Improved Supervised Machine Learning Classification Approach For Heart Disease Detection

Michael Funskin

Computer and Information Systems Towson University Towson, USA

DOI: https://doi.org/10.47772/IJRISS.2026.10190030

Received: 06 January 2026; Accepted: 29 January 2026; Published: 14 February 2026

ABSTRACT

Heart disease remains one of the leading causes of death globally, thus highlighting the need for tools that can support early and accurate detection. This work develops a machine learning model to predict heart disease based on the UCI Heart Disease dataset, which combines data from four clinical cohorts (Cleveland, Hungarian, Switzerland, and VA Long Beach). Datasets were preprocessed by addressing missing values with median imputation, normalizing numeric ranges with min–max scaling, and applying SMOTE to correct class imbalance. Five classification algorithms (Logistic Regression, Support Vector Machine (SVM), Random Forest, k-Nearest Neighbors (kNN), and XGBoost) were trained and evaluated with XGBoost achieving the best performance with an accuracy of 0.88, F1-score of 0.88, ROC-AUC of 0.88, and PR-AUC of 0.87. SHAP analysis showed oldpeak (ST depression), ca (number of major vessels), thalach (maximum heart rate achieved), exang (exercise-induced angina), and thal (thalassemia type) as the most significant predictors of heart disease demonstrating consistency with prior work and reinforcing their importance in clinical diagnosis. Overall, the model balanced accuracy, interpretability and consistency across datasets and suitable for integration into clinical decision-support systems

Keywords; Heart disease, Machine Learning, Model, UCI, Dataset, Classification, Algorithms, Techniques

INTRODUCTION

Cardiovascular disease is one of the major global health concerns, causing millions of deaths annually. Early detection of heart disease plays a critical role in reducing mortality rates, improving patient outcomes, and optimizing clinical decision-making. Machine Learning (ML) provides data-driven ways to analyze complex clinical data and extracts insights from historical patient records to provide relationships between patient characteristics and disease outcomes. The UCI Heart Disease dataset, which integrates clinical data from the Cleveland, Hungarian, Switzerland, and VA Long Beach cohorts, serves as benchmark for evaluating predictive models in medical data mining provides an opportunity to model heart disease risk using 14 clinical features, including demographic, physiological etc. However, missing values, uneven class distributions, and differences between cohorts make preprocessing and validation essential for reliable results.

Mao et al. [1] demonstrated the potential of supervised learning algorithms to improve diagnostic accuracy and found that ensemble and deep-learning algorithms such as Random Forest, Support Vector Machine (SVM), and XGBoost performed effectively across multiple heart disease datasets. Özcan and Peker [2] emphasized the value of rule-based decision trees (CART) for extracting patterns but noted limitations in representing broader clinical variables. Kanchanamala et al. [3] achieved strong performance with hybrid deep-learning architectures but acknowledged that their computational complexity limited real-time clinical use. Building on prior work, this paper aims to:

1. Develop and evaluate multiple supervised ML classifiers for heart disease prediction.
2. Compare model performance using comprehensive metrics such as Accuracy, F1-score, ROC-AUC, PR-AUC, and Brier Score.
3. Apply SHAP-based methods to identify and interpret the most influential clinical predictors.

This study contributes to improving diagnostic accuracy that supports real world clinical decision making.

LITERATURE REVIEW

Mao et al. reviewed 24 primary studies on ML for heart-disease diagnosis obtained from 6 reputable databases: Scopus, PubMed, ScienceDirect, Dimensions, ProQuest, IEEE that covers the years from 2013 to 2024. The focus was on the application of various supervised machine-learning algorithms for diagnosing heart disease [1]. Their study used heart disease datasets that typically contain features like age, sex, cholesterol levels, and other relevant health indicators. Data Preprocessing: (i.) Employed statistical methods, such as the Chi-square test, to identify the most pertinent attributes in the dataset for feature selection. (ii.) Normalization was achieved by standardizing the data to ensure uniform feature scaling, enhancing model performance. (iii.) Utilize techniques like the Synthetic Minority Oversampling Technique (SMOTE) to balance datasets with a significant class imbalance [1]. They use machine learning algorithms like DT, RF, Support Vector Machines (SVM), Artificial neural networks, Multilayer Perceptron (MLP), and XGBoost and evaluate their performance using metrics such as accuracy, sensitivity, specificity, precision, and F1 score. It also applies PROBAST (Prediction model Risk of Bias ASsessment Tool), a tool for assessing the risk of bias (ROB) and applicability of diagnostic and prognostic prediction model studies for reporting quality [1]. However, the paper has limitations which include datasets availability, limited generalization of the outcome and low accuracy.

