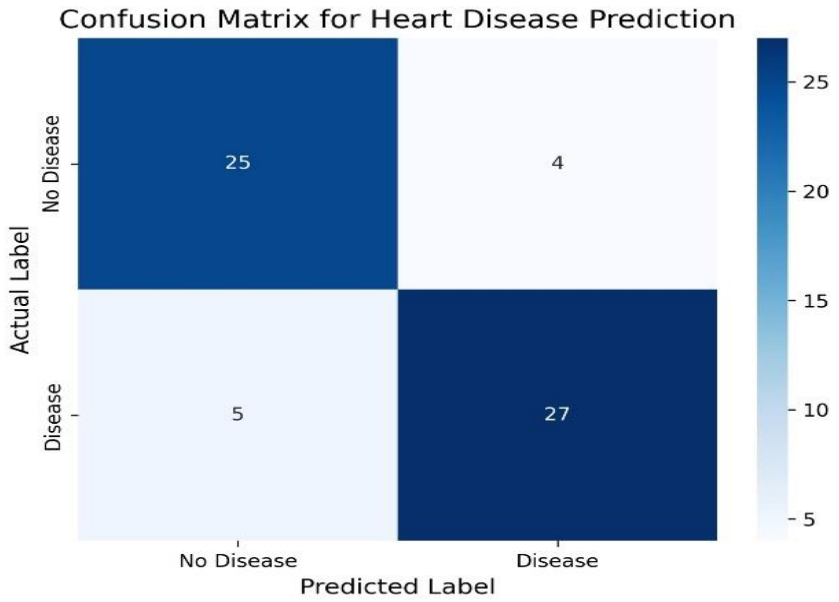

Fig. 2: Distribution of both heart disease-free and heart disease affected patients in the dataset.

Feature Importance Analysis

The feature importance was performed through the KNN model to determine which are most influential attributes affecting heart disease prediction. As shown in Fig. 3, features such as type of chest pain (cp), maximum heart rate (thalach) and number of major vessels (ca), and factor of depression (oldpeak) plays a crucial role in making predictions. This analysis helps in understanding the relationship between clinical attributes and heart disease, providing valuable insights for medical decision-making.

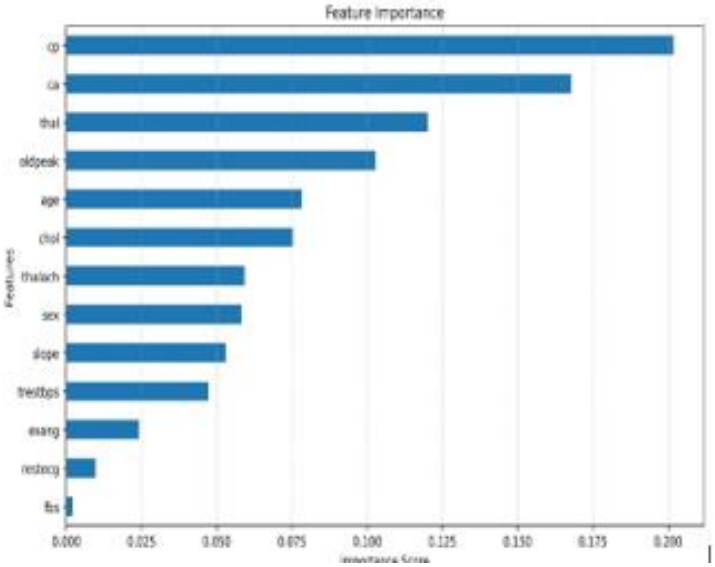

Fig.3: Feature importance analysis using KNN indicating the contribution of each attribute in predicting heart disease.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

The experimental results show that KNN achieved the highest accuracy among all models. However, Random Forest also demonstrated strong and consistent performance with an accuracy of around 88.5%, making it a reliable model for heart disease prediction.

