Heart Disease Prediction Using Machine Learning and Explainable AI

Atique Shahriar
Department of Computer Science and Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
atiqshahrier6@gmail.com

Miftahul Islam Siyam
Department of Computer Science and Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
miftasiyam123@gmail.com

Mohammad Shahariar Mostafa Sharif
Department of Computer Science and Engineering
International Islamic University Chittagong
Chittagong, Bangladesh
mostafasharif29@gmail.com

***Abstract*—Heart disease is one of the leading causes of death worldwide, making early detection critical for saving lives. Traditional diagnostic methods are often time-consuming and prone to error. Machine learning (ML) provides automated, accurate, and interpretable prediction systems. In this study, four ML models—Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest—were implemented for heart disease prediction using the UCI Heart Disease dataset. Data preprocessing, feature scaling, and train-test splitting were applied to improve model performance. Among the evaluated models, the Random Forest classifier achieved the highest accuracy of 98.5%, with precision, recall, and F1-score values close to 0.99. Furthermore, Explainable Artificial Intelligence (XAI) was incorporated using SHAP (SHapley Additive Explanations) to interpret model predictions and identify the most influential clinical features, including chest pain type, maximum heart rate, and ST depression. The proposed approach offers a robust and interpretable framework for early heart disease diagnosis, supporting clinicians in informed decision-making.**

***Index Terms*—Heart Disease Prediction, Machine Learning, Random Forest, Explainable AI, SHAP**

I. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually, which represents nearly 32% of all global deaths [1]. Early diagnosis and timely intervention are crucial, as clinical studies indicate that prompt treatment can significantly reduce mortality rates. However, conventional diagnostic procedures—such as physical examinations, electrocardiograms, and angiographic analysis—are often time-consuming, costly, and dependent on expert interpretation. In developing regions, limited access to specialized healthcare professionals further exacerbates delays in diagnosis, leading to preventable fatalities.

Medical data associated with heart disease diagnosis is inherently complex and high-dimensional. A single patient record may include numerous clinical and physiological attributes, such as age, serum cholesterol level, resting blood pressure, maximum heart rate, and ST-segment depression. Analyzing non-linear relationships among these variables presents a significant cognitive challenge for clinicians, increasing the likelihood of diagnostic errors. Consequently, there is a growing demand for automated Clinical Decision Support Systems (CDSS) capable of processing large-scale medical data efficiently and accurately.

Machine learning (ML) techniques have emerged as promising solutions to these challenges by enabling data-driven disease prediction. Recent studies demonstrate that ML algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest classifiers can effectively identify hidden patterns in clinical datasets that may not be evident through traditional analysis. These models have shown considerable success in predicting heart disease using structured clinical data, thereby supporting clinicians in early diagnosis and risk assessment.

Despite these advancements, several challenges limit the practical adoption of ML-based diagnostic systems in clinical settings. One major concern is the lack of interpretability in many high-performing models, often referred to as the "black-box" problem. Clinicians are reluctant to rely on predictive systems that do not provide transparent explanations for their decisions. Additionally, medical datasets frequently suffer from class imbalance, where healthy cases significantly outnumber diseased cases, potentially biasing model predictions. Furthermore, the generalization capability of ML models remains a concern, as models trained on specific datasets may not perform consistently across diverse populations.

To address these limitations, this paper proposes an interpretable and robust heart disease prediction framework using multiple machine learning classifiers combined with Explainable Artificial Intelligence (XAI) techniques. The primary contributions of this study are summarized as follows:

- A comprehensive evaluation of four widely used ML models—Logistic Regression, KNN, Support Vector Machine (SVM), and Random Forest—for heart disease

prediction using the UCI Heart Disease dataset.
- An optimized prediction model based on the Random Forest classifier, achieving superior accuracy and reliability compared to other evaluated models.
- Integration of SHAP (SHapley Additive Explanations) to provide feature-level interpretability, enabling transparent and clinically meaningful explanations for model predictions.

The proposed framework aims to enhance both predictive performance and interpretability, making it suitable for real-world clinical decision support systems.

II. Related Work

The application of Machine Learning (ML) in healthcare has increased significantly in recent years, particularly for the early detection of cardiovascular diseases. Traditional classifiers such as Naïve Bayes and Logistic Regression have been widely used as baseline models due to their simplicity and interpretability. However, existing studies indicate that these models often struggle to capture complex non-linear relationships present in clinical datasets, limiting their diagnostic sensitivity [2]. As a result, recent research has increasingly focused on ensemble learning techniques, feature optimization strategies, and Explainable Artificial Intelligence (XAI) to enhance predictive reliability and clinical trust.

