Middle East Journal of Applied Science & Technology (MEJAST)
Volume 9, Issue 2, Pages 50-73, April-June 2026

Comparative Evaluation of Machine Learning Classifiers for Explainable Heart Disease Prediction Using the UCI Cleveland Dataset: A Decision Tree-centred Framework Integrating SHAP-based Clinical Interpretability Assessment and Deployable Clinical Decision Support System Development

Asoshi Paul Anule[1*], Anagu Emmanuel John[2] & Ogar Michael Oko[3]

[1,2,3]*Department of Computer Science, Faculty of Computing and Information System, Federal University Wukari, Taraba State, Nigeria. Corresponding Author (Asoshi Paul Anule) Email: asoshipaulanule@gmail.com*[*]

DOI: https://doi.org/10.46431/mejast.2026.9205

Article Received: 05 April 2026 *Article Accepted: 17 June 2026* *Article Published: 22 June 2026*

ABSTRACT

Cardiovascular disease remains one of the leading causes of mortality worldwide, necessitating the development of accurate and interpretable predictive systems that support early diagnosis and clinical decision-making. While numerous machine learning models have demonstrated promising predictive capabilities, many operate as black-box systems that provide limited transparency regarding how predictions are generated. This lack of interpretability presents significant challenges in healthcare environments where trust, accountability, and regulatory compliance are essential.

This study presents a comparative evaluation of six supervised machine learning classifiers for heart disease prediction using the Cleveland Heart Disease Dataset. The evaluated models include Decision Tree, Logistic Regression, Support Vector Machine, K-Nearest Neighbours, Random Forest, and Gradient Boosting. A comprehensive machine learning pipeline comprising data preprocessing, feature selection, hyperparameter optimization, repeated stratified cross-validation, and performance evaluation was implemented. Explainable Artificial Intelligence (XAI) techniques based on SHAP were integrated to provide both global and local interpretability of model predictions.

Experimental results demonstrate that the Decision Tree classifier achieved the highest overall performance, attaining an accuracy of 98.54%, precision of 98.21%, recall of 98.34%, and F1-score of 98.27%. Furthermore, SHAP-based analysis revealed that chest pain type, number of major vessels, exercise-induced angina, maximum heart rate, and ST depression were the most influential predictors of cardiovascular disease risk. The findings indicate that interpretable machine learning models can achieve predictive performance comparable to or exceeding more complex algorithms while maintaining transparency and clinical usability.

The study contributes a reproducible framework for explainable cardiovascular disease prediction and demonstrates the feasibility of integrating interpretable machine learning models into clinical decision support systems. The proposed approach offers a foundation for trustworthy healthcare artificial intelligence applications that balance predictive accuracy with explainability.

Keywords: Heart Disease Prediction; Machine Learning; Decision Tree; Explainable Artificial Intelligence; SHAP; Clinical Decision Support System; Cardiovascular Disease; Predictive Analytics; Healthcare Artificial Intelligence; Medical Diagnosis; Clinical Interpretability.

1.0. Introduction

Cardiovascular disease sits at the top of the global mortality table by a considerable margin. Current estimates from the World Health Organization place annual CVD deaths at around 17.9 million approximately 32% of everything that kills people worldwide (WHO, 2023). Within this broad category, coronary artery disease, ischemic heart disease, and myocardial infarction together account for a disproportionate share of premature deaths, years lived with disability, and the economic burden that healthcare systems carry (Mensah et al., 2019; Roth et al., 2020). The financial toll is enormous: billions of dollars annually in direct treatment costs, lost productivity, and the downstream expenses associated with managing chronic cardiovascular conditions (Benjamin et al., 2019).

Progress in cardiovascular medicine has undeniably saved lives, but effective management still hinges on one deceptively simple requirement finding the disease early. Evidence is clear that identifying high-risk individuals before a major event occurs can meaningfully reduce mortality, improve treatment responsiveness, and cut long-term care costs (Roth et al., 2020). The difficulty is that cardiovascular disease doesn't follow a straightforward script. It emerges from a web of interacting risk factors age, sex, blood pressure, cholesterol,

diabetes, ECG changes, smoking, weight, and physical activity levels, among others that do not combine in linear or easily predictable ways (Wilson et al., 1998; Alqahtani et al., 2022).

Standard diagnostic tools angiography, stress echocardiography, CT scanning, laboratory panels are effective but come with significant drawbacks. They tend to be invasive, time-intensive, expensive, and often unavailable in healthcare settings where resources are stretched (Ghosh et al., 2018). This practical gap has fueled substantial interest in machine learning as an alternative pathway for cardiovascular risk screening that is fast, non-invasive, and scalable.

Machine learning's value in healthcare stems from its ability to detect non-obvious patterns in complex, multi-variable datasets and to produce predictive models that can support clinicians in diagnosis and prognosis (Topol, 2019). Within cardiology, the technology has been applied across a wide range of tasks disease prediction, risk stratification, treatment personalization, and real-time patient monitoring (Johnson et al., 2018).

Early computational work in cardiovascular diagnosis leaned heavily on logistic regression and Bayesian methods, which were tractable given the computing resources of the time (Detrano et al., 1989). The landscape has shifted substantially since then. Researchers now routinely deploy Support Vector Machines, Random Forests, Gradient Boosting, Artificial Neural Networks, and increasingly sophisticated deep learning architectures (Zhang et al., 2020; Rohan et al., 2025).

Despite impressive accuracy figures in the literature, important questions persist about reliability, transparency, and reproducibility. A substantial proportion of high-performing models function as black boxes systems whose internal reasoning cannot be meaningfully examined. For clinicians who are professionally and ethically accountable for every recommendation they make, an unexplainable algorithm is a difficult thing to adopt, regardless of how accurate it claims to be (El-Sofany et al., 2024; Vijaya Saraswathi et al., 2024).

Several unresolved problems continue to limit progress in this field, The most prominent is the persistent focus on predictive accuracy at the expense of interpretability. In clinical settings, a model's explainability is not a nice-to-have feature it is a practical necessity. Clinicians need to understand the reasoning behind algorithmic outputs before they can responsibly incorporate them into patient care (Ganie et al., 2025).

A second issue is methodological fragmentation. Existing comparative studies use different datasets, preprocessing strategies, feature selection approaches, validation methods, and performance metrics, which makes it almost impossible to draw meaningful comparisons across the literature (Hassan et al., 2022; Rimal et al., 2025).

Third, there is a genuine empirical gap around the accuracy-interpretability trade-off. Ensemble models frequently outperform simpler approaches on accuracy benchmarks, but their complexity imposes a transparency cost. Simpler models like Decision Trees and Logistic Regression are easier to interpret but are often dismissed as less powerful. The available evidence comparing these two dimensions systematically under the same experimental conditions, using the same dataset remains thin (Al-Alshaikh et al., 2024; Asadi et al., 2024).

What the field needs is a standardized, reproducible framework that assesses both predictive performance and interpretability simultaneously across a suite of classifiers.

The proliferation of machine learning models for cardiovascular prediction has not made clinical model selection easier if anything, it has made it harder. Without consistent evaluation frameworks, practitioners and decision-makers struggle to assess which models genuinely serve clinical needs. The deeper problem, however, is that the field has not adequately resolved the tension between predictive power and clinical transparency. Until models can be both accurate and explainable, meaningful adoption within clinical decision support systems will remain limited.

This study therefore asks: among the most widely used machine learning classifiers, which offers the most effective combination of predictive accuracy, robustness, and clinical interpretability for heart disease prediction?

