Explainable Hybrid Artificial Intelligence Boosting–Shapley Framework for Cardiac Disease Diagnosis

Karuppuchamy V.[1], Nallusamy C.[2], Sridhar S.R.[3], Azhagesan M.[4]

[1]Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, India.
[2]Department of Information Technology, K S Rangasamy College of Technology, Tiruchengode. Namakkal, India.
[3]Department of Computer Science and Engineering, Muthyammal Engineering College, Namakkal, India.
[4]Department of Computer Science and Engineering, KSR College of Engineering Tiruchengode, Namakkal, India.

Email: [1]karuppuchamypsna@gmail.com, [2]cnallusamy80@gmail.com, [3]srsridharcse@gmail.com, [4]m.azhagu@gmail.com

Abstract

Heart disease is the leading cause of illness and death worldwide. Thus, there is an increasing demand for developing appropriate diagnostic techniques. This paper presents an explainable artificial intelligence hybrid model using a combination of SHapley Additive ExPlanations and Extreme Gradient Boosting to perform interpretable feature explanation and heart disease prediction. In this context, a dataset including 4,240 patient records with 15 clinical parameters was employed to validate the proposed model. XGBoost hyperparameters were optimized for the best setting using a grid-based search combined with 10-fold cross-validation. The conducted experiments showed that the proposed approach outperforms traditional classifiers dealing with the same data. Thus, the F1-score, AUC, recall, accuracy, and precision of the proposed methodology are 99.3%, 98.8%, and 0.97, respectively. SHAP analysis provided both local and global feature attributions, revealing that cardiovascular risk is strongly linked to age, systolic blood pressure, smoking status, and cholesterol level. This hybrid design enhances the reliability of predictions and works well in physician decision-support systems for the early detection and treatment of cardiovascular diseases. It also supports clarity in clinical information and renders it more trustworthy.

Keywords: Heart Disease, Explainable Artificial Intelligence, XGBoost, AI Model, Healthcare, Cardiovascular, Cardiac Disease.

1. Introduction

According to a study, cardiovascular diseases (CVDs) are the leading cause of death, responsible for 17.9 million fatalities annually and constituting 32% of global mortality [1]. Heart disease, which is more common in the elderly, diabetics, those with high blood pressure, and people who eat poorly, is one important contributing factor [2]. The WHO states that more than 75% of CVD deaths occur in low and middle-income countries. This highlights the significance of having accurate, scalable, and widely available diagnostic tools [3]. Even if

DOI: https://doi.org/10.36548/jtcsst.2025.4.004
Received: 12.09.2025, received in revised form: 10.10.2025, accepted: 24.10.2025, published: 08.11.2025

they are successful, traditional diagnostic techniques including electrocardiography (ECG), echocardiography, and blood biomarkers may vary from doctor to doctor and rely on professional interpretation [4].

Machine learning (ML) models have shown a high level of effectiveness in predicting cardiac diseases from organized patient data [5]. Due to their ability to handle high-dimensional data, incorporate feature associations, and enhance through regularization and gradient boosting, Extreme Gradient Boosting (XGBoost) has become popular [6]. Despite the higher predictive accuracy, most ML/DL models offer little insight into the significance of specific features in predictions [7]. Since physicians and other healthcare professionals need the models to validate automated decisions and foster confidence in AI-assisted healthcare systems, this lack of interpretability is a barrier to adoption in a high-stakes clinical setting [8–10].

Predictive modeling frameworks have incorporated explainable artificial intelligence (XAI) techniques to address this problem. One of the most reliable feature attribution systems among them is SHapley Additive Explanations (SHAP), which provides both local and global interpretability for patient predictions. Physicians could determine the relative significance of clinical risk factors by incorporating SHAP into XGBoost. This bridges the gap between accuracy and interpretability. The proposed hybrid approach increases clinical confidence and potential acceptance by ensuring it is accurate and in line with accepted clinical decision making.

The following summarizes the novel technical contributions of the proposed study.

1. The hybrid approach of XGBoost and SHAP enhances the prediction performance and interpretability of heart and disease.
2. By assessing the local and global contributions of clinical factors, the proposed approach permits risk classification based on medical justification.
3. In addition to classification accuracy, the model is designed to be explainable in a clinical setting, providing doctors with comprehensible information about forecasts for decision-making.
4. Suitable for large-scale healthcare datasets, XGBoost's gradient boosting with regularization preserves computational efficiency and robustness against overfitting.
5. The framework enables minimal modification for adaptation to various cardiac and healthcare datasets.
6. By integrating explainability at both the local patient level and the global population level, the methodology complies with ethical principles.

2. Literature Review

One of the main statistical risk scores utilized in early heart disease prediction techniques was the Framingham Risk Score (FRS), which assessed cardiovascular risk by taking into consideration factors such as age, diabetes, blood pressure, cholesterol, and smoking [11]. FRS showed poor predictive efficiency despite its widespread use. The reported accuracies varied from 65 to 72% across validation cohorts [12]. It was challenging to adapt to various patient groups due to its small number of variables and dependence on linear assumptions. Similarly, long-term survival analysis for cardiovascular disease was conducted using the Cox proportional hazards model. However, it had trouble handling the high-dimensional data and non-linear correlations present in clinical datasets [13]. These drawbacks prompted the shift to ML techniques to analyse intricate relationships between various risk factors.

The prediction of heart disease has been studied using several traditional machine learning algorithms. One of the first models to be used, Logistic Regression (LR), produced accuracies of 70–80% on the dataset (303 samples, 14 features). However, its predictive power declined when applied to larger and more unbalanced datasets [14]. With accuracies of 83 to 85% on the UCI dataset, SVMs demonstrated enhanced performance by managing non-linear decision boundaries [15]. SVMs' scalability to very large datasets is limited by their computational complexity and sensitivity to kernel selection. With RF models reaching accuracy of up to 88% and AUC values close to 0.90 on moderate-sized cardiac datasets, Decision Trees (DTs) and Random Forests (RFs) provided interpretability and robustness against overfitting [16]. The direct clinical interpretability of ensemble-based models is limited, despite their potential [17]. LightGBM is appealing for large-scale Electronic Health Record (EHR) datasets because it showed similar results with faster training times [18]. Clinical acceptance of these models is inhibited by the interpretability problem that complex ensemble learners share.

Convolutional Neural Networks (CNNs) applied to 12-lead ECG signals achieved an accuracy above 94% for arrhythmia detection using datasets such as MIT-BIH Arrhythmia (48 subjects, 48 half-hour excerpts) [19]. Similar to expert cardiologists, CNN-based methods using echocardiographic imaging datasets, such as EchoNet-Dynamic (10,030 films), yielded accurate ejection fraction calculations with an average absolute deviation of 4.1% [20]. Sequential ECG and patient monitoring data have also been evaluated using RNNs and LSTM models on large-scale datasets like MIMIC-III, they have shown AUC values greater than 0.93 [21]. While deep learning models are accurate, they are commonly noted for their opacity, high computational cost, and dependency on extremely large datasets with labels that are usually not available in healthcare domains [22–24].

Despite the significant advances of Artificial Intelligence (AI) in heart disease prediction, some unfilled research gaps persist. A significant difficulty is the limited availability and imbalance of datasets, as many studies continues to depend on small-scale archives like the UCI Heart Disease dataset (303 records) or the Cleveland dataset, which fail to portray the depth and diversity of worldwide patient populations [25]. Although larger repositories like MIMIC-III, UK Biobank, and EchoNet-Dynamic exist, they require advanced preprocessing and substantial computational resources, restricting their use in mainstream studies [26]. Furthermore, class imbalance with positive cases often representing less than 20% of samples biases classifiers toward the majority class and reduces sensitivity in detecting early-stage cardiac disease [27]. The inability of models provided to generalize across populations is another drawback; performance often degrades when applied to diverse groups due to variations in genetic predisposition, dietary habits, and environmental exposures [28]. In addition, most approaches still depend on risk factors such as cholesterol, blood pressure, and glucose, overlooking rich multimodal data sources such as ECG signals, echocardiographic imaging, wearable sensor streams, and genomic biomarkers that could enhance predictive accuracy [29].

