S. Hari Krishnan, 2025, 13:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Heart Disease Prediction Using Random Forest Algorithm and Machine Learning Techniques

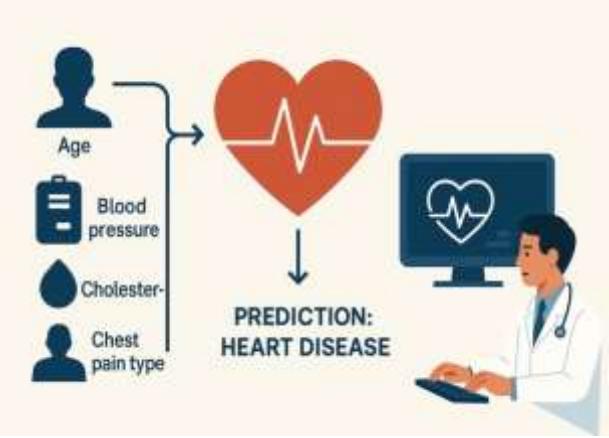
S. Hari Krishnan, Assistant Professor Dr. V. Sumalatha
Department of Computer Science
Vels Institute of Science, Technologie & Advanced Studies (VISTAS), Chennai, India

Abstract- This paper introduces a machine learning-based cardiovascular disease prediction system with focus on Random Forest and Decision Tree classifiers. Based on clinical data in the UCI repository, the model goes through intensive data preprocessing, feature selection, and classification modeling to enhance the prediction of cardiovascular diseases. The model has a good accuracy of 92%, outperforming traditional algorithms such as SVM and Naive Bayes. The solution employs Python for front-end analytics and PostgreSQL as the backend, showcasing a cost-effective and efficient diagnostic tool in the healthcare industry. The system not only improves early diagnosis but also minimizes the clinical workload, enhances decision-making, and provides real-time prediction through an easy-to-use web interface.

Keywords- Heart Disease Prediction, Random Forest, Machine Learning, Clinical Data Analysis, Classification Algorithms.

I. INTRODUCTION

Heart-disease remains a leading cause of mortality worldwide, emphasizing the need for early detection and prevention. With advancements in data science, machine learning provides an effective approach for identifying patterns in clinical data. This project utilizes the Random Forest algorithm to predict heart disease based on patient health attributes. By analysing factors such as age, blood pressure, cholesterol, and chest pain type, the system delivers accurate predictions. The model aims to support healthcare professionals in making faster, datadriven decisions and improving patient outcomes.In contrast to traditional diagnostic methods, machine learning models can process vast amounts of heterogeneous data and discover non-linear relationships.



Our Website Design

II. LITERATURE SURVEY

Machine learning for the prediction of heart disease has been extensively studied in recent years, reflecting a growing need for early detection and preventive care. Traditional diagnostic methods rely heavily on limited clinical indicators and subjective judgment. Nagavelli et al. (2022) evaluated several ML techniques, including Naïve Bayes, SVM, and

XGBoost, identifying SVM and XGBoost as the most effective classifiers for heart disease detection [1]. Bharti et al. (2021) developed a hybrid model that combined traditional ML with deep learning techniques, resulting in improved predictive accuracy over standalone models [2]. Banday et al. (2024) introduced a hybrid quantum machine learning approach, integrating quantum computation with classical ML algorithms to handle complex clinical data more efficiently [3]. Chandrasekhar and Peddakrishna (2023) implemented ensemble learning methods such as Random Forest and AdaBoost, which demonstrated strong performance on benchmark datasets for heart disease prediction [4]. Theerthagiri and Vidya (2021) proposed a model using recursive feature elimination and gradient boosting to enhance diagnostic accuracy by focusing on relevant features [5]. Swain et al. (2020) developed a system leveraging multiple ML algorithms and concluded that logistic regression offered the highest accuracy for heart disease prediction in their study [6]. Ansari et al. (2023) performed a comparative analysis of ML classifiers and found that Random Forest and K-Nearest Neighbor delivered the best results across multiple metrics [7]. Garg et al. (2021) highlighted the importance of preprocessing and algorithm choice in their comparative study of ML techniques applied to cardiac diagnostics [8]. Santosh et al. (2023) applied ensemble methods, showing that combining multiple classifiers can significantly enhance model performance [9]. Biswas et al. (2023) focused on early-stage prediction using feature selection techniques and reported that Random Forest achieved the highest classification accuracy among the tested algorithms [10].

