CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 937-955
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Coronary Heart Disease Risk Prediction under Class Imbalance Using XGBoost with SHAP-Based Interpretation

Siti Amiroch*[1], Fitri Nur Laili[1], Awawin Mustana Rohmah[1], and Dicka Yale Kardono[2]

[1]*Mathematics Department, Faculty of Science and Technology, Universitas Islam Darul 'Ulum, Lamongan, Indonesia*
[2]*Informatics Department, Faculty of Science and Technology, Universitas Islam Darul 'Ulum, Lamongan, Indonesia*

Abstract

Coronary heart disease (CHD) risk prediction is challenging because clinical data are heterogeneous and the response variable is imbalanced. This study develops an interpretable predictive framework for CHD risk using Extreme Gradient Boosting (XGBoost), median imputation, IQR-based winsorization, standardization, the Synthetic Minority Over-sampling Technique (SMOTE), bootstrap-based uncertainty assessment, and Shapley Additive Explanations (SHAP). The learning problem is formulated within a regularized empirical risk minimization framework, so the model is viewed as a statistical estimator rather than merely an algorithmic classifier. To avoid information leakage, train–test splitting is performed before any resampling, and SMOTE is applied only to the training data. The primary analysis is fixed *a priori* at an 80:20 stratified split, whereas 60:40 and 70:30 splits are treated as sensitivity analyses rather than model-selection devices. In the primary analysis, the model attains accuracy of 79.36%, precision of 27.88%, recall of 22.48%, F1-score of 24.89%, and ROC–AUC of 0.6502. The 95% bootstrap confidence interval for ROC–AUC is [0.6017, 0.6981]. SHAP analysis in probability space identifies `age`, `cigsPerDay`, `male`, `heartRate`, and `sysBP` as the most influential predictors. These results show that the proposed framework is mathematically well-structured and interpretable, but that its out-of-sample discrimination on this dataset is moderate rather than high.

Keywords: Coronary heart disease; Class imbalance; Extreme Gradient Boosting; SHAP; Bootstrap confidence interval

1. Introduction

Coronary heart disease (CHD) remains one of the leading contributors to global morbidity and mortality. Recent global reports indicate that cardiovascular diseases account for a substantial proportion of deaths worldwide and continue to impose a major public-health burden [1–4]. In clinical prevention, CHD risk is associated with multiple demographic, behavioral, and physiological factors, including age, smoking, blood pressure, diabetes, lipid profile, and body composition [5, 6]. This makes early risk prediction an important problem not only in medical decision support but also in predictive statistical modeling.

Traditional clinical risk scoring systems, including those derived from the Framingham Heart Study, provide an important foundation for cardiovascular risk assessment [6]. However, the increasing availability of multivariable health data has encouraged the use of machine-learning methods that can capture nonlinear relationships, variable interactions, and heterogeneous data structures more flexibly than many conventional statistical scoring rules. Among these methods,

*Corresponding author. E-mail: siti.amiroch@unisda.ac.id

Submitted: December 02, 2025 Reviewed: February 02, 2026 Accepted: April 11, 2026
DOI: https://doi.org/10.18860/cauchy.v11i1.37938

gradient boosting and, more specifically, Extreme Gradient Boosting (XGBoost), have attracted considerable attention because of their strong predictive flexibility, computational efficiency, and regularized tree-ensemble formulation [7, 8].

More broadly, boosting-based and machine-learning methods have been applied successfully in a range of biomedical and computational classification problems. Faroby et al. used extreme gradient boosting for the classification of IGF1R ligand compounds in the identification of herbal extracts [9]. Amiroch et al. employed machine-learning methods for predicting antiviral compounds targeting avian influenza viral proteins, showing the relevance of data-driven predictive modeling in molecular and biomedical contexts [10]. In addition, Muhammad et al. applied a boosting algorithm to classify COVID-19 variants [11], while Irawan and Jamhuri discussed broader machine-learning developments and research trends in a survey of past, present, and future directions of the field [12]. These studies reinforce the growing role of machine learning, including ensemble methods, in complex scientific prediction tasks.

Several previous studies have reported promising results for heart-disease prediction using XGBoost-based models [13, 14]. Nevertheless, many such studies remain primarily performance-oriented and focus mainly on reporting classification metrics. In imbalanced medical classification, such reporting can be misleading if the experimental design does not explicitly control for information leakage or if the test set is implicitly used to select the final model configuration. For journals with a strong mathematical orientation, it is therefore important that the predictive procedure be described not only as an implementation pipeline but also as a well-defined statistical learning framework with clear evaluation logic.

A central methodological issue in CHD prediction is class imbalance. In many clinical datasets, positive or high-risk cases form a minority relative to the non-risk class. This imbalance may bias a fitted classifier toward majority-class prediction, so that nominal accuracy appears acceptable even when sensitivity to the minority class is inadequate [15, 16]. For this reason, evaluation in imbalanced settings should not rely on a single performance measure. Accuracy, precision, recall, F1-score, and ROC–AUC provide complementary information about classifier behavior, especially when minority-class detection is clinically important [17–19].

Another challenge is interpretability. Although tree-ensemble methods often yield good predictive performance, they are frequently criticized as black-box models, which may limit trust and adoption in clinical decision support [20, 21]. To address this issue, explainable artificial intelligence (XAI) methods can be integrated into the predictive pipeline. Among them, SHAP provides a mathematically grounded attribution framework derived from cooperative game theory, allowing feature contributions to be examined at both the global and local levels [22, 23].

Based on these considerations, this study develops a CHD risk prediction framework using XGBoost, class-imbalance handling through SMOTE, bootstrap confidence intervals, and SHAP-based interpretation. The novelty claimed here is not the invention of a new classification algorithm, but the integration of regularized learning, leakage-aware experimental design, uncertainty quantification, and interpretable feature attribution into a single predictive modeling framework. The main objectives are: (i) to construct a statistically transparent CHD risk prediction model under class imbalance, and (ii) to identify the most influential predictors through a mathematically interpretable SHAP-based explanation.

2. Methods

This study develops a predictive framework for CHD risk modeling under class imbalance using XGBoost, SMOTE, bootstrap-based uncertainty assessment, and SHAP-based interpretation. The methodology consists of dataset specification, mathematical problem formulation, preprocessing and class rebalancing, model construction and hyperparameter optimization, performance evaluation, and interpretable explanation.

To present the methodology systematically, this section begins with the dataset and study variables, then introduces the mathematical formulation of the prediction problem, followed by

preprocessing, imbalance handling, model fitting, evaluation, and interpretability analysis.

2.1. Dataset and Study Variables

The study uses a public CHD dataset derived from the *Framingham Heart Study*, a well-known longitudinal cohort that has played an important role in cardiovascular risk modeling [6]. The working dataset contains $n = 4240$ observations and 16 variables, comprising 15 predictors and one binary response variable.[1] The target variable is `TenYearCHD`, defined in Eq. (1), where

$$Y = \begin{cases} 1, & \text{if the individual is classified as having 10-year CHD risk,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The predictors include demographic, behavioral, and clinical covariates: `male`, `age`, `education`, `currentSmoker`, `cigsPerDay`, `BPMeds`, `prevalentStroke`, `prevalentHyp`, `diabetes`, `totChol`, `sysBP`, `diaBP`, `BMI`, `heartRate`, and `glucose`. Table 1 summarizes the variable definitions.

Table 1: Attributes in the Framingham dataset.