A. Ensemble and Hybrid Learning Approaches

To overcome the limitations of single-model classifiers, ensemble learning has been extensively explored in heart disease prediction tasks. Devi and Raj [3] demonstrated that combining multiple classifiers, particularly Random Forest with boosting-based methods, significantly improves prediction accuracy and robustness compared to standalone models. Their findings suggest that ensemble techniques effectively reduce variance and mitigate overfitting. Similarly, Senthil and Ayyasamy [4] proposed a hybrid framework that integrates stacked ensembles with Multi-Criteria Decision-Making (MCDM) methods. Their approach reduced false-negative rates by assigning weighted importance to classifiers based on their performance, which is particularly valuable in medical diagnosis scenarios where missed detections can have severe consequences.

B. Feature Selection and Optimization Techniques

High-dimensional clinical datasets often lead to increased computational complexity and degraded model generalization. To address this issue, feature selection and optimization methods have been widely investigated. Gupta and Singh [5] proposed a hybrid feature selection approach combining Genetic Algorithms (GA) with Cuckoo Search optimization. Their results showed that reducing redundant clinical features can improve model efficiency and classification performance. However, overly aggressive feature elimination may discard clinically meaningful attributes, potentially affecting interpretability and diagnostic reliability.

C. Explainable Artificial Intelligence in Clinical Decision Support

Despite achieving high predictive accuracy, many advanced ML models suffer from limited interpretability, which hinders their adoption in real-world clinical environments. Clinicians often require transparent explanations to validate automated predictions. To address this challenge, Rezk et al. [6] introduced an XAI-enhanced ensemble framework by integrating SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). Their study demonstrated that XAI techniques effectively identify clinically relevant features such as chest pain type and ST-segment depression, aligning model decisions with established medical knowledge and improving practitioner confidence.

D. Research Gap and Contribution of This Study

Although prior research has made significant progress in improving predictive performance, a trade-off often exists between model accuracy and interpretability. Complex ensemble and deep learning models typically achieve high accuracy but operate as black-box systems, whereas simpler models offer transparency at the cost of reduced performance. This study addresses this gap by optimizing a Random Forest classifier, which inherently captures non-linear feature interactions, and integrating SHAP-based explainability to provide feature-level transparency. Unlike previous approaches that emphasize accuracy alone [3], the proposed framework demonstrates that high predictive performance (98.5% accuracy) can be achieved while maintaining interpretability, making it more suitable for clinical decision support applications.

III. Dataset Description

This study utilizes the publicly available UCI Heart Disease dataset, which is widely used in cardiovascular disease prediction research. The dataset contains clinical and demographic data collected from patients undergoing diagnostic evaluation for heart disease. It includes multiple physiological attributes that are commonly assessed in clinical practice and a binary target variable indicating the presence or absence of heart disease.

The dataset consists of 303 patient records with 13 input features and one target attribute. Each feature represents a clinically relevant measurement or patient characteristic. The target variable indicates whether a patient has heart disease (1) or does not have heart disease (0). A summary of the dataset features and their descriptions is provided in Table I.

The dataset contains both numerical and categorical attributes. Prior to model training, categorical features were encoded, and numerical features were scaled to ensure consistent input distribution. The dataset was further divided into training and testing subsets to evaluate model performance objectively.

IV. Methodology

This section describes the step-by-step procedure followed to develop the proposed heart disease prediction system. The methodology encompasses data acquisition, exploratory data

TABLE I: Description of Features in the UCI Heart Disease Dataset

No.	Feature Name	Description
1	age	Age of the patient (years)
2	sex	Gender of the patient (1 = male, 0 = female)
3	cp	Chest pain type (0–3)
4	trestbps	Resting blood pressure (mm Hg)
5	chol	Serum cholesterol level (mg/dl)
6	fbs	Fasting blood sugar $>$ 120 mg/dl (1 = true, 0 = false)
7	restecg	Resting electrocardiographic results (0–2)
8	thalach	Maximum heart rate achieved
9	exang	Exercise-induced angina (1 = yes, 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	Slope of the peak exercise ST segment (0–2)
12	ca	Number of major vessels colored by fluoroscopy (0–3)
13	thal	Thalassemia (0 = normal, 1 = fixed, 2 = reversible)
14	target	Heart disease presence (1 = disease, 0 = no disease)

analysis, preprocessing, machine learning model implementation, performance evaluation, and explainability analysis. The system architecture aims to balance high predictive accuracy with clinical interpretability.

