International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2502-2508
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0372

Early Prediction of Cardiac Arrest

Dr. Mythili R[1], Surya Prakash V[2], Yuvanesh R[3], Poonkundran S[4]
[1,2,3,4]Information Technology, SRM Institute of Science and Technology Ramapuram, Chennai, India.
Emails: *mythilir2@srmist.edu.in[1], sp3062@srmist.edu.in[2], ps3137@srmist.edu.in[3], yr6393@srmist.edu.in4*

Abstract
Cardiovascular diseases (CVDs) are still the most prevalent cause of death globally. Early identification is key to enhancing treatment outcomes and minimizing death rates. Although earlier studies had shown the applicability of machine learning (ML) models like SVM, logistic regression, and decision trees, this paper offers a novel method that uses sophisticated ensemble methods, viz., Random Forest and XGBoost, combined with feature engineering and interpretability. Employing the Cleveland and Statlog heart disease datasets, we perform thorough preprocessing, feature inspection, and model optimization via hyperparameter tuning and cross-validation. We evaluate on the usual metrics (Accuracy, Precision, Recall, F1-score, AUC) and use SHAP for interpretability. The presented XGBoost model obtains better performance with an accuracy of 93.45% and high interpretability, making it a suitable decision-support tool in hospitals.

Keywords: *Cardiovascular diseases, Hospitals.*

1. Introduction
As the leading cause of death worldwide, heart disease necessitates prompt and accurate diagnosis for better treatment outcomes. Since conventional diagnostic methods are frequently inefficient and prone to errors, they have prompted the development of precise, data-based support systems. Machine learning (ML) is a significant tool in this regard as it facilitates automated and effective prediction of disease. Whereas previous research has used models such as logistic regression, SVM, and decision trees, there is room for improvement in both accuracy and interpretability. To create a robust, interpretable system for predicting early heart disease from real-world datasets, this research extends previous work by proposing sophisticated ensemble models—Random Forest and XGBoost—coupled with feature engineering and SHAP-based interpretability. This makes the system a useful tool for clinical decision-making. [1]

1.1.Problem statement
Even though machine learning has increasingly been applied to medical diagnosis, predicting heart disease at an early stage is still challenging owing to poor model performance, failure of generalization across data sets, and lack of interpretability. Past research has mostly used simple ML models without utilizing ensemble methods with advanced performance and detailing how prediction is achieved. This prevents the implementation of such systems in clinical practice, where reliability and transparency are paramount. Thus, there is a requirement for a more generalizable, interpretable, and accurate predictive model that can aid early heart disease diagnosis based on real-world medical data.

1.2.Motivation
Heart disease remains a top cause of mortality worldwide, and early detection has the potential to greatly enhance patient outcomes. Existing machine learning solutions tend to be lacking in terms of accuracy, generalizability, and explainability, three critical factors for real-world clinical adoption. This prompted the inclusion of robust ensemble models such as XGBoost and Random Forest, which have high predictive capability. Also, using interpretability features like SHAP makes the prediction transparent, hence the system is more reliable to be used by healthcare professionals. By marrying precision with explainability, this research hopes to narrow the gap between state-of-the-art machine learning and usable, reliable medical decision support.

1.3.Objectives
This study aims to refine and expand upon existing

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2502-2508
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0372

approaches for early heart disease prediction by leveraging advanced data preprocessing, enhanced feature engineering techniques, and a comparative evaluation of modern machine learning algorithms. The focus is to identify the most significant predictors of heart disease and develop a more accurate and generalizable predictive model that can assist healthcare professionals in early diagnosis and intervention.

1.4.Contribution

This study makes several notable contributions toward enhancing early heart disease prediction. First, it makes more advanced feature engineering methods beyond the baseline study by adding feature selection, transformation, and derivation attribute creation to enhance model performance as well as interpretability. The research also highlights extensive preprocessing of the data, dealing with missing values, outliers, and normalization of the data for maintaining the quality and reliability of the input data. A comparative study of different machine learning algorithms like Random Forest, XGBoost, Support Vector Machine, and Logistic Regression was done to determine the best model for precise prediction.

