ISSN 2277-2685
IJESR/July-September. 2026/ Vol-16/Issue-3/184-190
Pallavi *et. al.,* / International Journal of Engineering & Science Research
Full Length Article

Heart Disease Prediction Using ML

Ms. A Jyothirmayi[1], Pallavi[2], Rithwika[3], Sniggdha[4]
[1]Assistant Professor; Department Of Electronics And Communication Engineering, Bhoj Reddy Engineering College For Women Hyderabad India.
[2,3,4]B. Tech Students; Department Of Electronics And Communication Engineering, Bhoj Reddy Engineering College For Women Hyderabad India.
Mail Id's; S.Pallavi6017@gmail.com[2], rithwikanakirekanti@gmail.com[3]
Accepted 29-06-2026
Author(s) Retains the Copyrights of This Article

Abstract
Heart disease remains one of the leading causes of death worldwide, emphasizing the need for accurate and timely diagnosis. Conventional diagnostic approaches primarily depend on clinical expertise, laboratory investigations, and imaging techniques, which may be time-consuming and less effective in identifying disease at an early stage. Recent advancements in Machine Learning (ML) have enabled the development of intelligent healthcare systems capable of analyzing large-scale medical data to support early and reliable disease prediction.
This project presents a Heart Disease Prediction System that employs machine learning algorithms to predict the likelihood of heart disease based on patient health parameters, including age, gender, blood pressure, cholesterol level, maximum heart rate, chest pain type, fasting blood sugar, and other clinical attributes. The dataset, obtained from a standard medical repository, is preprocessed through data cleaning, missing value handling, normalization, and feature selection to enhance prediction accuracy and model efficiency.
To identify the most effective predictive model, multiple machine learning algorithms, namely Logistic Regression, Decision Tree, and Random Forest, are implemented and evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Comparative analysis demonstrates the effectiveness of these algorithms in classifying patients based on the presence or absence of heart disease, with ensemble methods providing improved predictive performance.
The proposed system offers a fast, reliable, and cost-effective solution for the early detection of heart disease, enabling healthcare professionals to make informed clinical decisions and initiate timely treatment. Furthermore, the system can be integrated into smart healthcare platforms, wearable health monitoring devices, and mobile medical applications to facilitate continuous patient monitoring and preventive healthcare. Overall, this project demonstrates the significant role of machine learning in enhancing medical diagnosis, improving patient outcomes, and reducing mortality through early prediction and intervention.

Keywords: *Heart Disease Prediction, Machine Learning, Healthcare Analytics, Logistic Regression, Decision Tree, Random Forest, Medical Data Mining, Classification, Early Diagnosis, Predictive Analytics, Feature Selection, Disease Detection.*

Introduction

Heart disease is recognized as one of the major health concerns worldwide and continues to be a leading cause of illness and mortality. It comprises several cardiovascular conditions, including coronary artery disease, heart failure, and cardiac arrhythmias, which may result in serious complications or death if timely diagnosis and treatment are not provided. Global health statistics indicate that cardiovascular diseases account for a large proportion of annual deaths, emphasizing the importance of early identification and preventive healthcare measures.
The conventional diagnosis of heart disease mainly depends on clinical examinations and diagnostic procedures such as electrocardiograms (ECG), blood investigations, and angiography. Although these techniques provide reliable clinical information, they require skilled medical professionals for interpretation and are often associated with higher cost and longer diagnostic time. Moreover, the early symptoms of heart disease may not always be obvious, which can delay diagnosis and appropriate treatment.
Recent advancements in technology have enabled the application of Machine Learning (ML) in the healthcare sector for disease prediction and diagnosis. Machine learning algorithms are capable of processing large volumes of medical data, identifying hidden patterns, and generating accurate predictive models. By analyzing patient information such as age, blood pressure, cholesterol level, heart rate, and other clinical parameters, these algorithms can estimate the likelihood of heart disease at an early stage. The objective of this project is to develop a machine learning-based heart disease prediction system that provides accurate and

efficient predictions, thereby assisting healthcare professionals in making informed clinical decisions.