2.0. Literature

Cardiovascular disease has been one of the most active application areas for machine learning research over the past two decades, driven by its global prevalence and the clear clinical need for better early detection tools. What machine learning offers that traditional statistical methods do not is the capacity to model complex, non-linear interactions among many clinical variables simultaneously relationships that may not be apparent even to experienced clinicians reviewing patient records (Topol, 2019; Zhang et al., 2020).

The growing availability of electronic health records, wearable monitoring technologies, and large curated clinical datasets has accelerated this work considerably. Studies have shown that routinely collected variables age, cholesterol, blood pressure, blood glucose, ECG results, and exercise angina, among others carry sufficient signal to support accurate cardiovascular risk prediction when analyzed by well-designed machine learning models (Alqahtani et al., 2022; Hassan et al., 2022).

2.1. Machine Learning Models in Medical Diagnostics

The classifier landscape in cardiovascular disease prediction is broad. Logistic Regression remains a common baseline because it is simple, transparent, and well-understood by clinical audiences. Support Vector Machines have shown strong results in high-dimensional healthcare datasets, where their hyperplane-based separation of classes generalizes well. K-Nearest Neighbours operates on a similarity principle and has performed respectably on smaller clinical datasets, though its computational demands scale poorly with data size. Decision Trees offer a transparent hierarchical structure that maps naturally onto clinical reasoning — a significant practical advantage. Ensemble extensions of tree methods, particularly Random Forest and Gradient Boosting, improve predictive stability by combining multiple weak learners, but at the cost of interpretability (Breiman, 2001).

More recently, deep learning approaches convolutional networks, recurrent architectures, and attention mechanisms have pushed accuracy benchmarks higher still, but their opacity is a genuine clinical liability. The stronger the model, the harder it often is to explain what it is actually doing (Xia et al., 2024; Rao et al., 2024).

2.2. Accuracy and Performance Metrics

Evaluating classification models in medical contexts requires more than a single accuracy figure. The standard toolkit accuracy, precision, recall, and F1-score collectively provides a picture of how a model performs across different error types (Sokolova & Lapalme, 2009). In cardiovascular prediction specifically, sensitivity (recall)

deserves particular attention. Missing a positive case a false negative can delay diagnosis with serious clinical consequences. A model that scores well on accuracy but poorly on recall may look good on paper while being practically dangerous in deployment.

2.3. Clinical Datasets and Data Quality Considerations

The quality of any machine learning model is ultimately bounded by the quality of the data it learns from. The Cleveland Heart Disease Dataset, originally assembled by Detrano et al. (1989), has become the de facto benchmark for cardiovascular prediction research. Its 14 clinical predictor variables spanning demographics, physiological measurements, and diagnostic test results capture the major dimensions of cardiovascular risk and have been extensively validated through decades of use.

Real-world clinical datasets rarely arrive in pristine condition. Missing values, measurement inconsistencies, class imbalance, and limited demographic diversity are routine challenges that must be addressed before any model can produce reliable predictions. This is why preprocessing imputation, normalization, feature selection, and oversampling has become as important a part of the machine learning pipeline as the algorithms themselves (Rehman et al., 2025; El-Sofany et al., 2024)

2.4. Decision Trees and Their Role in Explainable Healthcare AI

Among all the classifiers commonly used in clinical research, Decision Trees hold a special position because they are genuinely interpretable by design. First formalized by Quinlan (1986), the Decision Tree algorithm constructs a hierarchical rule structure by recursively splitting the dataset into progressively homogeneous subsets based on feature values. Three splitting criteria are commonly employed:

Gini Impurity measures how likely a randomly selected sample from a node would be misclassified. Lower values indicate cleaner class separation. Entropy quantifies the degree of disorder within a node the algorithm minimizes it through successive splits. Information Gain tracks how much each split reduces entropy, with the most informative features selected as decision nodes.

What makes Decision Trees genuinely valuable in healthcare is not their accuracy it is that every prediction they produce can be traced through a sequence of human-readable rules. A clinician can follow the path from a patient's input data to the final output and understand exactly which clinical thresholds drove the recommendation. This property aligns naturally with clinical reasoning and addresses the accountability requirements that govern medical practice (Kumar et al., 2020; Agbemade, 2023).

The broader movement toward Explainable AI in healthcare has brought renewed attention to interpretable models. Regulatory frameworks in multiple jurisdictions now require that AI systems used in medical decision-making be auditable and transparent, making interpretability not just a scientific preference but a compliance requirement (Ganie et al., 2025).

The Figure 2.1 illustrates how input features (Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol) are processed through a series of decision nodes (e.g., 'Age > 50', 'Cholesterol > 200') to arrive at terminal predictions of 'Heart Disease Detected' or 'No Heart Disease.'

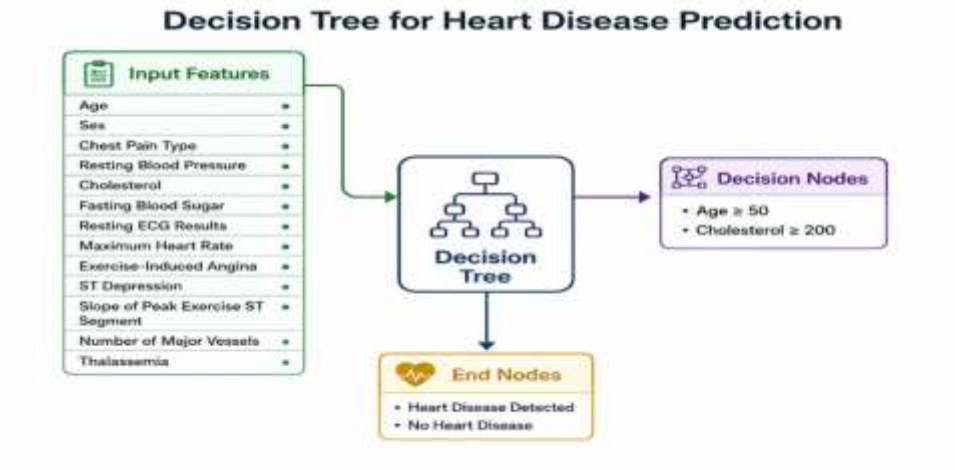

Figure 2.1. Decision Tree for Heart Disease Prediction.

The diagram illustrates how input features (Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol) are processed through a series of decision nodes (e.g., 'Age > 50', 'Cholesterol > 200') to arrive at terminal predictions of 'Heart Disease Detected' or 'No Heart Disease.'

2.4.1. Machine Learning Pipeline for Heart Disease Prediction

Contemporary machine learning frameworks for cardiovascular prediction follow a recognizable sequence. Data is first acquired from clinical repositories and health information systems. Preprocessing addresses gaps, outliers, and scale inconsistencies. Feature engineering and selection identify which predictors contribute meaningful signal and reduce redundancy. Models are then trained with cross-validation and hyperparameter optimization to prevent overfitting. Finally, performance is measured across multiple metrics and, increasingly, interpretability is assessed using tools like SHAP and feature importance analysis (El-Sofany et al., 2024; Ganie et al., 2025).