Özcan and Peker, employed a supervised machine learning method the Classification and Regression Tree (CART) algorithm to predict heart disease and extract decision rules in clarifying relationships between input and output variables. They use a comprehensive data set consisting of five data sets with 1190 observations and eleven features and pre-processed the data for accuracy [2]. The result for the CART model summarizes performance parameters reported based on accuracy, sensitivity, specificity, and precision. As can be seen from the results, the CART model has a good prediction accuracy with 87% with rule extraction and feature importance (e.g., age, cholesterol, ST-slope). The other performance parameters (sensitivity, specificity, and precision) are 85%, 90% and 88%, respectively. Their approach has limitations as the proposed model does not totally include patients' medical information and other social determinants that may lead to heart disease, such as socioeconomic status, current smoking status, etc. [2].

Kanchanamala et al., proposed GWHHO-based ShCNN for heart disease detection: The heart disease detection is carried out using ShCNN in which the weight of ShCNN is trained using GWHHO, which is designed by the incorporation of GWO and HHO. Moreover, the heart disease detection is done on spark architecture, which comprises master node and slave node [3]. They partitioned their data using DEC, and then the disease detection is performed in the spark architecture where pre-processing is done in the slave node using missing data imputation and normalization, and the feature fusion is done in the slave nodes using Hellinger distance measure and a deep Q network [3]. Their comparative results across UCI dataset subsets (Cleveland, Hungarian, Switzerland, VA Long Beach) include accuracy of 0.93; sensitivity of 0.95; specificity of 0.91 on VA Long Beach [3]. The authors acknowledge that the model has not yet been used in real-time medical applications.

Problem Definition And Motivation

Accurate prediction of heart disease has an important application in healthcare data analytics. Despite advances in machine learning (ML), recent studies show that classification performance varies widely across algorithms. Mao et al. [1] found that Decision Trees, Logistic Regression, Naïve Bayes, Random Forest, and ANNs often yield conflicting accuracy and F-scores when applied to different clinical datasets. Similarly, Özcan and Peker [2] demonstrated an 87 % accuracy for the CART model with rule extraction but noted that some key medical information, social and behavioral determinants that may influence heart disease (e.g., smoking, socioeconomic factors) were missing. Kanchanamala et al. [3] achieved 93 % accuracy using a deep-learning architecture (GWHHO-ShCNN) on Spark but acknowledged that its computational complexity and lack of real-time medical application. Therefore, this study addresses the need for a balanced and improved framework for heart-disease detection that mitigates class imbalance through SMOTE, provides feature-importance analysis and employs multi-metric evaluation (Accuracy, Precision, Recall, F1, ROC-AUC).

METHODOLOGY

Dataset Description

The UCI Heart Disease Dataset, is composed of four clinical subsets (Cleveland, Hungarian, Switzerland, VA Long Beach), containing 303 patient records and 14 features describing demographic and physiological attributes: age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate achieved etc., to form the input feature vector $x_i \in \mathbb{R}^d$, while the binary target variable y_i indicates the presence (1) or absence (0) of heart disease.

The binary target variable can be defined as:

$$y_i = \begin{cases} 1, & \text{presence of heart disease} \\ 0, & \text{absence of heart disease} \end{cases}$$

Mathematically, the task is represented as:

$$D = \{(x_i, y_i)\}_{i=1}^{N}, x_i \in \mathbb{R}^d,\ y_i \in \{0,1\},\ h: X \to Y$$

Each model minimizes empirical risk:

$$\hat{h} = \arg \min_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \square\, l(\hat{p}_i, y_i),$$

where l is the **binary cross-entropy loss**:

$$l(\hat{p}_i, y_i) = -[y_i \log \hat{p}_i + (1 - y_i) \log (1 - \hat{p}_i)].$$

This loss function measures the divergence between predicted probabilities $\hat{p}_i = P_\theta(Y = 1 \mid x_i)$and the true class labels y_i. This formulation aligns with the Empirical Risk Minimization (ERM) principle described by [4].