Literature Survey

S. Detrano et al., in their study titled *"International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease,"* presented a statistical and probabilistic approach for predicting heart disease using the Cleveland Heart Disease dataset. Their research demonstrated that clinical attributes such as age, blood pressure, and cholesterol are significant indicators in identifying heart disease. Although the proposed model established an important basis for data-driven diagnosis, its capability to represent complex relationships within medical data was limited.

K. Polat and S. Güneş investigated the application of Logistic Regression for heart disease prediction and reported that it is a simple, computationally efficient, and interpretable classification technique suitable for medical datasets. The study emphasized its effectiveness in binary classification problems while also noting that the algorithm has limitations in modeling nonlinear relationships among clinical features, which may reduce prediction performance in complex cases.

R. Chandra and S. Varma examined the use of Decision Tree algorithms for heart disease classification. Their findings indicated that Decision Trees effectively process both categorical and numerical attributes while providing a clear and understandable decision-making structure. The visual representation of decision rules makes the model easier to interpret; however, the study also identified overfitting as a major limitation, particularly when noisy or irrelevant data are present.

Breiman introduced the Random Forest algorithm as an ensemble learning technique that integrates multiple decision trees to improve classification performance. Research based on this approach demonstrated that Random Forest achieves better prediction accuracy and stronger generalization than individual classifiers. By combining the outputs of several trees, the algorithm minimizes overfitting and performs efficiently on datasets with a large number of features.

S. Parthiban and S. K. Srivatsa carried out a comparative study of various machine learning algorithms for heart disease prediction. Their analysis revealed that Logistic Regression provides a simple and interpretable model, Decision Trees offer better visualization of the decision process, and Random Forest delivers superior accuracy and robustness. The study concluded that ensemble learning techniques generally produce more reliable prediction results than individual classification models.

N. H. Abdul Wahab et al. evaluated different classification methods and highlighted the significance of data preprocessing and feature selection in enhancing prediction performance. Their research showed that appropriate handling of missing values together with data normalization considerably improves the effectiveness of machine learning algorithms used for heart disease prediction.

Based on the reviewed literature, it is evident that several machine learning techniques have been successfully applied to heart disease prediction, each possessing distinct strengths and limitations. Logistic Regression offers simplicity and interpretability, Decision Trees provide transparent decision-making, and Random Forest achieves higher predictive accuracy and robustness. Nevertheless, issues related to data quality, overfitting, and model optimization continue to present challenges, creating the need for more effective heart disease prediction systems.

Existing and Proposed Methods

Existing Methods

Heart disease has traditionally been diagnosed through clinical examinations and laboratory investigations performed by medical professionals. Physicians generally evaluate a patient's condition using diagnostic procedures such as electrocardiography (ECG), blood pressure measurement, cholesterol and blood glucose testing, together with an assessment of symptoms such as chest pain. Although these methods are clinically reliable, they require expert interpretation, multiple medical examinations, and considerable time before a diagnosis can be confirmed. Such limitations have encouraged the adoption of machine learning techniques to support faster and more objective disease prediction.

Traditional Machine Learning Approach

Conventional machine learning methods have been extensively employed for heart disease prediction by learning relationships between patient attributes and disease outcomes. Logistic Regression is one of the most widely adopted algorithms for binary classification because it estimates the probability of disease occurrence using clinical features. Its straightforward implementation and interpretability make it suitable for healthcare applications. However, the algorithm assumes a linear relationship between variables, which restricts its ability to model complex clinical patterns.

Decision Tree is another frequently used classification algorithm that represents decisions through a hierarchical tree structure. It can process both numerical and categorical variables while providing transparent decision rules that are easy for clinicians to understand. Nevertheless, Decision Trees often suffer from overfitting, particularly when trained on noisy datasets or datasets containing irrelevant attributes, resulting in reduced predictive performance.