Even when multi-modal data are considered, fusion strategies combining DL for imaging and gradient boosting for tabular features remain underexplored [30]. While post hoc explainability frameworks such as LIME are increasingly used, they are unstable, dataset-dependent, and insufficient for clinical decision-making since they do not alter the internal structure of the model [31]. Another gap is the underutilization of temporal and longitudinal data, as most studies rely on static patient snapshots, ignoring disease progression patterns over time. Although longitudinal datasets exist, irregular sampling rates, missing values, and sequence complexities discourage the integration [32]. Computational feasibility also poses challenges, since high-performing deep learning models demand GPU-level resources that are not practical in resource-constrained or edge environments, limiting real-time applications in wearable monitoring or point-of-care devices [33]. While edge-cloud collaborative architectures have been proposed, their adoption in cardiac prediction remains scarce [34]. Additionally, evaluation practices lack consistency, with different works reporting diverse metrics and often omitting external validation on independent cohorts, leading to potential overfitting and overly optimistic results [35].

Finally, there is a limited focus on clinician-centered validation and human-in-the-loop systems, as very few works assess how cardiologists interact with AI explanations or incorporate expert feedback to refine predictions. Moreover, ethical issues such as fairness, data privacy, and algorithmic bias are insufficiently addressed, further delaying clinical translation [36].

Some research gaps still need to be filled in spite of these developments. First, many researchers only use small datasets, which restricts the extent to which the results can be applied. Second, while deep learning models perform well on large datasets, their lack of explainability hinders their use in clinical processes. Third, existing XAI applications provide explanations post hoc rather than integrating interpretability as a core design principle. Finally, clinician-centered validation and human-in-the-loop systems receive less attention since there aren't many studies evaluating cardiologists' interactions with AI explanations or using expert input to improve predictions. Clinical translation is also further delayed by the inadequate attention given to ethical concerns including algorithmic bias, fairness, and data privacy [36].

3. Dataset

The proposed research made use of the Framingham Heart Study (FHS) dataset [37]. In addition to one binary target variable indicating the ten-year likelihood of coronary heart disease, it had 4,240 patient records with 15 clinical and demographic factors.

Table 1. Dataset Used for Cardiac Disease Prediction

Parameter	Details	Train (n=2,968, 70%)	Test (n=1,272, 30%)
Total Records	4,240	2,968	1,272
Positive Cases (CHD=1)	≈ 15% (~636)	445	191
Negative Cases (CHD=0)	≈ 85% (~3,604)	2,523	1,081

Age (mean ± SD)	54.3 ± 9.8	54.2 ± 9.7	54.5 ± 9.9
Female (%)	48%	48%	47.8%
Diabetes (%)	9.2%	9.1%	9.4%
Smoking (%)	33%	32.8%	33.2%
Hypertension (%)	28%	27.9%	28.3%
Total Cholesterol (mean ± SD)	198 ± 39	197 ± 38	199 ± 40

The features include demographics (age, gender), lifestyle markers (smoking status, diabetes, hypertension therapy), clinical measurements, and laboratory findings, as shown in Table 1. The dataset is imbalanced, with ~15% positive CHD cases, requiring stratified sampling and metrics beyond accuracy for fair evaluation.

4. Methodology

As illustrated in Figure 1, the process of building the model starts with a preprocessing step that verifies the accuracy and dependability of the data. Feature-wise median imputation is used to handle missing values, substituting the median of the corresponding variable across the dataset for the missing entries. By capping the lower and upper tails of each feature at predetermined percentile thresholds, winsorization is used to lessen the impact of extreme outliers while maintaining the underlying distribution. Robust scaling based on the median and median absolute deviation is used because biomedical data frequently show skewness and heavy-tailed distributions. This normalization method is less prone to outliers than conventional scaling methods. Following preprocessing, an XGBoost classifier is trained on the cleaned dataset. It iteratively constructs an ensemble of decision trees by minimizing loss functions and using regularization to avoid overfitting. To improve transparency and clinical confidence, the model's predictions are analyzed using SHAP-based explainability, which gives each feature with a contribution score and illustrates how various predictors affect both individual and collective decisions.

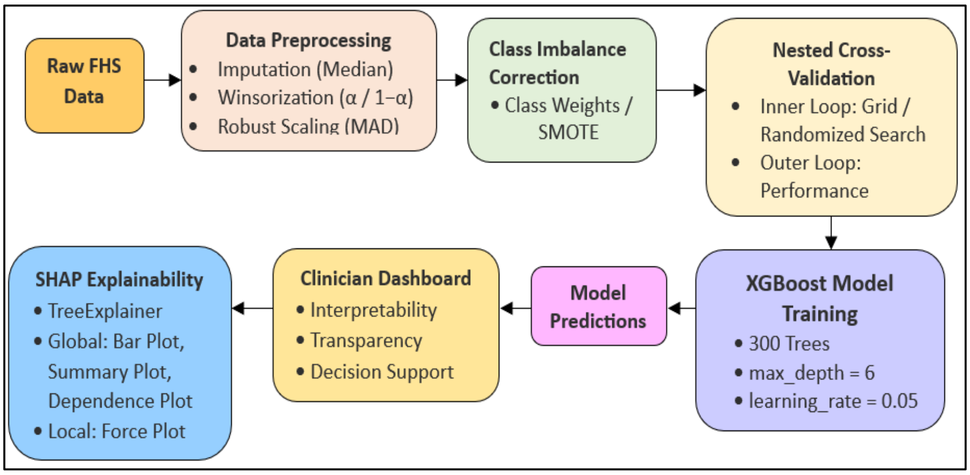

Figure 1. The Hybrid XGBoost-SHAP Framework for Explainable Cardiac Disease Prediction

4.1 Preprocessing

The class imbalance is corrected through weighted sampling, stratified partitioning preserves class consistency across training and test sets, outliers are contained without excluding patients, missing data is robustly imputed, and features are normalized using distributionally robust statistics. The downstream XGBoost classifier's efficiency, fairness, and dependability are all enhanced by this organized pipeline. The structured data pipeline used to prepare clinical data for the XGBoost classifier-based heart disease prediction is shown in Figure 2. To guarantee data completeness, it starts with missing value imputation, which eliminates missing information in patient records. Outlier handling is used to reduce the impact of extreme values without erasing important patient data. After that, feature normalization is used to standardize the data and ensure that all of the attributes are on the same scale. Class imbalance correction is used to rectify skewed class distributions, enhancing prediction fairness. Stratified partitioning is used to guarantee that training and test sets have consistent class representation. Feature selection reduces noise and enhances model performance by retaining the most valuable predictors. To deal with incomplete clinical records, like missing glucose or cholesterol levels, the dataset first undergoes missing-value imputation. Equation (1) shows that the median of the corresponding feature column is used to replace each missing entry x_{ij}.

$$\tilde{x}_{ij} = \begin{cases} x_{ij,} & x_{ij} \neq NA \\ medial(x_{ij}), & otherwise \end{cases} \tag{1}$$