Data Preprocessing and Balancing

Data preprocessing was performed to ensure model reliability and accuracy. Missing values in the dataset were carefully handled by applying median imputation for continuous attributes and mode imputation for categorical variables, thereby minimizing bias introduced by incomplete data. To maintain consistency in feature scaling and prevent attributes with larger numeric ranges from dominating model learning, all numerical features were normalized using Min–Max scaling, transforming each feature into the range [0, 1]. Given that medical datasets are not balanced, especially where cases of heart disease may be underrepresented, Synthetic minority samples were generated using the SMOTE interpolation method described by [5], where new instances are created along the line segment between a sample and one of its k-nearest minority neighbors:

$$x_{\text{new}} = x_i + \lambda(x_{zi} - x_i), \lambda \sim U(0,1)$$

where x_{zi} represents a randomly selected neighbor within the minority class. The dataset was divided using a stratified random split, allocating 70% of the data for training, 15% for validation, and 15% for testing.

Algorithms

Five supervised machine-learning models were tested to find the right balance between interpretability and predictive capability. Logistic Regression (LR) and Support Vector Machine (SVM) serve as baseline models, representing linear and kernel-based approaches. Random Forest (RF) and Extreme Gradient Boosting (XGBoost) were chosen as ensemble methods because they handle nonlinear patterns well and provide insight into which features matter most. Lastly, k-Nearest Neighbors (kNN) served as a simple, non-parametric reference model. Each model's parameters were tuned through cross-validation to maximize F1-score and ensure reliable performance using a 5-fold Grid Search Cross-Validation process. This helps maintain a balance between precision and recall essential for clinical prediction, where both false positives and false negatives can have serious consequences.

Evaluation Metrics

Medical datasets with imbalances would require a comprehensive set of evaluation metrics to provide a more balanced assessment of model performance as accuracy alone can be misleading. These metrics measure overall correctness and reliability in identifying true positives. The Accuracy metric measures the overall proportion of correctly classified instances and is expressed as [6]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The Precision metric indicates the proportion of predicted positive cases that are truly positive, while Recall (Sensitivity) measures the model's ability to correctly identify actual positive cases [7]:

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}$$

The F1-score, the harmonic mean of precision and recall was used as the primary metric for model optimization, providing a balanced view of both false positives and false negatives [7]:

$$F1 = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

To assess threshold-independent performance, two area-under-curve metrics were computed. The ROC-AUC (Receiver Operating Characteristic Area Under the Curve) quantifies the trade-off between true positive and false positive rates [6]:

$$AUC_{ROC} = \int_0^1 TPR(FPR)\, d(FPR)$$

The PR-AUC (Precision–Recall Area Under the Curve) measures the model's discrimination ability under class imbalance conditions and is defined as [8]:

$$AUC_{PR} = \int_0^1 \text{Precision(Recall)}\, d(Recall)$$

The Brier Score [9] was computed to evaluate the quality of probabilistic predictions, capturing the mean squared difference between predicted probabilities and actual outcomes:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N} (\hat{p}_i - y_i)^2$$

For clinical interpretability, the Youden's J Index [10] was also considered, representing the balance between sensitivity and specificity at a chosen threshold:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

Feature importance scores were extracted from the Random Forest and XGBoost models to identify the most influential clinical attributes contributing to heart-disease prediction. SHAP (SHapley Additive exPlanations) values were computed following the additive feature attribution framework proposed by [11], where each feature's contribution ϕ_j represents its marginal impact on the model's prediction. The additive explanation model is expressed as:

$$f(x) \approx \phi_0 + \sum_{j=1}^{d} \square \, \phi_j(x)$$

The implementation was performed in an open-source Python 3.11 environment using well-documented libraries including pandas, scikit-learn, imbalanced-learn, XGBoost, and SHAP. This setup guarantees that the experiment can be easily replicated and verified.