Ensemble Learning Approach

ISSN 2277-2685
IJESR/July-September. 2026/ Vol-16/Issue-3/184-190
Pallavi *et. al.,* / International Journal of Engineering & Science Research

To overcome the limitations of individual classifiers, ensemble learning techniques such as Random Forest have been introduced for heart disease prediction. Random Forest constructs numerous decision trees using different subsets of the training data and combines their predictions to generate the final output. This ensemble strategy minimizes overfitting and improves the model's ability to generalize to unseen data. Random Forest also performs well with high-dimensional datasets and generally achieves better predictive accuracy than single-tree models. However, these advantages are obtained at the cost of increased computational requirements and longer model training time.

Proposed Method

The proposed system introduces a machine learning framework for accurate and reliable heart disease prediction by integrating multiple classification algorithms. Unlike conventional approaches that depend mainly on manual diagnosis or a single predictive model, the proposed framework compares different machine learning algorithms before selecting the most suitable model based on performance. Clinical parameters including age, gender, blood pressure, cholesterol level, heart rate, and chest pain type are used as predictive features. The overall methodology consists of data acquisition, preprocessing, feature selection, model training, evaluation, and prediction. This systematic workflow improves prediction accuracy while ensuring better model generalization and consistency.

Block Diagram

Fig 3.1 Block diagram of heart disease

Fig 1 Block Diagram

The block diagram illustrates the complete workflow of the proposed heart disease prediction system. Initially, patient information including demographic and clinical parameters is collected as input. The acquired data then undergo preprocessing to remove inconsistencies, eliminate duplicate entries, convert categorical variables into numerical values, and normalize numerical attributes for improved learning performance.

Following preprocessing, the most relevant clinical features are identified during the feature selection stage, thereby reducing computational complexity while enhancing prediction accuracy. The selected features are subsequently supplied to three machine learning classifiers, namely Logistic Regression, Decision Tree, and Random Forest, which learn the underlying relationships within the dataset and generate prediction models. Finally, the trained model produces the prediction output by determining whether heart disease is present. Depending on the classification stage, the system may also estimate the severity level of the disease, thereby supporting clinical decision-making and early diagnosis.

Results and Discussion

Performance Metric

The performance of the proposed machine learning models is primarily evaluated using **accuracy**, which represents the proportion of correctly classified instances among the total number of test samples. Since the dataset used in this study contains a relatively balanced distribution of disease and non-disease cases, accuracy provides an appropriate measure for comparing the predictive performance of different classification models. Although accuracy serves as the main evaluation criterion in this work, additional performance measures such as precision, recall, F1-score, and the confusion matrix can provide a more comprehensive assessment and may be incorporated in future investigations.

Exploratory Data Analysis (EDA) Results

ISSN 2277-2685
IJESR/July-September. 2026/ Vol-16/Issue-3/184-190
Pallavi *et. al.,* / International Journal of Engineering & Science Research

