IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13258

HEART DISEASE PREDICTION USING RANDOM FOREST CLASSIFIER

Santhosh M[1], Dr. K. Santhi[2]

Student, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore.[1]

Professor, Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore.[2]

Abstract: Cardiovascular diseases continue to be one of the leading causes of mortality worldwide, placing a significant burden on global healthcare systems. Early detection of heart disease plays a crucial role in reducing complications, improving survival rates, and enabling timely medical intervention.

With the advancement of computational intelligence, machine learning techniques have emerged as effective tools for analyzing medical datasets. These techniques are capable of identifying hidden patterns and relationships that may not be immediately visible through manual assessment.

In this study, a predictive framework is developed using the Random Forest algorithm to assess heart disease risk. The model processes structured patient health records containing both demographic and clinical attributes relevant to cardiovascular evaluation [6].

A systematic methodology was implemented, including data preprocessing, feature optimization, supervised model training, and validation. These steps were performed to ensure data consistency, improve model efficiency, and enhance predictive accuracy.

The ensemble learning mechanism underlying the Random Forest classifier combines multiple decision trees to produce stable and reliable predictions. This approach reduces overfitting, improves generalization performance, and enhances classification robustness.

Experimental evaluation using standard performance metrics demonstrates that the proposed system achieves consistent and dependable results. The developed framework has the potential to function as an effective clinical decision-support tool, assisting healthcare professionals in identifying high-risk patients and supporting early preventive care strategies [9].

Keywords: Cardiovascular Risk Prediction, Ensemble Learning, Random Forest Model, Clinical Data Analysis, Supervised Classification, Predictive Healthcare, Data-Driven Diagnosis, Intelligent Medical Systems

I. INTRODUCTION

Cardiovascular diseases are among the most serious health challenges affecting populations across the world. Rapid urbanization, sedentary lifestyles, unhealthy dietary habits, and increased stress levels have contributed significantly to the growing number of heart-related disorders. As a result, early identification of individuals at risk has become a priority in modern healthcare systems.

Heart disease diagnosis typically involves the evaluation of multiple medical parameters, including blood pressure, cholesterol levels, chest pain characteristics, heart rate, and electrocardiogram results. The interpretation of these parameters requires professional expertise and careful clinical assessment. However, analyzing large volumes of patient data manually can be time-consuming and may introduce variability in decision-making.
In recent years, advancements in artificial intelligence and data analytics have transformed the healthcare domain.

Machine learning techniques, in particular, have demonstrated strong capability in extracting meaningful insights from complex medical datasets. These techniques can assist healthcare professionals by providing data-driven predictions that complement traditional diagnostic methods.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13258

Supervised learning algorithms are widely used for classification tasks in medical research[3]. Among them, ensemble learning methods have gained considerable attention due to their ability to improve predictive performance by combining multiple models. Ensemble techniques reduce variance and enhance generalization, making them suitable for healthcare applications where reliability is critical.

This study focuses on implementing the Random Forest algorithm for heart disease risk prediction. Random Forest constructs multiple decision trees using randomized feature subsets and aggregates their outputs to generate a final prediction. This mechanism enhances model stability and reduces the risk of overfitting compared to single-tree classifiers.

The dataset used in this research consists of structured patient records containing demographic and physiological attributes relevant to cardiovascular evaluation. A systematic pipeline including preprocessing, feature selection, model training, and performance evaluation was designed to ensure accurate and reliable predictions.

The primary objective of this research is to develop a dependable predictive framework that can assist clinicians in early detection of heart disease. By integrating machine learning techniques into healthcare decision-making, the proposed system aims to enhance diagnostic efficiency, reduce errors, and contribute to improved patient care outcomes.

II. LITERATURE REVIEW

Several studies have investigated the use of computational and statistical methods for predicting heart disease. Early research primarily relied on traditional statistical techniques such as probability-based models and regression analysis to estimate cardiovascular risk[2]. Although these approaches provided foundational insights, they were often limited in capturing complex relationships among multiple clinical variables.