No.	Attribute	Description
1	male	0 = Female, 1 = Male
2	age	Age (years)
3	education	Ordinal education category
4	currentSmoker	0 = Not a current smoker, 1 = Current smoker
5	cigsPerDay	Cigarettes smoked per day
6	BPMeds	0 = No blood pressure medication, 1 = Blood pressure medication
7	prevalentStroke	0 = No prior stroke, 1 = Prior stroke
8	prevalentHyp	0 = No hypertension, 1 = Hypertension
9	diabetes	0 = No diabetes, 1 = Diabetes
10	totChol	Total cholesterol (mg/dL)
11	sysBP	Systolic blood pressure (mmHg)
12	diaBP	Diastolic blood pressure (mmHg)
13	BMI	Body mass index (kg/m^2)
14	heartRate	Heart rate (beats/min)
15	glucose	Glucose (mg/dL)
16	TenYearCHD	Target: 0 = No 10-year CHD risk, 1 = 10-year CHD risk

2.2. Problem Formulation and Prediction Rule

Let the dataset be written as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, \tag{2}$$

where $x_i \in \mathbb{R}^p$ denotes the predictor vector and $y_i \in \{0, 1\}$ the corresponding binary CHD-risk outcome. Based on Eq. (2), the aim is to estimate the conditional probability

$$p(x) = \mathbb{P}(Y = 1 \mid X = x) \tag{3}$$

and to induce a binary decision rule based on that probability.

In this study, the predictive model is formulated within a regularized empirical risk minimization framework. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a score function and let $\sigma(z) = 1/(1 + e^{-z})$ denote the logistic link. The probability estimator is defined by

$$\hat{p}(x) = \sigma(f(x)). \tag{4}$$

Using Eq. (4), the learning problem is expressed as

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \sigma(f(x_i))) + \lambda \mathcal{R}(f) \right\}, \tag{5}$$

[1] https://www.kaggle.com/datasets/noeyislearning/framingham-heart-study/data

where $\ell(\cdot,\cdot)$ is the binary classification loss, $\mathcal{R}(f)$ is a regularization term, and $\lambda > 0$ is the complexity penalty. Eq. (5) is consistent with the gradient boosting perspective of additive function approximation and with the regularized tree-ensemble structure used by XGBoost [7, 8].

A binary classification rule is then defined by

$$\hat{y}(x) = \mathbf{1}\{\hat{p}(x) \geq \tau\}, \tag{6}$$

where $\tau \in (0,1)$ is a decision threshold. Unless otherwise stated, $\tau = 0.5$ is used in Eq. (6) for confusion-matrix-based metrics, while threshold-free discrimination is assessed through ROC–AUC.

2.3. Experimental Protocol and Data Preprocessing

All experiments were conducted in Python using Google Colaboratory. To avoid information leakage, the dataset was first partitioned into training and testing sets using stratified sampling. The primary analysis was fixed *a priori* at an 80:20 split, while 60:40 and 70:30 were used only as sensitivity analyses. All preprocessing operations that require parameter estimation were fitted using the training set only and then applied to the testing set. Random operations, including data splitting, resampling, and model fitting, were controlled by a fixed random seed.

Missing-value imputation. Missing values in numerical and ordinal variables, including `education`, `cigsPerDay`, `BPMeds`, `totChol`, `BMI`, `heartRate`, and `glucose`, were imputed using feature-wise medians computed on the training set. Binary variables were imputed by the most frequent category.

Outlier handling. Outliers were assessed using the interquartile range (IQR) criterion. For a variable x with quartiles Q_1 and Q_3, and $\mathrm{IQR} = Q_3 - Q_1$, values outside

$$[Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR}] \tag{7}$$

were flagged. The bounds in Eq. (7) were used to reduce the influence of extreme observations through winsorization based on training-derived cutoffs.

Feature scaling and encoding. Continuous variables were standardized using the z-score transformation

$$z = \frac{x - \mu}{\sigma}, \tag{8}$$

where μ and σ denote the training-set mean and standard deviation, respectively. Binary variables were retained in coded form, and the ordinal variable `education` was represented numerically according to its category order.

Correlation analysis. Pearson correlation coefficients were used descriptively to summarize marginal linear associations among predictors and between predictors and the target variable:

$$r_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{9}$$

Eq. (9) was used for interpretation rather than for automatic feature elimination.

2.4. Class Imbalance Handling via SMOTE

The target variable is imbalanced, with the positive class substantially smaller than the negative class. Since imbalance may bias a classifier toward majority-class prediction, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set only [15, 16]. Given

a minority observation x_i and one of its nearest minority neighbors x_{nn}, SMOTE generates a synthetic sample as

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i), \qquad \lambda \in (0, 1). \tag{10}$$

Thus, the generated observation in Eq. (10) lies on the line segment joining two minority-class observations and enriches the minority class without simply duplicating samples [15]. This interpolation is expected to improve the learned decision boundary in regions where minority-class observations are sparse.

Proposition 1 (Geometric validity of SMOTE). *Let x_i and x_{nn} be two minority-class observations. Then the SMOTE-generated point*

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i), \qquad \lambda \in (0, 1), \tag{11}$$

belongs to the line segment joining x_i and x_{nn}, and hence lies in the convex hull of minority-class observations.

Proof. Rewriting Eq. (11) gives

$$x_{\text{new}} = (1 - \lambda)x_i + \lambda x_{nn}. \tag{12}$$

Since $0 < \lambda < 1$, the coefficients $(1 - \lambda)$ and λ in Eq. (12) are nonnegative and sum to one. Therefore, x_{new} is a convex combination of x_i and x_{nn}, so it lies on the line segment between them and, in particular, in the convex hull of the minority sample set. □

2.5. XGBoost Model and Hyperparameter Optimization

Extreme Gradient Boosting (XGBoost) is a regularized tree-ensemble method that constructs an additive model by sequentially fitting decision trees to reduce predictive error [7, 8]. The fitted model is written as

$$f(x) = \sum_{t=1}^{T} f_t(x), \tag{13}$$

where each f_t denotes a regression tree. Its optimization objective is

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{p}(x_i)) + \sum_{t=1}^{T} \Omega(f_t), \tag{14}$$

where $\ell(\cdot)$ is the binary classification loss and $\Omega(f_t)$ is a regularization term controlling tree complexity [8]. Eq. (13) and Eq. (14) describe the additive and regularized structure used to stabilize the model and reduce overfitting in heterogeneous clinical data.

Hyperparameter optimization was performed using randomized search with five-fold stratified cross-validation on the training set. The search covered the parameters `n_estimators`, `learning_rate`, `max_depth`, `subsample`, `colsample_bytree`, `gamma`, `reg_lambda`, and `reg_alpha`. After selecting the best configuration according to cross-validated ROC–AUC on the training set, the final model for each split was refitted and evaluated on the corresponding held-out test set.

2.6. Evaluation Metrics

Model performance was evaluated using confusion-matrix quantities: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy, precision, recall, and

F1-score were computed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{17}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{18}$$

Eqs. (15)–(18) provide complementary information in imbalanced classification, where accuracy alone may be misleading [19]. To evaluate threshold-free discrimination, the area under the receiver operating characteristic curve (ROC–AUC) was also computed [17, 18].