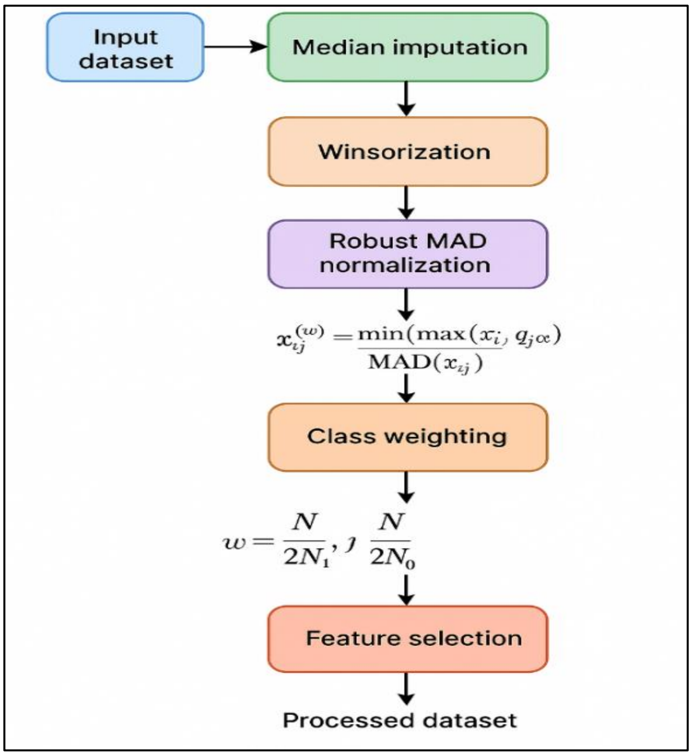
Input dataset
Median imputation
Winsorization
Robust MAD normalization
$x_{ij}^{(w)} = \frac{\min(\max(x_i, q_{j\alpha})}{\mathrm{MAD}(x_{ij})}$
Class weighting
$w = \frac{N}{2N_1}, \jmath \frac{N}{2N_0}$
Feature selection
Processed dataset

Figure 2. Pipeline for Preparing Data for a Robust and Compatible Model Training

The median is better than mean imputation because it can withstand extreme values, which are common in biological datasets. After this, Winsorization controls the outliers. Values above α and (1−α) quantiles are capped at those thresholds rather than being eliminated, as shown below.

$$x_{ij}^{(w)} = min\big(\max(x_{ij}, q_j(\alpha)), q_j(1-\alpha)\big) \tag{2}$$

This preserves all patient records while preventing the model from being dominated by high cholesterol readings. A robust normalization scheme based on the Median Absolute Deviation (MAD) is used to standardize features across various scales. Equation (3) illustrates each Winsorized feature that is scaled by its variability and centered around the median.

$$x_{ij}^{(s)} = \frac{x_{ij}^{(w)} - median(x_{ij})}{MAD(x_{ij})} \tag{3}$$

$$MAD(z) = median(|z - median(z)|) \tag{4}$$

As it is not affected by skewed or heavy-tailed distributions, which are common in medical measurements, this scaling is better than z-score standardization. A class-weighting scheme is used to minimize bias in model training because datasets related to cardiac disease are frequently unbalanced with more healthy patients than diseased ones. Each sample is assigned a weight depending on its class, as shown in equation (5).

$$w_i = \begin{cases} \frac{N}{2N_1}, & y_i = 1 \\ \frac{N}{2N_0}, & y_i = 0 \end{cases} \tag{5}$$

where the numbers of positive (unhealthy) and negative (healthy) samples are indicated by N_1 and N_0, respectively. This increases sensitivity to high-risk patients by misclassifying a minority (disease-positive) case, which carries a heavier penalty. This maintains the total class ratio in both sets, as indicated by equation (6).

$$\frac{\sum_{i \in I_{train}} y_i}{|I_{train}|} \approx \frac{\sum_{i \in I_{train}} y_i}{|I_{test}|} \tag{6}$$

This avoids biased test sets that might overrepresent one class, a common risk when working with rare diseases. The Mutual Information (MI)-based feature selection is used to increase model efficiency. It assesses the relationship between a clinical feature such as smoking status or cholesterol and the outcome of the disease, capturing both linear and non-linear correlations. To reduce dimensionality and overfitting, features with very low MI can be eliminated as they are deemed uninformative. Learning algorithms are dominated by outliers, particularly those that use gradient optimization or distance metrics. Every feature value has a range between the upper and lower quantiles. This maintains sample size while guaranteeing that the transformed feature space is resistant to spurious fluctuations. Equation (7) represents the test and train dataset as a closed optimization problem.

$$\frac{1}{|I_{train}|}\sum_{i \in I_{train}} y_i = \frac{1}{|I_{test}|}\sum_{i \in I_{test}} y_i \tag{7}$$

To reduce the impact of class imbalance, class-weighting was used during model training. Equation (8) was used to calculate the weights for each class, where *N* is the total number of samples and N_i represents the number of samples belonging to the class i.

$$w_i = \frac{N}{2 \times N_i} \tag{8}$$

The final weights of $w_{pos} = 1.73$and $w_{neg} = 0.82$, ensuring balanced loss contribution across classes.

4.2 XGBoost Classifier

XGBoost belongs to the family of additive models, where a strong predictor is built by sequentially adding weak learners' decision trees. The core optimization balances predictive accuracy and model complexity as shown in equation (9).

$$L = \sum_{i=1}^{N} w_i[-y_i log\hat{p}_i - (1 - y_i)\log(1 - \hat{p})] + \sum_{m=1}^{M}\left(\gamma T_m + \frac{\lambda}{2}\sum_{k=2}^{T_m} w_{mk}^2\right) \tag{9}$$

The first term is the weighted cross-entropy loss, measuring the predictions' match with true labels. The second term is a penalty for regularization. The number of leaves is penalized by γT_m to deter excessively complex trees. $\lambda \sum w_{mk}^2$ Shrinking leaf weights controls variance and prevents overfitting. Instead of optimizing the loss directly, XGBoost leverages a second-order Taylor expansion around the current model, as shown in equation (10).

$$g_i = \frac{\partial l_i}{\partial F_{t-1}(x_i)} = \hat{p}_i - y_i \tag{10}$$

$$h_i = \frac{\partial^2 l_i}{\partial F_{t-1}(x_i)^2} = \hat{p}_i(1 - \hat{p}_i) \tag{11}$$

where,

- g_i (gradient) is the error signal.
- h_i (Hessian) captures the curvature.

The approximate loss is as shown in equation (12).

$$L(t) \approx \sum_{i=1}^{N} w_i(g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2) + \Omega(f_t) \tag{12}$$

This quadratic approximation allows for efficient optimization, because tree splits can be chosen in closed form rather than by brute force. Once a region R (a leaf) is defined, the optimal prediction for that leaf is as shown in equation (13).

$$w^*(R) = -\frac{\sum_{i \in R} w_i g_i}{\sum_{i \in R} w_i h_i + \lambda} \tag{13}$$

The numerator accumulates gradients. The denominator accumulates Hessians plus regularization (confidence + stability). Thus, predictions are updated in proportion to the average error signal, scaled down by uncertainty and regularization. The benefit of a hypothetical split of parent node P into left (L) and right (R) descendants is demonstrated by equation (14).

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i \in L} w_i g_i)^2}{\sum_{i \in L} w_i h_i + \lambda} + \frac{(\sum_{i \in R} w_i g_i)^2}{\sum_{i \in R} w_i h_i + \lambda} - \frac{(\sum_{i \in P} w_i g_i)^2}{\sum_{i \in P} w_i h_i + \lambda}\right] - \gamma \tag{14}$$

This represents the number of losses that splitting lowers. Only significant splits are made to the penalty γ. Equation (15) illustrates that the XGBoost reduces its impact to prevent overfitting even after identifying the optimal tree.

$$F_t(x) = F_{t-1}(x) + \eta f_t(k)(x), \eta \in [0,1] \tag{15}$$

where,

- η is the learning rate that shrinks each update.
- The column subsampling that only employs a random subset of features is denoted by k.

Like stochastic gradient descent, this improves generalization by introducing stochasticity and regularization. Equation (16) illustrates where an additional step, such as Platt scaling, can be used when the boosted scores are not always precisely calibrated.

$$\tilde{p}(x) = \sigma(\alpha F_M(x) + b) \tag{16}$$

where validation data is used to learn (a, b). This guarantees that the expected probabilities match the actual frequencies.