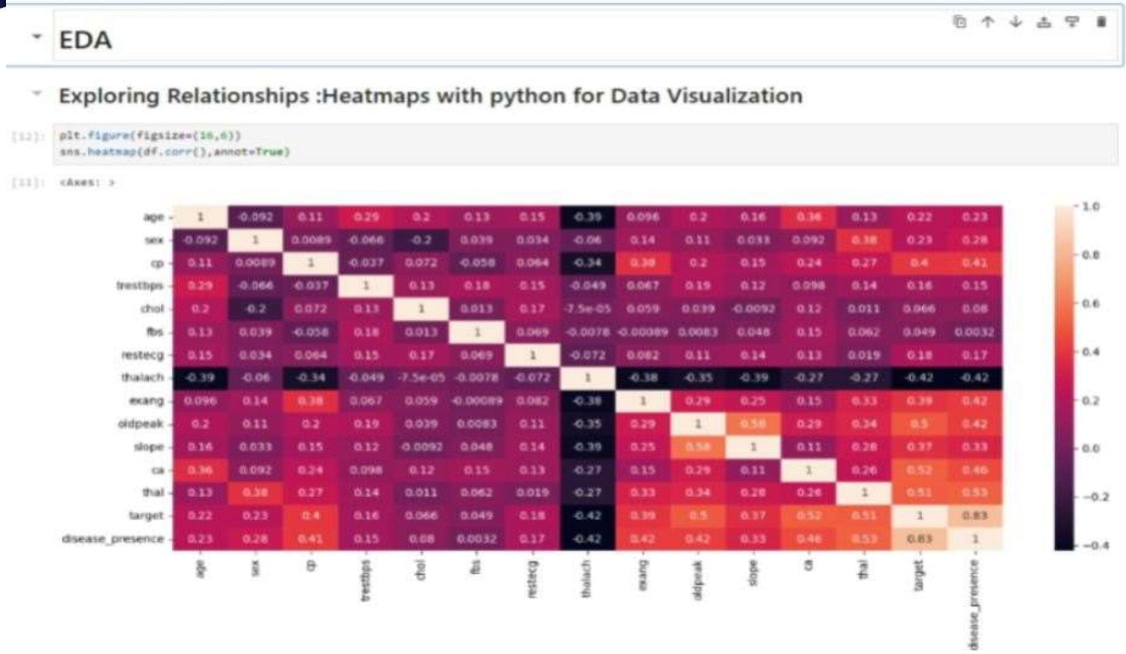

Fig 2 Heatmaps with python for data visualization

illustrates the correlation heatmap generated using Python. The visualization indicates that the variables target and disease_presence exhibit a strong positive correlation. Features including thalach, oldpeak, and ca show moderate relationships with the target variable, suggesting that these attributes contribute significantly to heart disease prediction. The heatmap assists in understanding the association among different variables and helps identify the most influential features for model development.

presents the distribution of patients with and without heart disease. Approximately 140 records correspond to patients diagnosed with heart disease, while nearly 160 records belong to individuals without the disease. This relatively balanced distribution supports reliable model training and evaluation.

The gender distribution shown in indicates that male patients constitute the majority of the dataset, with nearly 200 male records compared to approximately 95 female records. This observation reflects a noticeable imbalance in gender representation within the collected data.

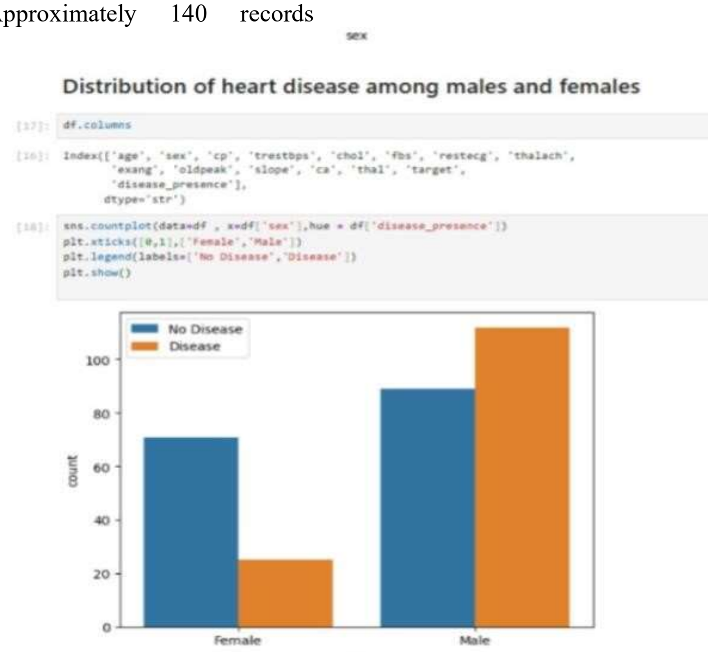

Fig 3 – Distribution of disease among males and females

compares disease occurrence between male and female patients. The results indicate that heart disease is more frequently observed among male patients. While most female participants belong to the non-disease category, the number of affected and unaffected male patients is comparatively similar, suggesting a higher prevalence of heart disease among males in the dataset.

The frequency of different chest pain categories is presented in Asymptomatic chest pain (CP = 4) appears most frequently, followed by non-anginal pain, atypical angina, and typical angina. This distribution demonstrates that asymptomatic chest pain is the predominant category among the patient records.

illustrates the relationship between chest pain type and disease status. Patients diagnosed with heart disease are predominantly associated with asymptomatic chest pain, whereas individuals without heart disease mainly exhibit non-anginal or atypical chest pain. This finding indicates that chest pain type is an important clinical feature for heart disease prediction.

