Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine-Learning Algorithms and Benchmarking Missing-Data Strategies for Predicting Angiographically Severe Coronary Artery Stenosis: A Retrospective Study

Amir Elhaei
Mazandaran University of Medical Sciences
Maryam Nabati
Mazandaran University of Medical Sciences
Jamshid Yazdani Charati
Mazandaran University of Medical Sciences
Kosar Hoseini
Mazandaran University of Medical Sciences
Maryam Farhadian
Hamadan University of Medical Sciences
Reza Ali mohammadpour
mohammadpour2002@gmail.com
Mazandaran University of Medical Sciences

Research Article

Keywords: coronary artery disease, Coronary artery stenosis, Missing data, Machine learning, Data imputation, Support vector machine, Model calibration, Model interpretability

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11337866/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Missing data is common in clinical datasets used to predict coronary artery disease (CAD) and can influence the relative performance of machine learning algorithms. This study evaluated and compared missing data handling strategies and machine learning models for predicting clinically significant coronary artery stenosis (70% or greater) using clinical variables.

Methods

In this retrospective single-center study, 623 patients undergoing coronary angiography were analyzed. We evaluated three missing-data strategies: indicator-based imputation, MICE-based iterative imputation, and k-nearest neighbor (KNN) imputation in combination with eight machine-learning algorithms, yielding 24 workflows. Model development used an 80/20 stratified train-test split, nested 5-fold cross-validation, randomized hyperparameter optimization, SMOTENC applied within training folds to handle class imbalance, and probability calibration. We selected the final model based primarily on mean outer cross-validated ROC-AUC, using mean outer Brier score as a tie-breaker. We derived the decision threshold from development-set out-of-fold predictions and applied it unchanged to the held-out test set. Model interpretation included SHAP analysis, permutation importance, and a surrogate decision tree.

Results

We selected the indicator-based imputation plus linear support vector machine (SVM) workflow, with a mean outer cross-validated ROC-AUC of 0.730 ± 0.069. On the test set, the model achieved a ROC-AUC of 0.759 (95% CI: 0.667–0.843) and a Brier score of 0.182. At the locked threshold of 0.6346, selected from development-set out-of-fold calibrated probabilities by maximizing balanced accuracy, balanced accuracy was 0.726, with a sensitivity of 75.3%, specificity of 70.0%, precision of 84.2%, and negative predictive value of 57.1%. SHAP analysis identified age, LVEF, WBC, WMA, and troponin among the variables with the greatest model-attributed importance. The surrogate decision tree reproduced the final model with 80.8% fidelity.

Conclusions

In this single-center cohort with substantial and heterogeneous missingness, predictive performance varied across missing-data and modeling strategies, while more complex learners did not demonstrate a discrimination advantage over a regularized linear SVM. The results support the feasibility of a calibrated, missingness-aware machine-learning framework for exploratory CAD risk stratification, but

external validation in larger, multicenter populations with different case mix, disease prevalence, and missingness patterns is required before clinical deployment.

1. Introduction

Coronary artery disease (CAD) remains a major cause of morbidity and mortality worldwide and continues to impose a substantial burden on healthcare systems [1, 2]. Among patients undergoing diagnostic evaluation, identifying clinically significant coronary artery stenosis is important because the extent of luminal narrowing contributes to subsequent diagnostic and therapeutic decision-making [3]. In the present study, angiographically reported severe coronary artery stenosis was operationally defined as a luminal diameter reduction of ≥ 70% [4]. Accordingly, distinguishing patients with versus without angiographically defined severe stenosis represents a clinically relevant binary prediction problem.

The anatomical assessment of coronary artery stenosis traditionally relies on cardiac imaging and, when clinically indicated, invasive coronary angiography [5, 6]. Although coronary angiography directly visualizes the coronary arteries, its invasive nature, procedural requirements, and resource demands limit its use as a general noninvasive assessment strategy [6, 7]. Noninvasive modalities such as coronary computed tomography angiography have expanded diagnostic options but also require specialized infrastructure, contrast administration in many settings, and expert interpretation [8, 9]. These considerations have motivated the development of prediction models based on routinely available clinical, physiological, and laboratory variables that may support risk stratification before definitive anatomical assessment [10]. Machine learning (ML) methods may be well suited to this context, as they can accommodate complex relationships among heterogeneous clinical predictors [11]. Previous studies have evaluated a range of approaches, including logistic regression, support-vector machines, random forests, gradient-boosting methods, and neural networks, for prediction of obstructive CAD [12, 13] as well as deep learning models applied to imaging data [14].

More recently, interpretable machine-learning models have also been investigated for long-term prognostic risk assessment in patients with CAD. In a prospective cohort of 705 hospitalized patients, a random survival forest model achieved a C-index of 0.710 (95% CI: 0.650–0.768) in the validation cohort for predicting 5-year major adverse cardiovascular events (MACE) [15].

However, the reported performance varies across datasets, predictor sets, outcome definitions, preprocessing procedures, and validation strategies, limiting direct comparisons between studies [16, 17].

An important methodological challenge in clinical prediction using routinely collected data is missing information [18]. Clinical variables may be unavailable because they were not measured, were not recorded, or were not available at the time of prediction [19]. Missing-data handling can therefore influence the distributions and relationships among predictors and, consequently, the performance of downstream prediction models [20]. Surveys of missing-data methods in ML further note that the suitability of an imputation strategy depends on the pattern of missingness and the characteristics of

the dataset under investigation [21]. In addition, unequal outcome-class distributions can affect model fitting and may make overall accuracy an inadequate measure of predictive performance [22]. Therefore, appropriate evaluation should consider discrimination, performance across both outcome classes, and calibration rather than relying on accuracy alone [23, 24]. These methodological considerations are particularly relevant in relatively small clinical datasets in which several preprocessing and modeling decisions may interact [25].

Despite the growing literature on ML-based CAD prediction, methodological components such as missing-data handling and model selection are frequently evaluated separately or under study-specific preprocessing and validation procedures [20, 26]. Consequently, differences in reported model performance may reflect not only the choice of learning algorithm but also differences in how incomplete data, class imbalance, model tuning, and validation are handled. Systematic reviews and meta-analytic evidence have likewise highlighted substantial methodological heterogeneity among ML studies of obstructive coronary artery disease and emphasized the importance of appropriate sample-size considerations, missing-data handling, feature-selection procedures, and rigorous validation [16]. A comparative evaluation in which alternative preprocessing and modeling strategies are assessed within a common experimental and validation framework can therefore provide more directly interpretable evidence about the practical consequences of these methodological choices. Accordingly, rather than proposing a new prediction algorithm, the present study uses a benchmarking framework to evaluate the binary prediction of angiographically reported severe coronary artery stenosis (≥ 70%) from clinical data. Three missing-data strategies namely indicator-based imputation [27], multiple imputation by chained equations (MICE) [28], and k-nearest-neighbors (KNN) imputation [29], were evaluated in combination with eight machine-learning algorithms representing different modeling paradigms. Model development incorporates class-imbalance-aware preprocessing [30] and nested cross-validation [31] to reduce information leakage and optimism during model selection and performance estimation. Calibration was additionally evaluated because discrimination alone does not characterize the reliability of the predicted probabilities [24]. Finally, post hoc explainability methods, including SHAP [32], permutation importance [33], and a surrogate decision tree [34], were used to characterize how clinical predictors contributed to the behavior of the selected model without using these explanations for model selection.

The study therefore addressed three methodological questions: (1) how alternative missing-data strategies affect predictive performance under the observed pattern of missing clinical data; (2) whether different machine-learning paradigms provide meaningfully different levels of discrimination, class-balanced performance, and calibration when evaluated under the same preprocessing and validation framework; and (3) how post hoc explainability methods characterize the contribution of clinical predictors to the behavior of the selected model. By evaluating these methodological choices within a common framework, this study aims to provide empirically grounded evidence for developing and evaluating clinical prediction models from incomplete structured healthcare data.

2. Methods

2.1 Study Design, Population, and Data Source

This retrospective study used medical records of adult patients with suspected coronary artery disease (CAD) who underwent diagnostic coronary angiography at Fatemeh Zahra (S) Hospital, Sari, Iran, a central referral cardiac hospital serving patients referred from multiple geographic areas. Between March 20, 2020, and March 20, 2021. All predictor variables included in the predictive modeling analyses were obtained before coronary angiography and before knowledge of the angiographic findings.

The study cohort was defined according to the following inclusion criteria: (1) age ≥ 18 years; (2) suspected CAD with diagnostic coronary angiography performed during the study period; (3) availability of a complete and interpretable angiographic report; and (4) inclusion of only the first angiographic examination when a patient had undergone multiple procedures.

Exclusion criteria during the screening phase included: records unrelated to the study objective; missing documentation in the archives; difficult or impossible access to the record; a mismatch between record information and required data; duplicate records; patients who were not hospitalized (i.e., received outpatient care); failure to perform angiography; unclear required information; incomplete records or documentation; and the absence of required laboratory results.

A total of 808 records were identified from the hospital archives and evaluated based on the inclusion and exclusion criteria. At this stage, 153 records were excluded for reasons documented in the archive review report. These included: 1 record involving a stab wound; 6 records with missing documentation; 3 records that were difficult to access; 9 records with data mismatches; 25 duplicate records; 39 records stored in inaccessible, high-level archive sections; 14 records of outpatients; 34 records lacking angiography data; 20 records with unclear information or status; 1 incomplete record; and 1 record lacking required laboratory results. Consequently, 655 records proceeded to the data cleaning and preparation stage.

During the cleaning phase, pre-established criteria were applied to manage the extent of missing data. Variables with more than 50% missing data were removed from the analytical set, and patients missing more than 50% of the required information were excluded from the analysis. Based on these criteria, 32 records were excluded. Additionally, binary variables with more than 95% of observations concentrated in a single category, patient identifiers, and variables that were duplicates, derived from, or redundant to the target variable were excluded prior to modeling.

Ultimately, out of the initial 808 records, 623 patients remained in the final dataset for statistical analysis and modeling; thus, a total of 185 records comprising 153 cases from the screening stage and 32 from the data cleaning stage were excluded from the study (Fig. 1). The thresholds for excessive variable-level and patient-level missingness were determined empirically after inspection of the observed missing-data distribution, with the objective of removing the least reliable variables or cases while retaining an adequate sample for predictive modeling.

Because this was a retrospective study based on the complete set of eligible archived records available during the predefined study period, no a priori sample size or statistical power calculation was performed. The aim of this study was to benchmark alternative missing-data strategies and machine-learning algorithms for predicting angiographically reported severe coronary artery stenosis (≥ 70%) using routinely available pre-angiography clinical data.

2.2 Angiographic Outcome Definition

Angiographic stenosis severity was determined from the qualitative visual assessment recorded by the attending specialist physician for each assessed coronary vessel rather than by independent quantitative coronary angiography (QCA).

Each assessed coronary vessel was classified into one of five severity categories: normal/no stenosis, mild stenosis (1–49%), moderate stenosis (50–69%), severe stenosis (70–99%), or complete occlusion (100%). At the patient level, the maximum stenosis severity observed among all assessed coronary vessels was used as the patient's overall angiographic severity.

The resulting five-category distribution was as follows: normal/no stenosis (n = 71, 11.4%), mild stenosis (n = 78, 12.5%), moderate stenosis (n = 50, 8.0%), severe stenosis (n = 245, 39.3%), and complete occlusion (n = 179, 28.7%).

For the primary binary prediction task, patient-level maximum stenosis severity was collapsed into two clinically defined categories: <70% stenosis, comprising patients classified as normal/no stenosis, mild, or moderate (n = 199, 31.9%), and ≥ 70% stenosis, comprising patients classified as severe or complete occlusion (n = 424, 68.1%). Thus, the positive class represented patients whose maximum angiographically assessed stenosis was at least 70%.