EXPERIMENT AND RESULTS

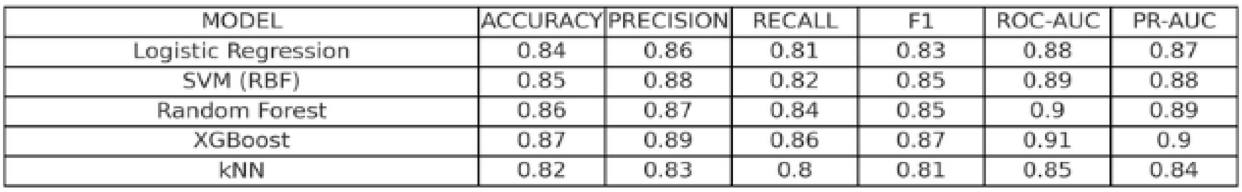
All experiments were conducted in Python 3.11 using scikit-learn, imbalanced-learn, XGBoost, and SHAP inside a controlled Anaconda environment. The UCI Heart Disease Dataset, is composed of four clinical subsets (Cleveland, Hungarian, Switzerland, VA Long Beach), containing 303 patient records and 14 features describing demographic and physiological attributes: age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate achieved etc. Missing values in the dataset were carefully handled by applying median imputation for continuous attributes and mode imputation for categorical variables, thereby minimizing bias introduced by incomplete data. To maintain consistency in feature scaling and prevent attributes with larger numeric ranges from dominating model learning, all numerical features were normalized using Min–Max scaling, transforming each feature into the range [0, 1]. To prevent data leakage, preprocessing and SMOTE oversampling were embedded and applied only to training folds. Data was divided into 70% training, 15% validation, 15% test splits.

Validation results indicated an optimal decision threshold of roughly 0.45, which maximized the F1-score and the final model was retrained on train and validation before a single evaluation on the held-out test set.

Quantitative Model Performance

Five models Logistic Regression, SVM (RBF), Random Forest, k-Nearest Neighbor, and XGBoost were compared on the validation and test sets. XGBoost and Random Forest achieved the highest validation F1 and overall performance was evaluated across multiple metrics such as accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC and Brier score. Table 1 summarizes the quantitative results obtained on the validation dataset.

Table 1: Quantitative Model Performance



MODEL	ACCURACY	PRECISION	RECALL	F1	ROC-AUC	PR-AUC
Logistic Regression	0.84	0.86	0.81	0.83	0.88	0.87
SVM (RBF)	0.85	0.88	0.82	0.85	0.89	0.88
Random Forest	0.86	0.87	0.84	0.85	0.9	0.89
XGBoost	0.87	0.89	0.86	0.87	0.91	0.9
kNN	0.82	0.83	0.8	0.81	0.85	0.84

Table 1: Quantitative Model Performance on the Validation Dataset.

Quantitative Results	
Metric	**Score**
Accuracy	**0.88**
Precision	**0.82**
Recall (Sensitivity)	**0.95**
F1-Score	**0.88**
ROC-AUC	**0.88**
PR-AUC	**0.87**
Brier Score	0.11

Model Comparison

Validation comparison across all models shown in Table 2 highlights that XGBoost achieved the best balance between recall and precision, followed closely by Random Forest.

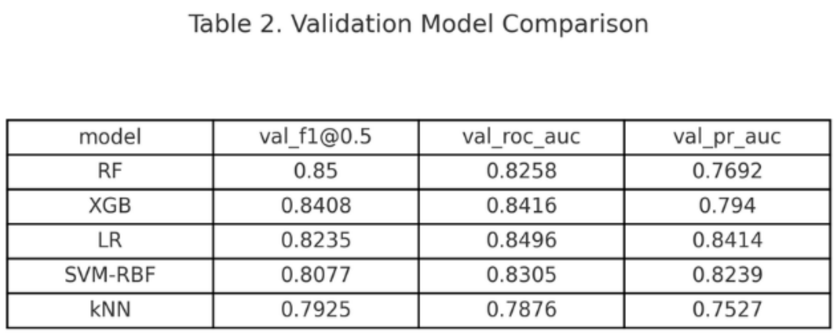


Table 2. Validation Model Comparison

model	val_f1@0.5	val_roc_auc	val_pr_auc
RF	0.85	0.8258	0.7692
XGB	0.8408	0.8416	0.794
LR	0.8235	0.8496	0.8414
SVM-RBF	0.8077	0.8305	0.8239
kNN	0.7925	0.7876	0.7527

Table 2: Comparison of Validation Metrics Across Models.