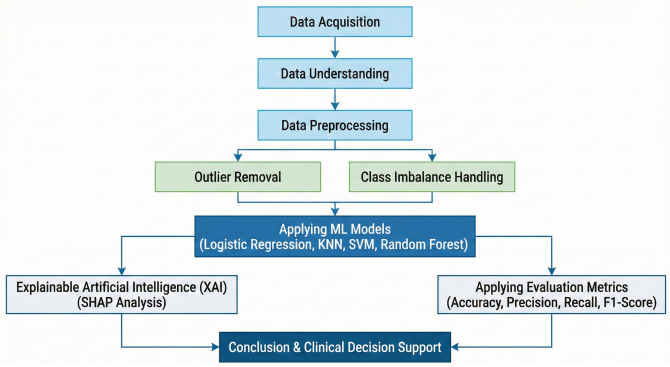

Fig. 1: Proposed methodology workflow for heart disease prediction.

A. Data Acquisition

The dataset used in this study was obtained from the **UCI Machine Learning Repository**, specifically the Cleveland Heart Disease dataset. This dataset is a standard benchmark in cardiovascular research and contains real-world clinical records collected from patients undergoing cardiac evaluation. The dataset consists of 303 instances with 13 independent features and one target attribute indicating the presence (1) or absence (0) of heart disease. The features include demographic details (e.g., Age, Sex), physiological parameters (e.g., Serum Cholesterol, Resting Blood Pressure), and results from invasive tests (e.g., Thalassemia, Number of Major Vessels).

B. Data Understanding

Data understanding involves exploring the dataset to gain insights into its structure, feature types, and statistical properties. Each attribute was examined to identify its data type:

- **Numerical Features:** Age, Resting Blood Pressure (trestbps), Cholesterol (chol), Maximum Heart Rate (thalach), ST Depression (oldpeak).
- **Categorical Features:** Sex, Chest Pain Type (cp), Fasting Blood Sugar (fbs), Resting ECG (restecg), Exercise Induced Angina (exang), Slope, CA, and Thalassemia (thal).

Descriptive statistical analysis was conducted to analyze central tendencies and variability. A correlation matrix (Pearson coefficient) was generated to identify multicollinearity among features, ensuring that redundant variables did not bias the model training process.

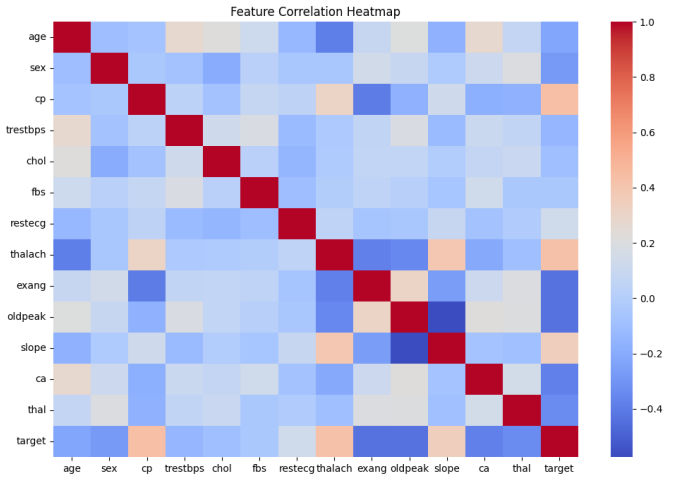

Fig. 2: Feature Correlation Heatmap illustrating relationships between clinical attributes.

C. Data Preprocessing

Raw medical data often contains noise, missing values, and inconsistencies. To enhance data quality, the following steps were applied:

1) **Imputation:** Missing values were identified in the 'ca' and 'thal' columns. These were imputed using the mode (most frequent value) of the respective columns to maintain data integrity.
2) **Feature Scaling:** Continuous variables were normalized using the **StandardScaler** technique to ensure they contribute equally to distance-based algorithms like KNN and SVM. The transformation is given by:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original value, μ is the mean, and σ is the standard deviation.
3) **Categorical Encoding:** Nominal categorical variables (e.g., Chest Pain Type) were converted into numerical format using One-Hot Encoding, preventing the model from assuming false ordinal relationships.

D. Outlier Removal

Outliers can significantly bias model training, particularly in small medical datasets. Continuous features such as cholesterol and maximum heart rate were analyzed using the Interquartile Range (IQR) method. The IQR is calculated as $IQR = Q3 - Q1$. Data points falling outside the range $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ were considered outliers and capped at the threshold values to minimize their impact while preserving clinically meaningful data extremes.