With the advancement of machine learning, classification algorithms such as Logistic Regression and Decision Trees became widely adopted in healthcare analytics. Logistic Regression offered simplicity and interpretability but struggled with nonlinear data patterns. Decision Tree models improved interpretability through rule-based structures; however, they were prone to overfitting and instability when applied to new datasets.

Support Vector Machines and Artificial Neural Networks were later introduced to enhance predictive performance[10]. Support Vector Machines demonstrated strong classification capability in high-dimensional feature spaces but required careful parameter tuning. Artificial Neural Networks were capable of modeling intricate feature interactions, yet they demanded large training datasets and significant computational resources, limiting their practical implementation in some clinical environments.

To address these limitations, ensemble learning techniques were proposed to improve model stability and generalization. Among these approaches, the Random Forest algorithm has gained significant attention due to its ability to combine multiple decision trees and reduce variance. Research findings consistently indicate that ensemble-based models provide higher reliability and improved diagnostic accuracy in structured medical datasets[7], making them suitable for heart disease risk prediction.

III. PROBLEM STATEMENT

Heart disease diagnosis involves analyzing multiple interrelated clinical parameters, including blood pressure, cholesterol levels, heart rate, chest pain characteristics, and other physiological indicators. The complexity of interpreting these variables simultaneously increases the difficulty of accurate risk assessment, especially in busy clinical environments[8]. Traditional diagnostic methods rely heavily on physician expertise and manual evaluation of medical reports. While clinical experience plays a crucial role, manual analysis may lead to inconsistencies, delayed decision-making, and potential human error when handling large volumes of patient data.

Existing automated prediction systems attempt to address these challenges; however, some models suffer from limitations such as overfitting, poor generalization to new datasets, and reduced interpretability. In healthcare applications, reliability and stability are essential because inaccurate predictions can significantly impact patient outcomes.

Therefore, there is a need for a robust, scalable, and accurate predictive framework capable of supporting early-stage heart disease detection. The objective of this research is to develop a dependable machine learning-based system that enhances diagnostic consistency, improves generalization performance, and assists healthcare professionals in making informed clinical decisions.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13258

IV. METHODOLOGY

The proposed heart disease prediction system using with an **Random forest classifier** was developed through a structured and sequential process. Each stage of the methodology ensures data reliability, model stability, and accurate prediction performance[6].

Step 1: Data Collection
The first step involved obtaining a structured clinical dataset containing anonymized patient records. The dataset includes demographic attributes (such as age and gender) and physiological parameters (**such as blood pressure, cholesterol, chest pain type, heart rate, and ECG results**). A binary target variable indicates whether heart disease is present or absent. The dataset was selected to support supervised classification tasks.

Step 2: Data Preprocessing
Raw medical data often contains missing values, inconsistencies, or outliers. In this stage, incomplete records were handled appropriately to maintain data quality. Numerical features were normalized to ensure uniform scaling, and categorical attributes were **Encoded are Handling Missing Value, Encoding Variable, Feature Scaling** into numerical form for compatibility with **machine learning algorithms**. Finally, the dataset was divided into training and testing sets to evaluate model generalization.

Step 3: Feature Selection
Not all attributes contribute equally to heart disease prediction. Feature importance analysis was performed to identify significant predictors. Redundant and weakly correlated features were removed to reduce model complexity. This step improves computational efficiency and enhances predictive accuracy. There are some feature selection are **Heart Rate, Serum cholesterol, Blood pressure, ECG, Thalassemia, Chest Pain Type.**

Step 4: Model Development
The predictive model was developed using the Random Forest algorithm. Random Forest constructs multiple decision trees using randomized subsets of training data and features. Each tree produces an individual prediction, and the final output is determined through majority voting. This ensemble strategy improves model stability and reduces overfitting.

Step 5: Model Evaluation
The trained model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. A confusion matrix was also analyzed to understand the distribution of correct and incorrect predictions. These evaluation measures ensure that the model performs reliably on unseen test data.

Step 6: Prediction and Deployment
In the final stage, the validated model was used to predict heart disease risk for new patient inputs. The system generates a binary classification output indicating potential disease presence. The framework can be integrated into clinical applications to assist healthcare professionals in decision-making.