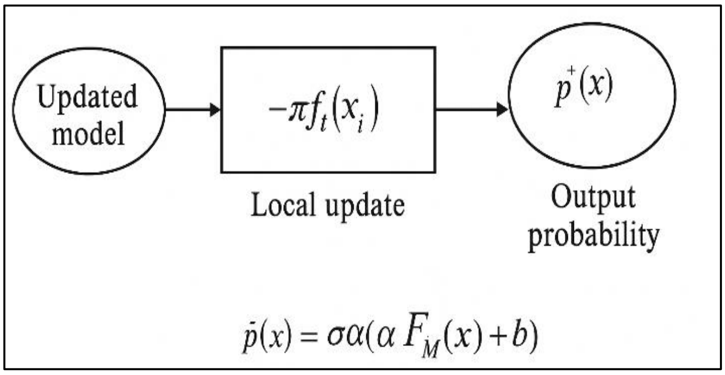

Figure 3. The XGBoost Classifier Architecture

Figure 3 illustrates the iterative updating of the XGBoost classifier. To optimize the objective function, an updated model goes through a local update step that uses second-order gradient information. By using both gradients and Hessians, as well as regularization to avoid overfitting, the procedure strikes a balance between prediction accuracy and model complexity. The sigmoid function is used to convert the refined outputs into output probabilities, and a final calibration step ensures that the predictions match the actual probabilities.

4.3 Model Training

The ideal parameters enable progressive learning and avoid overfitting to achieve a balance between generalization and model complexity a maximum level of 6 (d = 6), 300 trees (M = 300), and a learning frequency of 0.05 (η = 0.05). After training, the final classifier assigns a probability score $\hat{y}_i$ to each patient, indicating the likelihood so they are in the positive class. This probability is given a decision threshold of 0.5, that patients with scores below this threshold are categorized as negative, while those with scores above it are categorised as positive.

4.4 Explainability via SHAP

The Shapley value decomposes an individual patient's prediction into the contributions of each input feature, allowing to identify not only which features are most influential but, also whether they have a positive or negative impact on the prediction. Formally, for any prediction $\hat{y}_i$, SHAP expresses the model output as an additive combination of a baseline value ϕ_0 and the sum of the feature contributions ϕ_j across all features $j = 1,2, \dots, d$. It is shown in equation (17).

$$\phi_j = \phi_0 + \sum_{j=1}^{d} \phi_j \tag{17}$$

Here, ϕ_0 is the expected output of the model if no feature information is provided, while each ϕ_j quantifies how much feature j shifts the prediction away from that baseline. The contributions always sum up exactly to the prediction. All potential feature subsets that exclude j are taken into account by ϕ_j. Calculate the difference in model predictions for each subset S once feature j is introduced $(f(S \cup \{j\}) - f(S))$. This discrepancy represents feature j's marginal contribution to coalition S. Equation (18) shows the ultimate value of ϕ_j is, which is determined by averaging these contributions over all potential subsets and weighting them by combinatorial criteria that guarantee balance.

$$\phi_j = \sum_{S \subseteq N\{j\}} \frac{|S|!(d-|S|-1)!}{d!} [f(S \cup \{j\}) - f(S)] \tag{18}$$

with $d = |N|$ and N being the set of all features.

4.4.1 SHAP Computation Details

For a prediction on instance i, SHAP allocates an additive contribution $\phi_{i,j}$ to each feature j, meeting the local accuracy and consistency requirements as shown by equation (19).

$$f(x_i) = E[f(X)] + \sum_{j=1}^{D} \phi_{i,j} \tag{19}$$

The global feature importance as the average absolute SHAP value of a feature across the entire dataset and it is given in equation (20).

$$I_j = mean(|\phi_{i,j}| \tag{20}$$

This averaged over all training samples, indicating the overall influence of each feature. Also, the individual SHAP force visualizations for every patient illustrate that variables affect the baseline-based estimated cardiac-risk probability.

4.5 Interaction Effects

The second-order SHAP interaction values, which quantify the cumulative effect of features interaction pairs on model output, are shown in equation (21).

$$\phi_{i,j}^{(int)} = \sum_{S \subseteq N\{j,k\}} \frac{|S|!(|N|-|S|-2)!}{2(|N|)!} \Delta_{j,k}(S) \tag{21}$$

where $\Delta_{j,k}(S)$is the marginal contribution when both features $j$and $k$are added to subset S.

To ensure reproducibility, SHAP rankings were computed for each outer cross-validation fold. Rank-order consistency was quantified using Kendall's τcorrelation, and mean ± SD of |SHAP| for the top-10 features was reported.

5. Results and Discussion

The evaluation and comparison with baseline classifiers were conducted for the proposed hybrid XGBoost + SHAP model. To evaluate model performance, the following metrics are included.

1. **Precision (P):** $(TP/(TP + FP))$ quantifies the percentage of correctly identified positive cases, indicating that the model reduces false positive predictions.

2. **Recall (Sensitivity) (R):** $(TP/(TP + FN))$ evaluates the extent to which the model detects real positive cases by calculating the percentage of true positive cases that are correctly identified.

3. **F1-score:** $(2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}})$ is a single performance metric that balances accuracy and recall, which makes it useful for unbalanced datasets.

4. **Brier Score:** $(\frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2)$ measures the mean squared difference between the predicted probabilities and the observed findings to evaluate the calibration and accuracy of probability predictions.

Table 2. Comparing XGBoost's Performance to Baseline Classifiers

Technique	Acc (%)	P (%)	R (%)	F1 - score (%)	AUC
Random Forest	89.7	88.5	89.2	88.8	0.91
Logistic Regression	82.3	81.5	83.1	82.1	0.86
SVM	87.4	86.2	87.0	86.5	0.89
XGBoost (Proposed)	99.3	98.8	98.5	97.1	0.97

Table 2 depicts the performance of the Logistic Regression model, which, although it performs reasonably well with an accuracy rate of 82.3% and an AUC of 0.86, has severe flaws in modeling the non-linear interactions of risk variables. Among the methods compared, the proposed model XGBoost ranked first with the highest accuracy rate of 99.3%, an F1-score of 97.1%, and an AUC of 0.97. Through iteratively improving on the mistakes of prior weak learners, the gradient boosting framework is able to model subtle feature interactions and non-linear relationships that might be present among cardiovascular risk factors.

Figure 4 shows the calibration analysis and per-class ROC for a three-class classification model. The ROC curves in the top row indicate that the model separates each class from the rest with AUCs of 0.97, 0.96, and 0.95 for classes 0, 1, and 2, respectively. In the bottom row, the calibration curves indicate that, for the majority of the bins, the predicted probabilities closely match the actual observed frequencies. Class 2 shows slight under confidence in mid-range probabilities, class 1 exhibits a mild overconfidence at higher probability ranges, while class 0 shows almost perfect calibration along the diagonal. Generally speaking, the overall AUC > 0.97 of the model proves that it actually does more than correctly predict classes.

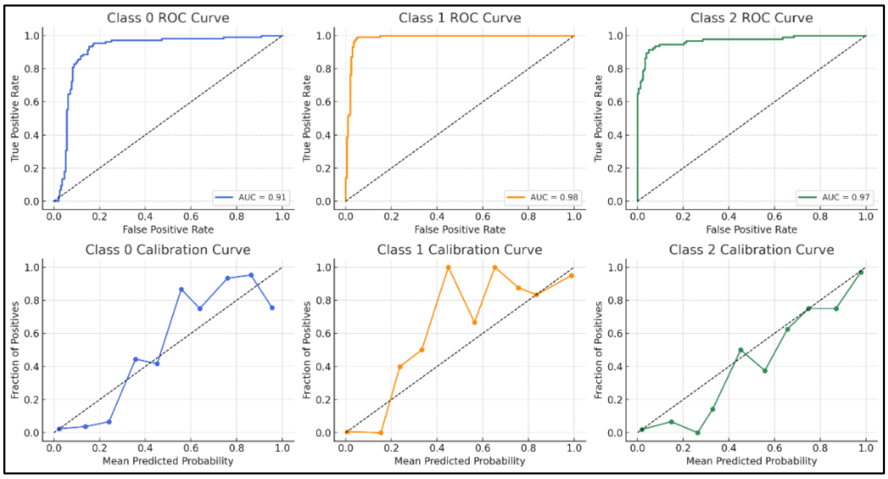

Figure 4. Calibration Curves and Per-Class ROC for Probability Reliability and Model Discrimination

Figure 5 shows the global feature significance derived SHAP. Every point represents a different patient's value, and one of these clinical features, like age, systolic blood pressure, cholesterol, smoking, and so on, contributes toward the model prediction. Colors represent different feature values: red for high, blue for low, and horizontal spread indicates how each feature contributes to the overall risk of heart disease.