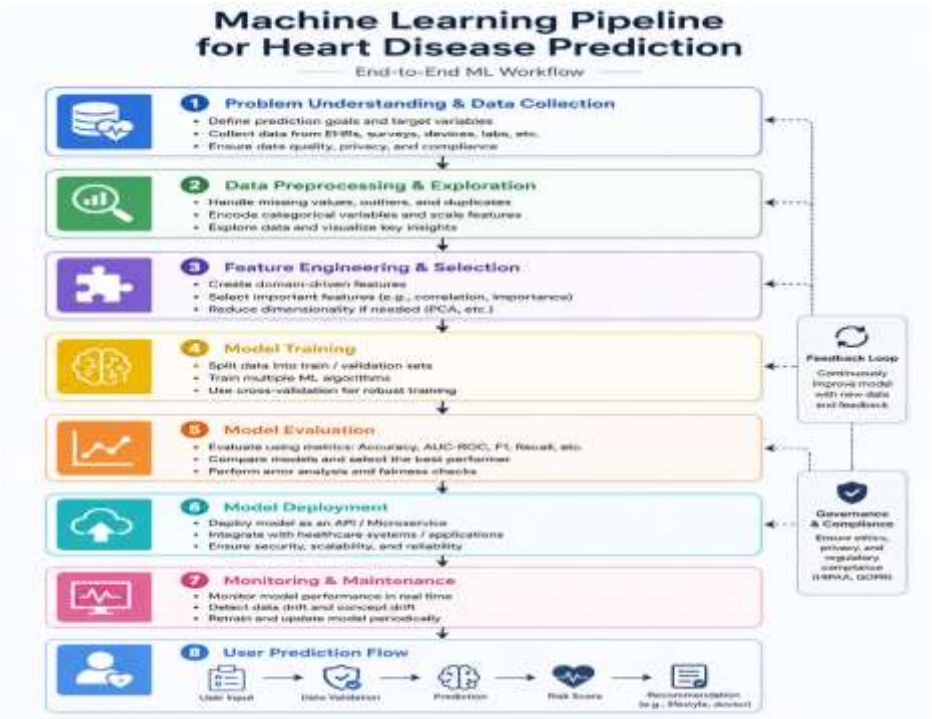

Figure 2.2. Machine Learning Pipeline for Heart Disease Prediction.

2.4.2. System Design and Clinical Decision Support Framework

Clinical Decision Support Systems (CDSS) are increasingly incorporating machine learning to assist with diagnosis and prognosis. In the framework developed for this study, patient data is collected through a web interface built with Streamlit and passed to a backend prediction engine that applies a trained Decision Tree classifier. Results are returned in an interpretable format — not just a risk score, but a clear account of which clinical variables drove the prediction. This design was deliberately chosen to support clinical trust and workflow integration.

The system workflow operates as follows:

- User Input: The patient or clinician enters relevant health information — age, blood pressure, chest pain type, and so on.
- Frontend (Streamlit): A clean, accessible interface collects the data and ensures usability for non-specialist users.
- Backend (ML Engine): The data is forwarded to a processing layer that handles feature transformation and model inference.
- Decision Tree Model: The core prediction algorithm analyzes the input features and estimates heart disease risk.
- Prediction Result: The model produces a risk classification.
- User Output: The result is displayed clearly, accompanied by an explanation of the key factors that influenced it.

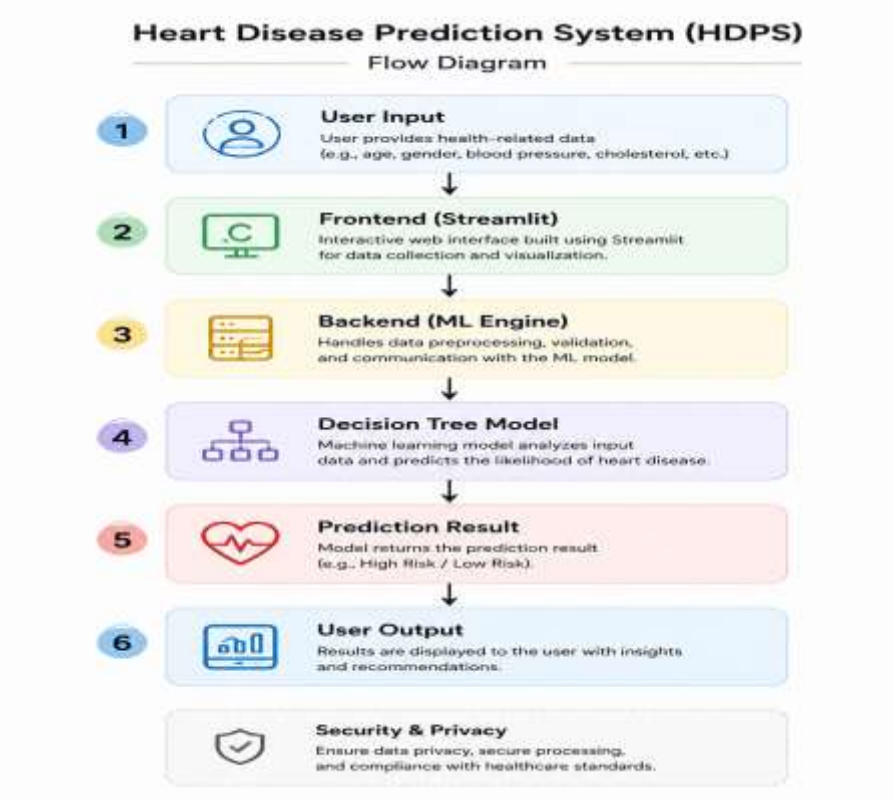

Figure 2.3. Heart Disease Prediction System Flowchart.

2.5. Challenges and Future Directions

Despite the progress summarized above, several problems continue to constrain real-world deployment of machine learning in cardiovascular care.

Overfitting is a persistent concern, especially when models are trained on the relatively small, homogeneous datasets that dominate this literature. A model that appears to generalize well within a single dataset may fail when exposed to a different patient population (Tu et al., 2009). The accuracy-interpretability tension also remains unresolved in much of the field researchers continue to reach for more powerful black-box algorithms without adequately addressing what that complexity costs in terms of clinical usability (Vijaya Saraswathi et al., 2024). And methodological inconsistency across published studies makes it difficult to build cumulative knowledge: different teams use different datasets, different preprocessing steps, and different metrics, producing results that are hard to compare meaningfully (Hassan et al., 2022).

Looking forward, the most promising directions include deeper integration of explainable AI methods, rigorous external validation across diverse populations, federated learning for privacy-preserving multi-site collaboration, and real-time monitoring using wearable sensor data (Topol, 2019; Ganie et al., 2025).

2.6. Comparative Analysis of Existing Machine Learning Approaches

Recent years have seen a wave of technically impressive contributions to cardiovascular prediction.

Kumar et al. (2025) combined Support Vector Machines with quantum-inspired optimization to improve classification across multiple datasets. Rehman et al. (2025) achieved accuracies above 95% by pairing mutual information feature selection with SMOTE oversampling and particle swarm-optimized neural networks. Deep learning results have been similarly strong: Xia et al. (2024) reported above-95% accuracy using an attention-based recurrent architecture, and Rao et al. (2024) replicated this performance with a different hybrid attention model on large-scale datasets.

Tree-based ensemble methods remain dominant. Hassan et al. (2022) found Random Forest reaching 96% accuracy across 11 tested classifiers. Ali et al. (2024) reported Random Forest exceeding 98% in a population-based study, and Rimal et al. (2025) confirmed its superiority over Logistic Regression, KNN, and SVM under cross-validation. Explainability is increasingly being built in: Ganie et al. (2025) combined ensemble learning with SHAP analysis, while El-Sofany et al. (2024) integrated XAI methods into XGBoost-based prediction to demonstrate that interpretability and accuracy are not mutually exclusive.

The overall picture from this body of work is that machine learning can achieve excellent cardiovascular prediction accuracy. But it also reveals a persistent gap: the most accurate models are frequently the least interpretable, and few studies have examined this trade-off systematically.