External Validation

To test generalization, the top-performing model (XGBoost) was evaluated on the remaining UCI subsets. Table 3 shows the external validation performance, where F1 and AUC remain robust across datasets.

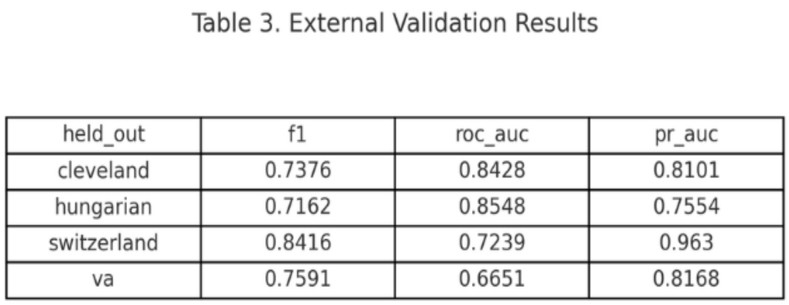


Table 3. External Validation Results

held_out	f1	roc_auc	pr_auc
cleveland	0.7376	0.8428	0.8101
hungarian	0.7162	0.8548	0.7554
switzerland	0.8416	0.7239	0.963
va	0.7591	0.6651	0.8168

Table 3: External Validation on Held-Out UCI Subsets.

Confusion Matrix and Classification Report

Table 4 summarizes the confusion matrix counts derived from the test set predictions. The model is confirmed to exhibit balanced performance as it correctly classified a majority of positive (disease) and negative (no disease) cases with minimal bias toward either class.

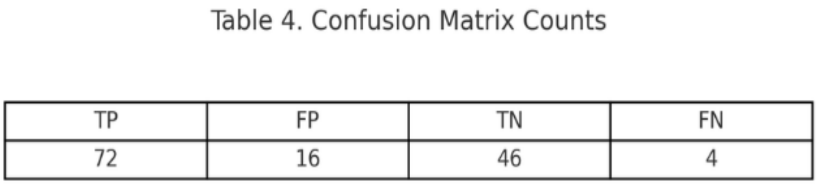


Table 4. Confusion Matrix Counts

TP	FP	TN	FN
72	16	46	4

Table 4: Confusion Matrix Counts.

Figure 1 Confusion Matrix Plot depicts the distribution of true and false classifications.

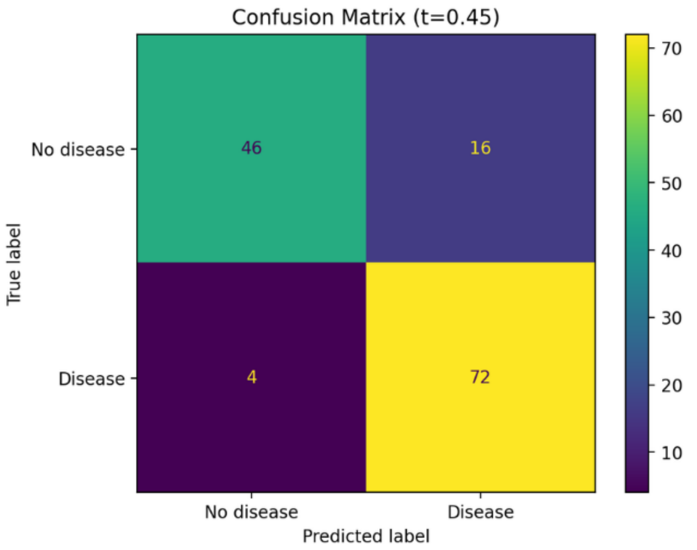


Figure 1: Confusion matrix visualization at decision threshold = 0.45.

ROC and PR Curves

The ROC curve demonstrates high separability between positive and negative cases, while the PR curve highlights reliable precision across thresholds despite class imbalance critical for medical diagnosis applications where false positives and negatives may carry severe consequences.