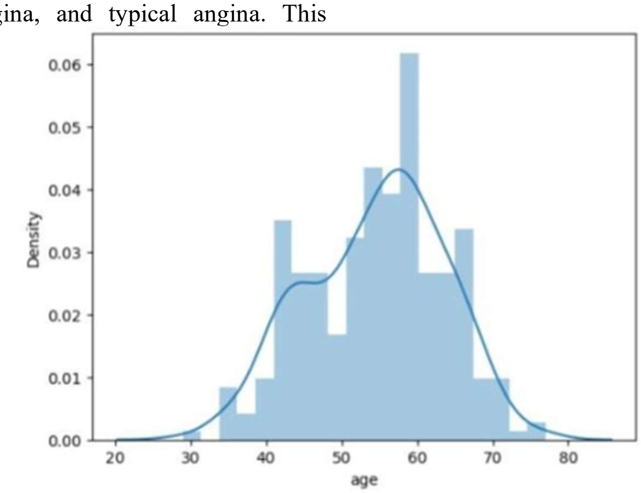

Fig 4 Distribution of Ages in Our Dataset

The age distribution displayed in follows an approximately normal pattern, with the majority of patients falling within the age range of 45 to 65 years. This observation suggests that middle-aged individuals represent the largest proportion of the dataset and therefore constitute the primary population considered during model development.

Results

The comparative performance of the implemented machine learning algorithms for binary classification is presented in Logistic Regression achieved the highest classification accuracy of 90%, followed by Random Forest with 86.7%, while Decision Tree obtained an accuracy of 80%. These results indicate that Logistic Regression produced the highest overall accuracy for the selected dataset, whereas Random Forest also demonstrated competitive predictive capability.

A sample patient record representing a 60-year-old male with asymptomatic chest pain and the corresponding clinical attributes was evaluated using the trained binary classification model. As illustrated in system predicted that the patient was affected by heart disease.

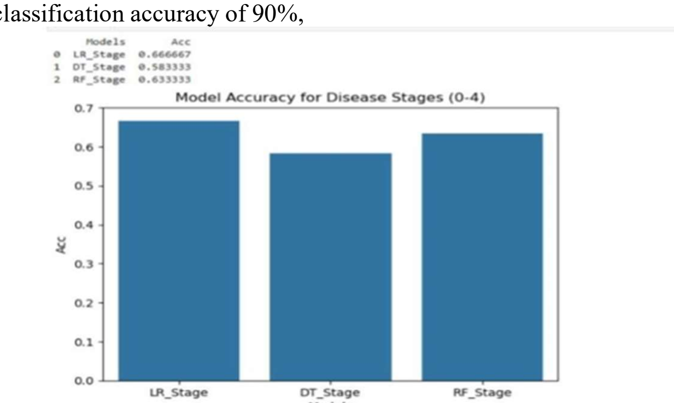

Fig 5 Accuracy of models in multi- classification

The multi-class classification results are present. When a new patient's clinical information was processed using the proposed two-stage prediction model, the final output classified the patient under Stage 4 (Critical) heart disease. This result indicates that the patient belongs to the most severe disease category based on the analysed clinical parameters.

Discussion

The experimental findings demonstrate that machine learning techniques can effectively support heart disease diagnosis by providing accurate and reliable predictions. In this study, Logistic Regression, Decision Tree, and Random Forest algorithms were implemented and evaluated for both binary and multi-class classification tasks. Each model exhibited distinct strengths and limitations when applied to the heart disease dataset.

Although Logistic Regression achieved the highest classification accuracy during experimentation, it occasionally produced incorrect predictions for certain patient records during testing. This observation suggests that overall accuracy alone should not be considered the sole indicator of model quality, particularly in healthcare applications where prediction reliability is of greater importance.