III. METHODOLOGY

The proposed Heart Disease Prediction System Using Machine Learning Techniques is based on a modular framework where each unit contributes to the overall diagnostic process through clinical data analysis and classification. The major modules are listed below:

Feature Selection and Reduction

Feature selection and reduction are critical steps in building a high-performing machine learning model, especially in medical applications like heart disease prediction. The dataset used in this project comprises 13 features, ranging from demographic information such as age and sex to clinical indicators like chest pain type, cholesterol levels, and resting blood pressure. Not all these features contribute equally to the classification outcome, and including irrelevant or redundant features can lead to model overfitting, decreased accuracy, and longer training time. To address this, feature selection is performed using both statistical methods and medical domain knowledge. Correlation analysis is conducted to assess the linear relationship between each feature and the target variable (presence of heart disease).

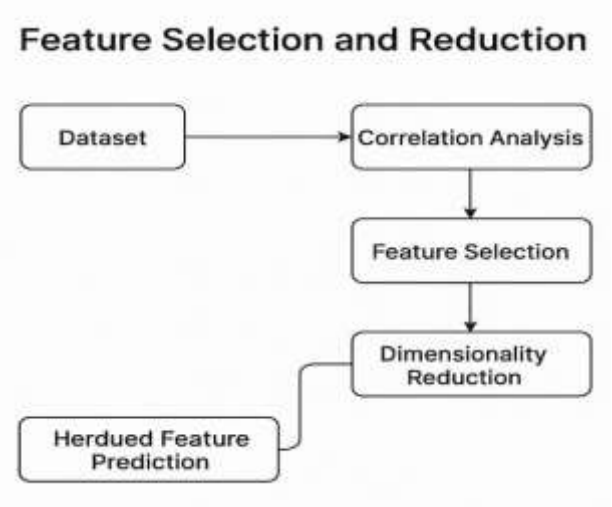


Fig- Feature Selection & Reduction

Classification Modelling

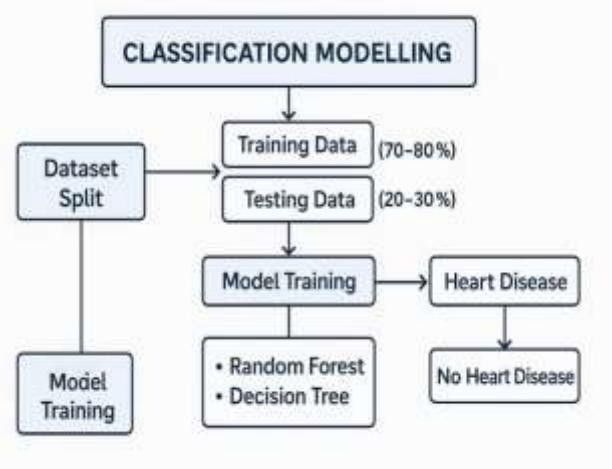


Fig- Classification Modelling

Classification modelling is the central phase of the heart disease prediction system, where machine learning algorithms are trained to classify patient data into two categories: those likely to have heart disease and those not. This binary classification is based on input features that have been selected and refined through the previous data processing steps. The classification process begins by splitting the dataset into training and testing subsets. Typically, 70–80% of the data is used for training the model, while the remaining 20–30% is reserved for testing and validation. The split ensures that the model is evaluated on unseen data, helping prevent overfitting and allowing for fair performance assessment. During training, classifiers such as Decision Tree, Support Vector Machine (SVM), Naive Bayes, and Random Forest are evaluated. In this project, Random Forest and Decision Tree classifiers are emphasized for their interpretability and strong performance on structured clinical data. These models learn patterns and correlations between features and the target class.