Comparison Between The Proposed Model And Previous Research

Authors	Methods	Accuracy (%)
Our study	KNN	91.8
Mohan et al. [3]	HRFLM	88.47
Amin et al. [4]	Naïve Bayes And Logistic regression	87.41
Latha & Jeeva [5]	NB, BN, RF, and MP	85.48
Patel et al. [9]	J48 with Reduced Error pruning Algorithm	56.76
Tomar & Agarwal [10]	Feature selection-based LSTSVM	85.59
Buscema et al. [11]	TWIST algorithm	84.14

CONCLUSION

In this research, the method of predicting heart disease system based on machine learning has been created and evaluated using several classification algorithms. The objective was to identify efficient and reliable models for accurately predicting heart disease using clinical data. Machine learning technologies, including linear regression algorithm, support vector machine algorithm (SVM), K-Nearest Neighbors (KNN), decision tree, and random forest, were tested and evaluated. To ensure the most effective model, data preprocessing methods including data cleaning and encoding, normalization and extra processing were applied. The model is evaluated using standard performance indicators, including accuracy, accuracy, recall, and ROC-AUC score. The experimental results show that KNN achieved the highest accuracy among all models. However, Although KNN achieved the highest accuracy, Random Forest remains a robust and reliable model due to its ability to handle complex feature interactions and provide stable performance across different data conditions. In addition, the importance analysis of characteristics showed that certain medical characteristics, such as the number of major vessels and the kind of chest pain and the maximum heart rate play a crucial factor in forecasting heart disease. Our research illustrates the effectiveness of group learning methods for handling complex medical and datasets enhancing prediction precision. The proposed model will assist medical experts in early detection and risk assessment, thus contributing to better clinical decision making and potentially lowering the death rate of cardiovascular diseases. In future research, the model will be able to further enhance with the incorporation of large-scale and varied data sets, advanced technology, deep learning, and the integration of real-time healthcare data into more resilient and scalable prediction systems.

The study is based on a publicly available dataset, which may not fully represent real-world clinical scenarios. The dataset size is limited, which may affect model generalization. Additionally, the model has not been validated on real-time clinical data.

Future work can include testing the model on larger and real-world datasets. Advanced techniques such as deep learning and explainable AI can be applied to improve performance and interpretability. The system can also be integrated into real-time healthcare applications. The use of cross-validation helps reduce overfitting and ensures consistent model performance.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

REFERENCES

1. N. Biswas, M. M. Ali, M. A. Rahaman, M. Islam, M. R. Mia, S. Azam et al., "Machine learning-based model to predict heart disease in early stage employing different feature selection techniques," BioMed Research International, vol. 2023, Article ID 6864343, pp. 1–15, 2023.
2. I. D. Mienye and Y. Sun, "Effective feature selection for improved prediction of heart disease," in Pan-African Artificial Intelligence and Smart Systems Conference, pp. 94–107, Springer, Cham, 2021.
3. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019.
4. M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol. 36, pp. 82–93, 2019.
5. C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of pre diction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, article 100203, 2019.
6. V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart dis ease prediction using machine learning techniques : a survey," International Journal of Engineering & Technology, vol. 7, no. 2.8, pp. 684–687, 2018
7. M. F. Rabbi, M. P. Uddin, M. A. Ali et al., "Performance evaluation of data mining classification techniques for heart dis ease prediction," American Journal of Engineering Research, vol. 7, no. 2, pp. 278–283, 2018.
8. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in 2017 IEEE Symposium on Computers and Communications (ISCC), pp. 204–207, Heraklion, Greece, 2017.
9. J. Patel, D. Tejal Upadhyay, and S. Patel, "heart disease prediction using machine learning and data mining technique," heart disease, vol. 7, no. 1, pp. 129–137, 2015.
10. D. Tomar and Agarwal, "Feature selection based least square twin support vector machine for diagnosis of heart disease," International Journal of Bio-Science and Bio-Technology, vol. 6, no. 2, pp. 69–82, 2014.
11. M. Buscema, M. Breda, and W. Lodwick, "Training with Input Selection and Testing (TWIST) Algorithm: A Significant Advance in Pattern Recognition Performance of Machine Learning," Journal of Intelligent Learning Systems and Applications, vol. 5, no. 1, article 27937, 2013.