Random Forest produced more stable and consistent predictions across both classification tasks. Its ensemble learning strategy enables the model to capture complex nonlinear relationships among clinical variables while reducing overfitting and improving generalization. These characteristics make Random Forest more dependable for heart disease prediction than individual classifiers.

Decision Tree provided an interpretable prediction model and simplified the decision-making process; however, its predictive performance was comparatively lower and it showed greater susceptibility to overfitting. Based on the overall experimental evaluation, Random Forest was selected as the final prediction model because of its consistent performance and dependable classification capability. The developed system successfully predicts both the presence of heart disease and its severity level, thereby providing practical support for early diagnosis and clinical decision-making.

Conclusion

The study titled "Heart Disease Prediction Using Machine Learning" successfully developed and evaluated a prediction system using three supervised machine learning algorithms, namely Logistic Regression, Decision Tree, and Random Forest. The proposed framework was designed to predict both the presence and severity of heart disease by utilizing patient clinical attributes. Before model development, the dataset underwent preprocessing and feature analysis to improve data quality and enhance the overall performance of the prediction models. Each algorithm was trained and assessed to determine its suitability for heart disease prediction. The experimental evaluation showed that although Logistic Regression achieved the highest classification accuracy on the selected dataset, Random Forest produced more stable and dependable predictions across both binary and multi-class classification tasks. Owing to its consistent performance and improved generalization capability, Random Forest was selected as the final prediction model for the proposed system.

The developed framework is capable of identifying whether a patient is affected by heart disease and, when disease is detected, estimating its severity level. Such predictions can support the early identification of cardiovascular conditions and provide valuable assistance to healthcare professionals during clinical decision-making. Overall, the findings demonstrate that machine learning techniques offer an effective approach for heart disease prediction, contributing to improved diagnostic accuracy and supporting intelligent healthcare applications.

Future Scope

The proposed heart disease prediction system offers several opportunities for further enhancement. The predictive performance of the model can be improved by training it on a larger and more diverse dataset containing patient records from different populations. The incorporation of advanced artificial intelligence techniques, including deep learning models, may further enhance prediction accuracy and enable the identification of more complex patterns within medical data. Future versions of the system can also include additional clinical parameters and real-time patient information to provide more comprehensive disease prediction.

Furthermore, the prediction model can be deployed as a web-based or mobile healthcare application, allowing users and medical professionals to access the system more conveniently. Integration with hospital information systems and electronic health record platforms would enable real-time clinical decision support and improve the practical applicability of the proposed framework in healthcare environments. These improvements have the potential to increase the effectiveness, accessibility, and clinical value of the heart disease prediction system.

References

[1] H. A. Al-Alshaikh, P. Prabu, R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. AlSagri, and A. A. AlSanad, “Comprehensive evaluation and performance analysis of machine learning in heart disease prediction,” *Scientific Reports*, vol. 14, Art. no. 7819, Apr. 2024.

[2] T. Banerjee and İ. Paçal, “A systematic review of machine learning in heart disease prediction,” *Turkish Journal of Medical Sciences*, vol. 49, no. 5, pp. 600–634, Sep. 2025, doi: 10.55730/1300-0152.2766.

[3] Md. M. Ahsan and Z. Siddique, “Machine learning-based heart disease diagnosis: A

systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, Art. no. 102289, Jun. 2022.
[4] Ch. Anwar Ul Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors*, vol. 22, no. 19, Art. no. 7227, Sep. 2022, doi: 10.3390/s22197227.
[5] R. Kumar *et al*., "A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions," *Frontiers in Artificial Intelligence*, 2025.
[6] A. R. Iqbal, S. J., and I. L. K., "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, 2022.
[7] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Mathematical Problems in Engineering*, Art. no. 8387680, 2021.
[8] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, Art. no. 88, 2023.
[9] P. Anbuselvan, "Heart disease prediction using machine learning techniques," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, pp. 515–518, 2020.
[10] R. Stonier, R. K. Gorantla, and K. Manoj, "Cardiac disease risk prediction using machine learning algorithms," *Healthcare Technology Letters*, vol. 11, no. 4, pp. 213–217, Nov. 2023.