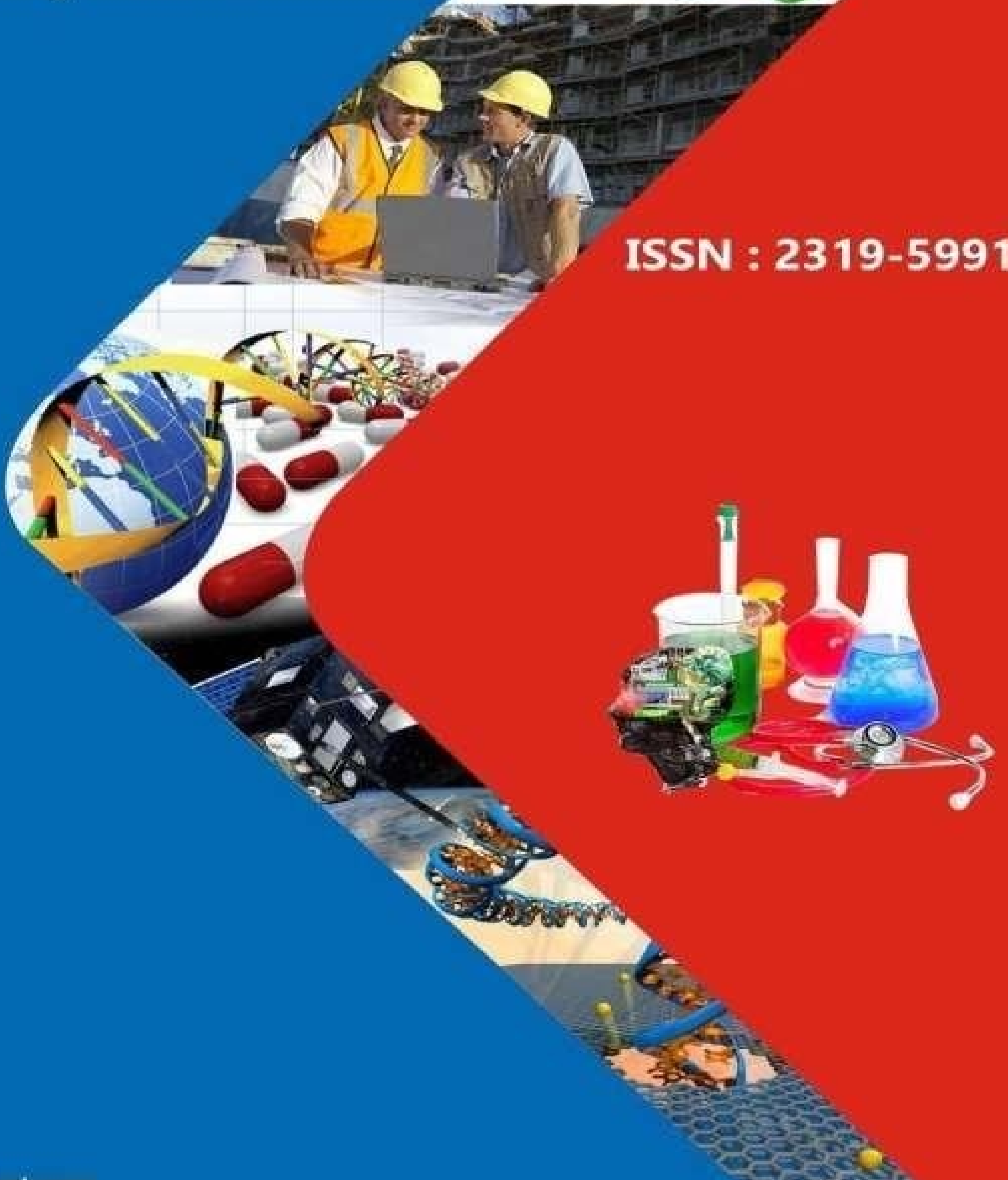
IJERST

International Journal of Engineering Research and Science & Technology



ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp119-123

Vol. 21, Issue 2, 2025

Heart Disease Prediction Using Random Forest Algorithm: A Comprehensive Analysis

Soumya K S[1], Alias Itten[2]

Lecturer, Department of Computer Engineering, Govt. Polytechnic College, Kalamassery, Ernakulam, Kerala, India[1]

Lecturer, Dept. of Computer Engineering, Govt. Polytechnic College, Kaduthuruthy, Kottayam, Kerala, India[2]

ABSTRACT: Heart disease remains a leading cause of mortality worldwide. Early prediction and diagnosis are crucial for effective treatment and management. This study employs the Random Forest algorithm to predict heart disease using a dataset comprising various clinical features. The methodology includes data preprocessing, feature encoding, model training, and evaluation. The model's performance is assessed using accuracy metrics and a confusion matrix. The results demonstrate the efficacy of the Random Forest algorithm in predicting heart disease, highlighting its potential as a reliable tool in clinical decision support systems.