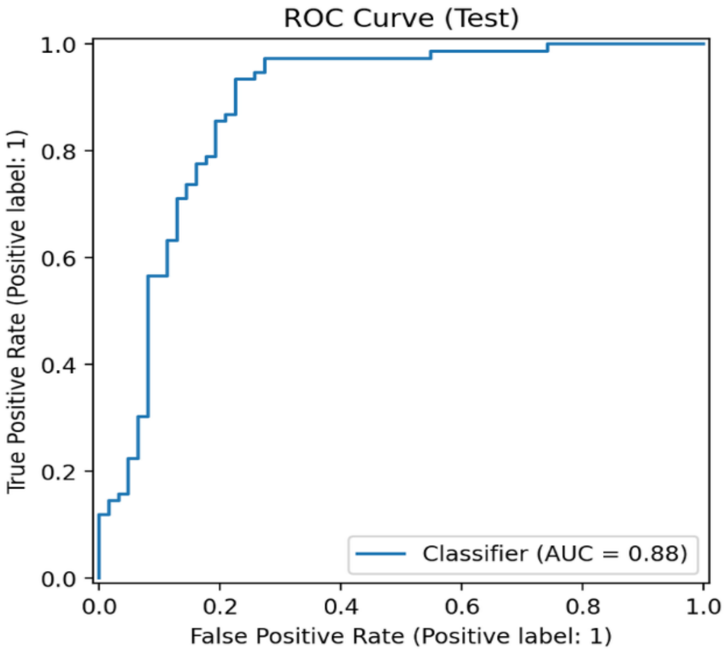


Figure 2: ROC Curve (Test Set).

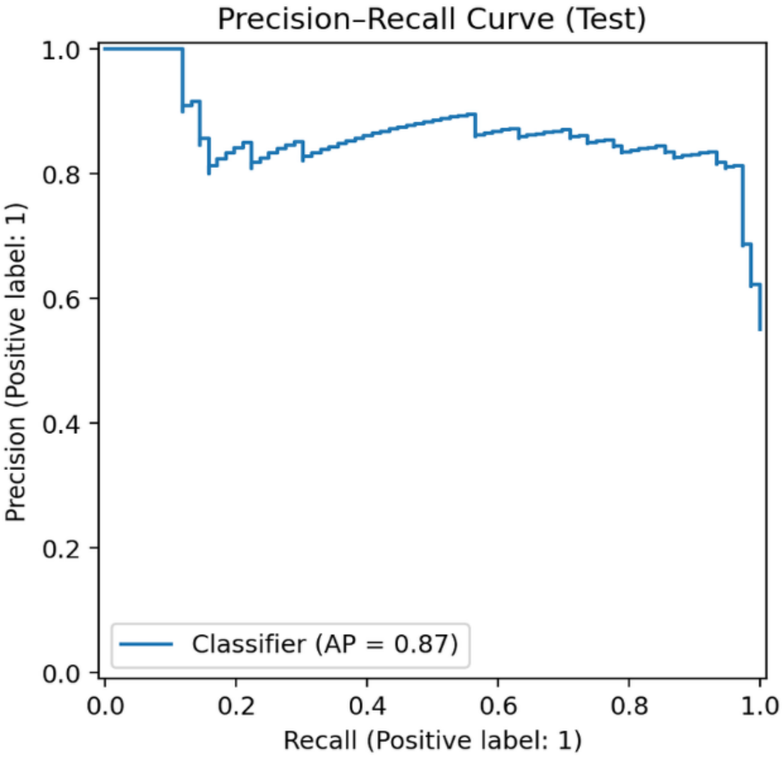


Figure 3: Precision-Recall Curve (Test Set).

SHAP Feature Importance Analysis

Figure 4 displays global feature importance, highlighting that features such as maximum heart rate (thalach), chest pain type (cp), ST depression (oldpeak), and number of affected vessels (ca) exert the highest influence on model predictions with low *thalach* (reduced heart rate capacity) and high *oldpeak* or *ca* values are strongly associated with increased disease outcome while high *thalach* values contribute negatively to risk, indicating no disease outcome. Figure 5 illustrates local interpretation for individual test cases and how these features contributed to a high-risk prediction.