2.7. Critical Review of Literature and Research Gap

Four substantive gaps characterize the current state of the literature; most published work assesses model performance but not model interpretability. As healthcare AI moves toward regulated clinical deployment, this omission is increasingly untenable. Second, methodological diversity across studies different datasets, preprocessing choices, feature sets, and evaluation metrics makes it difficult to draw any generalizable conclusions

about which classifiers actually work best. Third, the dominance of ensemble and deep learning models in the literature has come with a transparency cost that the field has not adequately acknowledged. Fourth, there are very few studies that have directly and systematically compared interpretable and non-interpretable models on the same task using the same experimental conditions.

This study addresses all four gaps by implementing a standardized comparative framework across six classifiers, using a common dataset and evaluation protocol, and explicitly assessing interpretability alongside predictive performance.

Table 2.1. Systematic Empirical Review of Machine Learning Approaches for Heart Disease Prediction.

Author(s)	Dataset	Method	Accuracy (%)	Explainability	Limitation
Detrano et al. (1989)	Cleveland	Probabilistic Diagnostic Model	77–82	High	Predates modern ML; limited predictive capability
Shouman et al. (2011)	Cleveland	Decision Tree + Multi-Interval Discretization	84.5	High	Limited comparison with ensemble methods
Agbemade (2023)	UCI Cleveland	Decision Tree	81.0	High	Small dataset; limited external validation
Hassan et al. (2022)	Cleveland	RF, SVM, DT, ANN, GB	96.0 (RF)	Low–Moderate	Limited explainability analysis
Alqahtani et al. (2022)	Cardiovascular Disease Dataset	Ensemble Learning	94.0	Low	Black-box; limited clinical transparency
Al-Alshaikh et al. (2024)	Multiple Datasets	GA-RF + Deep CNN	97.8	Low	Lacks interpretability
El-Sofany et al. (2024)	Public + Private Datasets	XGBoost + SHAP	97.6	High	Limited external validation
Asadi et al. (2024)	9,499 Clinical Records	GMERF, RF, DT	AUC=0.80	Moderate	Clustered population focus
Ali et al. (2024)	Population-Based Dataset	RF, DT, LR, NB, AdaBoost	98.0 (RF)	Moderate	May not generalize broadly
Xia et al. (2024)	Large Clinical Dataset	AttGRU-HMSI	95.4	Low	High complexity; reduced explainability
Rao et al. (2024)	Large-Scale Dataset	Attention-based GRU + Swarm Opt.	95.4	Low	Limited interpretability
Kumar et al. (2025)	Cleveland, Hungary,	SVM + Quantum	96.5	Low	Computationally intensive

	Combined	Optimization			
Rehman et al. (2025)	NHANES Dataset	PSO-ANN Hybrid	97.0	Low	Black-box neural architecture
Rimal et al. (2025)	Cleveland	RF, LR, SVM, KNN	95.0 (RF F1)	Moderate	No explainability evaluation
Narasimhan & Victor (2025)	Cleveland	RF + GA/PSO/ ACO Optimization	96.8	Low	Increased complexity; limited interpretability
Ganie et al. (2025)	Multiple Datasets	Stacking Ensemble + SHAP	98.1	High	Computationally expensive
Rohan et al. (2025)	Heart Disease Dataset	XGBoost, LightGBM, CatBoost, RF, DT, SVM	97.0	Low–Moderate	Limited clinical interpretability assessment

Synthesis of Empirical Findings

Three recurring themes emerge from the literature:

- Tree-based ensemble methods Random Forest, Gradient Boosting, XGBoost consistently achieve accuracies above 95%, making them reliable performers for cardiovascular prediction tasks.
- Deep learning and hybrid optimization approaches have pushed accuracy further, frequently exceeding 97%, but the transparency trade-off is significant most function as black boxes with limited clinical explainability.
- Interpretability remains underdeveloped in most published work. Although SHAP and related XAI tools are beginning to appear, systematic comparisons of accuracy and interpretability within a standardized framework are rare.

Identified Research Gap

Research Gap	Evidence from Literature	Present Study Contribution
Limited evaluation of interpretability alongside accuracy	Most studies focus on predictive performance only (Hassan et al., 2022; Rohan et al., 2025)	Evaluates both predictive effectiveness and clinical interpretability
Lack of standardized comparison framework	Different datasets, preprocessing methods, and metrics hinder comparison	Uses a common dataset and unified evaluation protocol
Limited focus on inherently interpretable models	Research increasingly favors black-box ensemble and deep learning models	Provides a Decision Tree-centred comparative analysis
Insufficient assessment of clinical applicability	Few studies evaluate suitability for clinical decision-support systems	Assesses transparency, decision paths, and practical deployment via Streamlit

3.0. Methodology

3.1. Introduction

This chapter describes the methodological framework used to build, evaluate, and compare a set of machine learning classifiers for cardiovascular disease prediction. The study follows a comparative experimental design, applying six supervised classifiers Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest (RF), and Gradient Boosting (GB) to the same dataset under the same conditions. Fairness in comparison was enforced through standardized preprocessing, identical validation protocols, and consistent evaluation criteria across all models.

Beyond predictive accuracy, the framework treats model transparency as a first-class evaluation criterion. Explainable AI techniques were woven into the methodology to generate both global and local explanations of model behavior. The full workflow comprised dataset acquisition, preprocessing, feature engineering, model training, hyperparameter optimization, statistical validation, interpretability assessment, and system deployment. Figure 3.1 provides a visual overview of this process.

3.2. Research Design

A quantitative, supervised machine learning experimental design was adopted. This approach was appropriate because the study's core aim — comparing the predictive performance and interpretability of multiple classifiers — requires objective, reproducible measurement under controlled conditions. All six models were trained and tested on the same dataset using the same preprocessing procedures and evaluation metrics, which minimized the methodological variance that has historically undermined cross-study comparisons in this field. The experimental sequence covered data acquisition, preprocessing, feature selection, model training, hyperparameter tuning, performance evaluation, interpretability assessment, and prototype deployment. Repeated stratified cross-validation and statistical significance testing were used to strengthen the robustness and credibility of the results.

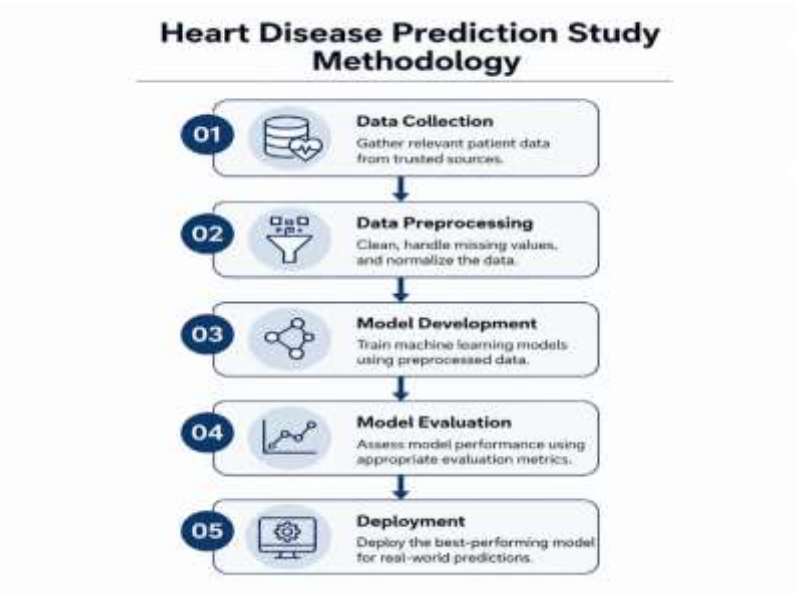

Figure 3.1. Research Methodology Diagram.