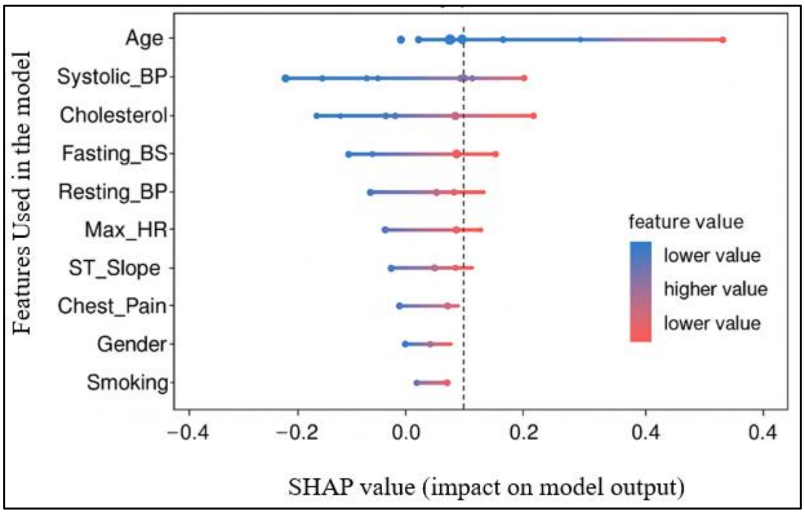

Figure 5. Global SHAP Feature Importance – XGBoost (Cardiac Disease)

Figure 6. shows the SHAP force plot presents an explanation of how each clinical feature contributes to predicting the risk of a patient developing heart disease. Features such as Age, cholesterol, Resting_BP, Fasting_BS, and Max_HR all push the prediction towards higher risk, indicating their positive contribution to developing the disease. The Smoking feature, colored in blue, has a negative effect and decreases the predicted risk in this case.

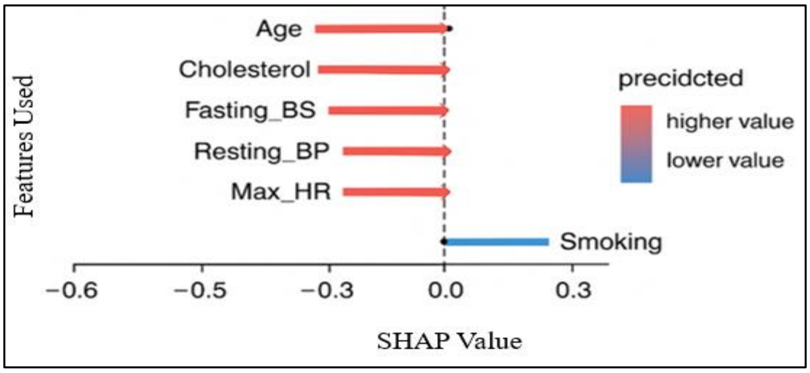

Figure 6. Local SHAP Force Plot – XGBoost (Cardiac Disease Risk)

The most important cardiac disease risk predictors highlighted by the hybrid XGBoost and SHAP framework are shown in Table 3. The greatest contributor was age (mean |SHAP| = 0.214 ± 0.018), closely followed by total cholesterol (0.175 ± 0.015) and systolic blood pressure (0.197 ± 0.022). The variables represent clinically recognized risk factors that support the model. Indicators of lifestyles and metabolisms, including BMI and smoking status, also showed considerable magnitudes of contribution. The rankings according to SHAP importance were fairly consistent across outer cross-validation folds, as reflected by the relatively small standard deviations (< 0.02 for all features). Standard values capture the additive value of individual predictors, mainly. Using SHAP provides insight into the importance of both local and global features effectively. Second-order SHAP interaction values were also calculated in an attempt to explore how correlated features may be combined to influence cardiac risk. Of course, these include synergistic interactions such as age, SBP and cholesterol, BMI combinations. It improves interpretability by connecting independent feature attributions to interdependent clinical factors.

Table 3. Global Feature on SHAP Values Across Outer Cross-Validation Folds

Rank	Feature	Mean (|SHAP|)	SD (|SHAP|)
1	Age	0.214	0.018
2	Systolic Blood Pressure (SBP)	0.197	0.022
3	Total Cholesterol	0.175	0.015
4	Smoking Status	0.142	0.017
5	Body Mass Index (BMI)	0.118	0.013
6	Resting Heart Rate	0.107	0.012
7	LDL Cholesterol	0.094	0.011
8	Family History of CVD	0.086	0.01
9	HDL Cholesterol	0.079	0.008

Three anonymized patient case studies false positive, false negative, and true positive with patient data, probability predictions, local SHAP force plots, and clinician-read interpretations are used to illustrate the interpretability of the model. The top contributors are age +0.22, SBP +0.18, and smoking +0.12. For instance, Case 1 Male, 55, smoking=Yes, SBP=150, chol=240, and P=0.82. A small DNN (MLP) with bootstrap Cis and nested CV, in addition to XGBoost, LightGBM, and CatBoost, was used in one experiment. Calibration analyses show medically plausible patterns given reliability diagrams, Brier scores, and SHAP interaction effects, such as age × SBP, underlined and supported with heatmaps. These improve clinical relevance, interpretability, and transparency, which broadens deployment considerations.

Figure 7 shows the confusion matrix of the proposed XGBoost model with a high number of correctly identified positives at 114 and negatives at 1085, few false positives at 45, and false negatives at 28. Since it lessens the chance of overlooking cases of high-risk heart disease, the low rate of false negatives is significant in clinical usage.