Proposition 2 (Interpretation of ROC–AUC)**.** *Let $\hat{p}(X^+)$ and $\hat{p}(X^-)$ denote the scores assigned to randomly selected positive and negative instances, respectively. If ties occur with probability zero, then*

$$\text{AUC} = \mathbb{P}\big(\hat{p}(X^+) > \hat{p}(X^-)\big). \tag{19}$$

Proof. The ROC–AUC is equivalent to the Mann–Whitney ranking probability, namely the probability that a randomly chosen positive observation receives a higher score than a randomly chosen negative observation. This follows from the equivalence between AUC and the normalized rank-sum statistic, which yields Eq. (19). □

2.7. Bootstrap Confidence Intervals for Discrimination Metrics

To quantify statistical uncertainty in the reported performance, nonparametric bootstrap confidence intervals were constructed on the held-out test set [24]. Let

$$\mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{test}}} \tag{20}$$

denote the test sample, and let $M(\mathcal{D}_{\text{test}})$ be a scalar performance functional such as ROC–AUC, accuracy, precision, recall, or F_1. For each bootstrap replication $b = 1, \ldots, B$, a resampled test set

$$\mathcal{D}_{\text{test}}^{*(b)} \tag{21}$$

was generated by sampling with replacement from $\mathcal{D}_{\text{test}}$, preserving the original sample size. The bootstrap statistic is

$$M^{*(b)} = M\left(\mathcal{D}_{\text{test}}^{*(b)}\right). \tag{22}$$

A two-sided percentile confidence interval at level $(1 - \alpha)$ is then

$$\text{CI}_{1-\alpha}(M) = \left[q_{\alpha/2},\, q_{1-\alpha/2}\right], \tag{23}$$

where q_u denotes the empirical u-quantile of $\{M^{*(b)}\}_{b=1}^{B}$. In this study, $B = 1000$ bootstrap replications and 95% confidence intervals were used. Eqs. (20)–(23) define the uncertainty-quantification procedure used in the primary analysis.

Proposition 3 (Bootstrap consistency for smooth performance functionals)**.** *Let $\hat{M}$ be a statistic computed from the test sample, where M is a sufficiently smooth functional of the empirical distribution. Then the nonparametric bootstrap distribution of $\hat{M}^*$ consistently approximates the sampling distribution of $\hat{M}$ as the test sample size tends to infinity.*

Proof. Let $\hat{F}_n$ denote the empirical distribution of the observed test sample. The nonparametric bootstrap draws resamples from $\hat{F}_n$, producing a bootstrap empirical distribution $\hat{F}_n^*$. By standard empirical process arguments, $\hat{F}_n$ converges to the underlying distribution, and the bootstrap reproduces the first-order fluctuation of $\hat{F}_n$ around that limit. If the statistic is a smooth functional of the empirical distribution, then the functional delta method yields bootstrap consistency. □

2.8. Model Interpretability Using SHAP

To complement predictive accuracy with interpretability, SHAP was used to attribute model outputs to individual features [20, 22]. SHAP provides both global interpretation, through ranking features by mean absolute contribution, and local interpretation, through decomposition of an individual prediction. Formally, the SHAP value for feature i is

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \Big[f(S \cup \{i\}) - f(S) \Big], \tag{24}$$

where N is the full feature set, S is a subset not containing feature i, and $f(S)$ denotes the model output under the feature coalition S [22]. In practice, SHAP values for tree-based models can be computed efficiently using TreeSHAP, which exploits the structure of tree ensembles [23]. In the present study, local and global explanations for the primary analysis were computed in probability space.

Proposition 4 (Additive decomposition of prediction in probability space)**.** *For a fixed observation x, if SHAP values are computed with model output on the probability scale, then*

$$\hat{p}(x) = \mathbb{E}[\hat{p}(X)] + \sum_{i=1}^{p} \phi_i. \tag{25}$$

Proof. This follows from the efficiency axiom of Shapley values. The total contribution of all features must equal the difference between the full model output and the baseline value. When the chosen model output is the predicted probability, one has

$$\sum_{i=1}^{p} \phi_i = \hat{p}(x) - \mathbb{E}[\hat{p}(X)], \tag{26}$$

which is equivalent to Eq. (25). □

3. Results and Discussion

This section presents the empirical findings together with their statistical and mathematical interpretation. In line with the methodological framework described earlier, the discussion emphasizes not only predictive performance but also leakage-aware experimental design, class-imbalance handling, uncertainty assessment, and SHAP-based explanation.

3.1. Exploratory Characteristics of the Dataset

The initial stage of the analysis examines the empirical structure of the dataset, including class distribution, missing values, and outlier behavior. Fig. 1 shows that the target variable `TenYearCHD` is highly imbalanced. The majority class (`TenYearCHD`=0) accounts for 84.81% of the observations, while the minority class (`TenYearCHD`=1) accounts for only 15.19%. Such

imbalance is methodologically important because empirical learning procedures may become biased toward majority-class prediction, thereby reducing sensitivity to positive CHD cases [15, 16].

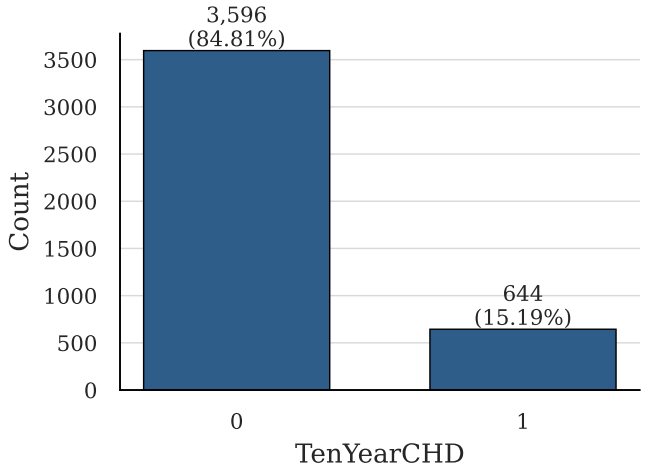


Fig. 1: Class distribution of the target variable (`TenYearCHD`).

Table 2 summarizes the number of missing values before imputation. The largest amount of missingness appears in `glucose`, followed by `education`, `BPMeds`, and `totChol`. From a statistical perspective, this motivates the use of robust imputation rather than case deletion, since excluding incomplete records would reduce the effective sample size and may distort the empirical distribution.

Table 2: Number of missing values before imputation.

Feature	Number of Missing Values
male	0
age	0
education	105
currentSmoker	0
cigsPerDay	29
BPMeds	53
prevalentStroke	0
prevalentHyp	0
diabetes	0
totChol	50
sysBP	0
diaBP	0
BMI	19
heartRate	1
glucose	388
TenYearCHD	0

The boxplot visualization in Fig. 2 indicates that several clinical variables exhibit substantial upper-tail dispersion, especially `glucose`, `sysBP`, `diaBP`, `totChol`, and `BMI`. These patterns justify the use of robust preprocessing such as median imputation and winsorization before model fitting. In particular, the winsorization step is based on the IQR bounds in Eq. (7).

After imputation, transformation, and encoding, the data were further examined through Pearson correlation analysis. Fig. 3 presents the heatmap of Pearson correlation coefficients among predictors and the response variable. Several predictors show positive marginal association with `TenYearCHD`, particularly `age`, `sysBP`, and `prevalentHyp`. This is consistent with established cardiovascular risk structure and with classical risk modeling derived from the Framingham framework [5, 6]. The correlations were computed according to Eq. (9).