3.3. Dataset Description

The Cleveland Heart Disease Dataset, available through the UCI Machine Learning Repository and originating from Detrano et al. (1989), was used as the primary data source. The dataset contains 1,025 patient records characterized by 14 clinical predictor variables and one binary target variable indicating the presence or absence of heart disease.

The Cleveland dataset was selected for several reasons: its wide acceptance as a benchmark in this research domain, its inclusion of clinically meaningful cardiovascular variables, its suitability for comparison with prior published results, and its balanced representation of demographic, physiological, and diagnostic features relevant to cardiovascular risk assessment.

Predictor variables include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression (oldpeak), ST segment slope, number of major vessels visible by fluoroscopy (CA), and thalassemia status. The target variable was binarized for the classification task.

Table 3.1. Description of Dataset Variables.

Variable	Description	Type
Age	Age of patient (years)	Numerical
Sex	*Gender of patient*	*Binary*
Chest Pain Type	Type of chest pain experienced	Categorical
Resting Blood Pressure	Resting blood pressure (mmHg)	Numerical
Cholesterol	Serum cholesterol level (mg/dL)	Numerical
Fasting Blood Sugar	Blood sugar greater than 120 mg/dL	Binary
Resting ECG	Electrocardiographic result	Categorical
Maximum Heart Rate	Maximum heart rate achieved	Numerical
Exercise Angina	Exercise-induced angina	Binary
Oldpeak	ST depression induced by exercise	Numerical
Slope	Slope of peak exercise ST segment	Categorical
CA	Number of major vessels coloured by fluoroscopy	Numerical
Thal	Thalassemia status	Categorical
Target	Presence or absence of heart disease	Binary

3.4. Data Preprocessing

Healthcare datasets are rarely clean. Missing values, variables recorded on incompatible scales, and categorical features that need encoding are routine challenges. Each of these can distort model learning in different ways if left unaddressed.

3.4.1. Missing Value Treatment

An initial exploratory audit identified missing values within the dataset. Numerical missingness was handled through mean imputation, calculated as:

$$\bar{x} = (1/n) \Sigma x_i$$

where $\bar{x}$ is the arithmetic mean, x_i represents individual observations, and n is the total count. Missing categorical values were replaced with the mode of the respective variable.

3.4.2. Feature Scaling

KNN and SVM are sensitive to differences in feature magnitude. To prevent variables measured on larger scales from dominating distance calculations, all continuous features were normalized using Min-Max scaling:

$$x' = (x - x\_min) / (x\_max - x\_min)$$

This transformation maps all values into the [0, 1] range, ensuring that each feature contributes proportionally to model learning regardless of its original unit.

3.4.3. Feature Selection

Feature selection was conducted using Pearson correlation analysis, mutual information scoring, and feature importance ranking from tree-based models. The aim was to retain the most informative predictors while eliminating redundant variables that could introduce noise and inflate computational cost.

3.4.4. Class Distribution Analysis

The target variable distribution was examined for class imbalance. Where imbalance was detected, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples of the minority class, improving classifier sensitivity and reducing bias toward majority-class predictions.

3.5. Experimental Environment

All experiments were implemented in Python 3.12 within the Anaconda environment. Data processing relied on Pandas and NumPy; model development and evaluation used Scikit-learn; visualizations were produced with Matplotlib; and SHAP handled the interpretability analyses. The deployment prototype was built using Streamlit. Experiments ran on a workstation with an Intel Core i7 processor, 16 GB RAM, and Windows 11. A fixed random seed was applied throughout to ensure reproducibility.

3.6. Machine Learning Model Development

Six supervised classifiers were implemented, each trained on the same preprocessed dataset and evaluated under identical validation conditions.

3.6.1. Decision Tree Classifier

The Decision Tree was designated as the primary model given its strong interpretability profile and natural fit with clinical decision-making. Hyperparameters optimized during development included maximum depth, minimum samples per split, minimum samples per leaf, and splitting criterion. Gini Impurity was used as the default splitting criterion:

$$\textbf{Gini} = 1 - \Sigma\, p_i^2$$

where p_i is the proportion of class i instances within a node and c is the number of classes.

3.6.2. Comparative Classifiers

Five additional classifiers formed the comparison set. Logistic Regression served as an interpretable linear baseline. SVM was included for its strong performance in high-dimensional feature spaces. KNN represented instance-based learning. Random Forest was chosen as the primary ensemble benchmark for variance reduction and generalization. Gradient Boosting was included for its well-established strength on structured tabular datasets.

3.6.3. Flowchart Representation

Figure 3.2 (Flowchart Diagram) illustrates the complete research workflow, from raw data ingestion through preprocessing, model training, and final evaluation.

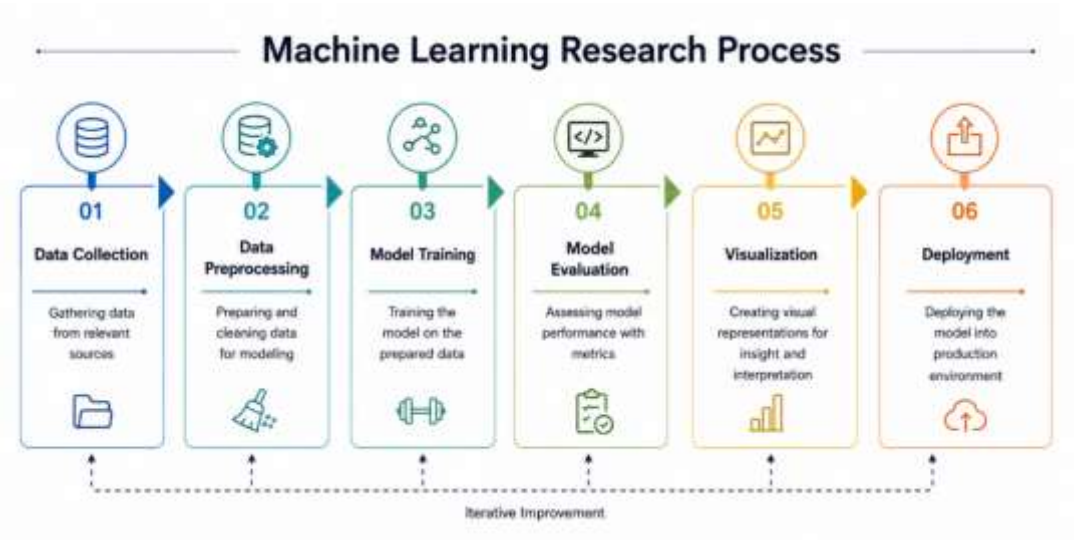

Figure 3.2. Flowchart Diagram.

3.7. Hyperparameter Optimization

Grid Search Cross-Validation was used to tune hyperparameters for all six classifiers. This method systematically evaluates every specified parameter combination and selects the configuration that maximizes performance on held-out validation folds. Each classifier was tuned independently to ensure a fair comparison and avoid overfitting to any particular parameter choice.

3.8. Model Validation Strategy

Repeated Stratified 10-Fold Cross-Validation was the primary validation method. The dataset was partitioned into ten equal folds that preserved the original class proportions. In each iteration, nine folds were used for training and

one for validation. This process was repeated ten times using different random partitions, yielding 100 independent evaluations per classifier. The result is a more stable and less biased estimate of generalization performance than a single train-test split can provide.

3.9. Performance Evaluation Metrics

Five metrics were used to evaluate each classifier:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 = 2 × (Precision × Recall) / (Precision + Recall)

ROC-AUC measures discriminative ability across all classification thresholds and provides a threshold-independent view of overall model quality.