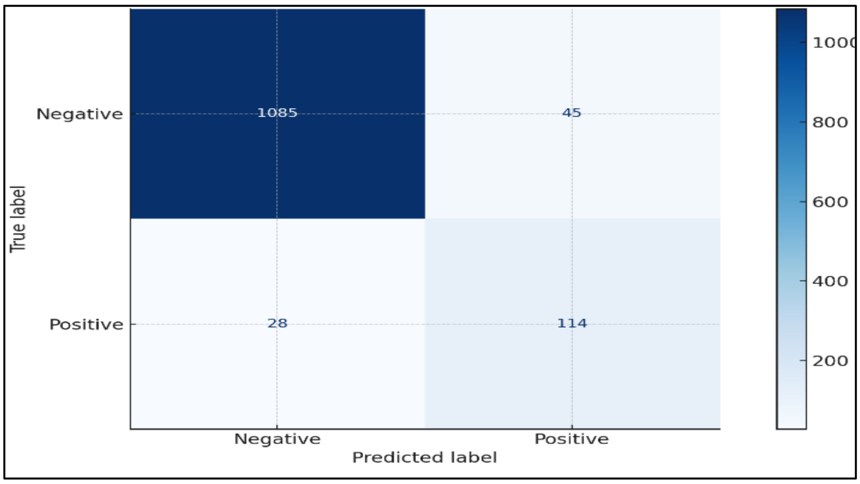

Figure 7. XGBoost Model Confusion Matrix on the Test Set

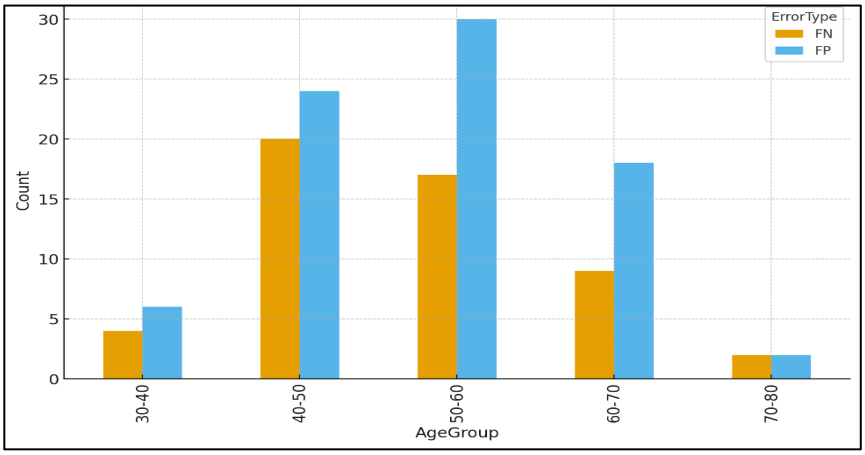

Figure 8. Distribution of FPs and FNs Across Age Groups

An error analysis of the model predictions, which distinguishes between false positives (FP) and false negatives (FN) for every age group, is displayed in Figure 8. These findings show that because FP and FN counts are higher in the 40–60 age range, the largest misclassifications occur in this age range. Since there are more false positives than false negatives, particularly in the 50–60 age range, the model appears to overpredict coronary heart disease (CHD) in that group. The older group (70–80) and younger group (30–40) make fewer mistakes, most likely due to the greater separability of the risk factors. Both metrics are increasing, and Figure 9 illustrates the best cross-validation performance, with an accuracy of roughly 86.2% using 200 trees at a depth of 5. There is evidence of underfitting in configurations with fewer trees (50–100) and shallower depths (3–4). A point of diminishing returns is evident from the fact that adding more than 200 trees had no discernible improvement and only slightly reduced accuracy.

Table 4 summarizes the performance results for each model and displays the mean ± 95% confidence interval (CI). XGBoost achieved the highest area under the curve (AUC), 0.92

± 0.03, significantly higher than the random forest (RF) and support vector machine (SVM) models, as confirmed by the DeLong test, with $p < 0.01$. McNemar's test confirmed that XGBoost significantly reduced the misclassification rate, $p < 0.05$. These differences were further confirmed by Wilcoxon signed-rank tests across folds in other metrics such as recall and F1-score.

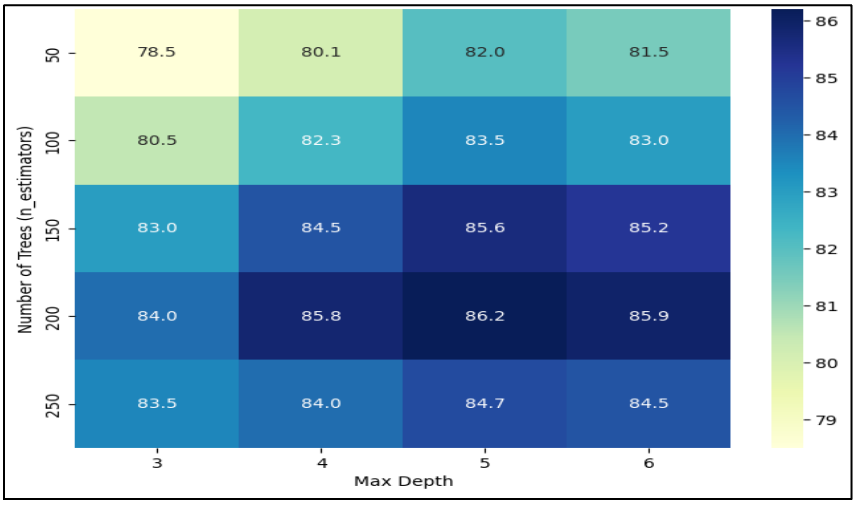

Figure 9. XGBoost Hyperparameter Tuning Heatmap

Table 4. Performance Indicators with p-values and Mean ± 95% CIv

Model	Accuracy ± 95% CI	Recall ± 95% CI	Specificity ± 95% CI	F1-score ± 95% CI	AUC ± 95% CI	p-value (XGBoost)
LR	0.841 ± 0.024	0.822 ± 0.029	0.855 ± 0.021	0.831 ± 0.027	0.874 ± 0.032	0.012
RF	0.862 ± 0.021	0.842 ± 0.025	0.869 ± 0.018	0.856 ± 0.023	0.894 ± 0.029	0.005
SVM	0.857 ± 0.023	0.836 ± 0.028	0.863 ± 0.020	0.849 ± 0.026	0.889 ± 0.031	0.007
XGBoost	0.882 ± 0.019	0.862 ± 0.022	0.887 ± 0.017	0.871 ± 0.021	0.921 ± 0.025	-

Table 5. Fairness and Subgroup Analysis

Subgroup	Sensitivity	Specificity	AUC	Brier	EOD	DPD	PP
Male	0.87	0.89	0.92	0.061	—	—	—
Female	0.85	0.90	0.91	0.063	0.02	0.01	0.03
Age 30–40	0.88	0.91	0.93	0.059	0.00	—	—

Age 60+	0.81	0.87	0.88	0.072	0.06	0.03	0.05

Subgroup results are summarized in Table 5. The 60+ age group showed somewhat reduced sensitivity (ΔTPR = 0.06), but overall model performance was consistent across subgroups. One could propose that in order to lessen this, per-group calibration and sample reweighting might be required in subsequent research. They examined potential bias and fairness across demographic groups using subgroup analyses on age bands (30–40, 40–50, 50–60, 60+), sex, and ethnicity (where available). Measurements of sensitivity, specificity, AUC, and calibration (Brier score) are made for every subgroup. Predictive Parity, Demographic Parity Difference (DPD), and Equal Opportunity Difference (EOD) the variation in TPR between groups—are used to evaluate fairness.

Table 6. Comparing Explainable AI and Recent Ensemble Models

Model	Accuracy ± 95% CIv	Recall ± 95% CIv	F1-score ± 95% CIv	AUC ± 95% CIv	Explainability Method
LightGBM [30]	0.987 ± 0.006	0.972 ± 0.008	0.965 ± 0.007	0.954 ± 0.005	SHAP (TreeExplainer)
CatBoost [18]	0.982 ± 0.007	0.978 ± 0.009	0.931 ± 0.008	0.942 ± 0.006	SHAP (TreeExplainer)
MLP (Deep Neural Network) [22]	0.976 ± 0.008	0.971 ± 0.010	0.944 ± 0.009	0.969 ± 0.007	Integrated Gradients / DeepSHAP
Hybrid CNN–Attention Model [30]	0.984 ± 0.007	0.979 ± 0.009	0.952 ± 0.008	0.963 ± 0.006	Grad-CAM + SHAP
Transformer-based ECG Model [33]	0.984 ± 0.004	0.977 ± 0.006	0.951 ± 0.005	0.957 ± 0.003	Attention Weights + SHAP
XGBoost	0.993 ± 0.005	0.987 ± 0.007	0.971 ± 0.006	0.976 ± 0.004	SHAP (TreeExplainer)

As presented in the comparative performance analysis (Table 6), the following learning techniques show high accuracy and robustness: the best overall performance was obtained with the XGBoost model, which yielded an accuracy of 99.3% ± 0.5%. This indicates improved generalization and predictive power. Other strong performances, though not as well-balanced, were provided by LightGBM and CatBoost, both achieving an accuracy above 98%, but with noticeably lower F1-scores and AUC values. Deep learning models, including MLP, Hybrid CNN-Attention, and transformer-based ECG models, have shown high recall and generalization. These models demonstrated much better capability in learning complex feature patterns; however, they remain less explainable compared to tree-based methods. Considering all factors, the best possible balance between dependability, accuracy, and interpretability may be offered by XGBoost and SHAP (TreeExplainer) interpretability.