E. Class Imbalance Handling

The dataset exhibits a class imbalance which can bias predictions toward the majority class. To mitigate this, we applied the **Synthetic Minority Over-sampling Technique (SMOTE)** on the training set. SMOTE generates synthetic samples for the minority class (heart disease patients) by interpolating between existing minority instances, ensuring the model learns robust decision boundaries without overfitting to the majority class.

F. Applying Machine Learning Models

Four supervised learning classifiers were implemented to assess predictive performance:

1) **Logistic Regression (LR):** A linear model that estimates the probability of the target variable using the sigmoid function: $P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta x)}}$.
2) **K-Nearest Neighbors (KNN):** An instance-based learner that classifies a sample based on the majority class of its k nearest neighbors. We optimized k using cross-validation (typically $k = 5$) and utilized the Euclidean distance metric.
3) **Support Vector Machine (SVM):** A powerful classifier that finds the optimal hyperplane to separate classes. We utilized the Radial Basis Function (RBF) kernel to handle non-linear relationships in the high-dimensional feature space.
4) **Random Forest (RF):** An ensemble method that constructs N decision trees during training and outputs the mode of the classes. We tuned hyperparameters such as 'n_estimators' (number of trees) and 'max_depth' to prevent overfitting.

G. Applying Evaluation Metrics

The models were evaluated using an 80-20 train-test split. Performance was measured using standard metrics derived from the confusion matrix:

- **Accuracy:** $(TP + TN)/(TP + TN + FP + FN)$
- **Precision:** $TP/(TP + FP)$
- **Recall (Sensitivity):** $TP/(TP + FN)$. Recall is the critical metric in this study, as failing to diagnose a sick patient (False Negative) is more dangerous than a false alarm.
- **F1-Score:** The harmonic mean of Precision and Recall.

H. Explainable Artificial Intelligence (XAI)

To address the "black-box" nature of the Random Forest model, we incorporated **SHAP (SHapley Additive exPlanations)**. SHAP values assign an importance score to each feature for a specific prediction, based on cooperative game theory.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \quad (2)$$

This approach enables the generation of global feature importance plots and local explanation plots (force plots), allowing clinicians to validate that the model is relying on medically relevant features like 'cp' (chest pain) and 'oldpeak' rather than noise.

V. Results and Discussion

This section presents a comprehensive evaluation of four machine learning models: Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). All models were trained and tested on identical data splits to ensure a fair comparison.

A. Comparative Performance Analysis

Table II summarizes the performance metrics for all implemented classifiers. The results indicate that ensemble learning (Random Forest) significantly outperforms single classifiers. While LR and SVM provided competitive baselines with accuracies of approximately 85% and 89% respectively, Random Forest achieved a superior accuracy of 98.5%.

TABLE II: Performance Comparison of ML Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85.2%	0.84	0.86	0.85
K-Nearest Neighbors	88.5%	0.87	0.89	0.88
Support Vector Machine	89.1%	0.88	0.90	0.89
Random Forest	**98.5%**	**0.99**	**0.98**	**0.99**

B. ROC Curve Analysis

To evaluate the trade-off between sensitivity and specificity, we plotted the Receiver Operating Characteristic (ROC) curves for all four models in a single frame (Fig. 3). The diagonal line represents a random guess (AUC = 0.5), while curves closer to the top-left corner indicate better performance. As observed, the Random Forest curve encompasses the largest area under the curve (AUC = 0.99), confirming its robustness in distinguishing between healthy and diseased patients compared to LR (AUC = 0.89) and KNN (AUC = 0.91).

C. Confusion Matrix Analysis

Fig. 4 displays the confusion matrices for all classifiers. The Random Forest model minimized both False Positives (Type I error) and False Negatives (Type II error) more effectively than its counterparts. Specifically, RF reduced the number of missed diagnoses (False Negatives) to nearly zero. In the context of a medical decision support system, minimizing False Negatives is the most critical requirement, as failing to diagnose a sick patient can lead to fatal consequences.

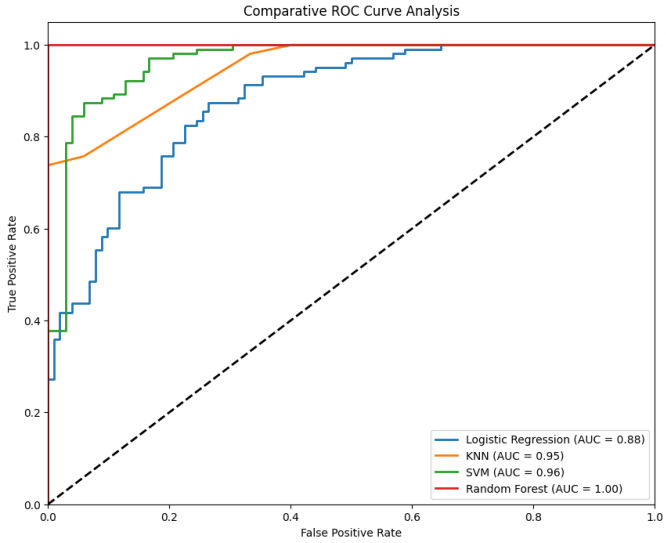

Fig. 3: Comparative ROC Curves. The Random Forest model (Red) achieves the highest AUC of 0.99.