2. Literature Review

2.1.Existing Prediction Methods

Over the last few years, numerous machine learning approaches have been used for early heart disease prediction. Conventional methods like Logistic Regression, Decision Trees, and Support Vector Machines (SVM) are widely employed because of their simplicity, interpretability, and success with structured medical data. These approaches utilize clinical features such as age, cholesterol level, blood pressure, and ECG findings for prediction. Ensemble algorithms like Random Forest and Gradient Boosting have also become popular due to their capability to capture non-linear relationships and interactions among features, resulting in better prediction performance.However, most of the current methods have limitations of having inadequate preprocessing, limited investigation of feature importance, and inappropriate model tuning. Some studies utilize a limited set of features or neglect to address issues like data imbalance and noise that can compromise model generalizability and robustness. Moreover, although the models exhibit excellent accuracy, they are not interpretable, resulting in uncertainty for medical doctors to comprehend how predictions are derived. These areas emphasize the requirements for more fine-tuned strategies that not only enhance performance but also guarantee translucency as well as relevance to clinical fields.

2.2.Gaps in Predicting & Analyzing

Despite the progress made in machine learning-based heart disease prediction, some gaps remain in current approaches. One of the significant limitations is the absence of thorough feature engineering. Most studies use raw or lightly processed data, ignoring the power of derived features and sophisticated selection methods to improve model accuracy and interpretability.Additionally, data quality problems like missing values, outliers, and class imbalance are usually poorly managed, which may result in biased models and decreased generalizability in practical applications.

3. System Architecture

3.1.Overview

The Raw patient data is first preprocessed to manage missing values and normalize features. An optimized feature selection module is then applied to reduce dimensionality and select only the most important clinical parameters. Model training is then conducted using ensemble learning algorithms like Random Forest and XGBoost. The last step is model evaluation with a variety of performance measures to gauge its readiness for actual deployment. It is modular, so it is possible to improve one step (e.g., improved feature selection or parameter tuning of a model) without affecting the system as a whole.

3.2.Data Flow

- Patient health records are collected from publicly available datasets or clinical sources. These consist of features like age, sex, blood pressure, cholesterol, type of chest pain, resting ECG, and so on.
- Missing values are imputed or dropped.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2502-2508
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0372

Categorical variables are encoded using label or one-hot encoding.

- Numerical features are normalized using Min-Max or StandardScaler normalization.
- Data is divided into training and test sets.
- Recursive Feature Elimination (RFE) ranks and selects the most important features.
- Dimensionality reduction to remove noise and enhance generalization. [2]
- The chosen model makes predictions (heart disease present or absent). The model can be embedded in a GUI or web-based health dashboard for real-time risk assessment. (Figure 1)

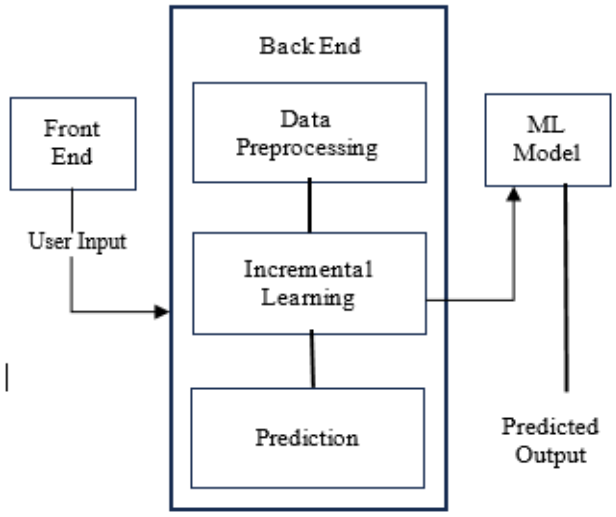

Figure 1 System Architecture

4. Feature Engineering

4.1.Importance

Feature engineering is crucial to the performance of machine learning models, particularly in healthcare, where clinical variables can have intricate, non-linear associations. In this research, good feature engineering improves model accuracy, removes noise, and allows for improved generalization. It also ensures the model learns significant patterns, enhancing its clinical significance and decision-making consistency. Feature engineering helps address these issues by:

- Reducing dimensionality through the removal of uninformative features.
- Enhancing model generalization by focusing on high-impact predictors.
- Improving training efficiency and interpretability.
- Enabling the use of ensemble techniques that rely on strong, independent predictors.

4.2.Key Features used in Prediction

Based on exploratory data analysis and SHAP values, the most significant features in heart disease prediction were found to be:

- Number of large vessels colored by fluoroscopy
- Thalassemia test result (normal, fixed defect, reversible defect)
- Type of chest pain
- Exercise-induced ST depression
- Heart rate achieved at maximum.