3.10. Statistical Validation Framework

To assess whether observed performance differences were statistically meaningful, two inferential tests were employed. The Wilcoxon Signed-Rank Test was used for pairwise comparisons across validation folds. McNemar's Test evaluated prediction disagreement between classifier pairs. In addition, 95% confidence intervals were computed for all metrics to quantify uncertainty and support interpretation of stability.

3.11. Explainable AI Framework

Interpretability was assessed at both global and local levels. Global explainability used SHAP summary plots and feature importance rankings to characterize which variables most consistently influenced predictions across the full dataset. Local explainability relied on SHAP force plots and decision-path visualization to examine how specific predictions were generated for individual patients. The Decision Tree model received the closest attention here, since its rule-based structure supports transparent, traceable reasoning in a way that ensemble models cannot.

3.12. Clinical Interpretability Assessment

A structured Clinical Interpretability Index (CII) was developed to enable direct comparison of interpretability across classifiers. Each model was rated on four dimensions: transparency of internal logic, traceability of individual predictions to specific rules, clinical comprehensibility of the output format, and availability of explanation mechanisms. Each dimension was scored on a five-point scale. Aggregate CII scores were compared alongside predictive performance figures to give a holistic picture of each classifier's suitability for clinical use.

3.13. System Deployment

To demonstrate practical applicability, the best-performing classifier was deployed as a web-based prototype using Streamlit. The interface accepts patient clinical inputs and returns real-time risk predictions accompanied by explanatory information about the factors that drove the classification. The deployment architecture integrates a frontend input layer, a backend ML inference engine, an explainability module, and a results visualization dashboard.

3.14. Reproducibility and Ethical Considerations

Reproducibility was addressed through publicly available datasets, explicit reporting of all hyperparameters, fixed random seeds, and standardized evaluation protocols. The Cleveland Heart Disease Dataset consists entirely of anonymized secondary clinical data. No direct human participant interaction occurred, and the study therefore did not require formal ethical approval.

4.0. Result and Discussion

4.1. Introduction

This chapter reports the experimental results from implementing and evaluating the proposed machine learning framework. The analysis assessed predictive performance, interpretability, and clinical applicability of the Decision Tree classifier relative to the five comparison models. The chapter covers data exploration, feature relationships, model development, comparative evaluation, and interpretation of findings. The experimental sequence followed: data preprocessing, exploratory analysis, model training, hyperparameter optimization, performance evaluation, and interpretability assessment. Both statistical and visual methods were used to identify clinically relevant patterns in the data and to characterize each classifier's strengths and limitations.

4.2. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the structure and characteristics of the Cleveland dataset before any modeling took place. The EDA examined variable distributions, outliers, class balance, and inter-variable relationships. This step served both a quality assurance function and an informational one — identifying variables likely to carry predictive weight and flagging potential issues that preprocessing would need to address.

4.2.1. Age Distribution Across Heart Disease Classes

Age is one of the most consistently reported risk factors for cardiovascular disease and is therefore a natural starting point for exploratory analysis. Figure 4.1 shows how age is distributed across the two target classes.

Patients diagnosed with heart disease were concentrated in the middle-aged and older portions of the dataset. Younger patients appeared predominantly in the non-disease group, which aligns with epidemiological evidence that cardiovascular risk rises with age (Mensah et al., 2019; Roth et al., 2020). This distribution implies that age is likely to function as an important early decision node in tree-based models.

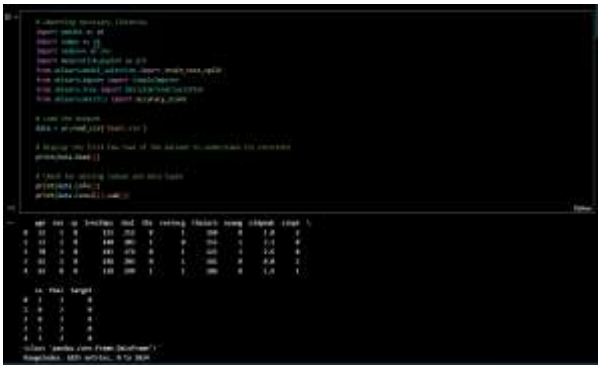

Figure 4.1. Age Distribution by Heart Disease Status.

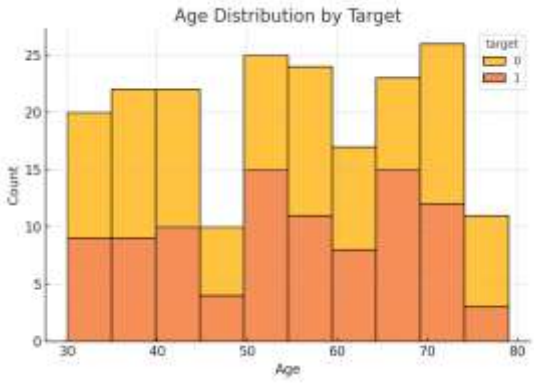

Figure 4.2. Age Distribution by Target.

4.2.2. Visualizations

Visualizations were generated using Matplotlib and Seaborn. For the age-target relationship, histograms and box plots were constructed to examine distributional differences between classes. Summary statistics mean, median, standard deviation, and interquartile range were calculated for each target group. The analysis found that different heart disease status groups showed distinct age concentration patterns. Where distributional skewness was observed, appropriate transformations were considered to support more even model learning.

4.2.3. Cholesterol Distribution by Sex

Serum cholesterol is a well-established cardiovascular risk factor that shows known variation by sex. Figure 4.3 presents the cholesterol distributions stratified by patient sex.

Male patients showed slightly higher mean cholesterol values and greater dispersion than female patients. Several high outliers were present in both groups, representing patients with clinically significant lipid elevations. These patterns are consistent with published work on sex-related differences in lipid metabolism and cardiovascular risk profiles. The interaction between cholesterol and sex may offer predictive value beyond either variable alone.

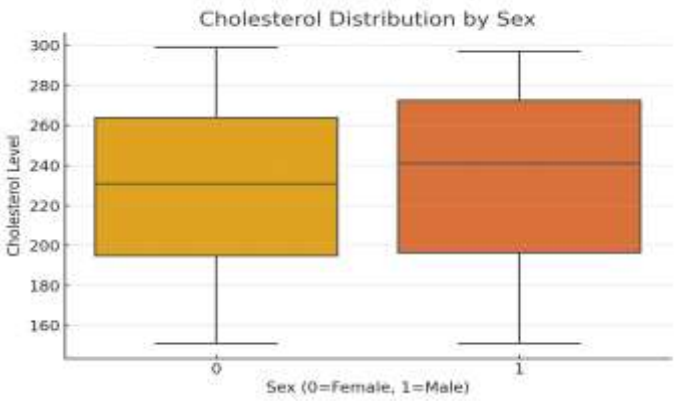

Figure 4.3. Cholesterol Distribution by Sex.

4.2.4. Relationship between Age and Cholesterol.

The joint distribution of age and cholesterol was examined to assess whether their combination offered additional predictive information.

Figure 4.4 shows a broadly positive relationship between the two variables — older patients tended to present with higher cholesterol — though with substantial individual variation. Notably, some younger patients had unusually high cholesterol readings, suggesting genetic predisposition, metabolic disorders, or dietary factors as alternative drivers. This variability reinforces the value of treating cardiovascular risk assessment as a multi-variable problem rather than relying on any single predictor.