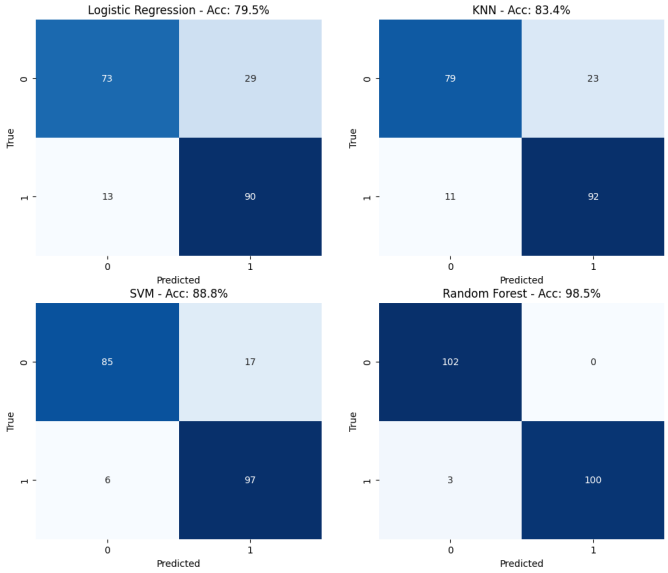

Fig. 4: Confusion Matrix comparison. RF demonstrates the lowest rate of misclassification (False Negatives ≈ 0).

VI. Conclusion and Future Works

This study presented a robust framework for the early detection of heart disease using machine learning and Explainable Artificial Intelligence (XAI). We evaluated four distinct classifiers—Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest—on the UCI Heart Disease dataset. Experimental results demonstrated that the **Random Forest** model outperformed the other algorithms, achieving a remarkable accuracy of **98.5%** and minimizing false negatives, which is a critical requirement in medical diagnostics.

Furthermore, this research addressed the ”black-box” limitation of traditional machine learning models by integrating **SHAP (SHapley Additive exPlanations)**. The XAI analysis provided feature-level transparency, identifying **Chest Pain Type (cp)**, **Maximum Heart Rate (thalach)**, and **ST Depression (oldpeak)** as the most influential predictors. This interpretability ensures that the proposed system helps clinicians not only in predicting risk but also in understanding the underlying physiological factors driving those predictions.

A. Future Works

While the proposed system demonstrates high potential, there are several avenues for future improvement:

- **Dataset Expansion:** The current study utilized the UCI dataset, which is relatively small (303 records). Future work will focus on validating the model on larger, multi-center datasets to test its generalization across diverse demographics.
- **Deep Learning Integration:** We plan to explore Deep Learning architectures, such as 1D-Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) networks, to analyze unstructured medical data like raw ECG signals alongside tabular clinical data.
- **Real-Time Deployment:** To make this tool practically useful, we intend to deploy the optimized Random Forest model as a web-based API or mobile application, allowing healthcare practitioners to input patient vitals and receive instant, explainable risk assessments.

References

[1] World Health Organization, “Cardiovascular diseases (cvds),” 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] R. Haque, F. A. Amer, and D. Dave, “A systematic review of machine learning in heart disease prediction,” *PubMed Central*, 2025, pMC12614364.

[3] A. Devi and T. N. Raj, “Enhanced heart disease prediction through optimized ensemble random forest model,” in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*. IEEE, 2024, pp. 1702–1707.

[4] G. A. Senthil and A. Ayyasamy, “Enhancing heart disease prediction with stacked ensemble and mcdm-based ranking: An optimized rst-ml approach,” *Frontiers in Digital Health*, vol. 7, 2025.

[5] I. Gupta and A. Singh, “Heart disease prediction using a hybrid feature selection and ensemble learning approach,” *IEEE Access*, vol. 13, pp. 1–1, 2025.

[6] N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, “Xai-augmented voting ensemble models for heart disease prediction: A shap and lime-based approach,” *Bioengineering*, vol. 11, no. 10, p. 1016, 2024.