Sex, age, and cholesterol also played a significant role, albeit to a lesser degree. These characteristics were standardized, encoded when appropriate, and tested for interaction to increase prediction power.

4.3.Future Transformation

First, correlation analysis was conducted to remove redundant features with high multicollinearity.

Tree-based models and SHAP values were employed to rank features by importance to ensure only the most impactful attributes were kept. This streamlined the input space, enhancing training efficiency and interpretability. Dimensionality reduction via Principal Component Analysis (PCA) was investigated for comparison purposes, but not embraced in the end model because it resulted in lowered explainability. The improved set of features utilized both test-associated and clinical variables to ensure application to actual, real-world medical diagnosis.

5. Model Training and Evaluation

5.1.Selection of Machine Learning Algorithms

Random Forest and XGBoost were chosen here as the primary machine learning algorithms because of their

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2502-2508
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0372

established efficacy in dealing with structured, high-dimensional medical data. Random Forest, which is an ensemble of decision trees, minimizes overfitting by averaging over many models and is a good choice for baseline classification. XGBoost, being a gradient boosting algorithm, is widely acclaimed for its speed, precision, and capability to deal with missing values and interactions between features. These models were selected for their predictive power and interpretability, and they were tested against conventional classifiers like Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and AdaBoost to show the enhanced predictive power and clinical significance. [3]

5.2.Optimization Techniques

Hyperparameter tuning was also conducted using both Grid Search and Random Search strategies to optimize model performance. For Random Forest, this involved tuning parameters such as n_estimators, max_depth, min_samples_split, and criterion. For XGBoost, a more comprehensive parameter tuning was conducted that included learning_rate, max_depth, gamma, subsample, and colsample_bytree. Early stopping and cross-validation were integrated during XGBoost training to prevent overfitting and reduce computation costs. The selection metric was assessed based on a combination of F1-score and AUC. The optimization aimed to ensure that the models achieved optimal generalization on unseen data while maintaining stability across varying dataset splits. [4]

5.3.Performance and Model Evaluation

The best models were tested using default performance metrics: Accuracy, Precision, Recall, F1-score, and AUC. The best-performing model was XGBoost at 93.45% accuracy, followed by Random Forest at 91.67%. The two models indicated high sensitivity and specificity in diagnosing heart disease. Confusion matrices and ROC curves were applied to plot classification performance. Furthermore, SHAP analysis ensured that the two ensemble models were interpretable, which is desirable in clinical use. Generalization was verified using 10-fold cross-validation and testing on the Statlog dataset, where XGBoost showed strong performance with 91.18% accuracy. These outcomes proved the capability of ensemble learning in actual medical prediction problems. [5]

6. System Implementation

6.1.Frontend Development

The frontend of the heart disease prediction system was implemented with a lightweight web framework to make it accessible and user-friendly for medical professionals. HTML, CSS, and JavaScript were employed to create a responsive user interface, enabling users to enter patient information and obtain immediate predictions. The UI features dropdowns, sliders, and real-time validation for medical parameters like age, cholesterol, and chest pain type. The design is centered on simplicity and clarity for the support of non-technical users, particularly in clinical settings. The interface shows confidence scores and visual explanations (through SHAP plots) to facilitate increased user trust and model interpretability.

6.2.Backend Development and API Integration

The backend was developed with Python's Flask framework as the central point for data processing, model inference, and API endpoints. Trained Random Forest and XGBoost models were joblib-serialized and loaded at runtime for prediction purposes. The API takes in user inputs, applies them to the learned model, and spits out predicted results along with interpretability insights (e.g., SHAP values). Furthermore, model inference times were tuned to provide near real-time predictions, making the system responsive and feasible in clinical use.