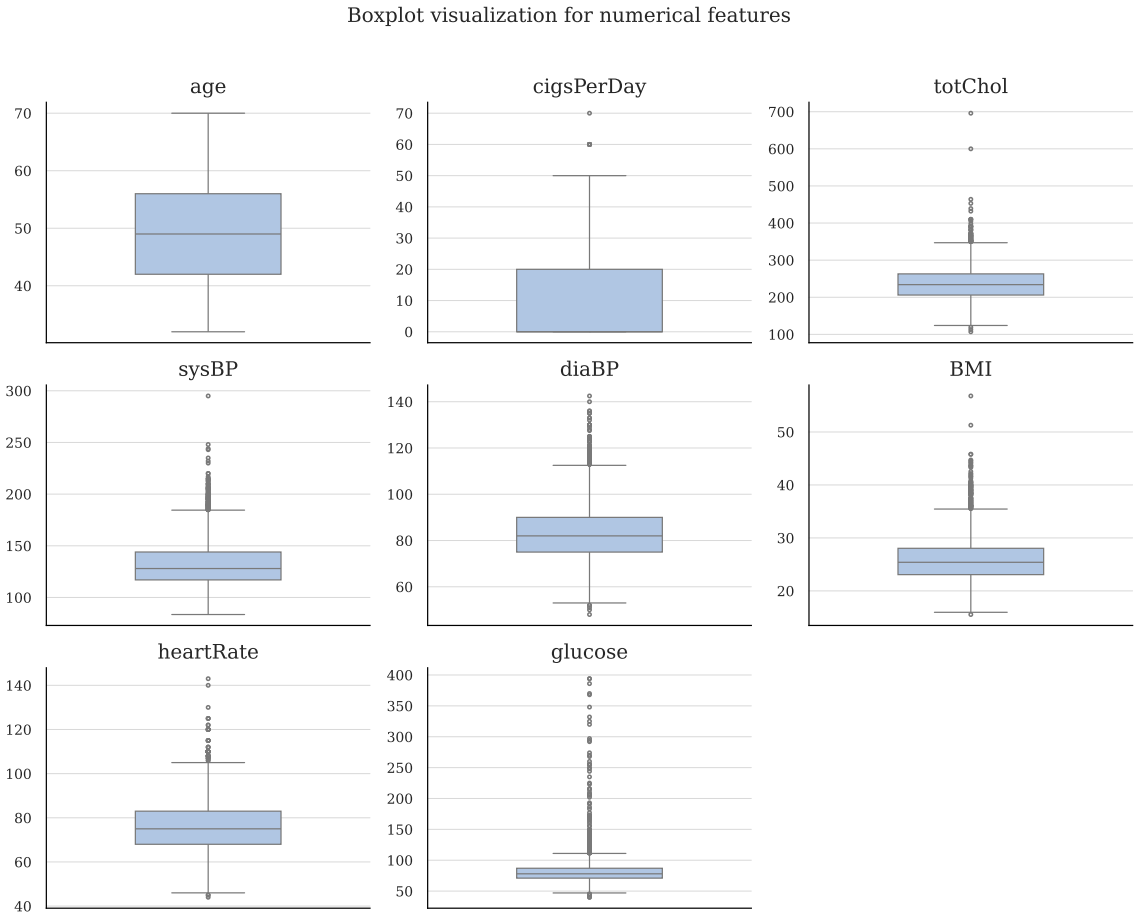


Fig. 2: Boxplot visualization for numerical features.

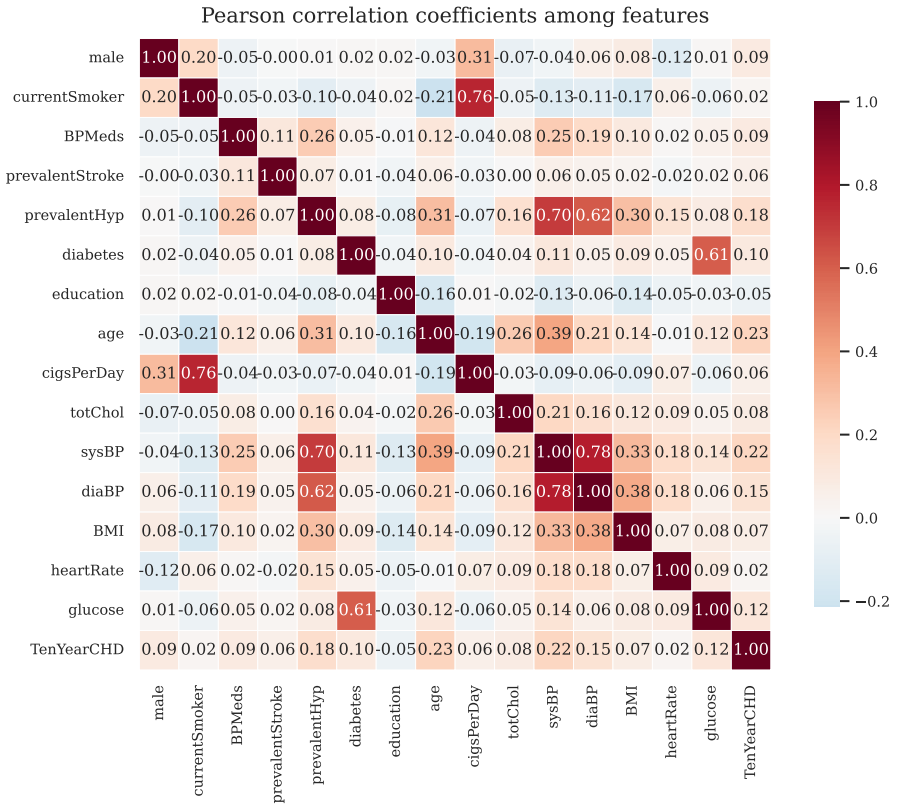


Fig. 3: Heatmap of Pearson correlation coefficients among features.

3.2. Leakage-Aware Split Design

The corrected experimental design uses three stratified train–test splits, but only the 80:20 split is the pre-specified primary analysis. Table 3 summarizes the sample sizes and class counts in each split. Since the split is performed before preprocessing and oversampling, the test set preserves the original class imbalance and remains suitable for out-of-sample evaluation.

Table 3: Train and test sample sizes with class counts for each split ratio.

Split Ratio	Role	Train Size	Test Size	Train Class 0	Train Class 1	Test Class 0	Test Class 1
60:40	Sensitivity	2544	1696	2158	386	1438	258
70:30	Sensitivity	2968	1272	2517	451	1079	193
80:20	Primary	3392	848	2877	515	719	129

This split logic is important for interpreting the results below. In particular, the 80:20 analysis is not selected because it achieves the best test-set score; rather, it is the designated primary analysis established before evaluating the test data. The 60:40 and 70:30 analyses are included only to assess sensitivity to training-sample proportion.

3.3. Preprocessing Effects and Class Rebalancing

For the primary 80:20 split, the original training set contains 2877 majority-class observations and 515 minority-class observations. After applying SMOTE to the training set only, the class distribution becomes balanced, as shown in Fig. 4. Because the test set is left untouched, this training-set-only rebalancing improves class representation during fitting without contaminating the out-of-sample evaluation.

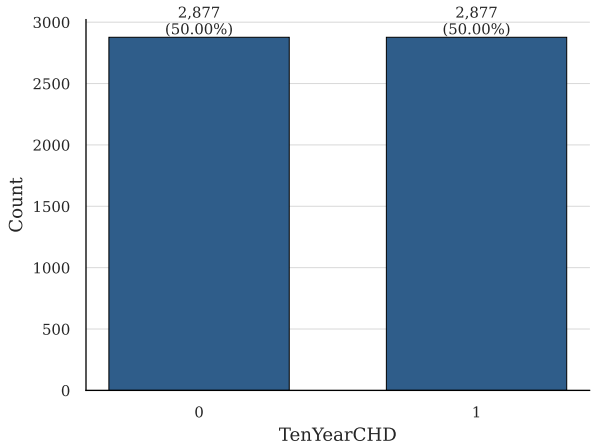


Fig. 4: Target class distribution in the primary training set after applying SMOTE.