2.3 Candidate Predictors and Data Preparation

The initial predictor set comprised 39 clinical variables covering demographic characteristics, symptoms, medical history, physiological measurements, laboratory findings, and cardiac variables. The candidate predictors consisted of 13 binary variables sex, smoking, dyspnea, stomach ache, chest pain, family history, troponin, hypertension (HTN), ischemic heart disease (IHD), hyperlipidemia/dyslipidemia (HLP.DLP), WMA, axis, and diabetes mellitus (DM) and 26 continuous variables: age, body mass index (BMI), systolic blood pressure, diastolic blood pressure, pulse rate (PR), respiratory rate (RR), erythrocyte sedimentation rate (ESR), white blood cell count (WBC), red blood cell count (RBC), hemoglobin (HB), hematocrit (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), platelet count, fasting blood sugar (FBS), blood urea nitrogen (BUN), creatinine, cholesterol, triglycerides (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), sodium (Na), potassium (K), heart rate (HR), and left ventricular ejection fraction (LVEF).

No feature selection procedure was used in the final predictive modeling pipeline. The complete candidate predictor set was retained to allow the different machine learning algorithms to operate on the same clinical information and to permit post hoc interpretation of the selected model.

Baseline demographic and clinical characteristics were summarized separately for patients with < 70% and ≥ 70% stenosis. Continuous variables were summarized using mean ± standard deviation when approximately normally distributed and median with interquartile range otherwise. Distributional assessment was based on the Kolmogorov–Smirnov test [35]. Between-group comparisons used the independent-samples Student's t-test or Mann–Whitney U test for continuous variables, as appropriate, and the chi-square test or Fisher's exact test for categorical variables when expected cell counts were small. All tests were two-sided, with $P < 0.05$ considered statistically significant. These descriptive comparisons were not used for predictor selection or model development.

2.4 Assessment of Missingness

Missing data patterns were examined before model development. For each candidate predictor, the number and percentage of observed and missing values were calculated across the complete cohort. The resulting missingness profile is presented in Additional file 1: Table S2.

Little's MCAR test [36] was used to assess whether the observed missing-data pattern was compatible with the assumption that data were missing completely at random (MCAR) [37]. Because Little's test cannot establish a missing-not-at-random (MNAR) mechanism, the missing-data mechanism was not assumed to be definitively MCAR. In addition, the study dataset contained both continuous and categorical clinical variables, for which the assumptions underlying global MCAR testing may be imperfect. Therefore, rather than prespecifying a single imputation method on the basis of the MCAR test, three alternative missing-data strategies were evaluated empirically within the same predictive modeling framework. All imputation procedures were fitted exclusively using the corresponding training partition within each model-development split. The fitted imputation transformations were then applied to the corresponding validation or held-out test partition without refitting on those data. This fold-local implementation was used to prevent information leakage.

2.5 Missing-Data Strategies

Three missing-data strategies were benchmarked:

1. indicator-based median/mode imputation;
2. iterative chained imputation using a Bayesian Ridge estimator; and
3. K-nearest-neighbors (KNN) imputation.

The three strategies were treated as alternative components of the complete predictive modeling pipeline rather than as preprocessing steps performed once before model development.

2.5.1 Indicator-Based Imputation

For each candidate feature, a binary missingness indicator was generated, with 1 representing a missing observation and 0 representing an observed value. Missing continuous variables were imputed using the median calculated from the corresponding training partition, whereas missing categorical variables were imputed using the training-partition mode. The missingness indicators were retained as additional predictors. This approach was intended to preserve potentially informative information contained in the occurrence of missingness itself rather than assuming that the absence of a recorded value was necessarily unrelated to the prediction target [27].

2.5.2 Iterative Chained Imputation

The second strategy used MICE-based iterative imputation implemented with scikit-learn's IterativeImputer. A Bayesian Ridge regression model was used as the estimator for the iterative imputation procedure. The algorithm was initialized using median imputation and was allowed to perform up to 30 iterations, with convergence defined using a tolerance of 1×10^{-3}. Posterior sampling was disabled (sample_posterior = False), resulting in a single deterministic imputed dataset for each training partition [28]. rather than classical multiple imputation with multiple completed datasets and Rubin's rules. Because the objective of this study was predictive benchmarking rather than statistical parameter inference, uncertainty from multiple imputed datasets was not propagated into the final predictive estimates. This distinction is important when interpreting the results and is considered in the study limitations.

2.5.3 K-Nearest-Neighbors Imputation

KNN imputation was used as a non-parametric alternative based on similarity between observations [38]. Five nearest neighbors (k = 5) were used to estimate missing continuous values. Neighbor distances were calculated using the available continuous information after standardization. Continuous variables were standardized using means and standard deviations estimated exclusively from the relevant training partition so that variables measured on larger numerical scales would not dominate the distance calculation. The resulting transformation was then applied to validation and test observations. Categorical variables were imputed using the mode of the corresponding training partition and subsequently encoded for model fitting. For all three strategies, preprocessing operations were incorporated into the modeling pipeline so that estimates required for imputation, scaling, and encoding were learned from training data only.

2.6 Train–Test Partitioning

After data cleaning, the complete cohort of 623 patients was divided once into a development set and a held-out test set using a stratified 80:20 split with random_state = 42. The development set contained 498 patients, including 339 patients with ≥ 70% stenosis and 159 patients with < 70% stenosis. This set was used exclusively for nested cross-validation, hyperparameter optimization, calibration during model development, threshold selection, and model selection. The held-out test set contained 125 patients, including 85 positive cases and 40 negative cases. It was reserved for the final evaluation of the selected model and was not used during model development or selection.

2.7 Machine-Learning Models

Eight machine-learning algorithms were evaluated: Logistic Regression (LR) [39];Decision Tree (DT) [40]; Random Forest (RF) [41]; Multi-Layer Perceptron neural network (MLP) [42]; XGBoost (XGB)[43]; Bagging (BG) [44]; Support Vector Machine (SVM) [45]; and LightGBM (LGBM) [46]. These algorithms were selected to represent different modeling paradigms, including regularized linear classification, individual decision trees, bagging and tree ensembles, gradient boosting, support-vector methods, and neural-network learning. Logistic Regression was implemented with L2 regularization using the liblinear solver. The regularization parameter C was tuned during hyperparameter optimization. The Decision Tree model provided a rule-based nonlinear classifier. Its maximum depth, minimum number of samples required for a split, minimum leaf size, and splitting criterion were included in the search space. Random Forest was implemented as an ensemble of decision trees using bootstrap aggregation and randomized feature selection. The number of trees, maximum tree depth, minimum leaf size, and number of features considered at each split were tuned. The MLP model used one- or two-hidden-layer architectures with ReLU or tanh activation functions. Hidden-layer configuration, L2 regularization, and initial learning rate were tuned, with early stopping enabled to limit unnecessary training and reduce overfitting. XGBoost was evaluated as a gradient-boosting tree model. The number of estimators, maximum tree depth, learning rate, subsampling fraction, column-subsampling fraction, minimum child weight, and L2 regularization were included in the search space. Bagging was implemented using decision trees as the base estimator. The number of base estimators, maximum number of samples, maximum number of features, and bootstrap setting were tuned. SVM was evaluated using both linear and radial-basis-function (RBF) kernels. The regularization parameter C was tuned for both kernels, while gamma was tuned for the RBF kernel. The final calibrated probabilities for SVM were obtained using the external calibration procedure described below rather than SVC's internal probability estimation. LightGBM was evaluated as a gradient-boosting tree model. The number of estimators, number of leaves, maximum depth, learning rate, minimum child samples, subsampling fraction, column-subsampling fraction, and L2 regularization were tuned (Additional file 1: Table S1).

2.8 Hyperparameter Optimization

The combination of three missing-data strategies and eight machine-learning algorithms resulted in 24 candidate workflows. Each workflow was optimized separately.

Hyperparameters were optimized using RandomizedSearchCV [47] with 15 sampled configurations. The search was based on predefined finite parameter grids rather than continuous probability distributions. The inner cross-validation procedure used stratified five-fold cross-validation, with ROC-AUC as the optimization criterion.

The hyperparameter search spaces were algorithm-specific because the tunable parameters differ substantially among the eight learning algorithms. For Logistic Regression and SVM, C values were selected from a 100-point logarithmically spaced grid generated using np.logspace(-3, 2, 100). The SVM search contained separate configurations for the linear and RBF kernels.

The principal hyperparameter search spaces were as follows: C for logistic regression; maximum depth, minimum samples per split, minimum samples per leaf, and splitting criterion for decision trees; number of estimators, maximum depth, minimum leaf size, and maximum features for random forests; hidden-layer architecture, activation function, L2 regularization, and initial learning rate for the MLP; number of estimators, maximum depth, learning rate, subsampling, column subsampling, minimum child weight, and L2 regularization for XGBoost; number of estimators, maximum samples, maximum features, and bootstrap for bagging; kernel, C, and gamma for the RBF kernel for SVM; and number of estimators, number of leaves, maximum depth, learning rate, minimum child samples, subsampling, column subsampling, and L2 regularization for LightGBMAll preprocessing and resampling operations were performed within the corresponding pipeline during model fitting. Thus, hyperparameter optimization was performed jointly with the preprocessing and balancing steps rather than after preprocessing the complete development dataset.

Unless explicitly included in the search space, model parameters were maintained at their specified fixed values or library defaults. The random seed was fixed at 42, and computational procedures were executed sequentially (n_jobs = 1) to improve reproducibility.

No model-level class weighting was combined with the resampling procedure. For applicable models, class_weight was therefore maintained at its default value of None.

2.9 Class-Imbalance Handling

The binary outcome contained 424 positive cases (68.1%) and 199 negative cases (31.9%). To reduce potential dominance of the majority class during model fitting, training data were resampled using the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC) [48].

SMOTENC was selected because the predictor set contained both continuous and categorical variables. It was implemented using the imbalanced-learn package [49] and incorporated into the modeling pipeline.

The critical requirement was that resampling occur only within training data. SMOTENC was therefore applied exclusively during the fitting stage of the relevant training partition and was never applied to an outer validation fold or to the held-out test set. Validation and test observations retained their original class distribution.

The number of nearest neighbors used by SMOTENC was fixed at k = 5, with a safety rule of min (5, n_minority − 1) to ensure that the selected value did not exceed the number of available minority-class observations in a given training partition. The random seed was fixed at 42.

Because the feature matrix contained both continuous and categorical variables, SMOTENC was the effective sampler used in the analyzed workflows. The implementation also contained a fallback to RandomOverSampler if a training partition contained fewer than two minority-class observations; this condition was not reached during the present analysis.

Missing values were imputed before SMOTENC was applied, followed by one-hot encoding and classification. Applying SMOTENC before one-hot encoding allowed categorical variables to retain their categorical representation during resampling. Resampling was restricted to the training data, and all preprocessing steps were implemented within the model pipeline.

2.10 Nested Cross-Validation Framework

Model development and candidate-workflow comparison were performed using nested stratified five-fold cross-validation [31] within the 498-patient development set.

In the outer loop, the development set was divided into five stratified folds. In each iteration, four folds were used for model development and the remaining fold was held out as the outer validation fold

Within each outer training partition, an inner five-fold stratified cross-validation procedure was used for hyperparameter optimization. The hyperparameter combination producing the highest mean inner ROC-AUC was selected for the corresponding training partition.

The complete workflow including imputation, resampling, encoding, hyperparameter optimization, and calibration was treated as the modeling unit. All preprocessing and resampling estimates were therefore learned exclusively from the relevant training partitions, without access to the outer validation fold.

The outer-fold predictions were pooled to form out-of-fold (OOF) predictions for the development set. These predictions were used in the subsequent development-stage calibration and threshold-selection procedures.

Because the outer-fold results were also used to select the final workflow, they should be interpreted as model-development and benchmarking results rather than as an unbiased estimate of the final selected model's performance after selection. Final performance of the locked selected model was assessed separately on the held-out test set.

2.11 Probability Calibration

Probability calibration was performed using sigmoid calibration (Platt scaling) [50] implemented with scikit-learn's CalibratedClassifierCV. This step was used to transform model decision scores into estimated probabilities of the positive class.

Calibration was performed within the model-development procedure using three-fold cross-validation nested within each outer training partition. Hyperparameter optimization remained nested within the corresponding calibration-training data. The calibration procedure therefore did not use the outer validation observations or the held-out test set when estimating calibration parameters.