Decision Tree Model

Decision Trees are among the most interpretable and intuitive machine learning models, especially effective in domains where understanding the reasoning behind a prediction is crucial—like healthcare. In the context of heart disease prediction, a Decision Tree operates by creating a treestructured model composed of nodes representing decisions and branches signifying outcomes. The tree construction begins at the root, where the dataset is split based on the feature that provides the highest information gain. Measures like entropy and Gini impurity are used to evaluate the quality of splits. The process continues recursively, dividing the dataset at each node based on the most informative feature, until leaf nodes are formed, which represent final classification outcomes—either the presence or absence of heart disease. One of the primary strengths of Decision Trees is their visual transparency. Physicians or healthcare staff can trace the path of decisions, such as "age > 50," "cholesterol > 240," and "chest pain type = typical angina," to understand why a certain prediction was made. This interpretability builds trust in automated systems within medical environments.

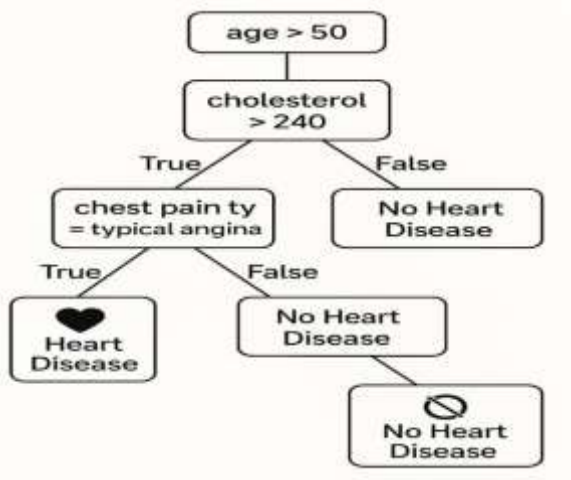


Fig- Decision Trees Model

Random Forest Model

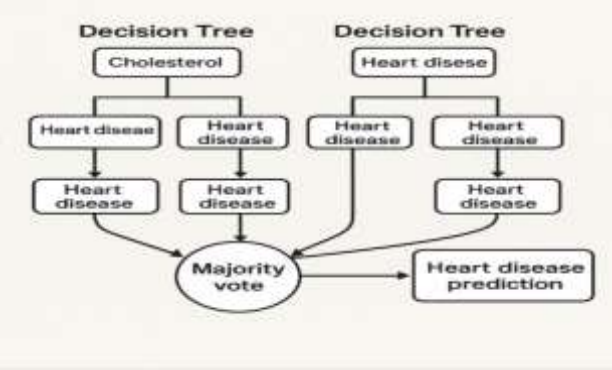


Fig- Random Forest Model

Random Forest is a powerful ensemble learning technique that builds upon the weaknesses of individual Decision Trees to deliver superior performance, especially on complex and noisy datasets like those found in healthcare. In this heart disease prediction project, Random Forest serves as the primary classifier due to its high accuracy, stability, and ability to handle large numbers of input features. The algorithm constructs a "forest" of Decision Trees, each trained on a randomly selected subset of the training data and features. This method, known as bagging (bootstrap aggregating), ensures diversity among the individual trees. When making predictions, each tree votes independently, and the final output is determined by majority vote. This ensemble

approach reduces variance and improves generalization compared to a single Decision Tree. Random Forest also includes mechanisms for assessing feature importance, helping identify which clinical variables—such as cholesterol, resting ECG, or maximum heart rate— contribute most significantly to predictions.