Proposition 1 gives the geometric interpretation of this procedure. In particular, Eqs. (10)–(12) show that SMOTE enriches minority support through convex interpolation rather than arbitrary extrapolation. Thus, the oversampling mechanism remains confined to the observed minority-region geometry of the training data.

3.4. Predictive Performance Across Primary and Sensitivity Analyses

To examine the effect of training-sample proportion on predictive performance, three train–test split ratios were evaluated: 60:40, 70:30, and 80:20. For each split, XGBoost hyperparameters were optimized using randomized search with stratified cross-validation on the training set.

3.4.1. Sensitivity Analysis: Split Ratio 60:40

For the 60:40 split ratio, the tuned XGBoost model attains accuracy of 80.13%, precision of 32.60%, recall of 28.68%, F1-score of 30.52%, and ROC–AUC of 0.6716. The selected hyperparameters

are presented in Table 4, while the evaluation metrics are summarized in Table 5. The confusion matrix and ROC curve are jointly displayed in Fig. 5.

Table 4: Optimal hyperparameters for the XGBoost model with a 60:40 split ratio.

Hyperparameter	Value
`n_estimators`	401
`learning_rate`	0.0798
`max_depth`	4
`subsample`	0.9766
`colsample_bytree`	0.7555
`gamma`	1.9392
`reg_lambda`	2.9895
`reg_alpha`	1.7897

Table 5: Model evaluation on the testing set for the 60:40 split ratio.

Metric	Value
Accuracy	80.13%
Precision	32.60%
Recall	28.68%
F1-Score	30.52%
ROC–AUC	0.6716

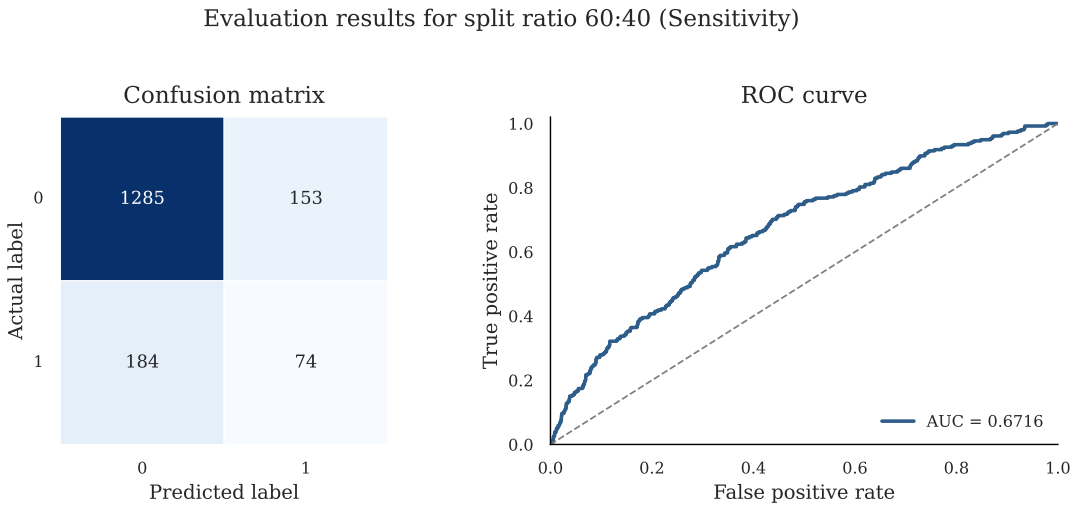


Fig. 5: Evaluation results of the XGBoost model for the 60:40 split ratio (sensitivity analysis): confusion matrix and ROC curve.

These results indicate only moderate discrimination. Although the ROC–AUC is the largest among the three examined splits, this does not alter the inferential role of the 80:20 primary analysis, because the sensitivity analyses are not used for model selection.

3.4.2. Sensitivity Analysis: Split Ratio 70:30

For the 70:30 split ratio, the model attains accuracy of 78.62%, precision of 28.42%, recall of 26.94%, F1-score of 27.66%, and ROC–AUC of 0.6648. The tuned hyperparameters are presented in Table 6, while the model evaluation on the testing set is summarized in Table 7. The confusion matrix and ROC curve are jointly displayed in Fig. 6.

The 70:30 split yields performance of similar scale to the 60:40 split. This suggests that, after correcting the evaluation design, the predictive ability is relatively stable across moderate changes in split proportion, but remains far from the level of strong discrimination.

Table 6: Optimal hyperparameters for the XGBoost model with a 70:30 split ratio.

Hyperparameter	Value
`n_estimators`	698
`learning_rate`	0.0105
`max_depth`	5
`subsample`	0.9314
`colsample_bytree`	0.9317
`gamma`	0.3974
`reg_lambda`	3.6450
`reg_alpha`	1.4137

Table 7: Model evaluation on the testing set for the 70:30 split ratio.

Metric	Value
Accuracy	78.62%
Precision	28.42%
Recall	26.94%
F1-Score	27.66%
ROC–AUC	0.6648

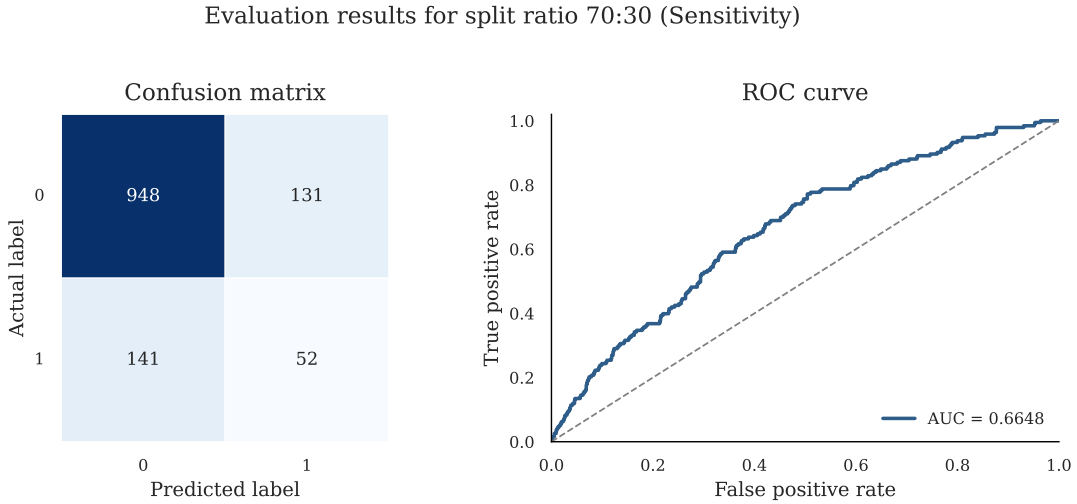


Fig. 6: Evaluation results of the XGBoost model for the 70:30 split ratio (sensitivity analysis): confusion matrix and ROC curve.

3.4.3. Primary Analysis: Split Ratio 80:20

For the pre-specified primary 80:20 split, the model achieves accuracy of 79.36%, precision of 27.88%, recall of 22.48%, F1-score of 24.89%, and ROC–AUC of 0.6502. The optimal hyperparameters are presented in Table 8, while the model performance on the test set is summarized in Table 9. The confusion matrix and ROC curve are jointly displayed in Fig. 7.

Table 8: Optimal hyperparameters for the XGBoost model with the primary 80:20 split ratio.