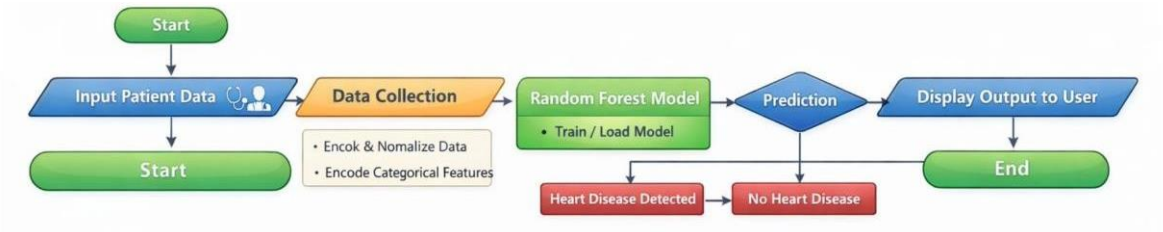

Fig:1.1: (System Flow Diagram of HEART DISEASE PREDICTION USING RANDOM FOREST CLASSIFIER)

V. RESULTS AND DISCUSSION

Experimental evaluation revealed that the ensemble-based model achieved consistent and stable predictive performance[1]. The aggregation of multiple trees reduced variance and minimized overfitting, leading to improved generalization on unseen data.

Balanced precision and recall values indicate the model's ability to correctly identify both highrisk and low-risk cases. Misclassification rates were comparatively low, demonstrating the effectiveness of ensemble learning in healthcare analytics applications[8].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13258