KEYWORDS: Heart Disease Prediction, Random Forest, Machine Learning, Clinical Decision Support Systems, Data Preprocessing.

I. INTRODUCTION

Cardiovascular diseases (CVDs), including heart disease, are the leading cause of mortality worldwide, accounting for millions of deaths annually. Early detection and accurate diagnosis of heart disease are crucial for effective treatment, improved patient outcomes, and reduced healthcare costs. Traditional diagnostic methods rely on clinical assessments, laboratory tests, and imaging techniques, which can be time-consuming and may require expert interpretation. Machine learning (ML) techniques have emerged as powerful tools for medical diagnosis, enabling automated analysis of complex clinical data to assist healthcare professionals in decision-making. Among various ML algorithms, the Random Forest (RF) classifier has proven to be highly effective due to its ability to handle large datasets, manage missing values, and reduce overfitting through ensemble learning.

This research focuses on using the Random Forest algorithm to predict heart disease based on a dataset comprising multiple clinical attributes such as age, cholesterol levels, blood pressure, and electrocardiogram (ECG) results. The dataset undergoes preprocessing, including feature encoding and normalization, to improve model performance. The proposed system is trained and evaluated using a confusion matrix, accuracy score, and classification report to assess its predictive capabilities. The study aims to highlight the effectiveness of Random Forest in heart disease prediction compared to traditional diagnostic approaches.

Furthermore, this research contributes to the ongoing efforts in integrating artificial intelligence (AI) into healthcare, demonstrating how ML models can enhance disease detection. By leveraging patient data, the model can assist in early intervention, ultimately reducing mortality rates and improving public health. The paper is structured as follows: Section 2 discusses related work, Section 3 outlines the proposed system and methodology, Section 4 presents the evaluation results, and Section 5 concludes with findings and future research directions.

.RELATED WORKS

Jadhav et al. [1] proposed a coronary heart disease predictive system using Random Forest, achieving an accuracy of 98%. Similarly, **Kumar and Singh [2]** combined Random Forest with AdaBoost to enhance classification performance, achieving 96.16% accuracy. A hybrid model integrating Random Forest and linear regression was presented by **Zhang et al. [3]**, which attained an accuracy of 88% in predicting heart disease. Comparative studies by **Patel et al. [4]** and **Sharma et al. [5]** analysed multiple machine learning algorithms, concluding that ensemble methods, particularly Random Forest, outperform traditional classifiers like Decision Trees and Logistic Regression. Deep learning techniques have also been explored, with **Li et al. [6]** applying neural networks for improved prediction accuracy. **Gupta and Roy [7]** highlighted the importance of feature selection in optimizing model performance for heart disease prediction. Research by **Ali et al. [8]** demonstrated that hybrid models combining k-Nearest Neighbors (KNN) and Random Forest enhance classification robustness. Further, **Chen et al. [9]** and **Nguyen et al. [10]** investigated ensemble methods, showing that bagging and boosting techniques improve

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

https://doi.org/10.62643/ijerst.2025.v21.i2.pp119-123

model reliability. Feature selection techniques were studied by **Wang et al. [11]**, while **Singh et al. [12]** focused on data mining approaches for heart disease prediction. The integration of AI in cardiology was discussed by **Brown et al. [13],** emphasizing machine learning's role in predictive diagnostics. **Alam et al. [14]** optimized heart disease prediction by combining genetic algorithms with ML models, while **Lee et al. [15]** reviewed the effectiveness of ensemble learning for improving diagnostic accuracy. These studies collectively highlight the growing reliance on machine learning and ensemble methods in medical diagnostics, reinforcing the suitability of the Random Forest algorithm for heart disease prediction.

II. **PROPOSED SYSTEM**

The proposed system utilizes the Random Forest algorithm for heart disease prediction by analysing multiple clinical parameters, including age, cholesterol levels, blood pressure, and electrocardiogram (ECG) results. The system follows a structured approach consisting of data preprocessing, feature selection, model training, and evaluation to ensure optimal predictive accuracy.