Hyperparameter	Value
`n_estimators`	291
`learning_rate`	0.0188
`max_depth`	6
`subsample`	0.8486
`colsample_bytree`	0.9425
`gamma`	0.6092
`reg_lambda`	0.6102
`reg_alpha`	0.8803

Table 9: Model evaluation on the testing set for the primary 80:20 split ratio.

Metric	Value
Accuracy	79.36%
Precision	27.88%
Recall	22.48%
F1-Score	24.89%
ROC–AUC	0.6502

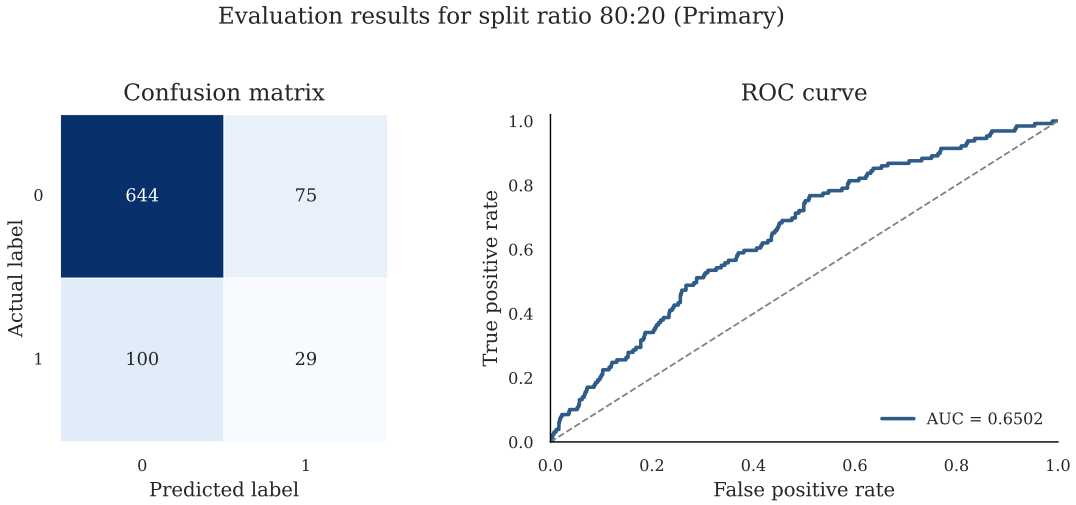


Fig. 7: Evaluation results of the XGBoost model for the primary 80:20 split ratio: confusion matrix and ROC curve.

The confusion matrix in Fig. 7 shows TN = 644, FP = 75, FN = 100, and TP = 29. Thus, while the classifier correctly identifies a large fraction of the majority class, it misses a substantial number of minority-class cases at the default threshold $\tau = 0.5$ in Eq. (6). In clinical screening terms, this means that sensitivity remains limited.

Table 10: Comparison of XGBoost model performance across the primary and sensitivity analyses.

Split Ratio	Role	Accuracy	Precision	Recall	F1-Score	ROC–AUC
60:40	Sensitivity	80.13%	32.60%	28.68%	30.52%	0.6716
70:30	Sensitivity	78.62%	28.42%	26.94%	27.66%	0.6648
80:20	Primary	79.36%	27.88%	22.48%	24.89%	0.6502

Table 10 shows that the three split ratios yield broadly comparable results. Importantly, the primary analysis is not identified as the best-performing split; rather, it is the split fixed in advance for formal reporting. Under this corrected leakage-aware protocol, the empirical performance is clearly more modest than the overly optimistic values reported in the earlier draft.

By Proposition 2, the primary ROC–AUC of 0.6502 means, through Eq. (19), that approximately a randomly selected CHD-risk individual receives a higher predicted score than a randomly selected non-risk individual with probability 0.6502. This corresponds to moderate threshold-independent discrimination, not strong separation between the classes.

3.5. Statistical Reliability of the Primary Model

Because point estimates alone do not quantify uncertainty, the primary 80:20 analysis should be interpreted together with the bootstrap procedure described in the Methods section. Table 11 reports percentile-based 95% confidence intervals based on 1000 nonparametric bootstrap resamples of the held-out primary test set. These intervals are constructed according to Eq. (22) and Eq. (23).

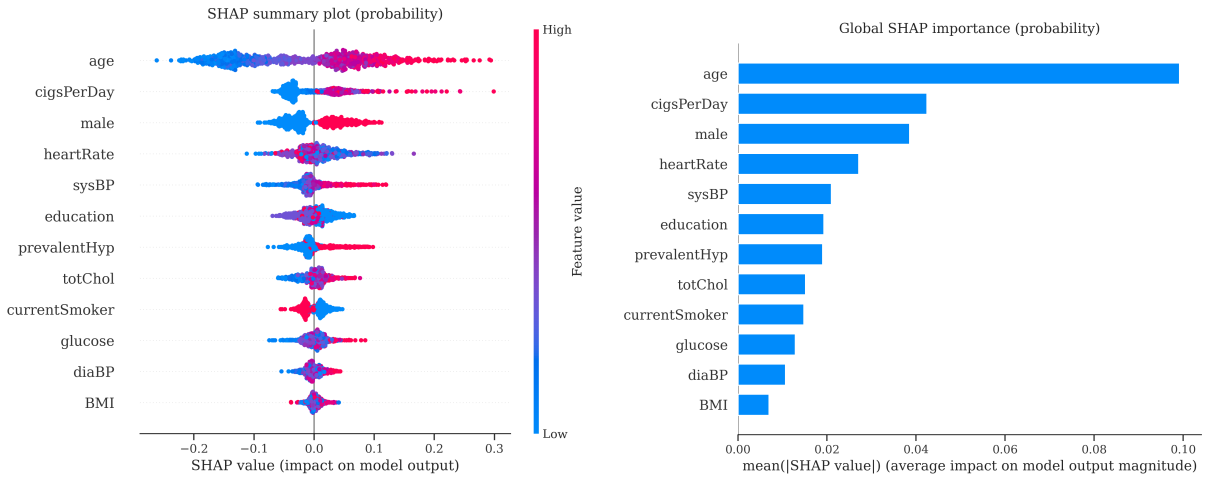
Table 11: Bootstrap percentile 95% confidence intervals for the primary 80:20 model.

Metric	Point Estimate	Lower 95% CI	Upper 95% CI
Accuracy	0.793632	0.766509	0.821934
Precision	0.278846	0.186257	0.360360
Recall	0.224806	0.150721	0.291060
F1-Score	0.248927	0.165840	0.318006
ROC–AUC	0.650192	0.601739	0.698068

The bootstrap interval for ROC–AUC remains well below values typically associated with strong discrimination. Likewise, the intervals for precision, recall, and F1-score are fairly wide and centered at relatively low values. Thus, Proposition 3 supports not merely the computational use of bootstrap intervals, but also the substantive conclusion that the primary model should be interpreted cautiously. The model is statistically coherent, yet its predictive effectiveness on this dataset is limited.

3.6. SHAP-Based Global and Local Interpretation

After fixing the 80:20 split as the primary analysis, interpretability is examined using SHAP in probability space. The global interpretation of the model is presented in Fig. 8, which combines the SHAP summary plot and the mean absolute SHAP bar plot. The summary plot in Fig. 8(a) illustrates both the relative importance of features and the direction of their effects on the model predictions across observations. Meanwhile, Fig. 8(b) provides the overall ranking of features based on mean absolute SHAP values, which are computed from the feature contributions in Eq. (24).



(a) SHAP summary plot. **(b)** SHAP bar plot of mean absolute SHAP values.