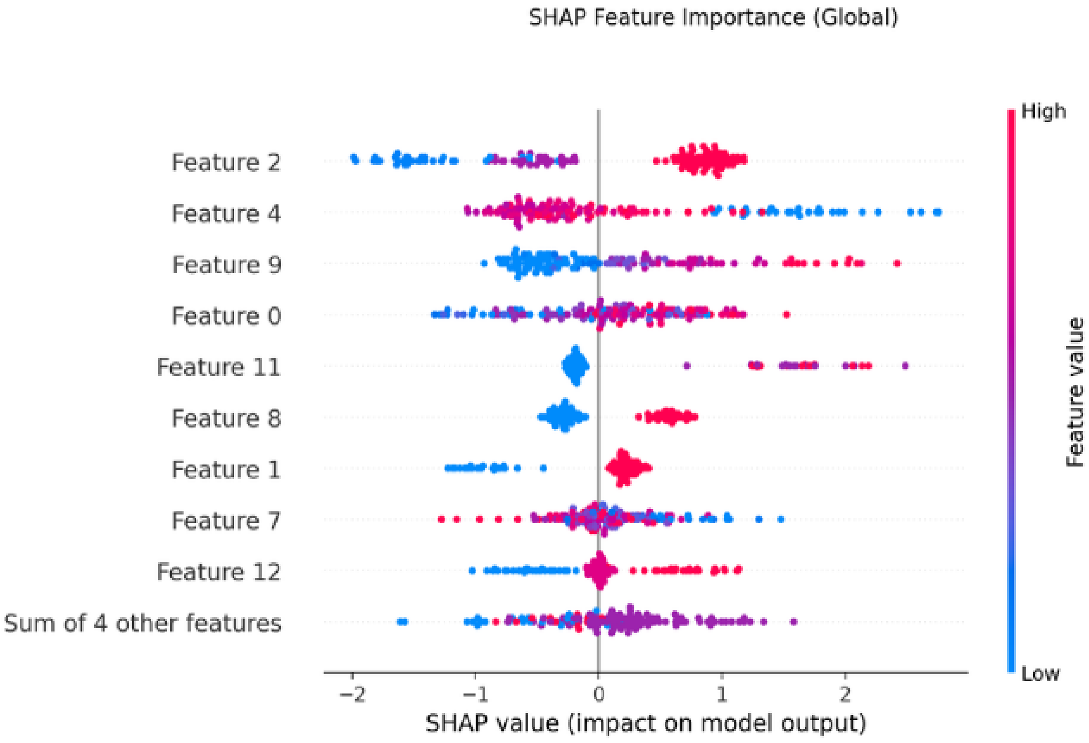


Figure 4: Global SHAP Feature Importance.