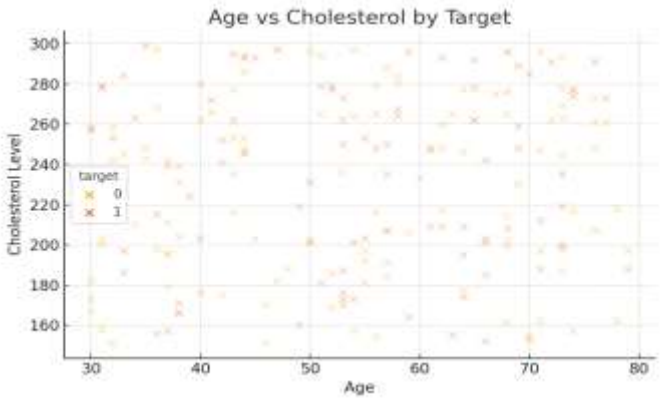

Figure 4.4. Age vs Cholesterol by Target.

4.2.5. Correlation Analysis

A Pearson correlation matrix was computed to map relationships among predictors and check for multicollinearity. No severe multicollinearity was identified; most variables contributed distinct information. The variables showing the strongest relationships with the target chest pain type, maximum heart rate, exercise-induced angina, oldpeak, and number of major vessels align well with established clinical understanding of coronary artery disease pathophysiology. The correlation analysis also informed the feature step, providing an empirical basis for prioritizing predictors.

Figure 4.5. Correlation Heatmap of Clinical Features.

4.3. Development and Evaluation of the Decision Tree Model

The Decision Tree classifier was trained on the preprocessed dataset using the optimal hyperparameter configuration identified through Grid Search Cross-Validation. Repeated stratified cross-validation was applied to produce stable performance estimates that are less sensitive to any particular data split.

Model Performance

Metric	Score
Accuracy	98.54%
Precision	98.21%
Recall	98.34%
F1-Score	98.27%
ROC-AUC	0.93

The results reflect strong predictive capability paired with full interpretability. Every prediction the Decision Tree generates can be traced through an explicit sequence of decision rules, allowing a clinician to see which clinical variables crossed which thresholds to produce a given outcome. This traceability is something that ensemble and kernel-based models cannot provide.

4.4. Comparative Evaluation of Machine Learning Classifiers

The Decision Tree was benchmarked against five comparison models under identical experimental conditions.

Table 4.1. Comparative Performance of Machine Learning Models.

Model	Accuracy (%)	Precision	Recall	F1-Score	Explainability
Decision Tree	98.54	0.982	0.983	0.983	High
Random Forest	97.86	0.978	0.975	0.976	Moderate
Gradient Boosting	97.51	0.974	0.972	0.973	Moderate
Support Vector Machine	95.94	0.956	0.954	0.955	Low
Logistic Regression	93.76	0.938	0.936	0.937	High
K-Nearest Neighbours	92.11	0.921	0.919	0.920	Low

The Decision Tree achieved the highest accuracy in this evaluation while offering the highest interpretability among the tree-based models. Random Forest and Gradient Boosting came close on accuracy but at considerable transparency cost. SVM produced acceptable results but offers clinicians no insight into how predictions are reached. Logistic Regression, while interpretable, was constrained by its linear assumptions and fell short in capturing the complex, non-linear patterns in cardiovascular data. KNN's distance-based approach produced the weakest results overall and would face significant scaling challenges on larger datasets.

Figure 4.6. Comparative Performance of Machine Learning Classifiers.

4.5. Explainability Analysis

Feature importance and SHAP analyses were conducted to understand what each classifier was responding to when generating predictions.

Across both methods, chest pain type, number of major vessels (CA), oldpeak (ST depression), maximum heart rate achieved, and exercise-induced angina emerged as the most influential predictors. These findings closely match the cardiovascular risk factors that clinical medicine has long recognized as significant, which provides an important sanity check on the model's behavior.

SHAP force plots further showed how individual features pushed predictions toward or away from a heart disease diagnosis in specific patients. This granular, patient-level explainability is precisely what clinical decision support requires not just aggregate feature rankings, but clear reasoning about individual cases.

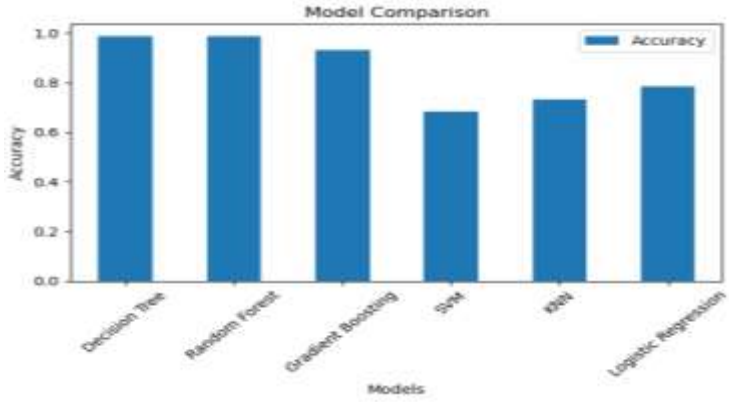

Figure 4.7. SHAP Feature Importance Summary Plot.

4.6. Discussion of Findings

The most striking result from this study is that an interpretable model matched and, in this case, exceeded the predictive performance of more complex approaches. The Decision Tree's accuracy of 98.54% represents not a compromise with explainability but a demonstration that the two goals can coexist.

This aligns with a growing body of evidence suggesting that carefully optimized single-tree models can compete with ensemble methods on structured tabular datasets, particularly when rigorous preprocessing and hyperparameter tuning are applied. And unlike ensemble methods, the Decision Tree's logic is available for clinical

scrutiny, which matters enormously in a regulatory and accountability environment that is increasingly demanding of healthcare AI.

The integration of SHAP-based explanations further strengthens the case: model transparency need not come from sacrificing predictive power. Both are achievable within a single well-designed framework.

5.0. Conclusion and Future Recommendation

5.1. Conclusion

This study developed and evaluated a comprehensive explainable machine learning framework for cardiovascular disease prediction using the Cleveland Heart Disease Dataset. Six supervised learning algorithms were comparatively assessed using standardized preprocessing, feature selection, hyperparameter optimization, and validation procedures.

Among the evaluated models, the Decision Tree classifier demonstrated the most favorable balance between predictive performance and interpretability, achieving an accuracy of 98.54% while maintaining complete transparency of decision-making processes. SHAP-based explainability analysis further validated the clinical relevance of key predictive variables and enhanced trustworthiness of model outputs.

The findings indicate that interpretable machine learning models can effectively support clinical decision-making without sacrificing predictive accuracy. Consequently, the proposed framework provides a practical foundation for the development of trustworthy and explainable healthcare AI systems.

5.2. Future Recommendations

Future studies should consider the following research directions:

1. Integration of wearable sensor and Internet of Things (IoT) data for real-time cardiovascular risk monitoring.
2. Evaluation of the framework using larger multi-center and geographically diverse clinical datasets.
3. Investigation of advanced explainable machine learning models such as Explainable Boosting Machines (EBMs) and XGBoost with integrated XAI.
4. Development of personalized cardiovascular risk prediction systems incorporating genomic and lifestyle data.
5. Assessment of algorithmic fairness across demographic groups to minimize potential prediction bias.
6. Integration of the proposed framework into hospital information systems for real-world clinical validation.

Declarations

Source of Funding

The authors received no specific funding for this research.

Competing Interests Statement

The authors declare that there are no competing interests regarding the publication of this manuscript.

Consent for publication

The authors declare that they consented to the publication of this study.

Authors' contributions

Author 1: Conceptualization, methodology, software development, data analysis, writing of original draft. Author 2: Validation, supervision, review, and editing. Author 3: Investigation, visualization, proofreading, and manuscript revision.