Fig. 8: Global SHAP interpretation of the primary model in probability space: (a) SHAP summary plot showing feature impact and direction across observations, and (b) SHAP bar plot showing global feature importance based on mean absolute SHAP values.

As shown in Fig. 8(a) and Table 12, `age` is the most influential feature in the primary model. Other major contributors include `cigsPerDay`, `male`, `heartRate`, and `sysBP`. This ranking is clinically plausible and consistent with established CHD risk structure [5, 6]. The appearance of `education` and `prevalentHyp` among the more influential variables suggests that both socioeconomic and clinical factors contribute to the fitted decision structure.

By Proposition 4, the SHAP values for each observation decompose the predicted probability according to Eq. (25) into a baseline term plus additive feature contributions. This is particularly useful here because the explanations are on the probability scale, so the waterfall plots can be interpreted directly as upward or downward changes in the model-estimated CHD risk.

Table 12: Top ten features ranked by mean absolute SHAP value for the primary 80:20 model.

Feature	Mean Absolute SHAP Value
age	0.099162
cigsPerDay	0.042393
male	0.038551
heartRate	0.027117
sysBP	0.020962
education	0.019283
prevalentHyp	0.018986
totChol	0.015132
currentSmoker	0.014771
glucose	0.012838

To provide local explanation, waterfall plots are shown for the observations with the highest and lowest predicted CHD risk in the primary analysis.

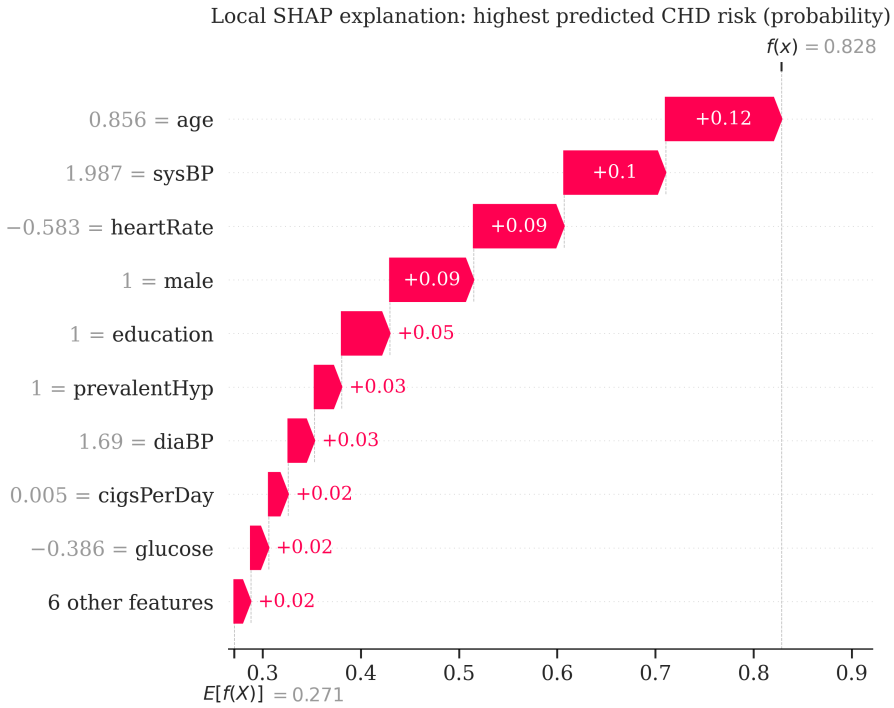


Fig. 9: SHAP waterfall plot for the observation with the highest predicted CHD risk in the primary 80:20 analysis.

The high-risk waterfall plot shows that a relatively small number of features dominate the final predicted probability. In particular, older age and smoking-related variables are among the strongest upward contributors. Conversely, the low-risk waterfall plot shows that favorable values of these same variables, together with more favorable hemodynamic and metabolic features, pull the predicted probability downward. Hence, even though the overall classifier exhibits only moderate discrimination, the fitted model remains interpretable in a clinically coherent manner.

3.7. Overall Discussion

The overall findings indicate that the proposed framework is methodologically stronger than a routine benchmark-style implementation, but empirically more modest than the earlier draft suggested. Once the evaluation pipeline is corrected so that train–test splitting precedes over-sampling and the primary split is fixed in advance, the resulting performance is substantially lower than the earlier optimistic values. This is not a weakness of the corrected analysis; on the

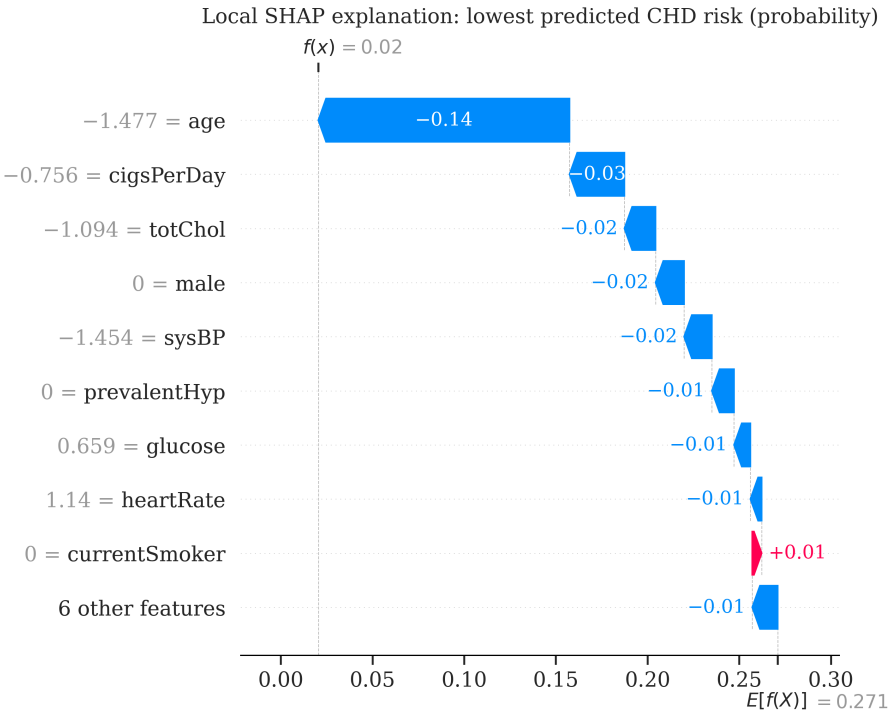


Fig. 10: SHAP waterfall plot for the observation with the lowest predicted CHD risk in the primary 80:20 analysis.

contrary, it demonstrates the importance of leakage-aware experimental design in imbalanced clinical prediction.

From a mathematical and statistical perspective, several conclusions can be drawn. First, the class imbalance present in the original data justifies the use of SMOTE, and Proposition 1 together with Eq. (11) and Eq. (12) clarifies why the resulting classifier improves minority-class representation without extrapolating outside the observed minority region. Second, Proposition 2 provides a probabilistic interpretation of ROC–AUC through Eq. (19), showing that the primary model ranks positive cases above negative ones with probability approximately 0.6502. Third, Proposition 3 justifies bootstrap-based uncertainty assessment, and the resulting intervals based on Eq. (23) confirm that the reported performance should be interpreted as moderate rather than strong. Fourth, Proposition 4 ensures through Eq. (25) that the SHAP-based explanations are locally faithful to the predicted probabilities.