The ensemble=False configuration was used. Under this configuration, cross-validation was used to generate out-of-fold predictions for fitting the sigmoid calibrator, after which the selected base estimator

was refitted on the complete training partition. The final calibrated model therefore corresponded to a single selected set of model hyperparameters.

Sigmoid calibration was selected as a parametric calibration approach appropriate for a dataset of this size [51] and was used consistently across candidate workflows to maintain comparability.

Calibration performance was assessed using the Brier score and calibration curves [24]. The Brier score was calculated from predicted probabilities and the observed binary outcomes.

2.12 Model Selection

After completion of the nested cross-validation procedure for all 24 candidate workflows, model selection was based primarily on the mean outer ROC-AUC across the five outer folds.

The mean outer Brier score was used as the secondary ranking criterion, followed by candidate name as a deterministic final tie-breaker. The selection procedure was performed entirely within the development set.

The workflow selected by this procedure was the linear-kernel SVM combined with indicator-based imputation. Its mean outer ROC-AUC was 0.7299 (SD = 0.069). The regularization parameter selected in the outer folds was C = 0.002848 in four folds and C = 0.001789 in the remaining fold. For the final refit on the complete development set, the modal value, C = 0.002848, was used with the linear kernel.

The held-out test set was not used to select the model, hyperparameters, imputation strategy, calibration method, or decision threshold.

2.13 Decision Threshold Selection

Because ROC-AUC evaluates discrimination independently of a particular classification threshold, threshold-dependent performance was evaluated separately.

For the selected SVM workflow, the classification threshold was determined using the calibrated OOF probabilities generated from the development set. Balanced accuracy, defined as the arithmetic mean of sensitivity and specificity, was used as the threshold-selection criterion [52]. The threshold search evaluated balanced accuracy across candidate thresholds derived from the OOF predictions. The threshold yielding the maximum balanced accuracy was selected. In the event of a tie, specificity and then the absolute distance from 0.5 were used as deterministic tie-breakers.

The resulting threshold was 0.6346 and was subsequently locked before evaluation of the held-out test set.

The locked threshold was applied once to the test-set calibrated probabilities to generate binary predictions. No test-set outcome information was used during threshold selection.

2.14 Performance Measures

Predictive performance was evaluated using complementary measures of discrimination, calibration, and threshold-dependent classification.

Discrimination was assessed using the area under the receiver operating characteristic curve (ROC-AUC) and average precision (AP; also referred to as PR-AUC).

Calibration was assessed using the Brier score and logarithmic loss (log loss).

Threshold-dependent classification performance included accuracy, balanced accuracy, sensitivity, specificity, positive predictive value (precision), negative predictive value (NPV), F1 score, and Matthews correlation coefficient (MCC) [53].

ROC-AUC was used as the primary model-selection metric because it evaluates ranking discrimination across classification thresholds. Balanced accuracy and the other threshold-dependent measures were used to characterize classification behavior at the prespecified locked threshold.

All performance measures were calculated for the development-set OOF predictions and for the final held-out test-set predictions.

2.15 Bootstrap Confidence Intervals

Ninety-five percent confidence intervals for the test-set performance measures were estimated using a non-parametric stratified bootstrap procedure with 2,000 resamples [54].

For each bootstrap iteration, observations were resampled with replacement separately within the two true outcome classes, thereby preserving the observed class composition of the test set. Performance metrics were recalculated using the frozen final model, fitted preprocessing pipeline, calibration procedure, and locked classification threshold. No model refitting or hyperparameter re-optimization was performed within the bootstrap iterations.

The 2.5th and 97.5th percentiles of the bootstrap distributions were used as the lower and upper limits of the 95% confidence intervals.

These confidence intervals quantify sampling variability in the held-out test set conditional on the fitted model, preprocessing pipeline, calibration, and locked threshold. They do not propagate uncertainty arising from model selection, hyperparameter optimization, imputation, resampling, or threshold selection.

2.16 Pairwise Comparison of ROC-AUC

Paired DeLong tests were used as secondary analyses to compare the test-set ROC-AUC of the prespecified selected workflow with the other CV-qualified candidate workflows [55].

The comparisons were performed using predictions generated by the final fitted pipelines on the same held-out test observations, thereby accounting for the paired nature of the ROC curves.

Because multiple candidate workflows were compared with the selected workflow, the resulting P-values were adjusted using the Holm–Bonferroni [56] procedure to control the family-wise error rate.

These comparisons were performed only after the model-selection lock and were not used to select the final model, calibration method, or decision threshold.

2.17 Post Hoc Interpretability

Post hoc interpretability analyses were performed after the model, calibration procedure, and decision threshold had been locked. The purpose of these analyses was to characterize how the selected model used the available predictors rather than to perform feature selection or modify the fitted model. Three complementary approaches were used: SHAP values; permutation importance; and a shallow surrogate decision tree. All three analyses were applied to the final selected linear-kernel SVM.

2.17.1 SHAP Analysis

SHAP (Shapley Additive explanations) values were used to characterize the contribution of individual predictors to the calibrated probability of the positive class [32].

The analysis was performed on the imputed feature representation before one-hot encoding so that explanations could be expressed at the level of the original clinical variables rather than being fragmented across dummy variables. We internally represented categorical predictors numerically for compatibility with the SHAP masking procedure. SHAP analysis was performed only after completion of model selection and final test-set evaluation had been initiated. It had no role in model selection, hyperparameter optimization, threshold selection, or feature selection.

The Permutation Explainer implementation from the SHAP library was used [32]. A background set of 30 randomly selected training observations was used, and explanations were generated for a randomly selected subset of 50 test observations using a fixed random seed. The explained quantity was the calibrated predicted probability of the positive class.

Global feature importance was summarized using the mean absolute SHAP value across the explained test observations. Local explanations were also generated to illustrate how individual predictors contributed to particular predictions.

SHAP values were interpreted as model-attribution measures rather than causal effects. In particular, correlated clinical variables and the possibility of implausible feature combinations during independent masking were considered when interpreting the results.

2.17.2 Permutation Importance

Permutation importance was used as a complementary model-agnostic measure of predictor contribution [33].

For each predictor, its values were randomly permuted in the held-out test data while the remaining variables were left unchanged. The resulting decrease in ROC-AUC was used to quantify the predictor's contribution to model discrimination. Each feature was permuted 10 times, and the mean change in ROC-AUC was recorded.

Permutation importance was calculated only after the final model had been locked and therefore did not influence model selection or model fitting.

Because permutation of individual predictors can interact with correlations among clinical variables and can produce combinations that are not representative of the observed joint distribution, permutation importance was interpreted as a descriptive measure of predictive contribution rather than a causal effect.

For WMA, particular caution was required because approximately 48.3% of observations were missing and the selected Indicator strategy retained a missingness indicator. Consequently, the apparent importance of WMA may partly reflect information associated with whether the variable was recorded rather than the physiological wall-motion status alone.

2.17.3 Surrogate Decision Tree

A shallow surrogate decision tree was used to provide an approximate, human-readable representation of the decision behavior of the selected SVM [57].

The surrogate tree had a maximum depth of three and was trained to reproduce the binary predictions of the final SVM rather than the observed clinical outcome. The surrogate was fitted using development-set data and evaluated on the held-out test set using the same locked classification threshold of 0.6346.

Surrogate fidelity was defined as the proportion of test observations for which the surrogate and the final SVM produced the same binary prediction. This measure therefore represents agreement with the fitted SVM and should not be interpreted as diagnostic accuracy.

For interpretability, standardized split thresholds were converted back to the corresponding original clinical measurement units using the training-set means and standard deviations. The resulting approximate decision rules were used only to illustrate the behavior of the fitted model and were not treated as independent clinical decision rules.

2.18 Statistical Analysis and Reproducibility

All analyses were performed using Python version 3.14.6. The computational environment consisted of an AMD Ryzen 5 4600H processor with 16 GB RAM. To promote reproducibility and avoid variability associated with parallel execution, model-development procedures were executed using a single CPU worker (n_jobs = 1), with the random seed fixed at 42 throughout the workflow.

The principal software packages and versions used were NumPy 2.5.1[58], pandas 3.0.5 [59], scikit-learn 1.9.0 [60], imbalanced-learn 0.14.2 [49], SciPy 1.18.0 [61], statsmodels 0.14.6 [62]XGBoost 3.4.0 [43], LightGBM 4.6.0 [46], SHAP 0.52.0 [32], pyreadstat 1.3.6, cloudpickle 3.1.2, matplotlib 3.11.1[63], and seaborn 0.13.2 [64].

The complete nested cross-validation and model-comparison procedure required approximately 13 hours, 10 minutes, and 55 seconds of wall-clock time when executed on a single CPU core.

All preprocessing transformations, imputation procedures, resampling operations, hyperparameter optimization, and calibration steps were restricted to the appropriate training partitions during model development. The held-out test set remained untouched until the model-selection lock had been completed.

No external validation cohort was available in the present study. Accordingly, the held-out test set represents an internal evaluation of the selected model rather than external validation or independent multicenter validation. An overview of the complete analytic workflow is provided in Fig. 2.

Generative AI assistance was used to support the development, debugging, and refinement of the Python code used in the statistical analysis. The final code and analytical procedures were reviewed and validated by the authors.

3. Results

3.1. Baseline Characteristics

A total of 623 patients were included in the analysis. Of these, 424 (68.1%) had clinically significant coronary artery stenosis (≥ 70%), whereas 199 (31.9%) had stenosis below 70%. Baseline demographic and clinical characteristics of the two groups are summarized in Table 1.

Patients with stenosis ≥ 70% were older than those with stenosis < 70% (61.18 ± 10.93 vs. 56.51 ± 11.07 years; $p < 0.001$) and were more frequently male (58.87% vs. 48.74%; $p = 0.02$). Smoking was more prevalent among patients with stenosis ≥ 70% (33.33% vs. 20.81%; $p = 0.01$), as were diabetes mellitus (38.68% vs. 28.64%; $p = 0.02$) and positive troponin (39.59% vs. 13.54%; $p < 0.001$). BMI also differed between the groups ($p = 0.04$), although the median value was slightly lower in the ≥ 70% group.

Several laboratory and hemodynamic measurements also differed between groups. Patients with stenosis ≥ 70% had higher pulse rate, erythrocyte sedimentation rate, white blood cell count, fasting blood sugar, blood urea nitrogen, and creatinine concentrations, with p-values ranging from < 0.001 to 0.02. Serum sodium was lower in the ≥ 70% group ($p = 0.012$). Left ventricular ejection fraction was lower among patients with stenosis ≥ 70% (median 48% vs. 50%; $p < 0.001$), and regional wall motion abnormality was substantially more frequent (82.61% vs. 42.03%; $p < 0.001$).

No statistically significant between-group differences were observed for dyspnea, stomach ache, family history, hypertension, previous ischemic heart disease, diastolic blood pressure, respiratory rate, red blood cell count, hemoglobin, hematocrit, MCV, MCH, MCHC, platelet count, cholesterol, triglycerides, HDL, LDL, potassium, or axis (all $p > 0.05$).