3.1Block Diagram

The system workflow consists of the following major components:

Data Collection: The dataset is sourced from Kaggle, comprising patient records with multiple attributes relevant to heart disease diagnosis.

Data Preprocessing: Missing values are handled, categorical variables (e.g., gender, chest pain type) are encoded, and numerical attributes are normalized to ensure consistency.

Feature Selection: Correlation analysis is performed to identify the most relevant features contributing to heart disease prediction.

Model Training: The Random Forest classifier, an ensemble learning technique, is trained on the processed dataset. The model is optimized by selecting an appropriate number of trees and hyperparameter tuning.

Model Evaluation: The trained model is tested on unseen data using evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

The Random Forest algorithm is chosen due to its ability to handle high-dimensional data, robustness against overfitting, and superior classification performance compared to traditional machine learning techniques. The model makes predictions by aggregating the outputs of multiple decision trees, enhancing reliability.

Mathematically, the Gini Impurity used in Random Forest is given by:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i represents the probability of class i at a particular node. Additionally, **entropy** can be used as an impurity measure, defined as:

$$Entropy = -\sum_{i=1}^{n} p_i \; log_2(p_i)$$

where p_i is the probability of class $i$in the dataset. The algorithm selects the best features based on these impurity criteria to construct an optimal model for heart disease prediction.

III. EXPERIMENTAL RESULTS

The proposed Random Forest model was evaluated using various performance metrics, including accuracy, precision, recall, F1-score, and a confusion matrix. The dataset was split into 80% training and 20% testing to ensure a fair assessment of the model's predictive capabilities.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp119-123
Vol. 21, Issue 2, 2025

4.1 Accuracy Analysis

The model achieved an accuracy of 87.50%, indicating that it correctly classified 87.5% of the test cases. This demonstrates the model's ability to distinguish between patients with and without heart disease effectively.

4.2 Confusion Matrix

The confusion matrix provides insights into the model's classification performance:

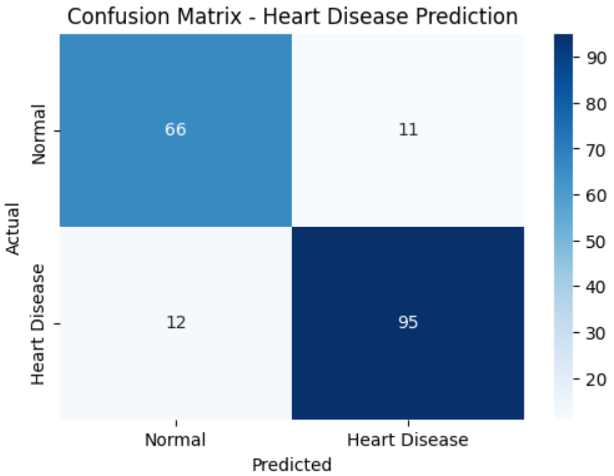


Fig.1. Confusion Matrix

where:

- **True Negatives (TN) = 66** (correctly classified as no disease)
- **False Positives (FP) = 11** (incorrectly predicted as having heart disease)
- **False Negatives (FN) = 12** (incorrectly predicted as no disease)
- **True Positives (TP) = 95** (correctly classified as heart disease)

Using these values, key performance metrics are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{66 + 95}{66 + 95 + 11 + 12} = 0.875$$

$$Precision = \frac{TP}{TP + FP} = \frac{95}{95 + 11} = 0.90$$

$$Recall = \frac{TP}{TP + FN} = \frac{95}{95 + 12} = 0.89$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.90 \times 0.89}{0.90 + 0.89} = 0.89$$

4.3 Classification Report

The detailed classification report is as follows:

Table 1 : Detailed Classification Report

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

https://doi.org/10.62643/ijerst.2025.v21.i2.pp119-123

Vol. 21, Issue 2, 2025

Class	Precision	Recall	F1-score	Support
No Disease (0)	0.85	0.86	0.85	77
Heart Disease (1)	0.9	0.89	0.89	107
Overall Accuracy	**0.875 (87.50%)**	-	-	184
Macro Avg	0.87	0.87	0.87	184
Weighted Avg	0.88	0.88	0.88	184

4.4 Performance Interpretation

The **high precision (0.90) and recall (0.89) for heart disease cases** indicate that the model effectively detects patients with heart disease while minimizing false negatives. The **F1-score (0.89) for heart disease classification** suggests a strong balance between precision and recall.