Taken together, these results show that the proposed model is not merely a high-performing classifier but a regularized and interpretable predictive framework whose behavior can be analyzed through class geometry, discrimination theory, uncertainty quantification, and cooperative-game-based feature attribution. At the same time, the empirical results indicate that the model should not yet be described as a clinically ready screening instrument. Potential future improvements include threshold optimization, cost-sensitive learning, class-weighted boosting, probability calibration, and external validation on independent cohorts.

4. Conclusion

This study developed an interpretable framework for coronary heart disease (CHD) risk prediction under class imbalance using XGBoost, leakage-aware preprocessing, training-set-only SMOTE, bootstrap-based uncertainty assessment, and SHAP. The problem was formulated within a regularized empirical risk minimization framework through Eq. (5) and Eq. (6), so the analysis was positioned as predictive statistical modeling rather than routine algorithm benchmarking.

The primary analysis was fixed *a priori* at an 80:20 stratified split, while 60:40 and 70:30 were

treated as sensitivity analyses only. Under this corrected evaluation protocol, the primary model achieved accuracy of 79.36%, precision of 27.88%, recall of 22.48%, F1-score of 24.89%, and ROC–AUC of 0.6502. The 95% bootstrap confidence interval for ROC–AUC, constructed according to Eq. (23), was [0.601739, 0.698068], indicating moderate rather than strong discrimination.

Methodologically, SMOTE was interpreted as a convex interpolation mechanism through Eq. (11) and Eq. (12), while SHAP provided additive feature attribution directly on the probability scale through Eq. (25). The most influential predictors in the primary analysis were `age`, `cigsPerDay`, `male`, `heartRate`, and `sysBP`.

Overall, the study shows that the proposed framework is mathematically structured, interpretable, and methodologically transparent, but that its predictive performance on this dataset remains limited. The model is therefore better viewed as a careful baseline for imbalanced CHD prediction than as a clinically ready decision tool.

CRediT Authorship Contribution Statement

Siti Amiroch: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing–Original Draft, Visualization. **Fitri Nur Laili**: Methodology, Software, Validation, Data Curation, Visualization, Writing–Review & Editing. **Awawin Mustana Rohmah**: Validation, Visualization, Writing–Review & Editing, Supervision. **Dicka Yale Kardono**: Software, Validation, Visualization, Writing–Review & Editing, Supervision.

Declaration of Generative AI and AI-assisted Technologies

The authors declare that generative AI and AI-assisted technologies were used to support English-language editing, paraphrasing, and improvement of text clarity. All scientific content, methodological decisions, data analysis, and interpretation of results were performed by the authors, who reviewed and approved the final manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding. The authors thank the providers of the public Framingham Heart Study dataset hosted on Kaggle for making the data accessible for research and educational purposes.

Data and Code Availability

The dataset analyzed in this study is publicly available from the Kaggle repository.[2] The code used to preprocess the data, train the XGBoost model, estimate bootstrap confidence intervals, and generate SHAP analyses is available from the corresponding author upon reasonable request.

References

[1] World Health Organization. *Cardiovascular Diseases (CVDs)*. WHO Fact Sheet. Accessed 2026-01-29. 2023. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] World Heart Federation. *World Heart Report 2023*. Accessed 2026-01-29. 2023. https://world-heart-federation.org/resource/world-heart-report-2023/.

[2] https://www.kaggle.com/datasets/noeyislearning/framingham-heart-study/data

[3] Gregory A. Roth, George A. Mensah, Catherine O. Johnson, et al. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019". In: *Journal of the American College of Cardiology* 76.25 (2020), pp. 2982–3021. DOI: 10.1016/j.jacc.2020.11.010.

[4] Salim S. Virani et al. "Heart Disease and Stroke Statistics—2024 Update: A Report From the American Heart Association". In: *Circulation* (2024). DOI: 10.1161/CIR.0000000000001209.

[5] Donna K. Arnett, Roger S. Blumenthal, Michelle A. Albert, et al. "2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease". In: *Circulation* 140.11 (2019), e596–e646. DOI: 10.1161/CIR.0000000000000678.

[6] Ralph B. Sr. D'Agostino, Ramachandran S. Vasan, Michael J. Pencina, et al. "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study". In: *Circulation* 117.6 (2008), pp. 743–753. DOI: 10.1161/CIRCULATIONAHA.107.699579.

[7] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[8] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[9] Mohammad Hamim Zajuli Al Faroby, Siti Amiroch, Bernadus Anggo Seno Aji, and Avriono Aritonang. "Classification of IGF1R Ligand Compounds for Identification of Herbal Extracts Using Extreme Gradient Boosting". In: *Jurnal Informatika* 16.3 (2022), pp. 139–150. DOI: 10.26555/jifo.v16i3.a23286.

[10] Siti Amiroch, Mohammad Isa Irawan, Imam Mukhlash, Mohammad Hamim Zajuli Al Faroby, and Chairul Anwar Nidom. "Machine Learning for the Prediction of Antiviral Compounds Targeting Avian Influenza A/H9N2 Viral Proteins". In: *Symmetry* 14.6 (2022), p. 1114. DOI: 10.3390/sym14061114.

[11] Izzudin Muhammad, Imam Mukhlash, Mohammad Jamhuri, Mohammad Iqbal, and Mohammad Isa Irawan. "Classification of COVID-19 Variants Using Boosting Algorithm". In: *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. IEEE. 2022, pp. 29–34. DOI: 10.23919/EECSI56542.2022.9946452.

[12] Mohammad Isa Irawan and Mohammad Jamhuri. "State of the Art of Machine Learning: An Overview of the Past, Current, and the Future Research Trends in the Era of Quantum Computing". In: *AIP Conference Proceedings*. Vol. 2641. 1. AIP Publishing LLC. 2022, p. 040009. DOI: 10.1063/5.0131848.

[13] K. Budholiya, S. K. Shrivastava, and V. Sharma. "An Optimized XGBoost Based Diagnostic System for Effective Prediction of Heart Disease". In: *Journal of King Saud University–Computer and Information Sciences* 34.7 (2022), pp. 4514–4523. DOI: 10.1016/j.jksuci.2020.10.013.

[14] A. Handika Permana, F. Rakhmat Umbara, and F. Kasyidi. "Klasifikasi Penyakit Jantung Tipe Kardiovaskular Menggunakan Adaptive Synthetic Sampling dan Algoritma Extreme Gradient Boosting". In: *Building of Informatics, Technology and Science (BITS)* 6.1 (2024), pp. 499–508. DOI: 10.47065/bits.v6i1.5421.

[15] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[16] Haibo He and Edwardo A. Garcia. "Learning from Imbalanced Data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[17] John A. Swets. "Measuring the Accuracy of Diagnostic Systems". In: *Science* 240.4857 (1988), pp. 1285–1293. DOI: 10.1126/science.3287615.

[18] Tom Fawcett. "An Introduction to ROC Analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[19] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In: *PLOS ONE* 10.3 (2015), e0118432. DOI: 10.1371/journal.pone.0118432.

[20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[21] Julia Amann, Alessandro Blasimme, Effy Vayena, Daniel Frey, and Viola I. Madai. "To Explain or Not to Explain? Artificial Intelligence Explainability in Clinical Decision Support Systems". In: *PLOS Digital Health* 1.2 (2022), e0000016. DOI: 10.1371/journal.pdig.0000016.

[22] Scott M. Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. arXiv preprint. 2017. DOI: 10.48550/arXiv.1705.07874.

[23] Scott M. Lundberg et al. "From Local Explanations to Global Understanding with Explainable AI for Trees". In: *Nature Machine Intelligence* 2.1 (2020), pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

[24] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC, 1993.