Table 1
Baseline clinical and demographic characteristics of the study population, stratified by stenosis severity

Feature		Severity		Total	P
		< 70% (n = 199)	≥ 70% (n = 424)		
Age		56.51 ± 11.07	61.18 ± 10.93	59.68 ± 11.18	< 0.001[a]
Sex	Male	97 (48.74%)	249 (58.87%)	346 (55.63%)	0.02[c]
	Female	102 (51.26%)	174 (41.13%)	276 (44.37%)	0.02[c]
Smoking: yes		31 (20.81%)	105 (33.33%)	136 (29.31%)	0.01[c]
Troponin: positive		13 (13.54%)	116 (39.59%)	129(33.2%)	< 0.001[c]
BMI		27.73 (25.11, 31.24)	26.99 (24.80, 29.48)	27.34(24.83,30.12)	0.04[b]
Dyspnea: yes		48 (24.12%)	93 (21.93%)	141 (22.63%)	0.54[c]
Stomach ache: yes		12 (6.03%)	19 (4.48%)	31 (4.98%)	0.43[c]
Chest pain: yes		157 (78.89%)	362 (85.38%)	519 (83.31%)	0.05[c]
family history: yes		71 (38.80%)	168 (43.75%)	239 (42.15%)	0.28[c]
HTN: yes		104 (52.26%)	237 (55.90%)	341 (54.74%)	0.44[c]
IHD: yes		38 (19.10%)	83 (19.58%)	121 (19.42%)	0.91[c]
DM: yes		57 (28.64%)	164 (38.68%)	221 (35.47%)	0.02[c]
Systole		125 (120, 131)	130 (120, 140)	130(120,140)	0.08[b]
Diastole		80(70, 80)	80 (70, 84)	80(70,80)	0.11[b]
PR		75 (70, 80)	76 (70, 82)	76(70,80)	0.02[b]

Note: Continuous variables are reported as mean ± SD (normally distributed) or median (IQR) (non-normally distributed); categorical variables as n (%). Percentages are calculated from non-missing observations per variable. Between-group comparisons used the independent-samples t-test (a), Mann–Whitney U test (b), or Fisher's exact test (c). All tests were two-sided; $p < 0.05$ was considered statistically significant. BMI, body mass index; HTN, hypertension; IHD, ischemic heart disease; DM, diabetes mellitus; PR, pulse rate; RR, respiratory rate; ESR, erythrocyte sedimentation rate; WBC, white blood cell; RBC, red blood cell; HB, hemoglobin; HCT, hematocrit; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; FBS, fasting blood sugar; BUN, blood urea nitrogen; HDL, high-density lipoprotein; LDL, low-density lipoprotein; Na, sodium; K, potassium; HR, heart rate; LVEF, left ventricular ejection fraction; WMA, wall motion abnormality.

Feature	Severity		Total	P
	< 70% (n = 199)	≥ 70% (n = 424)		
RR	18 (15, 18)	18 (16, 18)	18(16,18)	0.14[b]
ESR	14(8.50, 25)	20 (10, 31)	17(10,28)	0.01[b]
WBC	6800 (5600, 7850)	7900 (6400, 9900)	7400(6100,9400)	< 0.001[b]
RBC	4.34 ± 0.57	4.30 ± 0.58	4.31 ± 0.58	0.35[a]
HB	12.67 ± 1.94	12.77 ± 1.97	12.74 ± 1.96	0.54[a]
HCT	38.42 ± 4.59	38.42 ± 4.57	38.42 ± 4.58	0.99[a]
MCV	91.20 (87.30, 94.60)	91.25 (87.12, 94.60)	91.20(87.10,94.60)	0.716[b]
MCH	29.90 (28.00, 31.50)	30.25 (28.40, 31.98)	30.20(28.30,31.80)	0.080[b]
MCHC	32.60 (31.40, 34.00)	32.90 (31.60, 34.10)	32.80(31.50,34.10)	0.110[b]
PLATLET	210000 (180000, 246000)	215000 (181000, 261500)	213000 (180750,256250)	0.239[b]
FBS	106 (93.25, 126)	114.50 (95.00, 163.25)	111(95,148)	0.005[b]
BUN	14.50 (12, 17)	15.20 (13.00, 20.00)	15(12.35,19)	0.002[b]
Creatinine	0.97 (0.80, 1.16)	1.00 (0.90, 1.20)	1.0(0.87,1.20)	0.010[b]
Cholesterol	152 (127.50, 182.50)	160.00 (131.00, 190.50)	157(130,189)	0.134[b]

Note: Continuous variables are reported as mean ± SD (normally distributed) or median (IQR) (non-normally distributed); categorical variables as n (%). Percentages are calculated from non-missing observations per variable. Between-group comparisons used the independent-samples t-test (a), Mann–Whitney U test (b), or Fisher's exact test (c). All tests were two-sided; $p < 0.05$ was considered statistically significant. BMI, body mass index; HTN, hypertension; IHD, ischemic heart disease; DM, diabetes mellitus; PR, pulse rate; RR, respiratory rate; ESR, erythrocyte sedimentation rate; WBC, white blood cell; RBC, red blood cell; HB, hemoglobin; HCT, hematocrit; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; FBS, fasting blood sugar; BUN, blood urea nitrogen; HDL, high-density lipoprotein; LDL, low-density lipoprotein; Na, sodium; K, potassium; HR, heart rate; LVEF, left ventricular ejection fraction; WMA, wall motion abnormality.

Feature		Severity		Total	P
		< 70% (n = 199)	≥ 70% (n = 424)		
Triglyceride		146 (100.50, 193.50)	147.50 (104.25, 205.75)	147(101,203.50)	0.355[b]
HDL		44 (38, 50.)	43.00 (37.00, 49.00)	43(38,50)	0.136[b]
LDL		80 (62, 104)	89.00 (66.00, 112.25)	87(65,110)	0.061[b]
Na		140 (137, 142)	139.00 (136.00, 141.00)	139(136,141)	0.012[b]
K		4.10 (3.80, 4.30)	4.10 (3.90, 4.40)	4.1(3.80,4.40)	0.298[b]
Axis	Abnormal	15 (9.7)	48 (12.7)	63 (11.8)	0.33[c]
	Normal	140(90.3)	330 (87.3)	470(88.2)	
HR		70 (60, 75)	75 (60, 80)	75(60,75)	0.008[b]
LVEF		50 (48.50, 53)	48 (40, 50)	50(40,53)	< 0.001[b]
WMA	Abnormal	29 (42.03%)	209 (82.61%)	238 (73.91%)	< 0.001[c]
	Normal	40 (57.97%)	44 (17.39%)	84 (26.09%)	

Note: Continuous variables are reported as mean ± SD (normally distributed) or median (IQR) (non-normally distributed); categorical variables as n (%). Percentages are calculated from non-missing observations per variable. Between-group comparisons used the independent-samples t-test (a), Mann–Whitney U test (b), or Fisher's exact test (c). All tests were two-sided; $p < 0.05$ was considered statistically significant. BMI, body mass index; HTN, hypertension; IHD, ischemic heart disease; DM, diabetes mellitus; PR, pulse rate; RR, respiratory rate; ESR, erythrocyte sedimentation rate; WBC, white blood cell; RBC, red blood cell; HB, hemoglobin; HCT, hematocrit; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; FBS, fasting blood sugar; BUN, blood urea nitrogen; HDL, high-density lipoprotein; LDL, low-density lipoprotein; Na, sodium; K, potassium; HR, heart rate; LVEF, left ventricular ejection fraction; WMA, wall motion abnormality.

3.2 Missing-Data Pattern

Across the 623 patients, missingness varied substantially across candidate predictors. The highest proportions of missing observations were observed for WMA (48.3%), troponin (37.6%), smoking (25.5%), and left ventricular ejection fraction (22.8%), (Additional file 1: Figure S2). Little's MCAR test yielded $\chi^2 = 2031.620$ with 1,931 degrees of freedom ($P = 0.055$). This result did not provide sufficient evidence to reject the MCAR assumption; however, because a nonsignificant MCAR test does not

establish a missing-at-random or missing-not-at-random mechanism, the missing-data strategies were compared empirically within the predictive modeling framework.

3.3 Benchmark Performance in Nested Cross-Validation

All 24 candidate workflows were evaluated on the development set (n = 498), and the complete benchmark results are presented in Additional file 1: Table S3. The highest mean outer ROC-AUC was observed for the indicator × SVM workflow (0.730 ± 0.069), which also had the lowest mean outer Brier score (0.190).

The highest mean outer ROC-AUC was observed for the indicator × SVM workflow (0.730 ± 0.069), which also had the lowest mean outer Brier score among the candidate workflows (0.190). The next were obtained for MICE × Random Forest (0.726 ± 0.073), Indicator × Logistic Regression (0.722 ± 0.059), MICE × LightGBM (0.721 ± 0.073), and KNN × Random Forest (0.720 ± 0.066). The indicator × SVM workflow was therefore selected for final model fitting and held-out test evaluation.

Across the three missing data strategies, SVM and Logistic Regression generally showed consistent discrimination, with mean outer ROC-AUC values ranging from approximately 0.707 to 0.730. Tree-based ensemble methods showed variable performance across imputation strategies, with mean outer ROC-AUC values generally ranging from approximately 0.689 to 0.726. The single Decision Tree had the lowest discrimination across all three missing-data strategies, with mean outer ROC-AUC values ranging from 0.625 to 0.665 and an MCC of 0 at the reference threshold of 0.5.

The choice of missing-data strategy also altered model performance within the same algorithm. Across the evaluated workflows, switching between Indicator, MICE, and KNN imputation produced differences in mean outer ROC-AUC of up to approximately 0.04 for a given algorithm. Thus, imputation strategy contributed materially to the observed benchmark performance (Fig. 3).

Values represent mean outer ROC-AUC across the five outer folds on the development set (n = 498). The workflow with the highest mean outer ROC-AUC was Indicator × SVM (0.730). Decision Tree showed the lowest mean outer ROC-AUC under all three missing-data strategies. The heatmap summarizes the observed interaction between machine-learning algorithm and missing-data strategy.

3.4 Performance of the Selected Model on the Held-Out Test Set

The held-out test set contained 125 patients, including 85 with stenosis ≥ 70% and 40 with stenosis < 70%. The selected model achieved a test-set ROC-AUC of 0.759 (95% CI: 0.667–0.843) and a Brier score of 0.182 (95% CI: 0.164–0.200). Average Precision was 0.864 (95% CI: 0.806–0.919), and log loss was 0.542 (95% CI: 0.501–0.586). At the locked threshold, accuracy was 0.736, balanced accuracy was 0.726, sensitivity was 0.753, specificity was 0.700, MCC was 0.433, and F1 score was 0.795. Based on the corresponding confusion matrix, precision was 0.842 and NPV was 0.571(Table 2).

The test-set confusion matrix comprised 28 true negatives, 12 false positives, 21 false negatives, and 64 true positives. The corresponding sensitivity and specificity were 75.3% and 70.0%, respectively.

Table 2
Nested CV reference vs. held-out test performance of the selected winner.

Metric	Development-set OOF	Held-out test set (95% CI)
Mean outer Roc AUC	0.730	0.759 (0.667, 0.843)
Average Precision	0.834	0.864(0.806, 0.919)
Mean outer Brier	0.190	0.182 (0.164, 0.200)
Log loss	0.560	0.542(0.501, 0.586)
Accuracy	0.703	0.736 (0.656, 0.808)
Balanced Accuracy	0.578	0.726 (0.64, 0.805)
Sensitivity	0.920	0.753 (0.659, 0.835)
Specificity	0.233	0.700 (0.550, 0.825)
MCC	0.212	0.433 (0.268, 0.586)
Precision	0.719	0.842 (0.78, 0.904)
NPV	0.617	0.571(0.472, 0.681)
F1	0.809	0.795 (0.726, 0.854)

The OOF column reports development-set out-of-fold performance at the fixed reference threshold of 0.5. The test-set column reports performance on the held-out internal test set (n = 125; 85 positive and 40 negative) at the locked threshold of 0.6346. Threshold-dependent metrics are not directly comparable between the two columns because different thresholds were used. Test-set 95% confidence intervals were obtained from 2,000 stratified bootstrap resamples.

The pre-calibration and post-calibration ROC curves were indistinguishable, with an identical ROC-AUC of 0.759 (Fig. 4).

Because sigmoid calibration is a monotonic transformation of the model scores for a fixed fitted model, the ranking of test-set observations was preserved the precision–recall curves were indistinguishable before and after calibration, with an identical average precision of 0.864 (95% CI: 0.806–0.919). The horizontal reference line corresponds to the positive-class prevalence of approximately 0.68 (Additional file 1: Figure S1).

The calibration plot shows the relationship between predicted probabilities and observed outcome frequencies in eight quantile-based groups. The test-set Brier score was 0.182 (Fig. 5).

The matrix contains 28 true negatives, 12 false positives, 21 false negatives, and 64 true positives. Sensitivity was 0.753 and specificity was 0.700 (Additional file 1: Figure S2).

3.5 Threshold Selection

The development-set-derived decision threshold was 0.6346, corresponding to the maximum balanced accuracy obtained from the development-set out-of-fold predictions under the prespecified selection rule (Fig. 6).

The curves were calculated from the development-set out-of-fold predictions of the selected Indicator × SVM workflow. The locked threshold of 0.6346 corresponded to the maximum balanced accuracy under the prespecified selection rule.

3.6 Secondary Pairwise Comparison of ROC-AUC

As a secondary analysis, the selected workflow was compared on the held-out test set with the other CV-qualified candidate workflows using paired DeLong tests. After Holm adjustment for multiple comparisons, none of the pairwise differences in ROC-AUC reached statistical significance (all adjusted $p > 0.05$). These comparisons were secondary and were not used to alter the prespecified model selection or replace the selected workflow.

3.7 SHAP-Based Feature Importance

Age had the largest mean absolute SHAP value (0.0421), followed by LVEF (0.0323), BMI (0.0183), WBC (0.0181), WMA (0.0169), Troponin (0.0154), FBS (0.0146), BUN (0.0132), ESR (0.0131), and Na (0.0124) (Additional file 1: Table S4.).

The SHAP summary plot shows the distribution and direction of feature contributions to calibrated predicted probabilities. Age had the largest overall contribution, and higher age values were generally associated with higher predicted probability. Lower LVEF values were also associated with higher predicted probability in the evaluated sample (Additional file 1: Figure S3).

3.7.1 Permutation Importance

WBC showed the largest mean decrease in ROC-AUC (0.0243), followed by WMA (0.0175), Troponin (0.0173), Age (0.0127), Diastole (0.0097), Smoking (0.0093), and Sex (0.0083). The resulting ranking differed from the SHAP ranking; for example, WBC ranked first by permutation importance but fourth by SHAP, whereas Age ranked first by SHAP but fourth by permutation importance (Additional file 1: Figure S4).

Bars represent the mean decrease in test-set ROC-AUC after permuting each feature over 10 repetitions; error bars represent ± 1 SD. WBC, WMA, Troponin, Age, and Diastole produced the largest mean decreases in ROC-AUC. These values represent model-based importance measures and should not be interpreted as causal effects.

3.7.3 Surrogate Decision Tree

The surrogate achieved 80.8% agreement with the SVM predictions on the held-out test set. The approximation primarily used LVEF, WBC, ESR, FBS, and Age. The detailed surrogate decision rules are provided in Additional file 1: Table S5.

The reported thresholds are standardized values (z-scores), and the corresponding rules represent an approximation of the SVM decision function rather than the exact decision boundary.

4. Discussion

4.1. Principal Findings

This study evaluated 24 candidate workflows combining three missing-data strategies with eight machine-learning algorithms for the binary classification of clinically significant coronary artery stenosis (≥ 70%) in a single-center retrospective cohort with incomplete clinical data. The highest mean outer ROC-AUC was obtained by the indicator × SVM workflow (0.730 ± 0.069), with a mean outer Brier score of 0.190. the best-performing workflow in the nested cross-validation benchmark combined indicator-based missing-data handling with a linear SVM. On the held-out internal test set (n = 125; 85 positive and 40 negative), the locked Indicator × SVM model achieved an ROC-AUC of 0.759 (95% CI: 0.667–0.843) and a Brier score of 0.182 (95% CI: 0.164–0.200) [10]. At the locked threshold of 0.6346, selected from the development-set out-of-fold probabilities, balanced accuracy was 0.726, with sensitivity of 75.3%, specificity of 70.0%. Consequently, the test-set ROC-AUC of 0.759 represents the performance of the frozen final model on previously unseen observations within the same study cohort and should be regarded as internal validation rather than external validation.

In this dataset, increasing model complexity did not yield a distinct advantage over the regularized linear SVM model. This finding aligns with the results of a systematic review indicating that, for clinical prediction models, there is no evidence of machine learning methods outperforming logistic [65]. The effect of the missing-data strategy was also evident. For a given machine-learning algorithm, changing the imputation strategy produced differences in mean outer ROC-AUC of up to approximately 0.04. The best-performing workflow used indicator-based imputation, in which missingness indicators were retained alongside median or mode imputation. One possible explanation is that the fact that a measurement was missing may contain information related to clinical workflow or patient assessment [66]. However, this interpretation cannot be established from the present data, and informative missingness may reflect clinical decision processes rather than the underlying disease state [27]. In addition, the present implementation of MICE was a single deterministic iterative imputation rather than a multiple-imputation procedure with Rubin's rules [67]. Accordingly, the observed performance of indicator imputation should be interpreted as a dataset-specific empirical finding rather than evidence that indicator-based imputation is generally superior to MICE or KNN.

The surrogate decision tree reproduced the final SVM's binary predictions with 80.8% fidelity on the test set. In particular, the high contribution of WMA warrants cautious interpretation because WMA had substantial missingness (48.3%) and was retained in the indicator-based pipeline. Part of its model-attributed importance may therefore reflect the missingness pattern itself. More broadly, the nested cross-validation design, fold-specific preprocessing, training-only resampling, out-of-fold probability generation for threshold selection, and selection lock before test evaluation provided safeguards against common forms of information leakage. Secondary paired DeLong comparisons [55] did not identify statistically significant differences between the selected workflow and the other CV-qualified candidate workflows after Holm adjustment [56] (all adjusted $p > 0.05$). Thus, while the indicator × SVM workflow had the highest mean outer ROC-AUC and was selected according to the prespecified ranking criterion, the observed performance differences should not be interpreted as definitive evidence that it is superior to all other candidate workflows.

4.2. Clinical Implications

The operating characteristics of the selected model provide an initial indication of its predictive behavior, but they do not establish clinical utility or readiness for implementation. At the locked threshold of 0.6346, the model achieved 75.3% sensitivity and 70.0% specificity on the held-out internal test set, with a precision of 84.2% and an NPV of 57.1%. Therefore, a negative prediction in this study should not be interpreted as sufficient evidence to exclude clinically significant stenosis or to withhold further clinical evaluation.

The variables receiving the greatest model-attributed importance, including Age, LVEF, WMA, Troponin, WBC, ESR, and FBS, were routinely available before coronary angiography. However, the importance estimates should not be interpreted as evidence that these variables have independent causal effects on angiographically assessed stenosis. The differences between SHAP and permutation rankings also illustrate that feature importance depends on the interpretability method and the relationships among predictors.

The surrogate decision tree offers a simplified representation of the final SVM's prediction structure. However, the surrogate achieved only 80.8% fidelity on the test set and therefore does not reproduce the complete SVM decision function. Its rules are best regarded as an exploratory approximation rather than as a clinical decision rule.

4.3. Limitations

First this was a single-center, retrospective study with a relatively small development set (n = 498) and a held-out test set of 125 patients. No external validation cohort was available; therefore, the reported performance represents internal validation and may not generalize to populations with different case mix and missingness patterns.

Second, the reference outcome (stenosis severity) was based on qualitative visual assessment by the attending physician rather than quantitative coronary angiography (QCA). Visual assessment may differ

from QCA-based measurement. Such measurement variability may have introduced outcome misclassification and may have affected estimates of model performance. Relatedly, the < 70% group is clinically heterogeneous, comprising both angiographically normal individuals and patients with mild-to-moderate, non-obstructive atherosclerotic disease.

Third, the missing-data mechanism could not be established with certainty. Little's MCAR test yielded p = 0.055, but this does not establish an MCAR or exclude an MNAR mechanism [36, 37]. In addition, MICE was implemented as a single deterministic iterative imputation without Rubin's rules, so imputation uncertainty was not fully propagated [68]; formal multiple imputation could be evaluated in future studies.

Fourth, class imbalance was addressed using SMOTENC within the training folds [48]. However, resampling-based imbalance correction may adversely affect probability calibration and does not necessarily improve discrimination, as demonstrated in previous research using SMOTE with logistic regression [69]. Decision-curve and net-benefit analyses were not performed. Secondary paired DeLong comparisons of test-set ROC-AUC did not identify statistically significant differences after Holm adjustment [56]. However, these non-significant findings do not establish equivalence between candidate models [70], particularly given the modest test-set size (n = 125).

Fifth, the interpretability methods have inherent limitations. SHAP was based on permutation masking, which may generate feature combinations that are not representative of the observed clinical data [71]. Permutation importance may be difficult to interpret when predictors are correlated, because permuting an individual feature can disrupt dependencies among predictors and affect the resulting importance estimates [72]. Accordingly, the reported feature-importance results should be interpreted as descriptive, model-specific attributions rather than causal effects.

Sixth, the hyperparameter search used 15 randomized iterations per candidate [47]. Although the search spaces were predefined and reported in Table S1, this limited search may not have fully explored the configuration space of more complex models such as XGBoost and LightGBM. Consequently, alternative hyperparameter configurations could potentially yield somewhat different performance estimates.

4.4 Future Directions

Several directions follow from the present findings.

First, the influence of missing-data handling should be examined in larger cohorts and across datasets with different missingness patterns. Future studies could compare indicator-based methods with fully specified multiple-imputation approaches that propagate imputation uncertainty [28], as well as sensitivity analyses under alternative missing-data assumptions.

Second, external validation should be performed in geographically and clinically distinct populations [73]. Such validation should assess not only discrimination but also calibration, threshold-dependent performance, and the stability of the selected predictors. Decision-curve and net-benefit analyses would

further clarify whether the model provides clinically meaningful benefit across relevant threshold probabilities [74, 75].

Third, the interpretability analysis could be extended using methods that more explicitly account for predictor dependence. Conditional or dependence-aware explanation methods could therefore provide more robust assessments of predictor contribution[71].

Fourth, the surrogate model should be evaluated in independent datasets rather than only by its fidelity to the original SVM. An 80.8% fidelity indicates that the surrogate captures a substantial portion of the SVM's binary prediction pattern but does not demonstrate equivalent diagnostic performance or clinical validity. Future studies should therefore assess whether simplified surrogate rules remain stable across institutions.

Finally, future work should investigate the stability of threshold selection and performance estimates under repeated resampling or bootstrap-based model development [76, 77].

5. Conclusion

In this study, we systematically benchmarked three missing-data strategies and eight machine-learning algorithms for the binary classification of clinically significant coronary artery stenosis (≥ 70% stenosis). The selected workflow was a linear Support Vector Machine (SVM) combined with Indicator-based imputation, which achieved a mean outer cross-validated ROC-AUC of 0.730 and a test ROC-AUC of 0.759 (95% CI: 0.667–0.843). At the out-of-fold-selected threshold of 0.6346, the model achieved a balanced accuracy of 0.726, with sensitivity of 75.3%, specificity of 70.0%, and a Brier score of 0.182. However, the NPV was 57.1%, indicating that a negative prediction should not be interpreted as sufficient to exclude clinically significant stenosis in this high-prevalence setting. Accordingly, the model should not be considered a stand-alone diagnostic test. Overall, the findings indicate that a regularized linear model combined with a missing-indicator strategy and calibrated estimates can provide a practical predictive framework for identifying clinically significant coronary artery stenosis under incomplete clinical data. Nevertheless, external validation in larger, multicenter cohorts with different case mixes and disease prevalences is essential to assess generalizability before any clinical deployment [73]. Differences is missingness patterns should also be examined, as they may affect model performance and calibration.

Abbreviations

AP	average precision
BG	bagging
CAD	coronary artery disease
CI	confidence interval

DT	decision tree
KNN	k-nearest neighbors
LGBM	LightGBM
LR	logistic regression
MCC	Matthews correlation coefficient
MICE	multiple imputation by chained equations
MLP	multi-layer perceptron
NPV	negative predictive value
RBF	radial basis function
RF	random forest
ROC-AUC	area under the receiver operating characteristic curve
SD	standard deviation
SHAP	SHapley Additive exPlanations
SMOTENC	Synthetic Minority Over-sampling Technique for Nominal and Continuous
SVM	support vector machine
WBC	white blood cell count
WMA	wall motion abnormality
LVEF	left ventricular ejection fraction
XGB	XGBoost

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Ethics Committee of Mazandaran University of Medical Sciences (ethics approval code: IR.MAZUMS..REC.1403.20325). The requirement for informed consent was waived because of the retrospective design and use of archived clinical records.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analyzed during the current study are not publicly available because they contain patient-level clinical information and are subject to institutional and ethical restrictions. De-identified data may be available from the corresponding author upon reasonable request and subject to approval by the relevant institutional authorities.

Supplementary Information

Additional_file_1 contain Supplementary Table S1-S5 and Supplementary Figure S1-S4.

Competing interests

The authors declare that they have no competing interests.

Funding

The creation of this study and manuscript was not supported financially by grants, funding, or any other financial means.

Authors' contributions

AE: Conceptualization, methodology, formal analysis, software, data curation, and writing original draft; MN: Clinical supervision and interpretation of the clinical implications of the findings; JY: Critical review and substantive revision of the manuscript; KH: Data acquisition and data curation; MF: Critical review and substantive revision of the manuscript; RA: Conceptualization, methodology, and ahraagreed to be accountable for their respective contributions and for appropriately addressing questions related to the accuracy and integrity of the work. All the authors gave comments and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the support and cooperation of the Vice-Chancellor for Research and Technology at Mazandaran University of Medical Sciences and Dr. Mohammad Ali Hosseinian, President of Fatemeh Zahra (S) Hospital. The authors also thank the hospital archive staff for their assistance in accessing the archived medical records and for providing clarification regarding the archival records and retrieval process.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this manuscript, the authors used ChatGPT (GPT-5.6 Luna; OpenAI) to assist with the development, debugging, and refinement of Python code used in the data analysis. DeepSeek(V4-pro) was used to assist with manuscript organization and structural development. Claude (Sonnet 5 Anthropic) was used for English-language translation, grammar correction, and linguistic

editing. All AI-assisted outputs were critically reviewed, verified, and edited by the authors. The authors take full responsibility for the accuracy, originality, integrity, interpretation, and final content of the manuscript. No generative AI tool was used to generate, modify, or fabricate research data, statistical results, or scientific conclusions.

References

1. Rittiphairoj T, Bulstra C, Ruampatana C, Stavridou M, Grewal S, Reddy CL, et al. The economic burden of ischaemic heart diseases on health systems: a systematic review. BMJ Glob Health. 2025;10:e015043. https://doi.org/10.1136/bmjgh-2024-015043.
2. Martin SS, Aday AW, Allen NB, Almarzooq ZI, Anderson CAM, Arora P, et al. Circulation. 2025;151:e41–660. https://doi.org/10.1161/CIR.0000000000001303. 2025 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association.
3. Vrints C, Andreotti F, Koskinas KC, Rossello X, Adamo M, Ainslie J, et al. 2024 ESC guidelines for the management of chronic coronary syndromes. Eur Heart J. 2024;45:3415–537. https://doi.org/10.1093/eurheartj/ehae177.
4. Neumann F-J, Sousa-Uva M, Ahlsson A, Alfonso F, Banning AP, Benedetto U, et al. 2018 ESC/EACTS Guidelines on myocardial revascularization. Eur Heart J. 2019;40:87–165. https://doi.org/10.1093/eurheartj/ehy394.
5. Lawton JS, Tamis-Holland JE, Bangalore S, Bates ER, Beckie TM, Bischoff JM, et al. 2021 ACC/AHA/SCAI guideline for coronary artery revascularization: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2022;145:e4–17. https://doi.org/10.1161/CIR.0000000000001039.
6. Gurav A, Revaiah PC, Tsai T-Y, Miyashita K, Tobe A, Oshima A, et al. Coronary angiography: a review of the state of the art and the evolution of angiography in cardio therapeutics. Front Cardiovasc Med. 2024;11:1468888. https://doi.org/10.3389/fcvm.2024.1468888.
7. Dewey M, Biavati F, Marchese A, Rossini R. Coronary computed tomography angiography is the new reference standard for the diagnosis of coronary artery disease: pros and cons. EuroIntervention. 2024;20:e1490–2. https://doi.org/10.4244/EIJ-E-24-00055.
8. Abdelrahman KM, Chen MY, Dey AK, Virmani R, Finn AV, Khamis RY, et al. Coronary Computed Tomography Angiography From Clinical Uses to Emerging Technologies. J Am Coll Cardiol. 2020;76:1226–43. https://doi.org/10.1016/j.jacc.2020.06.076.
9. Korosoglou G, Giusca S. Contrast agent volume in coronary computer tomography angiography—where are the limits? Quant Imaging Med Surg. 2021;11:4511–3. https://doi.org/10.21037/qims-21-488.
10. Lee H-G, Park S-D, Bae J-W, Moon S, Jung CY, Kim M-S, et al. Machine learning approaches that use clinical, laboratory, and electrocardiogram data enhance the prediction of obstructive coronary artery disease. Sci Rep. 2023;13:12635. https://doi.org/10.1038/s41598-023-39911-y.

11. Wan S, Wan F, Dai X. Machine learning approaches for cardiovascular disease prediction: A review. Arch Cardiovasc Dis. 2025;118:554–62. https://doi.org/10.1016/j.acvd.2025.04.055.
12. Zhang Z, Shao B, Liu H, Huang B, Gao X, Qiu J, et al. Construction and validation of a predictive model for coronary artery disease using extreme gradient boosting. J Inflamm Res. 2024;17:4163–74. https://doi.org/10.2147/JIR.S464489.
13. Leiherer A, Schnetzer L, Mink S, Muendlein A, Bermeitinger B, Saely C, et al. Predicting coronary artery stenosis using machine learning to reduce unnecessary angiographies. Atherosclerosis. 2025;407. https://doi.org/10.1016/j.atherosclerosis.2025.119535.
14. Tu L, Deng Y, Chen Y, Luo Y. Accuracy of deep learning in the differential diagnosis of coronary artery stenosis: a systematic review and meta-analysis. BMC Med Imaging. 2024;24:243. https://doi.org/10.1186/s12880-024-01403-4.
15. Jiang Z, Zhou H, Liu Y, Yin H, Zhu J, Xiong X, et al. Development and validation of an interpretable machine learning model for predicting 5-year major adverse cardiovascular events in patients with coronary artery disease. BMC Cardiovasc Disord. 2026;26:749. https://doi.org/10.1186/s12872-026-06103-1.
16. Guo J, Bao C, Gong L, Zhang Z. Comparing machine learning models and traditional approaches for predicting obstructive coronary artery disease: a systematic review and meta-analysis. Front Cardiovasc Med. 2026;13:1892900. https://doi.org/10.3389/fcvm.2026.1892900.
17. Banerjee T, Paçal İ. A systematic review of machine learning in heart disease prediction. Turk J Biol. 2025;49:600–34. https://doi.org/10.55730/1300-0152.2766.
18. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338. https://doi.org/10.1136/bmj.b2393.
19. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inf Assoc. 2013;20:144–51. https://doi.org/10.1136/amiajnl-2011-000681.
20. Nijman SW, Leeuwenberg AM, Beekers I, Verkouter I, Jacobs JJ, Bots ML, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol. 2022;142:218–29. https://doi.org/10.1016/j.jclinepi.2021.11.023.
21. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data. 2021;8:140. https://doi.org/10.1186/s40537-021-00516-9.
22. Ventura V das, de GJ CMV, de Almeida JM, Pessoa BP, Polanczyk CA, do Nascimento GF et al. Beyond Area Under the Receiver Operating Characteristic Curve: Evaluating Predictive Performance Metrics Under Class Imbalance in Real-World Clinical Data. JMIR Form Res. 2026;10:e86379. https://doi.org/10.2196/86379
23. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA. 2017;318:1377–84. https://doi.org/10.1001/jama.2017.12126.

24. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.
25. Riley RD, Pate A, Dhiman P, Archer L, Martin GP, Collins GS. Clinical prediction models and the multiverse of madness. BMC Med. 2023;21:502. https://doi.org/10.1186/s12916-023-03212-y.
26. Bektaş J, Ibrikçi T, Özcan İ. The impact of imputation procedures with machine learning methods on the performance of classifiers: An application to coronary artery disease data including missing values. Biomed Res. 2018;29:2780–5. https://doi.org/10.4066/biomedicalresearch.29-18-199.
27. Groenwold RH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Can Med Assoc J. 2012;184:1265–9. https://doi.org/10.1503/cmaj.110977.
28. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45:1–67. https://doi.org/10.18637/jss.v045.i03.
29. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inf Decis Mak. 2016;16:74. https://doi.org/10.1186/s12911-016-0318-z.
30. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221–32. https://doi.org/10.1007/s13748-016-0094-0.
31. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91.
32. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al. editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017. pp. 4765–74.
33. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20:1–81.
34. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. ACM Comput Surv. 2018;51:1–42. https://doi.org/10.1145/3236009.
35. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. J Am Stat Assoc. 1951;46:68–78. https://doi.org/10.1080/01621459.1951.10500769.
36. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. J Am Stat Assoc. 1988;83:1198–202. https://doi.org/10.1080/01621459.1988.10478722.
37. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92. https://doi.org/10.1093/biomet/63.3.581.
38. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17:520–5. https://doi.org/10.1093/bioinformatics/17.6.520.
39. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. Mach Learn Res. 2008;9:1871–4. https://doi.org/10.5555/1390681.1442794

40. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Belmont, CA: Wadsworth; 1984.
41. Breiman L, Random Forests, editors. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324
42. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6. https://doi.org/10.1038/323533a0.
43. Chen T, Guestrin C, XGBoost:. A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016. pp. 785–94. https://doi.org/10.1145/2939672.2939785
44. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40. https://doi.org/10.1007/BF00058655.
45. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1007/BF00994018.
46. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30. Long Beach, CA: Curran Associates, Inc.; 2017. pp. 3146–54.
47. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281–305. https://doi.org/10.5555/2503308.2188395.
48. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Synth Minor -Sampl. 2002;16:321–57. https://doi.org/10.1613/jair.953.
49. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18:1–5.
50. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Cambridge, MA: MIT Press; 2000. https://doi.org/10.7551/mitpress/1113.003.0008.
51. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning - ICML '05. Bonn, Germany: ACM Press; 2005. pp. 625–32. https://doi.org/10.1145/1102351.1102430
52. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition. IEEE; 2010. pp. 3121–4. https://doi.org/10.1109/ICPR.2010.764
53. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:6. https://doi.org/10.1186/s12864-019-6413-7.
54. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
55. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45. https://doi.org/10.2307/2531595.
56. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70. https://doi.org/10.2307/4615733.

57. Martens D, Baesens B, Van Gestel T, Vanthienen J. Comprehensible credit scoring models using rule extraction from support vector machines. Eur J Oper Res. 2007;183:1466–76. https://doi.org/10.1016/j.ejor.2006.04.051.
58. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.
59. McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. Austin, TX: SciPy; 2010. pp. 51–6. https://doi.org/10.25080/Majora-92bf1922-00a
60. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
61. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
62. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. Austin, TX; 2010. pp. 92–6. https://doi.org/10.25080/Majora-92bf1922-011
63. Hunter JD, Matplotlib. A 2D graphics environment. Comput Sci Eng. 2007;9:90–5. https://doi.org/10.1109/MCSE.2007.55.
64. Waskom M. seaborn: statistical data visualization. J Open Source Softw. 2021;6:3021. https://doi.org/10.21105/joss.03021.
65. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.
66. Sisk R, Lin L, Sperrin M, Barrett JK, Tom B, Diaz-Ordaz K, et al. Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. J Am Med Inf Assoc. 2021;28:155–66. https://doi.org/10.1093/jamia/ocaa242.
67. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30:377–99. https://doi.org/10.1002/sim.4067.
68. Rubin DB. Multiple imputation for nonresponse in surveys. New York; 1987. https://doi.org/10.1002/9780470316696.
69. Van den Goorbergh R, Van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inf Assoc. 2022;29:1525–34. https://doi.org/10.1093/jamia/ocac093.
70. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. BMJ. 1995;311:485. https://doi.org/10.1136/bmj.311.7003.485.
71. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell. 2021;298:103502. https://doi.org/10.1016/j.artint.2021.103502.

72. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9:307. https://doi.org/10.1186/1471-2105-9-307.
73. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14:49–58. https://doi.org/10.1093/ckj/sfaa188.
74. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Mak. 2006;26:565–74. https://doi.org/10.1177/0272989X06295361.
75. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352. https://doi.org/10.1136/bmj.i6.
76. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(19960229):361–87. https://doi.org/10.1002/(SICI)1097-0258.)15:4%253C361::AID-SIM168%253E3.0.CO;2-4.
77. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

Figures

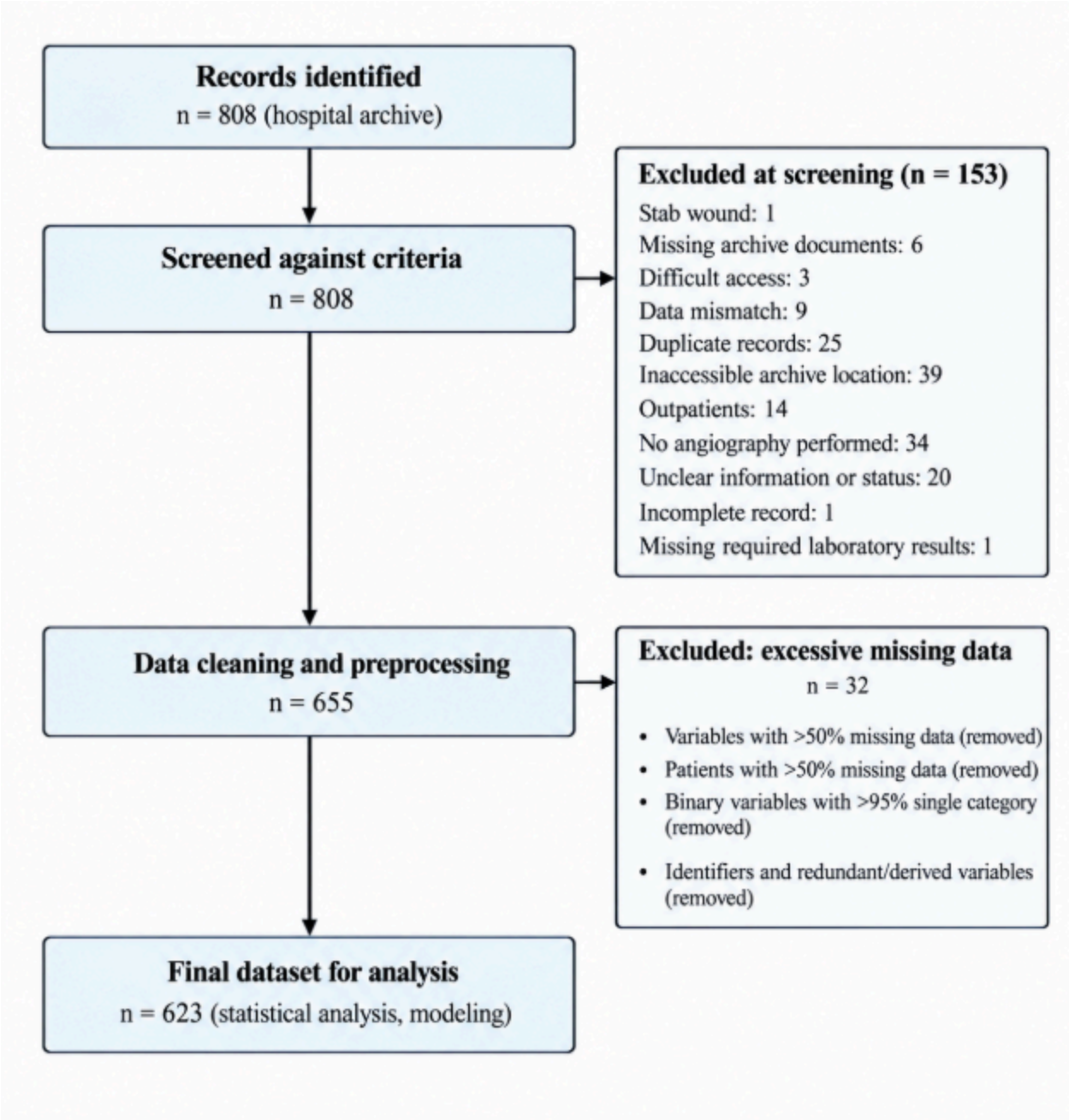


Figure 1

Flow diagram of patient selection and exclusion process. Of 808 records identified from the hospital archive, 623 patients were included in the final analysis.

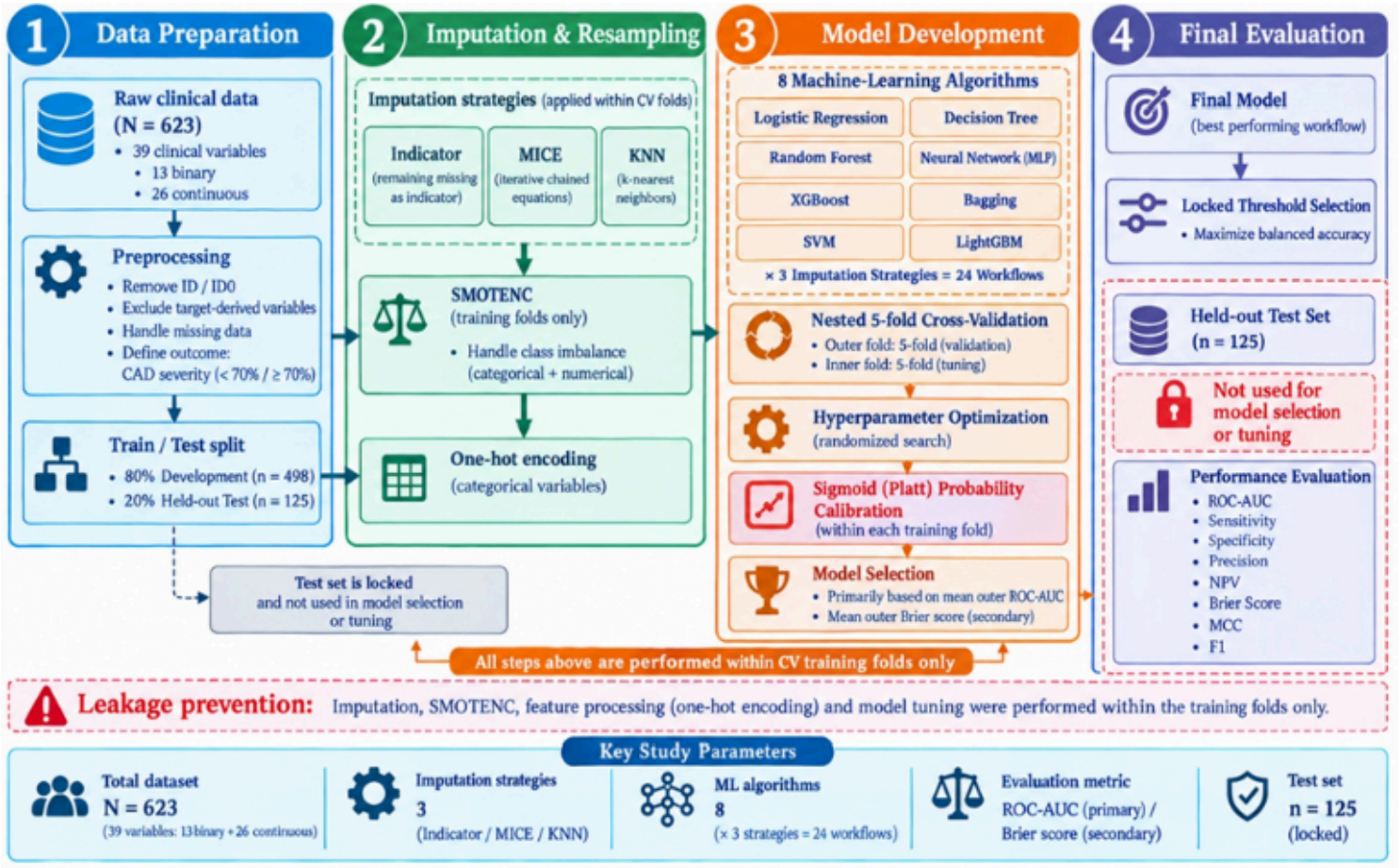


Figure 2

Overview of the analytic pipeline for predicting clinically significant (≥70%) coronary artery stenosis.

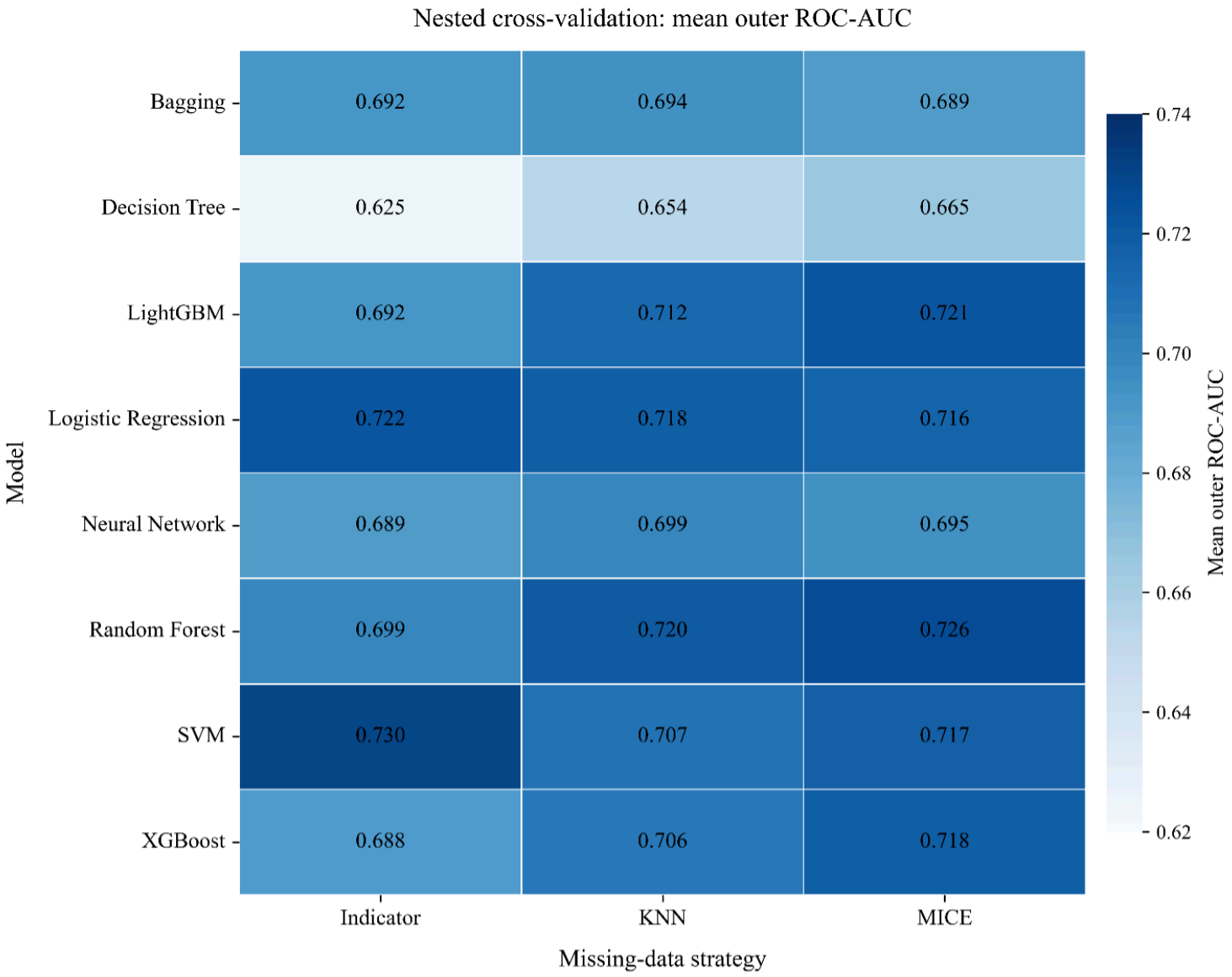


Figure 3

Heatmap of mean outer ROC-AUC across the 24 candidate workflows

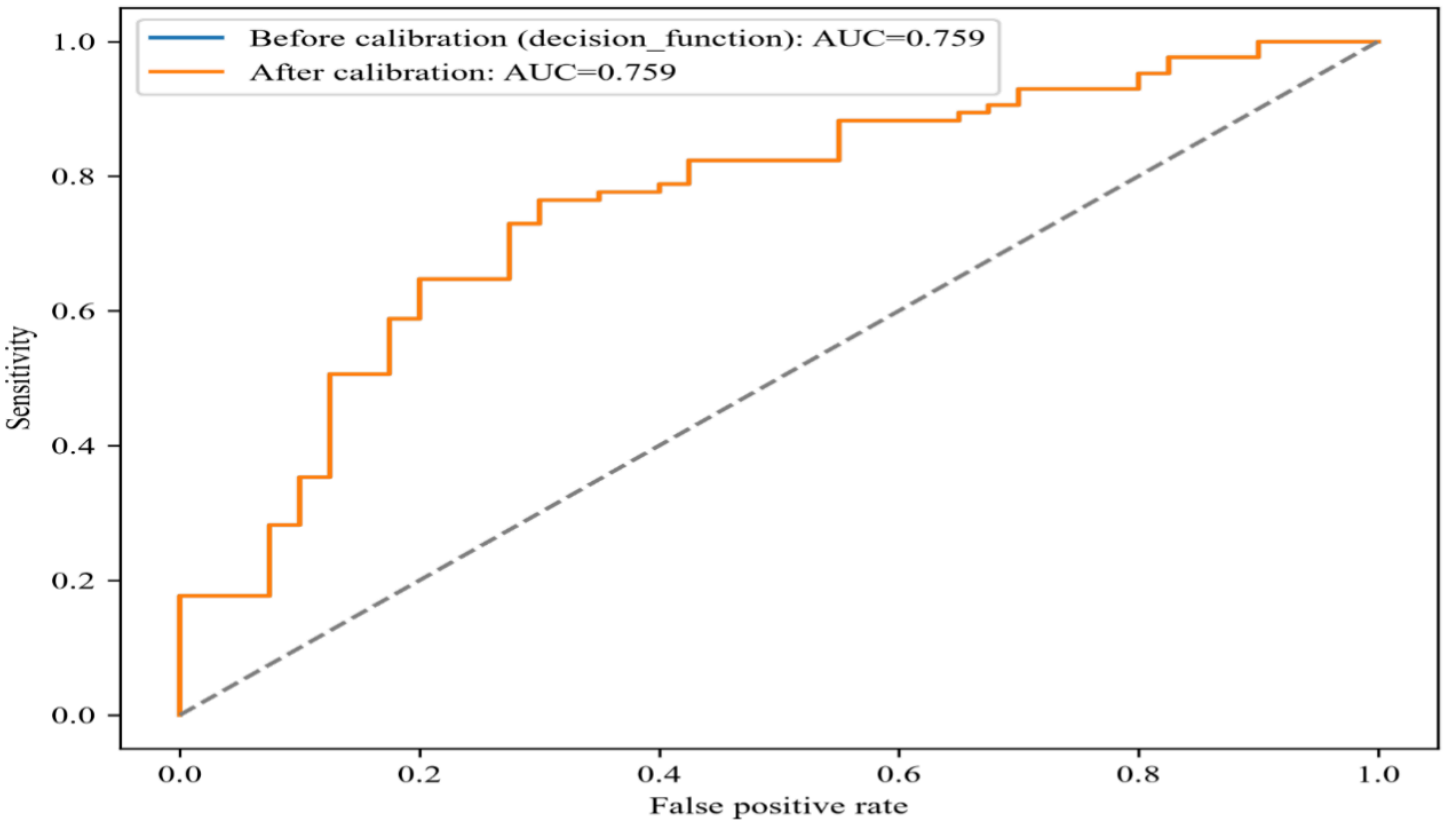


Figure 4

ROC curves of the selected Indicator × SVM workflow on the held-out test set before and after Platt calibration

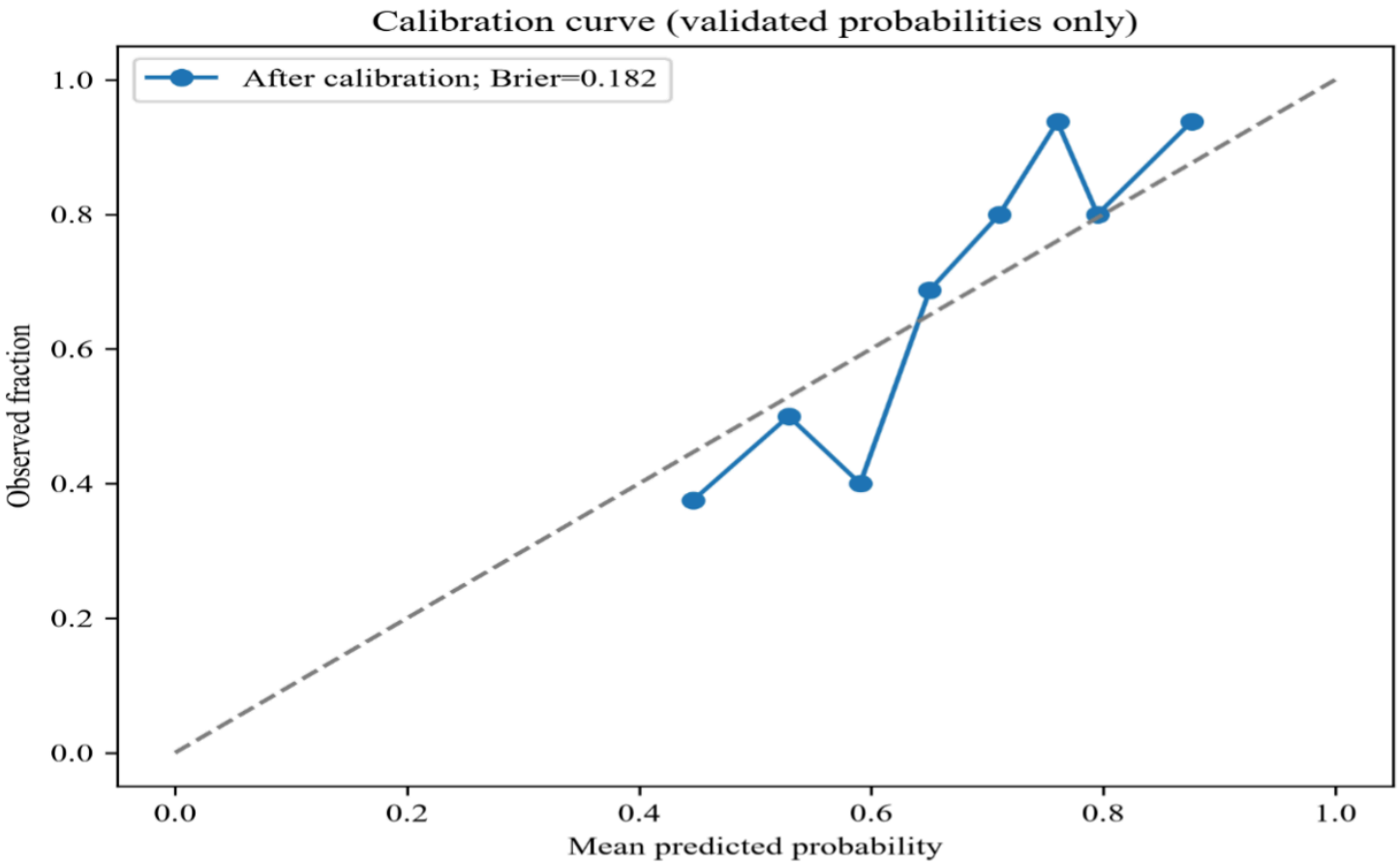


Figure 5

Calibration curve of the selected Indicator × SVM workflow on the held-out test set after Platt calibration.

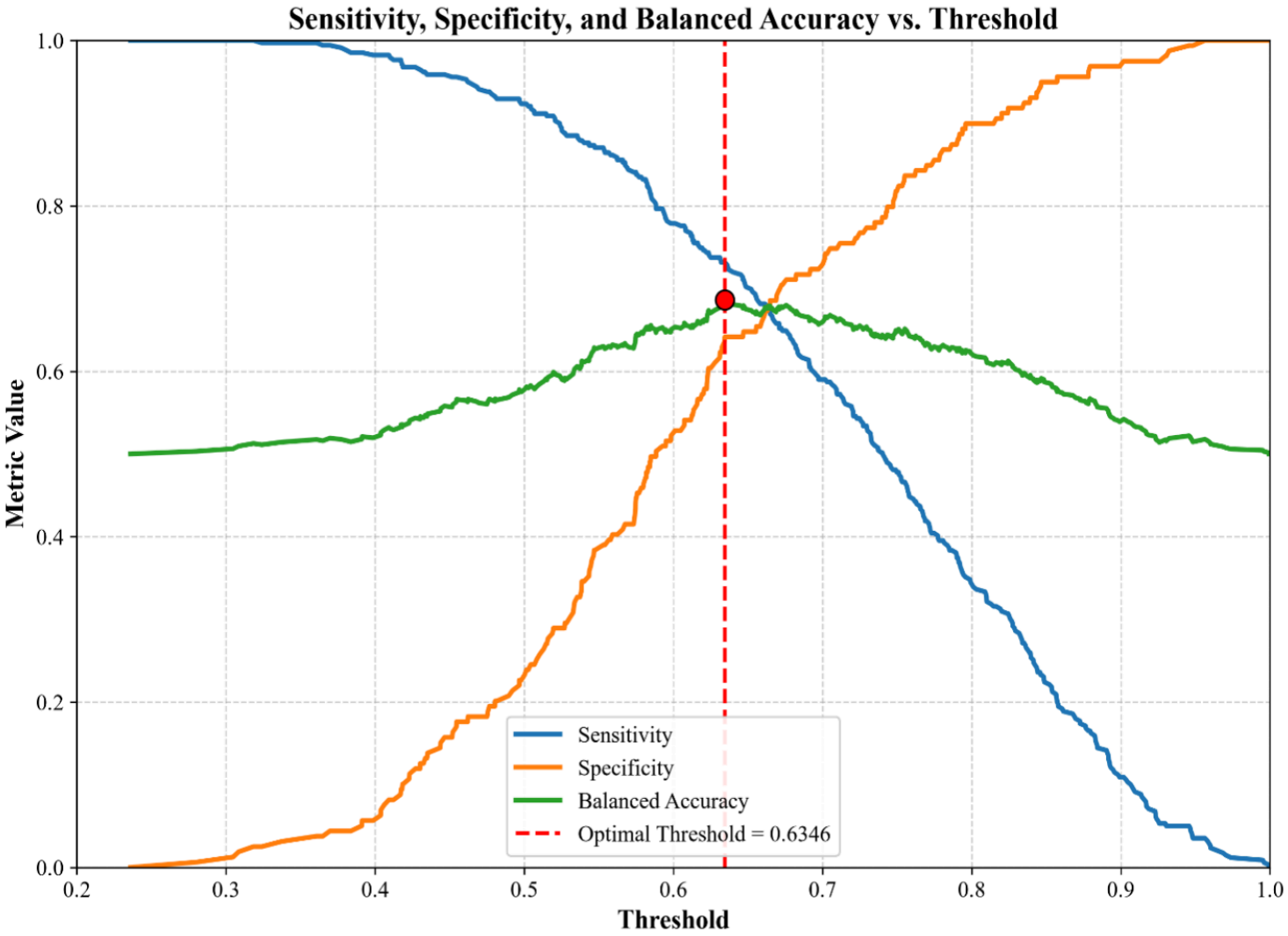


Figure 6

Sensitivity, specificity, and balanced accuracy as a function of the decision threshold.

The curves were calculated from the development-set out-of-fold predictions of the selected Indicator × SVM workflow. The locked threshold of 0.6346 corresponded to the maximum balanced accuracy under the prespecified selection rule.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1SupplementaryInformation.docx