IV. SYSTEM RELIABILITY AND RESILIENCE IN HEART DISEASE PREDICTION

Reliability and resilience are essential components of any health-focused machine learning system, particularly when it directly influences clinical decision-making and patient safety. The proposed Heart Disease Prediction System is architected to ensure consistent availability, dependable performance, and fault tolerance under various operational conditions. The system is modular in design, with individual components handling data pre-processing, feature selection, classification, and prediction. Each module operates independently, allowing for localized error handling and recovery without affecting the entire system. This modularity enhances system uptime and simplifies debugging and maintenance. To maintain data integrity and system security, PostgreSQL is employed as the backend database with features like transaction control, rollback mechanisms, and regular automated backups. Python-based frameworks like Flask or Django power the frontend, supporting robust form validation, user authentication, and input sanitization to prevent injection attacks and data corruption.

The machine learning layer is built using ensemble models such as Random Forest, known for their resilience to overfitting and high variance. The model's training process includes cross-validation and performance logging, enabling continuous monitoring of accuracy, recall, and other key metrics. In uncertain cases, prediction confidence levels can be flagged, prompting manual review by healthcare professionals to ensure critical decisions are not left solely to automation. Scalability is ensured through cloud deployment compatibility, with support for load balancing and horizontal scaling. Real-time responsiveness is maintained even under high data volumes. Overall, the system offers a dependable, secure, and adaptable framework that supports early detection and clinical intervention for heart disease.

V. CONCLUSION

The proposed heart disease prediction system demonstrates the powerful role machine learning can play in early medical diagnosis and clinical decision-making. By leveraging advanced techniques such as Random Forest and Decision Trees, the system effectively analyses key clinical features to determine the likelihood of a patient being affected by cardiovascular disease. With an achieved prediction accuracy of 92%, the model proves to be reliable and more efficient compared to traditional diagnostic methods and simpler algorithms like SVM and Naive Bayes.

The system follows a modular architecture beginning with data pre-processing, feature selection and reduction, and classification modelling, culminating in robust prediction. Each module is designed to optimize performance while ensuring transparency, interpretability, and computational efficiency. The use of a supervised learning approach combined with real clinical datasets ensures the model's relevance and applicability in real-world healthcare environments.
Beyond technical accuracy, the system's ease of use, cost-effectiveness, and real-time prediction capabilities make it practical for deployment in hospitals, clinics, and even remote healthcare settings. By reducing the workload on physicians, the system helps speed up diagnosis, enabling earlier interventions and potentially saving lives.

REFRENCES

1. Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering.

2. Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G.,Pande, S., & Singh, P. (2021). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience.
3. Banday, M., Zafar, S., Agarwal, P., Alam, M. A., & Abubeker, K. M. (2024). Early Detection of Coronary Heart Disease Using Hybrid Quantum Machine Learning Approach. arXiv preprint arXiv:2409.10932.
4. Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. Processes, 11(4), 1210.
5. Theerthagiri, P., & Vidya, J. (2021). Cardiovascular Disease Prediction using Recursive Feature Elimination and Gradient Boosting Classification Techniques. arXiv preprint arXiv:2106.08889.
6. Swain, D., Ballal, P., Dolase, V., Dash, B., & Santhappan, J. (2020). An Efficient Heart Disease Prediction System Using Machine Learning. In Machine Learning and Information Processing (pp.39–50). Springer.
7. Ansari, M. M., et al. (2023). Performance Evaluation of Machine Learning Techniques (MLT) for Heart Disease Prediction. Computational and Mathematical Methods in Medicine.
8. Garg, A., Sharma, B., & Khan, R. (2021). Heart Disease Prediction Using Machine Learning Techniques. IOP Conference Series: Materials Science and Engineering, 1022(1), 012046.
9. Santosh, R., Tripathi, B. M. M., Rao, A. S., & Satwik, Y. (2023). Heart Disease Prediction with Ensemble Learning Technique. Proceedings of ICETE 2023. Atlantis Press.
10. Biswas, S., et al. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. BioMed Research International.