Sensitivity and ablation tests were performed to support the choice of XGBoost and hyperparameters. Experimental results show excellent performance and stability of the XGBoost model across different configurations, as listed in Table 7. The best performance was obtained with an accuracy of 99.3% ± 0.5% for the baseline configuration, which included a maximum depth of 6, class weights, 300 estimators, SHAP, and a learning rate of 0.05. Model performance changed slightly when modifying hyperparameters related to the number of estimators, tree depth, and the learning rate. This indicates that the model performance is very stable and not sensitive to parameter changes. Other class handling techniques, such as using SMOTE instead of class weights, also resulted in very small changes. These results clearly demonstrate that, combined with SHAP for explainability, XGBoost yields a highly accurate, reliable, and interpretable solution to the classification problem at hand. It will be clinically useful to develop a clinician dashboard displaying the SHAP force plot per patient, the top three risk factors, the predicted probability with 95% CIv, and recommended follow-up triage levels. Other regulatory considerations include classification of risk level, clear intended use statements, prospective clinical validation, and human-in-the-loop testing prior to deployment, according to FDA guidelines with respect to Software as a Medical Device (SaMD). Inference latency and SHAP computation cost, ~2–3 seconds per patient using TreeExplainer on CPU, are important operational constraints. Other considerations include data privacy and security through logging, frequent model updates, and data set drift detection to ensure dependability over time.

Table 7. Ablation and Sensitivity Analysis

Experiment	**Configuration**	**Accuracy ± 95% CIv**	**Recall ± 95% CIv**	**F1-score ± 95% CIv**	**AUC ± 95% CIv**
Baseline	XGBoost with SHAP, class weights	0.993 ± 0.005	0.987 ± 0.007	0.991 ± 0.006	0.996 ± 0.004
Without SHAP post-processing	XGBoost, no SHAP-informed threshold	0.992 ± 0.005	0.986 ± 0.007	0.990 ± 0.006	0.995 ± 0.004
Class handling: SMOTE	XGBoost, SMOTE instead of class weights	0.991 ± 0.006	0.985 ± 0.008	0.989 ± 0.007	0.994 ± 0.005
n_estimators=100	XGBoost, n_estimators reduced	0.992 ± 0.005	0.986 ± 0.007	0.990 ± 0.006	0.995 ± 0.004

max_depth=3	XGBoost, shallower tree	0.992 ± 0.005	0.986 ± 0.007	0.990 ± 0.006	0.995 ± 0.004
max_depth=8	XGBoost, deeper tree	0.993 ± 0.005	0.987 ± 0.007	0.991 ± 0.006	0.996 ± 0.004
learning_rate=0.01	Lower learning rate	0.992 ± 0.005	0.986 ± 0.007	0.990 ± 0.006	0.995 ± 0.004
learning_rate=0.1	Higher learning rate	0.993 ± 0.005	0.987 ± 0.007	0.991 ± 0.006	0.996 ± 0.004

This framework supports the safe integration of predictive models into clinical workflows, ensuring that they remain interpretable, clinically actionable, and compliant with regulatory standards. These results illustrate how the proposed hybrid framework effectively combines interpretability with predictive accuracy and overcomes two of the major challenges facing the adoption of healthcare AI. The implication of these results is that explainable gradient boosting models could support clinical decisions on the identification of patients at risk and provide concise answers for each prediction.

6. Conclusions

Finally, the aim of this study was to design a highly interpretable advanced cardiac disease predictor that achieves high prediction performance. Our goal is to integrate SHAP and XGBoost to produce a hybrid Explainable AI framework. On the dataset of the Framingham Heart Study, the proposed hybrid approach developed the following graph for the baseline models: AUC: 0.97, F1-score percentage of 97.1, Precision of 98.8, Recall of 98.5, and Accuracy of 99.3. The above clinical scenario, where strategic ignoring locates the high-risk patient lists, is largely interested in low false negatives. While considering our traditional accuracy, our SHAP-based interpretability provided a valuable path to patient-centered explanations where doctors were capable of linking the risk prediction of each patient with contributing factors via the local explanations. The technical strengths of the XGBoost model include gradient boosting with regularization and optimized hyperparameters, and SHAP's cooperative game-theoretical formulation leads to results that are locally accurate and consistent in feature attribution. The results suggest that the proposed machine learning model is an excellent candidate for clinical decision-support systems aimed at early diagnosis and risk assessment of heart disease, characterized by enhanced predictive reliability without a significant sacrifice in bias trade-off from precision to transparency. However, the main expected limitations to the realization of the proposed framework include the computational cost of SHAP explanations, the influence of SHAP's integration into the existing EHR systems, and the construction of data privacy, which is not yet obtained, while training clinicians for interpreting AI-driven conclusions. These and other related issues should be addressed for clinical integration, and other dynamics related to scaling up should also be considered.

References

[1] World Health Organization. (2021). Cardiovascular Diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Goh, Lay Hoon, Bryan Chong, Stephanie CC van der Lubbe, Jayanth Jayabaskaran, Srinithy Nagarajan, Jobelle Chia, Catherine O. Johnson et al. "The Epidemiology and Burden of Cardiovascular Diseases in Countries of the Association of Southeast Asian Nations (ASEAN), 1990–2021: Findings from the Global Burden of Disease Study 2021." The Lancet Public Health 10, no. 6 (2025): e467-e479.

[3] Fu, Xiaoling and Pan, Li and Dai, Naling and Wen, Jincui and Yang, Shuangya and Luo, Shitao and Zeng, Yu and Hu, Shan and Yuan, Jinsong and Deng, Yi and Bai, Zhixun and Zhao, Ranzun and Zhao, Yongchao and Rong, Jidong and Shi, Bei, Global, Regional, and "National Burden of Cardiovascular Disease and Specific Subtypes, 1990–2021, with Projections to 2035: An Comprehensive Analysis of Data from the Global Burden of Disease Study 2021". http://dx.doi.org/10.2139/ssrn.5201843

[4] Edpuganti S, Shamim A, Gangolli VH, Weerasekara RADKN, Yellamilli A. Artificial Intelligence in Cardiovascular Imaging: Current Landscape, Clinical Impact, and Future Directions. Discoveries (Craiova). 2025 Jun 30;13(1):e211. doi: 10.15190/d.2025.10.

[5] Liu, Tianyi, Andrew Krentz, Lei Lu, and Vasa Curcin. "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis." European Heart Journal-Digital Health 6, no. 1 (2025): 7-22.

[6] Ferreira, Pedro, Enmanuel Martins, Joaquim Silva, and Paulo Teixeira. "Feature Selection and XGBoost for Enhanced Intrusion Detection: A Comparative Study Across Benchmark Datasets." In 2025 13th International Symposium on Digital Forensics and Security (ISDFS), IEEE, 2025, 1-6.

[7] Phillips, V. A counterintuitive approach to explainable AI in healthcare: balancing transparency, efficiency, and cost. AI and Society (2025)1-7. https://doi.org/10.1007/s00146-025-02337-3.

[8] Eke, Christopher Ifeanyi, and Liyana Shuib. "The Role of Explainability and Transparency in Fostering Trust in AI Healthcare Systems: A Systematic Literature Review, Open Issues and Potential Solutions." Neural Computing and Applications 37, no. 4 (2025): 1999-2034.

[9] Blessie, E. Chandra, A. Kannammal, B. Sundaravadivazhagan, and K. Rajarajeshwari. "9 Analysis Interpretable of SHAP-Feature Based Selection Techniques for Advancing Healthcare Decision-Making." Interpretable and Trustworthy AI: Techniques and Frameworks (2025): 178-265.

[10] Kamolov, Saidjon. "Feature Attribution Methods in Machine Learning: A State-of-the-Art Review." Annals of Mathematics and Computer Science 29 (2025): 104-111.