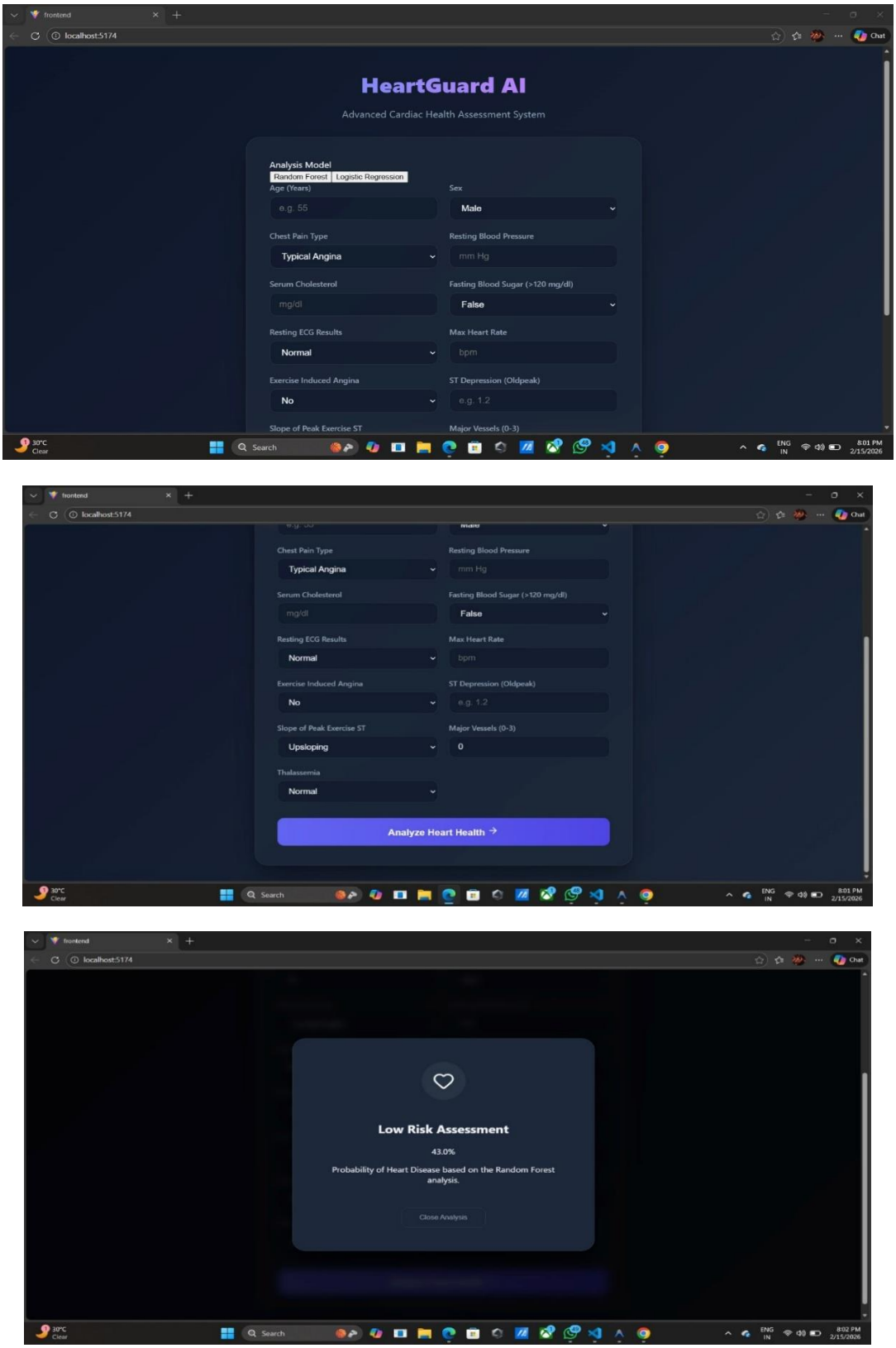

Fig: 1.2: (Food Nutrition Analyzer)

The experimental results indicate that the predictive model developed using the Random Forest algorithm achieved stable and consistent classification performance on the testing dataset. The ensemble structure, which aggregates multiple decision trees, contributed to improved generalization and reduced overfitting compared to single-model classifiers.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13258

Evaluation metrics such as accuracy, precision, recall, and F1-score demonstrated balanced performance in identifying both positive and negative heart disease cases. The confusion matrix analysis further confirmed that the number of misclassified instances was relatively low, indicating reliable predictive capability.

VI. CONCLUSION

This study presented a machine learning-based framework for early prediction of heart disease using the Random Forest algorithm. The proposed system followed a structured pipeline including data preprocessing, feature selection, supervised training, and performance evaluation to ensure reliability and consistency. By leveraging ensemble learning principles, the model effectively reduced overfitting and improved generalization when applied to unseen patient data.

VII. FUTURE WORK

Future studies may also explore the integration of deep learning architectures and compare their performance with the Random Forest algorithm to identify optimal modeling strategies. Furthermore, deploying the system as a web-based or mobile healthcare application would enable real-time risk assessment and broader accessibility. Integration with electronic health record systems could enhance practical usability and support continuous monitoring of patient health. These advancements would contribute to the development of a more comprehensive and scalable clinical decision-support framework.

REFERENCES

[1]. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[2]. R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.
[3]. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[4]. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
[5]. World Health Organization, "Cardiovascular Diseases (CVDs) Fact Sheet," WHO Global Health Reports, 2023.
[6]. Kaggle, "Heart Disease Dataset," Online Repository for Machine Learning Research.
[7]. Manasvi, N., Sreekala, P., Aishwarya, S., & Jawalkar, A.*Heart Disease Prediction Using Random Forest Classifier.* Journal of Cardiovascular Disease Research, Vol. 14 No. 2, 2023. DOI: 10.48047/.
[8]. N. Nasution, M. A. Hasan, and F. B. Nasution, "Predicting Heart Disease Using Machine Learning: An Evaluation of Logistic Regression, Random Forest, SVM, and KNN Models on the UCI Heart Disease Dataset," *IT Journal Research and Development*, vol. 9, no. 2, pp. 140–150, 2025, doi:10.25299/itjrd.2025.17941.
[9]. R. A. Jamadar, A. Garje, T. Bhorde, and V. Jadhav, "Heart Disease Prediction Based on Optimized Random Forest Model Using Machine Learning," *International Journal of Scientific Research in Science and Technology*, vol. 8, no. 3, pp. 337–340, 2021.
[10]. S. Noori Mohammad Ali and N. M. Ahmed, "Comparing Machine Learning Models for Cardiovascular Disease Prediction," *Journal of Pioneering Medical Sciences* vol. 2, no. 1, pp. 140–408, 2025, doi:10.47310/jpms2025140408.