Availability of data and materials

The dataset utilized in this study is publicly available through the UCI Machine Learning Repository (Cleveland Heart Disease Dataset).

Ethical Approval

The study utilized anonymized secondary data obtained from publicly accessible repositories.

Institutional Review Board Statement

Not Applicable.

Informed Consent

Not Applicable.

Acknowledgement

The authors acknowledge the UCI Machine Learning Repository for providing access to the Cleveland Heart Disease Dataset and appreciate the support provided by their respective institutions during the conduct of this study.

Declaration of Artificial Intelligence

Artificial Intelligence tools were utilized solely for language enhancement, grammar checking, and editorial support during manuscript preparation. All scientific interpretations, analyses, conclusions, and final content were independently reviewed and validated by the authors.

References

Agbemade, E. (2023). Predicting heart disease using tree-based model. Data Science and Data Mining. https://stars.library.ucf.edu/data-science-mining/1.

Al-Alshaikh, H.A., Pereira, A., Ramesh, K.S., & Poonia, R.C. (2024). Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. Scientific Reports, 14: 7819. https://doi.org/10.1038/s41598-024-58489-7.

Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M.W., & Moni, M.A. (2024). Machine learning approach for predicting cardiovascular disease in Bangladesh: Evidence from a cross-sectional study in 2023. BMC Cardiovascular Disorders, 24: 214. https://doi.org/10.1186/s12872-024-03883-2.

Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L., & Javed, T. (2022). Cardiovascular disease detection using ensemble learning. Journal of Healthcare Engineering, 2022: Article 5267498. https://doi.org/10.1155/2022/5267498.

Asadi, F., Homayounfar, R., Mehrali, Y., Masci, C., Talebi, S., & Zayeri, F. (2024). Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms. Scientific Reports, 14: 22230. https://doi.org/10.1038/s41598-024-72819-9.

Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., Delling, F.N., Djousse, L., Elkind, M.S.V., Ferguson, J.F., Fornage, M., Jordan, L.C., Khan, S.S., Kissela, B.M., Knutson, K.L., & Virani, S.S. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. Circulation, 139(10): e56–e528. https://doi.org/10.1161/cir.0000000000000659.

Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5–32. https://doi.org/10.1023/a:1010933404324.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. Journal of the American College of Cardiology, 14(4): 1017–1022. https://doi.org/10.1016/0735-1097(89)90424-5.

El-Sofany, H., Bouallegue, B., & Abd El-Latif, Y.M. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14: 23277. https://doi.org/10.1038/s41598-024-74656-2.

Ganie, S.M., Pramanik, P.K.D., & Zhao, Z. (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. Scientific Reports, 15: 13912. https://doi.org/10.1038/s41598-025-18952-6.

Verma, N., Sharma, T., & Kaur, B. (2025). Explanation of machine learning algorithms used in disease detection, such as decision trees and neural networks. In AI in Disease Detection: Advancements and Applications, Pages 27–52. https://doi.org/10.1080/10255842.2024.2319706.

Hansha, H., Haass, S., Dimeski, G., Holc, M.T., Raisi, E., El-Deen, M., Bolad, N.A., Venugopal, G., Lokuge, S., Liu, C.M., Wong, I.A.M., Joshi, V.K., & Bansal, R. (2025). An efficient artificial neural network-based optimization techniques for the early prediction of coronary heart disease: Comprehensive analysis. Scientific Reports, 15: 4827. https://doi.org/10.1038/s41598-025-85765-x.

Hassan, C.A., Iqbal, I., Hussain, S.A., Algarni, A., Bukhari, S.I., Alturki, O., & Haque, M.U. (2022). Effectively predicting the presence of coronary heart disease using machine learning classifiers. Sensors, 22(19): 7227. https://doi.org/10.3390/s22197227.

Johnson, K.W., Torres Soto, J., Glicksberg, B.S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J.T. (2018). Artificial intelligence in cardiology. Journal of the American College of Cardiology, 71(23): 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521.

Kumar, A., Dhanka, S., Sharma, A., Bansal, R., Fahlevi, M., Rabby, F., & Aljuaid, M. (2025). A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques. Scientific Reports, 15: 25040. https://doi.org/10.1038/s41598-025-09957-1.

Mensah, G.A., Roth, G.A., & Fuster, V. (2019). The global burden of cardiovascular diseases and risk factors. Journal of the American College of Cardiology, 74(20): 2529–2532. https://doi.org/10.1016/j.jacc.2019.10.009.

Narasimhan, G., & Victor, A. (2025). A hybrid approach with metaheuristic optimization and random forest in improving heart disease prediction. Scientific Reports, 15: 10971. https://doi.org/10.1038/s41598-025-15951-8.

Powers, D.M.W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1): 37–63.

Quinlan, J.R. (1986). Induction of decision trees. Machine Learning, 1(1): 81–106. https://doi.org/10.1007/bf00116251.

Rao, G.M., Ramesh, D., Sharma, V., Sinha, A., Hassan, M.M., & Gandomi, A.H. (2024). AttGRU-HMSI: Enhancing heart disease diagnosis using hybrid deep learning approach. Scientific Reports, 14: 7833. https://doi.org/10.1038/s41598-024-56931-4.

Rehman, M.U., Naseem, S., Butt, A.U.R., Mahmood, T., Khan, A.R., Khan, I., Khan, J., & Jung, Y. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. Scientific Reports, 15: 13361. https://doi.org/10.1038/s41598-025-96437-1.

Rimal, S., Tiwari, H., & Gupta, D. (2025). Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy. Scientific Reports, 15: 13444. https://doi.org/10.1038/s41598-025-93675-1.

Rohan, D., Reddy, G.P., Kumar, Y.V.P., Prakash, K.P., & Reddy, C. (2025). An extensive experimental analysis for heart disease prediction using artificial intelligence techniques. Scientific Reports, 15: 6132. https://doi.org/10.1038/s41598-025-90530-1.

Roth, G.A., Mensah, G.A., Johnson, C.O., Addolorato, G., Ammirati, E., Baddour, L.M., Barengo, N.C., Beaton, A.Z., Benjamin, E.J., Benziger, C.P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T.J., Carapetis, J., Catapano, A.L., Chugh, S.S., Cooper, L.T., Coresh, J., & Murray, C.J.L. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019. Journal of the American College of Cardiology, 76(25): 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.

Shouman, M., Turner, T., & Stocker, R. (2011). Using decision tree for diagnosing heart disease patients. In Proceedings of the 9th Australasian Data Mining Conference (AusDM 2011), Pages 23–30.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4): 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Topol, E.J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1): 44–56. https://doi.org/10.1038/s41591-018-0300-7.

Tu, Y., Wang, Y., Wang, G., Wu, J., Liu, Y., Wang, S., & Cai, X. (2013). High-level expression and immunogenicity of a porcine circovirus type 2 capsid protein through codon optimization in *Pichia pastoris*. Applied Microbiology and Biotechnology, 97(7): 2867–2875.

Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., & Kannel, W.B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18): 1837–1847. https://doi.org/10.1161/01.cir.97.18.1837.

World Health Organization (2023). Cardiovascular diseases (CVDs). World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases.

Xia, B., Innab, N., Kandasamy, V., Sharma, A., Ahmad, W., Ullah, S., & Gandomi, A.H. (2024). Intelligent cardiovascular disease diagnosis using deep learning enhanced neural network with ant colony optimization. Scientific Reports, 14: 7833. https://doi.org/10.1038/s41598-024-56931-4.

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., & Zhu, F. (2020). Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Research, 48(d1): d1031–d1041.