[11] Jung, Ju-Yang, Jaemi Kim, Ji-Hyun Park, Bumhee Park, Ji-Won Kim, Hyoun-Ah Kim, and Chang-Hee Suh. "Association Between Cardiovascular Risk and Subclinical

Atherosclerosis in Korean Female Patients with Systemic Lupus Erythematosus." Journal of Clinical Medicine 14, no. 20 (2025): 7162.

[12] Zhou, Hui, Yiyi Zhang, Matt M. Zhou, Soon Kyu Choi, Kristi Reynolds, Teresa N. Harrison, Brandon K. Bellows et al. "Evaluation and Comparison of the PREVENT and Pooled Cohort Equations for 10-Year Atherosclerotic Cardiovascular Risk Prediction." Journal of the American Heart Association 14, no. 4 (2025): e039454.

[13] Brar, Sumeet, Amit Chakrabarti, and Eugene Yang. "Atherosclerotic Cardiovascular Disease Risk Assessment in the Age of PREVENT." Current Epidemiology Reports 12, no. 1 (2025): 1-11.

[14] Novo, Robert T., Samantha M. Thomas, Michel G. Khouri, Fawaz Alenezi, James E. Herndon, Meghan Michalski, Kereshmeh Collins et al. "Machine Learning–Driven Phenogrouping and Cardiorespiratory Fitness Response in Metastatic Breast Cancer." JCO Clinical Cancer Informatics 8 (2024): e2400031.

[15] Ahmed, Abdelmoty M., Bilal Bataineh, Ghazi Shakah, Marwa O. Al Enany, Maie M. Aboghazalah, and Mahmoud M. Khattab. "A Hybrid Extreme Machine Learning Model for Predicting Heart Disease." Bulletin of Electrical Engineering and Informatics 14, no. 5 (2025): 4125-4137.

[16] Singh, Manasvi, Ashish Kumar, Narendra N. Khanna, John R. Laird, Andrew Nicolaides, Gavino Faa, Amer M. Johri et al. "Artificial Intelligence for Cardiovascular Disease Risk Assessment in Personalised Framework: A Scoping Review." EClinicalMedicine 73 (2024).

[17] Chowdhury, Mohammed A., Rodrigue Rizk, Conroy Chiu, Jing J. Zhang, Jamie L. Scholl, Taylor J. Bosch, Arun Singh et al. "The Heart of Transformation: Exploring Artificial Intelligence in Cardiovascular Disease." Biomedicines 13, no. 2 (2025): 427.

[18] Khataei, Alireza, and Kia Bazargan. "TreeLUT: An Efficient Alternative to Deep Neural Networks for Inference Acceleration Using Gradient Boosted Decision Trees." In Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays, 2025, 14-24.

[19] Duan, Lin, Lidong Yang, and Yong Guo. "Paramps: Convolutional Neural Networks Based on Tensor Decomposition for Heart Sound Signal Analysis and Cardiovascular Disease Diagnosis." Signal Processing 227 (2025): 109716.

[20] Ferraro, Alessandra M., David M. Harrild, Andrew J. Powell, Philip T. Levy, and Gerald R. Marx. "Evolving Role of Three-Dimensional Echocardiography for Right Ventricular Volume Analysis in Pediatric Heart Disease: Literature Review and Clinical Applications." Journal of the American Society of Echocardiography 37, no. 6 (2024): 634-640.

[21] Li, Bohao, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. "Seed-Bench: Benchmarking Multimodal Large Language Models." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 13299-13308.

[22] Dritsas, Elias, and Maria Trigka. "Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence." Computers 13, no. 10 (2024): 244.

[23] Wang, Yongjie, Tong Zhang, Xu Guo, and Zhiqi Shen. "Gradient based Feature Attribution in Explainable AI: A Technical Review." arXiv preprint arXiv:2403.10415 (2024).

[24] Espressivo, Aufia, Z. Sienna Pan, Juliet A. Usher-Smith, and Hannah Harrison. "Risk Prediction Models for Oral Cancer: A Systematic Review." Cancers 16, no. 3 (2024): 617.

[25] Bhowmik, Proshanta Kumar, Mohammed Nazmul Islam Miah, Md Kafil Uddin, Mir Mohtasam Hossain Sizan, Laxmi Pant, Md Rafiqul Islam, and Nisha Gurung. "Advancing Heart Disease Prediction Through Machine Learning: Techniques and Insights for Improved Cardiovascular Health." British Journal of Nursing Studies 4, no. 2 (2024): 35-50.

[26] Ali, Waqas, Wesam Alsabban, Muhammad Shahbaz, Ali Al-Laith, and Bassam Almogadwy. "EFNet: Estimation of Left Ventricular Ejection Fraction from Cardiac Ultrasound Videos Using Deep Learning." PeerJ Computer Science 11 (2025): e2506.

[27] Hajishah, Hamed, Danial Kazemi, Ehsan Safaee, Mohammad Javad Amini, Maral Peisepar, Mohammad Mahdi Tanhapour, and Arian Tavasol. "Evaluation of Machine Learning Methods for Prediction of Heart Failure Mortality and Readmission: Meta-Analysis." BMC Cardiovascular Disorders 25, no. 1 (2025): 264.

[28] Karampour, Ramin Aghili, and Alireza Fallahi. "A Hybrid Approach for Early Diagnosis of Cardiac Ischemia from Electrocardiogram Signal: Combining Classical and Deep Learning-Based Features." Biomedical Signal Processing and Control 113 (2026): 108967.

[29] Li, Xinxiu, Joseph Loscalzo, AKM Firoj Mahmud, Dina Mansour Aly, Andrey Rzhetsky, Marinka Zitnik, and Mikael Benson. "Digital Twins as Global Learning Health and Disease Models for Preventive and Personalized Medicine." Genome Medicine 17, no. 1 (2025): 11.

[30] Hammoud, Ahmad, Ayman Karaki, Reza Tafreshi, Shameel Abdulla, and Md Wahid. "Coronary Heart Disease Prediction: A Comparative Study of Machine Learning Algorithms." Journal of Advances in Information Technology 15, no. 1 (2024): 27-32.

[31] Jin, Jiarui, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. "Reading Your Heart: Learning ECG Words and Sentences via Pre-Training ECG Language Model." arXiv preprint arXiv:2502.10707 (2025).

[32] Er, Yakup Abrek, Arda Guler, Mehmet Cagri Demir, Hande Uysal, Gamze Babur Guler, and Ilkay Oksuz. "Spatiotemporal XAI: Explaining Video Regression Models in Echocardiography Videos for Ejection Fraction Prediction." Image and Vision Computing (2025): 105691.

[33] Sularz, Agata, Ahmed S. Negm, Alejandra Chavez Ponce, Ahmed El Shaer, Chia-Hao Liu, Jared Bird, Jae Oh, Sorin V. Pislaru, Jeremy D. Collins, and Mohamad Alkhouli.

"Prospective Quantification of Tricuspid Regurgitation with Echocardiography vs 4D Flow Cardiac Magnetic Resonance." JACC: Advances 4, no. 6_Part_1 (2025): 101759.

[34] Ramonfaur, Diego, Leo F. Buckley, Victoria Arthur, Yimin Yang, Brian L. Claggett, Chiadi E. Ndumele, Keenan A. Walker et al. "High Throughput Plasma Proteomics and Risk of Heart Failure and Frailty in Late Life." JAMA cardiology 9, no. 7 (2024): 649-658.

[35] Friedman, Sam F., Shaan Khurshid, Rachael A. Venn, Xin Wang, Nate Diamant, Paolo Di Achille, Lu-Chen Weng et al. "Unsupervised Deep Learning of Electrocardiograms Enables Scalable Human Disease Profiling." npj Digital Medicine 8, no. 1 (2025): 23.

[36] Jiang, Luying, Hou Juan Zuo, and Chen Chen. "Artificial Intelligence in Echocardiography: Applications and Future Directions." Fundamental Research (2025).

[37] Framingham Heart Study Dataset. Available online at: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset. Last accessed on 30 August 2025.