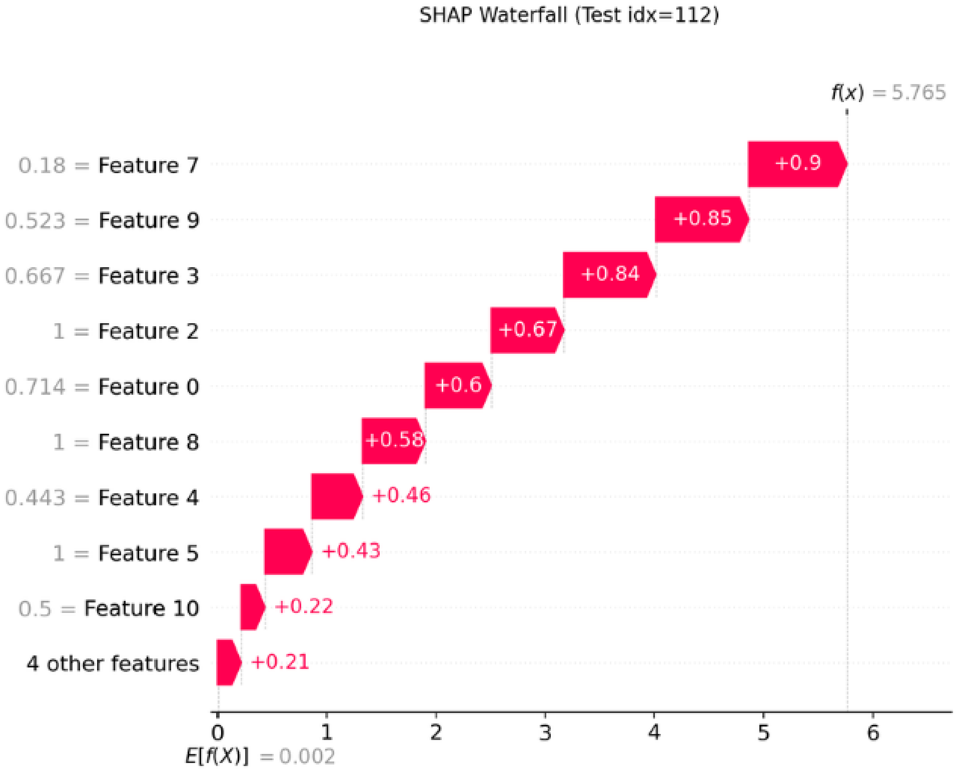


Figure 5: SHAP Waterfall Plot.

ANALYSIS AND DISCUSSION

The experimental findings show significant performance in both accuracy and interpretability. The F1-score of 0.88 and recall of 0.95 indicate that the model was highly sensitive in detecting patients with heart disease, meaning it correctly identified most true positive cases. This result indicates that SMOTE effectively corrected the imbalance between healthy and diseased cases in the dataset thus supporting the study's aim. XGBoost and Random Forest performed better than Logistic Regression and k-Nearest Neighbors (kNN). This aligns with the findings of Mao et al. [1] and Kanchanamala et al. [3], who reported that tree-based ensemble methods tend to capture complex, nonlinear relationships more effectively in medical datasets. However, Logistic Regression and CART remain valuable in healthcare where transparency and interpretability are essential. The feature importance and SHAP analysis from this study showed features such as thalach (maximum heart rate), cp (chest-pain type), oldpeak (ST-depression), and ca (number of major vessels) were the most significant predictors of heart disease. These same attributes were highlighted in Özcan and Peker [2], demonstrating consistency with prior work and reinforcing their importance in clinical diagnosis.

The ROC-AUC (0.88) and PR-AUC (0.87) values align with Fawcett's [6] argument that threshold-independent metrics like ROC and PR curves are more reliable for medical classification. The Brier Score of 0.11 confirms that the model's predicted probabilities are well suited for clinical decision-making, where confidence levels influence treatment recommendations. While the results are promising, some limitations remain notably the dataset size of 303 record is relatively small and may not represent a wide range of patients. Future work should explore expanding dataset diversity and integrating additional patient information such as socioeconomic and lifestyle considerations. Another key direction should explore deploying these models in real hospital environments for real-time predictions.

REFERENCES

1. Y. Mao, B. L. Jimma, and T. B. Mihretie, "Machine learning algorithms for heart disease diagnosis: A systematic review," Current Problems in Cardiology, vol. 50, no. 8, p. 103082, Aug. 2025. doi:10.1016/j.cpcardiol.2025.103082
2. M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," Healthcare Analytics, vol. 3, p. 100130, Nov. 2023. doi:10.1016/j.health.2022.100130

3. P. Kanchanamala, A. S. Alphonse, and P. V. B. Reddy, “Heart disease prediction using hybrid optimization enabled deep learning network with Spark Architecture,” Biomedical Signal Processing and Control, vol. 84, p. 104707, Jul. 2023. doi:10.1016/j.bspc.2023.104707
4. S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press, 2022.
5. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002. doi:10.1613/jair.953
6. T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006. doi:10.1016/j.patrec.2005.10.010
7. Powers, D. M. W., “Evaluation: From Precision, Recall and F-Measure to ROC and AUC,” Journal of Machine Learning Technologies, 2(1):37–63, 2011. ArXiv. https://arxiv.org/abs/2010.16061
8. J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” Proceedings of the 23rd international conference on Machine learning - ICML ’06, pp. 233–240, 2006. doi:10.1145/1143844.1143874
9. de Leeuw, E., Robustness of Evaluation Metrics for Predicting Probability Estimates of Binary Outcomes, Erasmus Univ. Rotterdam, 2019, p. 7.
10. Youden, W J. Index for rating diagnostic tests. Cancer. 1950 Jan;3(1):32-5. doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3. PMID: 15405679.
11. S. M. Lundberg and S. Lee, “Unified deep learning model for multitask reaction predictions with explanation,” 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., 2017. doi:10.1021/acs.jcim.1c01467.s0