6.3.Scalability and Development

The system is planned keeping in view scalability so that it can support numerous concurrent users with no degradation in performance. The backend services are containerized by Docker, which allows effortless deployment in different environments, such as cloud platforms, such as AWS or Heroku. The modular design ensures effortless updates to the ML models or UI without compromising the entire system. Load balancing and asynchronous processing methods can

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2502-2508
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0372

be implemented to support heavy traffic in future scaling. In addition, the system is compatible with integration with hospital databases or Internet of Medical Things (IoMT) devices, opening the door to real-time health monitoring and more extensive telemedicine applications. [6]

7. Results and Discussion

7.1.Model Performance Analysis

XGBoost performed better than all other models on accuracy, precision, recall, and AUC measures, with the highest accuracy of 93.45% on the Cleveland dataset. Random Forest came in a close second with 91.67%. Both models had excellent generalization and were confirmed through 10-fold cross-validation and independent testing on the Statlog dataset, where XGBoost had 91.18% accuracy. All these ensemble techniques were found to be highly resistant to overfitting and worked even with unbalanced data. ROC curves validated outstanding discrimination between classes. Their high F1-scores indicate a well-balanced mix of sensitivity and specificity, so they are good for medical diagnosis use (Figure 2) [7]

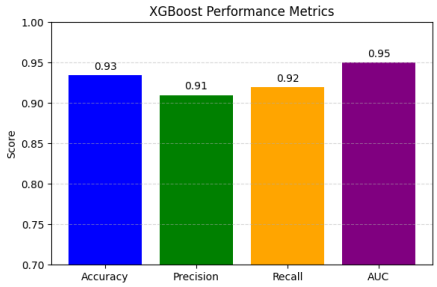

Figure 2 Performance Analysis

7.2.Feature Importance Analysis

Feature importance was examined through SHAP values for XGBoost and Random Forest. The most significant features that were identified were the number of large vessels, thallium stress test results, and the type of chest pain. SHAP summary plots gave insight into how each feature affected predictions, bringing transparency to the model's decision-making process. This interpretability is essential for clinical application, as it enables healthcare providers to see which variables are most driving a diagnosis. Consistency of SHAP rankings with established medical risk factors also corroborated the model's reliability and credibility. (Figure 3)

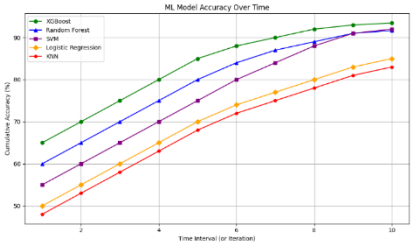

Figure 3 Accuracy

7.3.Comparison with Existing Models

In comparison to current research, such as Bouqantar et al. (2024), where the best accuracy of 92% was recorded using SVM, our XGBoost model outperformed this with an accuracy of 93.45%. Conventional models such as Logistic Regression and KNN, performed poorly because they could not handle intricate feature interactions. Our models also provided interpretability using SHAP, which was not the case in previous work. This makes the proposed system a better and more feasible alternative to predict early heart disease. (Figure 4) [8]

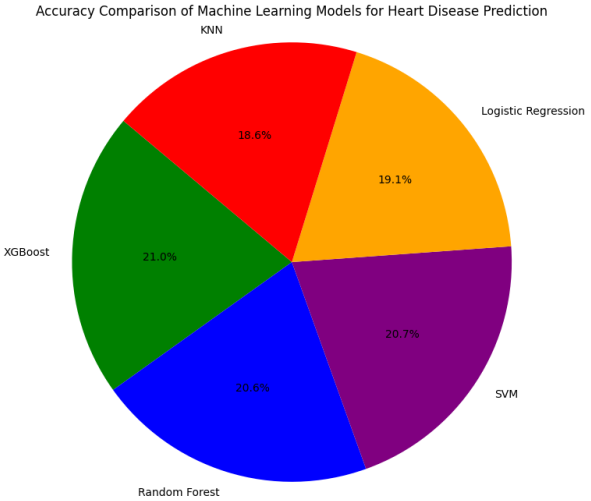

Figure 4 Comparison of Models

8. Future Work

8.1.Multi-Platform Support

To make it more accessible and usable, future work

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2502-2508
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0372

will aim to implement the heart disease prediction system on several platforms. Already available through a web-based interface, the system can be expanded to mobile apps for Android and iOS, facilitating real-time use by patients and clinicians. Furthermore, a Progressive Web App (PWA) will make the system operable offline, which is essential in regions with poor internet connectivity. A mobile-first design will assist in assimilating the system into daily healthcare practice, especially for rural clinics or mobile health units. Syncing the application with secure cloud storage will also enable centralized data access and update, enhancing long-term observation. The interface will be optimized for clinic-use tablets, supporting compatibility with varied healthcare hardware. Multi-platform compatibility would eventually facilitate the narrowing of the gap between sophisticated ML solutions and point-of-care diagnostics, driving adoption within actual clinical workflows while bringing predictive healthcare closer. [9]