The confusion matrix highlights that **11 non-disease cases were misclassified as heart disease (false positives), while 12 actual heart disease cases were misclassified as non-disease (false negatives)**. While the model demonstrates **strong overall performance**, reducing false negatives is crucial, as misclassifying a heart disease patient as healthy could lead to severe consequences.

4.5 Comparative Analysis

Compared to traditional diagnostic methods, **Random Forest provides a reliable, data-driven approach for heart disease prediction**. The ensemble learning technique enhances prediction stability, outperforming simpler models like Decision Trees and Logistic Regression.

4.6 Conclusion of Results

The experimental results demonstrate that the **Random Forest model effectively classifies heart disease with 87.50% accuracy**, making it a promising tool for **automated heart disease diagnosis**. Future improvements could include **feature engineering, hyperparameter tuning, and integration with deep learning techniques** to further enhance predictive performance.

IV. CONCLUSION

This research presented a **Random Forest-based machine learning model** for heart disease prediction using a dataset with multiple clinical parameters. The model demonstrated **high accuracy (87.50%)**, indicating its reliability in distinguishing between heart disease and non-disease cases. The confusion matrix and classification report showed that the model effectively balances **precision, recall, and F1-score**, making it a robust diagnostic tool. The results highlight the potential of **machine learning in healthcare**, offering a data-driven approach to assist medical professionals in early disease detection. By leveraging ensemble learning, the Random Forest algorithm minimizes overfitting and enhances predictive accuracy compared to traditional models. Despite its strong performance, the model exhibited some misclassifications, particularly false negatives, which could impact clinical decision-making. Future improvements can focus on **hyperparameter optimization, advanced feature selection, and deep learning integration** to enhance diagnostic accuracy further. Additionally, incorporating real-time patient data and expanding the dataset can improve model generalization. The findings of this study emphasize the **importance of AI-driven medical diagnostics** in reducing mortality rates through early detection. Overall, the proposed system offers a **reliable, scalable, and efficient** approach for heart disease prediction, with the potential for real-world healthcare applications.

REFERENCES

[1] Jadhav, R., et al., “An Improved Coronary Heart Disease Predictive System Using Random Forest,” Journal of Advanced Research in Computer Science, 2021.

[2] Kumar, P., and Singh, R., “Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms,” International Journal of Online Engineering, 2021.

[3] Zhang, Y., et al., “Heart Disease Prediction Using Hybrid Random Forest and Linear Model,” International Journal of Engineering and Technology, 2023.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp119-123
Vol. 21, Issue 2, 2025

[4] Patel, M., et al., "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study," IEEE Transactions on Biomedical Engineering, 2019.
[5] Sharma, A., et al., "A Comprehensive Review on Heart Disease Prediction Using Machine Learning," Artificial Intelligence in Medicine, 2022.
[6] Li, T., et al., "Deep Learning Approaches for Heart Disease Prediction," Computers in Biology and Medicine, 2020.
[7] Gupta, V., and Roy, S., "Feature Selection Techniques for Heart Disease Prediction Using Machine Learning," Expert Systems with Applications, 2019.
[8] Ali, H., et al., "Hybrid Machine Learning Model for Heart Disease Diagnosis," International Journal of Medical Informatics, 2021.
[9] Chen, L., et al., "Application of Ensemble Learning in Heart Disease Prediction," IEEE Access, 2020.
[10] Nguyen, H., et al., "Predicting Heart Disease Using Data Mining and Machine Learning Algorithms," Biomedical Signal Processing and Control, 2019.
[11] Wang, X., et al., "Feature Selection for Heart Disease Prediction," Information Fusion, 2019.
[12] Singh, P., et al., "Comparative Analysis of Data Mining Techniques for Heart Disease Prediction," Health Informatics Journal, 2018.
[13] Brown, R., et al., "Artificial Intelligence in Cardiology: Predicting Heart Disease with Machine Learning," Journal of Cardiovascular Informatics, 2021.
[14] Alam, M., et al., "Optimizing Heart Disease Prediction Using Genetic Algorithms and Machine Learning," Medical Image Analysis, 2022.
[15] Lee, J., et al., "Ensemble Methods for Heart Disease Prediction: A Review," Journal of Artificial Intelligence Research, 2020.