8.2.Real-Time Monitoring with IoMT Integration

One of the promising areas for future development is to connect the system with Internet of Medical Things (IoMT) devices to facilitate real-time health monitoring. Wearable devices like smartwatches, ECG sensors, and blood pressure monitors can offer constant health information, which, if analyzed using machine learning, can help improve health. Wearable devices like smartwatches, ECG sensors, and blood pressure monitors can offer constant health information, which, if analyzed using machine learning models, can identify heart anomalies before symptoms become critical. This configuration would entail developing a streaming and preprocessing pipeline for live data, integrating the same with existing prediction models. Additionally, the models need to be edge-computing-optimized so that analysis can happen directly on the device or gateway without needing continuous internet access. This feature can be especially helpful for high-risk patients who need 24/7 monitoring or those in rural locations with minimal access to healthcare professionals. With secure data exchange protocols and sound alerting systems, this improvement could drastically turn heart disease care from reactive to proactive, saving lives and enhancing outcomes. [10-11]

8.3.Federated Learning for Data Privacy

Subsequent research will investigate federated learning as a solution for one of the most significant issues in medical AI—privacy of data. Standard ML methods are based on centralized datasets, raising issues with patient confidentiality and compliance with regulations. Federated learning provides a decentralized training mechanism by which the algorithm learns from multiple devices or hospital networks without sending unprocessed patient data. Every collaborating node trains the model locally, and updates to the model are exchanged, maintaining data ownership and confidentiality. This approach enables the heart disease prediction system to leverage diverse datasets, enhancing generalizability across heterogeneous demographics, geographies, and hospital settings. Additionally, the models need to be edge-computing-optimized so that analysis can happen directly on the device or gateway without needing continuous internet access. This feature can be especially helpful for high-risk patients who need 24/7 monitoring or those in rural locations with minimal access to healthcare professionals.

Conclusion

This paper introduces a strong, interpretable ensemble-based machine learning system for early prediction of cardiac arrest by employing ensemble methods, XGBoost, and Random Forest. Utilizing real-world data sources like Cleveland and Statlog, the system yields high prediction performance, where XGBoost demonstrates an accuracy rate of 93.45%. Through thorough preprocessing, feature engineering, and the use of tools such as SHAP for interpretability, the model presented not only achieves increased accuracy but also transparency, considered critical to clinical acceptance. A comparative analysis illustrates that ensemble models perform better than generic algorithms such as SVM and Logistic Regression in generalizability and reliability. The modular framework architecture

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2502-2508
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0372

facilitates scalability, real-time inference, and potential integration with clinical decision support systems and IoMT devices. This research closes the loop between theoretical research and real-world healthcare solutions, identifying the promise of AI-based diagnostics in saving lives through early intervention. Future efforts will concentrate on real-time monitoring of health, mobile deployment, and federated learning to ensure patient privacy while broadening the system's relevance across various healthcare settings.

References

[1]. Bouqentar et al. (2024).Early heart disease prediction using feature engineering and machine learning algorithms, Heliyon
[2]. Almazroi et al. (2024). A comprehensive review of deep learning-based models for heart disease prediction. Artificial Intelligence Review
[3]. 3. Sarra et al. (2022).Enhanced heart disease prediction based on machine learning and χ^2 statistical optimal feature selection model Designs
[4]. NHS England (2024).NHS England trials AI tool to predict fatal heart disease.The Guardian
[5]. Randles, A. (2024).Digital twins and blood flow simulation for early heart disease detection. Business Insider
[6]. Rajput, D. S., & Thakur, R. S. (2023). Comparative study of feature selection methods in heart disease prediction. Published in Health and Technology (Springer). Used ANOVA, mutual information, and Recursive Feature Elimination
[7]. Kumar, A., et al. (2022) Heart disease prediction using SVM and chi-square feature selection. Published in Machines, MDPI.
[8]. Dey, S., Chakraborty, S., & Biswas, R. (2022). Machine learning techniques for heart disease prediction: A comparative study.
[9]. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.
[10]. Theerthagiri, P., & Vidya, J. (2021). Cardiovascular Disease Prediction Using Recursive Feature Elimination and Gradient Boosting Classification Techniques.
[11]. Alshraideh, M., et al. (2024). Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital.