International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 792-807
https://doi.org/10.70917/ijcisim-2026-3650
Article

A Feature-Optimized Hybrid Supervised Ensemble Learning Framework for Early Heart Attack Prediction Using Artificial Intelligence and Machine Learning

Piyush A Patel[1], Tulsidas V Nakrani[2]

[1,2]Department of Computer Application MCA, Sankalchand Patel College of Engineering, Sankalchand Patel University, Gujarat, India

Abstract: Background: While machine-learning models capture the complex interactions of clinical variables, small cardiovascular datasets may exhibit optimistic validation, data leakage, and mislabeling of outcomes.
Objective: This study sought to implement and assess a leakage-controlled, hybrid, supervised ensemble to classify the presence of coronary heart disease in the UCI Cleveland cohort. Even though the proposed title uses "heart attack prediction," the endpoint of this study is the determination of prevalent coronary disease at evaluation, and not the prediction of a future myocardial infarction.
Methods: The Cleveland dataset housed in the public domain consists of 303 samples and 13 features. The original 0–4 categorical variable of diagnostic severity was transformed into a binary variable with 0 for no disease and 1 for any diagnosis (1–4). The median and mode were imputed. Standardization and one-hot encoding were fitted only to training folds. Logistic regression, support vector machines, random forests, XGBoost, and LightGBM were optimized through stratified five-fold cross-validation. Optimized, weighted soft voting and logistic-regression stacking utilized out-of-fold probabilities. Evaluation of performance was conducted with an untouched 20% test set, and through two iterations of five-fold cross-validation. Discrimination, calibration, and uncertainty bootstrap were assessed, along with confusion-matrix metrics, paired tests, SHAP analysis, and permutation importance.
Results: The cohort consisted of 164 observations without disease, 139 with disease, and had six missing observations. From the 61-observation test set, stacking attained an accuracy of 91.8% with sensitivity of 92.9% and specificity of 90.9%, with an F1 score of 0.912, an MCC score of 0.836, and ROC-AUC of 0.960. This resulted in 30 true negatives, 3 false positives, 2 false negatives, and 26 true positives. For weighted voting, Brier score and ROC-AUC showed the best results with 0.0849 and 0.961, respectively. When comparing the rest of the models, the greatest mean ROC-AUC (0.898±0.042) was attained with weighted voting, but results of the seven models were comparable (Friedman p=0.628). The most important variables for this dataset were the number of major vessels, the chest-pain type, and the thallium-test type, and the induced angina and the ST depression and the maximum heart rate.
Conclusion: The use of heterogeneous ensembles on this small dataset proved discrimination and clinically relevant sensitivities. However, tuned single models have not shown any statistically relevant results. Before
Keywords: — heart-attack prediction; coronary heart disease; cardiovascular risk; artificial intelligence; machine learning; ensemble learning; stacking classifier; explainable AI; clinical decision support.

1. Introduction

Cardiovascular disease is still considered the leading cause of premature death and the most significant cause of burden on health services. Within the spectrum of cardiovascular disease, coronary atherosclerosis and acute myocardial infarction are the most prominent contributors [1]. Modern recommendations on prevention and management of chest pain have therefore centered around structured assessment of risk, testing, and intervention [2],

[3]. Existing systems such as SCORE2 and the PREVENT system of the American Heart Association, utilize specified longitudinal endpoints to estimate the risk of future events for a defined population [4], [5]. These systems exemplify one of the most fundamental constructs of clinical prediction systems. The model targets the prediction horizon, the population to which the system will apply, and the context within which the model will be used, must all be clearly defined. Even if the constituents of coronary disease are clinically associated, a model that is built on cross-sectional data of coronary angiography cannot be described as validated forecast of possible future myocardial infarction. There is also significant clinical value in differentiating between myocardial injury, infarction, and stable coronary disease and between myocardial infarction and overall cardiovascular risk [6].

Conventional modeling can be improved upon by artificial intelligence (AI) and machine learning (ML) by building into models the feature interactions and the non-linearities and complexities of high-dimensional datasets. The meta-analyses that are at our disposal have shown promising discrimination in many cardiovascular contexts with support-vector and boosting algorithms, yet significant heterogeneity exists in the study populations, endpoints, and especially the quality of the validations (9). Common observations in the literature regarding the reviews have included small and single center samples, insufficient representation of the social and ethnic demographics, large (or even full) absence of methodological description of missing data and very poor external validations (10), (13), (14). In addition to the aforementioned problems with small benchmark datasets, a serious risk of analytical leakage is present. If the test set is used in any way (e.g., feature selection, data imputation, data rescaling, data preparation for sampling and probability calibration), there can be an illusory inflation of the performance of the models. The only acceptable way to solve this problem is to completely integrate all methods of learning into the cross-validation procedures that guarantee the model performance, and leave the test set completely intact.

Building models with single classifiers provides a range of diverse inductive approaches. For example, logistic regression offers a parsimonious linear model of probability, whereas a support vector machine offers a model with a non-linear decision boundary that is implemented via a kernel function. On the other hand, random forests introduce a non-linear model with a high degree of accuracy via an ensemble of decorrelated decision trees, and both XG Boost and Light GBM provide models that capture interactions and non-linearities of a step function; however, all of the approaches model the residual error in a sequential way. All of the aforementioned techniques have their limitations. The interaction and the variety of the aforementioned approaches are important for the practice of ensemble learning. For example, in an ensemble of models built upon probabilistic outcomes, soft voting achieves the desired outcome, while stacking builds a model of meta-learning that is derived from the out-of-ensemble predictions (11), (12), (13). Nevertheless, it is important to remember that not all diverse and ensemble models are better than uncorrelated models. For example, an undersized meta-learning set combined with base learner models that are redundant and learned with too much precision

Equally important is post hoc interpretation. Transformations can be applied to features, and several methods can account for these transformations. For example, SHAP values describe how features affect a tree model after transformations, and permutation importance can measure the change to predictive power and discrimination if a variable is altered [47], [48]. These methods are beneficial for error analysis, and clinical plausibility but do not address causation. For example, the proximity of a variable to the diagnostic process may be the reason a variable, such as a variable from an angiographic process or a variable from an exercise test, is a strong predictor. This would be the case if changing the variable would not change the risk of a particular disease. Therefore, expectations and data generation process would be tied for a particular explanation.

This implementation study aims to address five distinct questions. First, how accurately do supervised models classify the UCI Cleveland coronary disease endpoint? Second, does heterogeneous ensemble methods surpass tuned base learners? Third, which of the input variables affect the prediction, and how does the explanation make an impact? Lastly, do imbalance addressing methods, optimization, and calibration impact the prediction and explanation? What is the explanation is necessary for research stage decision support? From these questions, the authors identified four hypotheses. These hypotheses stated that heterogeneous ensemble methods reach greater classification performance (H1), that better classification performance is reflected as better sensitivity, F1 score, ROC-AUC, and Matthews Correlation Coefficient (H2), that clinically relevant variables associated with exercise, symptoms, and vessels would be impactful to the prediction (H3), and that class-based manipulation with optimization would enhance the detection of the minority class and cause minimal overfitting (H4). The main contributor of this study was a comparison that was fully traceable and utilized several methodologies including, but not limited to, fold-contained preprocessing, out-of-fold ensemble building, a completely untouched holdout sample, repeated cross-validation, calibration, and paired statistical testing and utilized two distinctly different explanatory methodologies.

2. Materials and Methodology

2.1 Research Design

This study adopted a quantitative, computational, predictive-modelling approach. The workflow was built around five supervised classifiers, an optimally tuned weighted soft-voting ensemble, and a logistic-regression stacking ensemble. The analysis was built around TRIPOD+AI for predictive-model reporting and PROBAST+AI for risk-of-bias assessment [35],[36]. The primary analysis was conducted with a stratified 80:20 train-test split, with all model selection performed in the training partition. The analysis was reinforced by repeated stratified cross-validation on the training data. The design captured performance within the system, rather than the potential for the model to be generalized to other hospitals, or the prediction of future events.

2.2 Dataset, Participants, and Clinical Endpoint

The dataset of interest, heart disease dataset, was accessed from the University of California, Irvine Machine Learning Repository (UCI) [7]. The UCI repository, with its combination of four original databases, has the Cleveland subset, the most analyzed dataset which contains 303 records. The current work benefited from the available processed variables from Cleveland and the descriptions provided by Detrano et al. [8]. For the remaining records, thirteen variables were collected. In the original dataset, the diagnosis variable had a 0 to 4 scale. A score of 0 meant no evidence of the disease, whereas scores of 1-4 meant evidence of the disease, which, as per the repository description, meant a narrowing of a vessel by 50% or more of its diameter. In a binary sense, scores of 1-4 were given a 1, while a score of 0 was given a 0. The main clinical outcome of interest was coronary heart disease, and the study was not concerned with the prediction of a future myocardial Infarction.

Table 1. Dataset variables and operational definitions.

Variable	Type	Unit/coding	Clinical meaning
age	Numeric	years	Age at assessment
sex	Categorical	0=female; 1=male	Recorded sex
cp	Categorical	1–4	Chest-pain type: typical, atypical, non-anginal, asymptomatic
trestbps	Numeric	mm Hg	Resting systolic blood pressure
chol	Numeric	mg/dL	Serum cholesterol
fbs	Categorical	>120 mg/dL	Fasting blood-sugar indicator
restecg	Categorical	0–2	Resting electrocardiographic category
thalach	Numeric	beats/min	Maximum achieved heart rate
exang	Categorical	0=no; 1=yes	Exercise-induced angina
oldpeak	Numeric	mm	ST depression induced by exercise relative to rest
slope	Categorical	1–3	Slope of peak exercise ST segment
ca	Numeric/count	0–3	Major vessels colored by fluoroscopy
thal	Categorical	3, 6, 7	Normal, fixed defect, reversible defect
diagnosis	Outcome	0 vs 1–4	Binary absence/presence of angiographic coronary disease

All 303 records were included because there were no duplicates. The distribution of the classes was 164 negatives for the disease (54.1%) and 139 positive (45.9%). The only missing data was seen in two instances for thal

and four for ca. The repository provides de-identified secondary data under an open license. Because of this, there was no need to contact participants, and there was no need to carry out tasks to de-identify the data. Confirm the institutional requirements, though, for the analysis of secondary public data before submission.

2.3. Data Preprocessing and Leakage Control

Median imputation followed by standardization was used for the numeric variables (age, resting blood pressure, cholesterol, maximum heart rate, ST depression, and number of major vessels). Categorical variables underwent mode imputation followed by one-hot encoding. The dataset was split into 242 training records (131 negative, 111 positive) and 61 independent test records (33 negative, 28 positive) by stratification and a random seed of 42. The remaining test set was used for final reporting. In all operations, no duplicate removal, imputation statistics, scaler, encoder, resampling methods, thresholds, weights for assembling, or calibration methods were influenced by the test set labels.

Table 2. Data-preprocessing and feature-transformation procedures.

Stage	Implementation	Result/rationale
Duplicates	Exact-row inspection	No duplicates identified
Missing numeric values	Median imputation within fold	ca: 4 missing
Missing categorical values	Most-frequent imputation within fold	thal: 2 missing
Numeric scale	StandardScaler within fold	Applied to 6 numeric predictors
Categorical representation	One-hot encoding	Applied to 7 categorical predictors
Split	Stratified 80:20	242 training; 61 tests
Feature selection	Clinical retention + diagnostic review	All 13 original predictors retained
Leakage prevention	Pipeline and OOF predictions	Test set untouched until final evaluation

2.4 Class Imbalance and Feature Review

The outcome imbalance was mild. Original tuned pipelines were compared with class weighting and the application of SMOTE only in training folds [45]. Logistic-regression SMOTE improved repeated-fold MCC from 0.675 to 0.683. In contrast, support-vector class weighting and SMOTE left the original tuned configuration unchanged. Oversampling provided no significant improvement, and in a small dataset oversampling can cause more variance. Therefore, the final ensembles employed the tuned base pipelines, and were not forced to implement SMOTE. All 13 predictors were included. With the small sample size, extremely selective approaches would be unstable and may remove context that is clinically relevant. Diagnostic interpretation relied more on correlation, mutual information, SHAP, and permutation analyses rather than pruning based on the test set.

2.5 Base Classifiers and Hybrid Ensemble Formulation

The tested models consisted of logistic regression, radial-basis function support vector classification, random forest, XGBoost, and LightGBM. These models feature the linear, margin-based, bagging, and gradient-boosting families [41]–[44]. Each model provided an estimate of p_m(y=1|x_i). Let D={ (x_i,y_i)} be the data set for i=1,…,N, where the input feature x_i∈R^p and the output label y_i∈{0,1} indicates the absence/presence of coronary disease. Given M base classifiers, weighted voting was defined as:

$$P(y=1|x_i) = [\Sigma_{m=1}^{M} w_m\, p_m(y=1|x_i)] / [\Sigma_{m=1}^{M} w_m], \quad w_m \geq 0, \quad \Sigma_{m=1}^{M} w_m = 1 \qquad (1)$$

Voting weights defined were non-negative and were determined based on training out-of-fold probabilities. The calibrated final voting weights were set to 0.6291 for logistic regression, 0.2773 for SVM, 0.0050 for random forest, 0.0886 for XGBoost, and 0.000005 for LightGBM. Out-of-fold was fit to sigmoid calibration. The threshold τ=0.598 was applied, and it maximized the balanced accuracy of the training out-of-fold:

$\hat{y}_i = 1$ when $P(y=1|x_i) \geq \tau$; otherwise $\hat{y}_i = 0$ (2)

In stacking, the meta feature vector z_i=[p_1(x_i),…,p_M(x_i)] was constructed exclusively of out-of-fold base probabilities. This was estimated by the L2-regularized, class-weighted logistic function:

$P(y=1|z_i)= \sigma (\beta_0 + \Sigma_{m=1}^{M} \beta_m z_{im})$, $\sigma(a)=1/(1+e^{-a})$

2.6 Hyperparameter Optimisation and Experimental Environment

In the case of randomised hyperparameter search, the procedure consisted of five-fold stratified cross-validation,

The experiments take place in an environment with these specifications: OS: Linux, Python version: 3.13.5, scikit-learn: 1.8.0, XGBoost: 3.1.3, LightGBM 4.6.0, imbalanced-learn: 0.14.1, SHAP: 0.50.0, NumPy: 2.3.5, pandas: 2.2.3, Matplotlib: 3.10.8, SciPy: 1.17.0. The host has 56 logical processors and 4 GB of RAM. All the splits, along with the stochastic estimators, used random state 42 to ensure replicability, which still retains sampling uncertainty.

Table 3. Hyperparameter optimization results.

Model	Best CV AUC	Selected parameters
Logistic Regression	0.908	solver=liblinear; penalty=l2; class weight=balanced; C=0.2848035868435802
Support Vector Machine	0.908	kernel=rbf; gamma=0.001; class_weight=None; C=21.54434690031882
Random Forest	0.899	n_estimators=500; min_samples_split=8; min_samples_leaf=4; max_features=sqrt; max_depth=8; class_weight=balanced
XGBoost	0.898	subsample=0.75; scale_pos_weight=1.18; reg_lambda=2; reg_alpha=0; n_estimators=80; min_child_weight=5; max_depth=4; learning_rate=0.05; colsample_bytree=1.0
LightGBM	0.888	subsample=0.75; reg_lambda=0.5; reg_alpha=0; num_leaves=31; n_estimators=80; min_child_samples=10; max_depth=3; learning_rate=0.02; colsample_bytree=0.75; class_weight=balanced

2.7 Evaluation, Calibration, Statistics, and Explainability

The evaluation metrics were accuracy, precision, sensitivity (recall), specificity, negative predictive value, F1 score, the area under the ROC curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), Matthew's correlation coefficient (MCC), balanced accuracy, Brier score, training time, and inference time. The confusion matrix contained the notations of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The core definitions and equations are as follows:

Accuracy=(TP+TN)/(TP+TN+FP+FN); Sensitivity=TP/(TP+FN) (5)

Specificity=TN/(TN+FP); Precision=TP/(TP+FP) (6)

F1=2·Precision·Recall/(Precision Recall) (7)

MCC=(TP·TN−FP·FN)/√[(TP+FP) (TP+FN) (TN+FP) (TN+FN)] (8)

ROC-AUC focused on rank discrimination, while PR-AUC focused on positive-class retrieval, and Brier score focused on the accuracy of probabilities. Sigmoid calibration was assessed for weighted voting [38], [54]. Two repeats of five-fold stratified cross-validation resulted in ten paired fold observations for each model. Friedman test on ROC-AUC followed by Wilcoxon signed-rank tests with Holm correction assessed all differences. Stratified bootstrap confidence intervals were used to evaluate weighted voting uncertainty. A paired McNemar test compared classification errors of the strongest base model. SHAP was used to

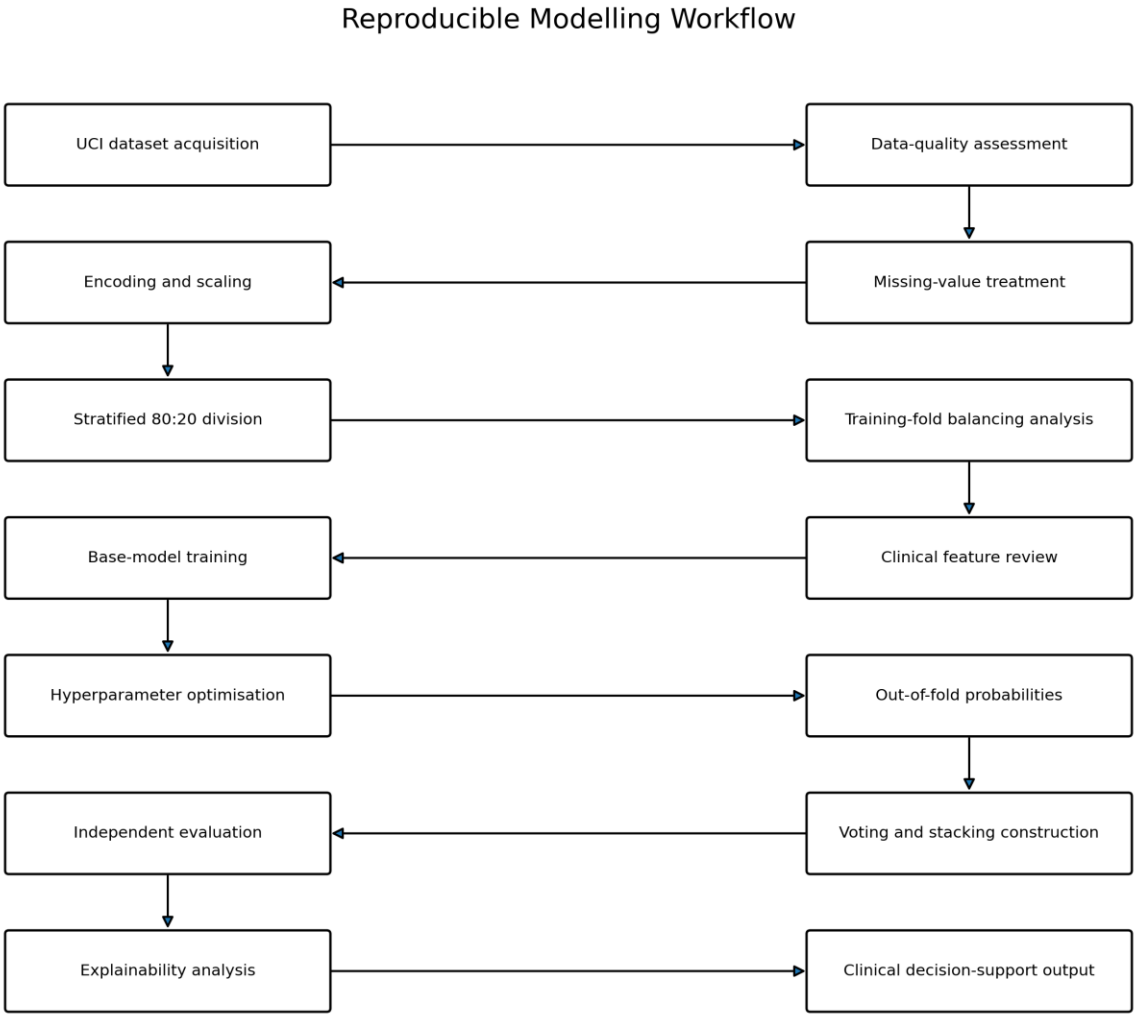

Figure 1. Leakage-controlled research workflow from public-data acquisition to calibrated coronary-disease classification.

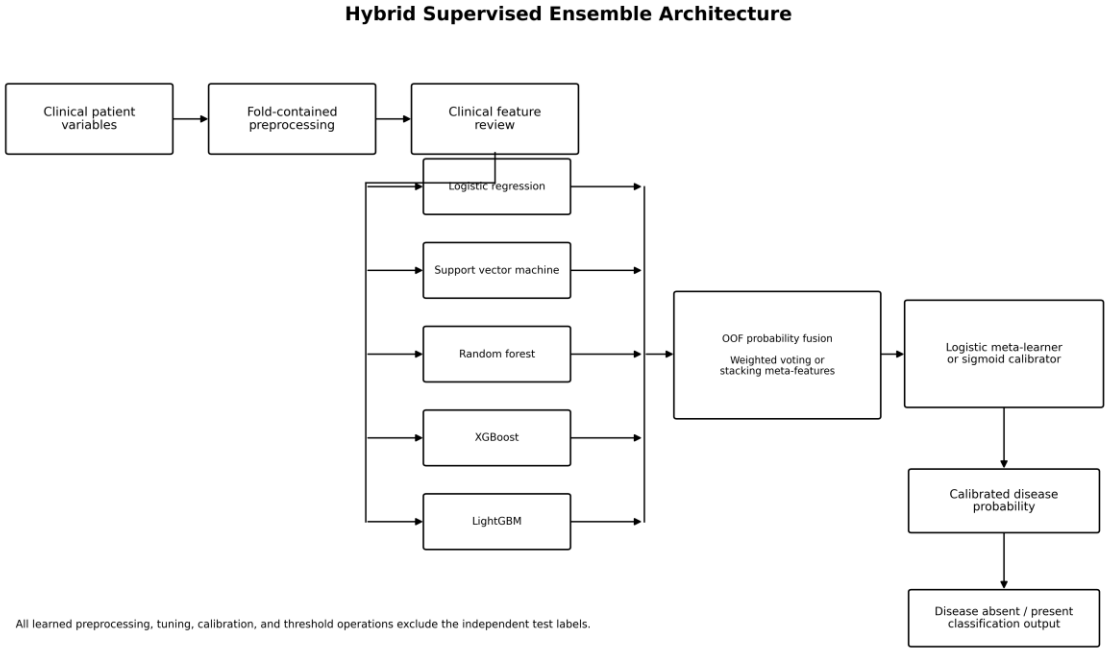

Figure 2. Hybrid supervised ensemble architecture. OOF denotes out-of-fold probability generation.

3. Results

3.1 Dataset Characteristics

There were 303 total records with 164 records that did not have and 139 records that did have angiographically defined coronary disease. Missingness impacted only 2.0% of records. It was resolved within the relevant training folds. Participants that had disease were on average older and their average age was 56.63 compared to 52.59 for participants without the disease. Disease positive participants also had a mean resting blood pressure of 134.57 and non-disease participants had a mean resting blood pressure of 129.25 mm Hg. Disease positive participants also had a lower maximum heart rate of 139.26 beats/min compared with 158.38 beats/min, and greater exercise induced ST depression of 1.57 compared to 0.59 mm. In the disease group the mean number of major vessels was 1.14, and in the disease negative group the mean was 0.27. Asymptomatic chest pain coding, exercise induced angina, and reversible thallium defects were also more in the positive class of the study. These associations are solely descriptive.

Table 4. Selected descriptive characteristics of the analytical cohort.

Characteristic	No disease	Disease
Age, years	52.59±9.51	56.63±7.94
Resting BP, mm Hg	129.25±16.20	134.57±18.77
Cholesterol, mg/dL	242.64±53.46	251.47±49.49
Maximum heart rate	158.38±19.20	139.26±22.59
ST depression	0.59±0.78	1.57±1.30
Major vessels, n	0.27±0.63 (n=161)	1.14±1.02 (n=138)
Male sex	92/164 (56.1%)	114/139 (82.0%)
Asymptomatic chest pain	39/164 (23.8%)	105/139 (75.5%)
Exercise angina	23/164 (14.0%)	76/139 (54.7%)
Reversible thal defect	28/163 (17.2%)	89/138 (64.5%)

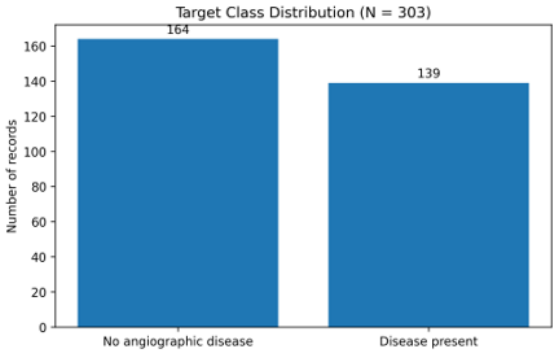

Figure 3. Target-class distribution.

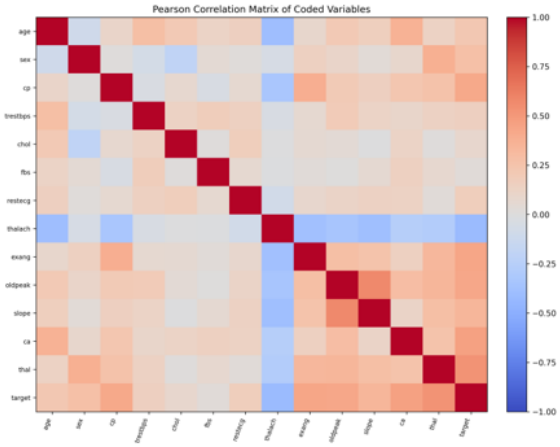

Figure 4. Numeric-feature correlation matrix.

3.2. Repeated Cross-Validation Performance

soft voting achieved the best mean ROC-AUC (0.898±0.042) and mean PR-AUC (0.899±0.042) scores in repeated cross-validation. Weighted soft voting was followed in order by logistic regression (ROC-AUC 0.895±0.043), SVM (0.894±0.049), stacking (0.892±0.039), random forest (0.891±0.047), XGBoost (0.889±0.036), and LightGBM (0.874±0.038). Logistic regression had the highest mean balanced accuracy (0.817), followed closely by weighted voting with 0.815. Therefore, in regards to means, the ensemble method had a very small rank advantage from the metric chosen and was fair compared to fold-to-fold variation.

Table 5. Repeated stratified cross-validation performance (mean±SD across 10 paired folds).

Model	ROC-AUC	PR-AUC	Sensitivity	Specificity	F1	MCC	Brier
LightGBM	0.874±0.038	0.886±0.036	0.756±0.097	0.801±0.073	0.758±0.057	0.563±0.091	0.151±0.016
Logistic Regression	0.895±0.043	0.892±0.048	0.788±0.089	0.847±0.094	0.801±0.070	0.642±0.128	0.131±0.030
Random Forest	0.891±0.047	0.895±0.043	0.774±0.088	0.813±0.083	0.776±0.061	0.592±0.109	0.138±0.022
Stacking Ensemble	0.892±0.039	0.894±0.039	0.797±0.072	0.809±0.087	0.789±0.054	0.609±0.102	0.136±0.029
Support Vector Machine	0.894±0.049	0.895±0.048	0.783±0.086	0.839±0.081	0.794±0.064	0.629±0.115	0.132±0.030
Weighted Soft Voting	0.898±0.042	0.899±0.042	0.788±0.089	0.843±0.102	0.799±0.071	0.639±0.131	0.130±0.028
XGBoost	0.889±0.036	0.895±0.034	0.784±0.079	0.824±0.074	0.787±0.048	0.613±0.085	0.137±0.021

3.3 Independent Test Performance

On the untouched 61-record test set, stacking ensemble reported the highest scores of classification accuracy (0.918), balanced accuracy (0.919), F1-score (0.912), and MCC (0.836). Stacking ensemble identified 26 of the 28 positive records and 30 out of the 33 negative records. Weighted voting and random forest, as well as stacking ensemble, recorded the same results, that is 28 TN, 5 FP, 2 FN, and 26 TP. After stacking ensemble, weighted voting achieved the highest ROC-AUC score, which was 0.961, and the lowest Brier score, which was 0.0849. It was concluded that weighted voting achieved slightly better ranking and probability accuracy compared to stacking ensemble. However, stacking ensemble achieved better thresholder classification. Logistic regression and SVM missed two positive cases and generated six false positives. False positives were reduced to four by both XGBoost and LightGBM, with the tradeoff of one additional false negative.

Table 6. Independent test-set performance.

Model	Acc.	Prec.	Sens.	Spec.	F1	ROC-AUC	PR-AUC	MCC	Brier	Train s	Infer ms
Logistic Regression	0.869	0.812	0.929	0.818	0.867	0.958	0.941	0.745	0.089	0.76	2.85
Support Vector Machine	0.869	0.812	0.929	0.818	0.867	0.957	0.946	0.745	0.089	2.06	3.24
Random Forest	0.885	0.839	0.929	0.848	0.881	0.949	0.931	0.775	0.104	15.46	22.75
XGBoost	0.885	0.862	0.893	0.879	0.877	0.955	0.949	0.770	0.093	3.12	3.71
LightGBM	0.885	0.862	0.893	0.879	0.877	0.946	0.940	0.770	0.111	3.63	3.41
Weighted Soft Voting	0.885	0.839	0.929	0.848	0.881	0.961	0.947	0.775	0.085	1.13	35.97
Stacking Ensemble	0.918	0.897	0.929	0.909	0.912	0.960	0.943	0.836	0.089	7.00	41.25

Table 7. Independent test-set confusion matrices and negative predictive value.

Model	Threshold	TN	FP	FN	TP	NPV
Logistic Regression	0.500	27	6	2	26	0.931
Support Vector Machine	0.500	27	6	2	26	0.931
Random Forest	0.500	28	5	2	26	0.933
XGBoost	0.500	29	4	3	25	0.906
LightGBM	0.500	29	4	3	25	0.906
Weighted Soft Voting	0.598	28	5	2	26	0.933
Stacking Ensemble	0.663	30	3	2	26	0.938

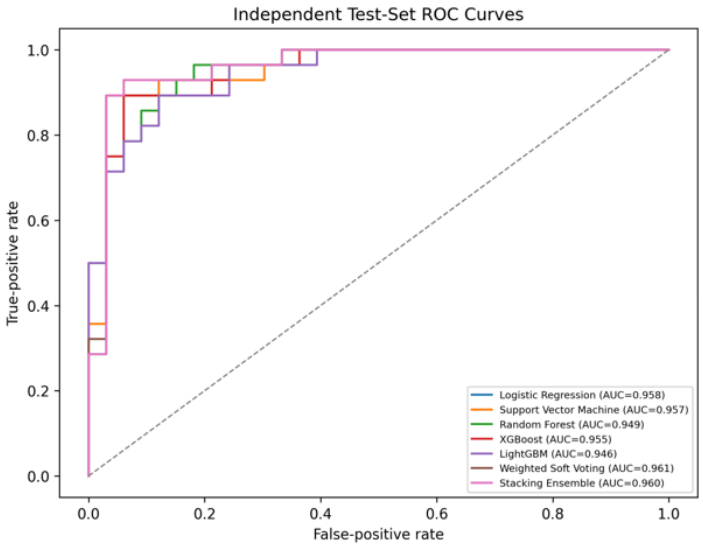

Figure 5. Test-set ROC curves

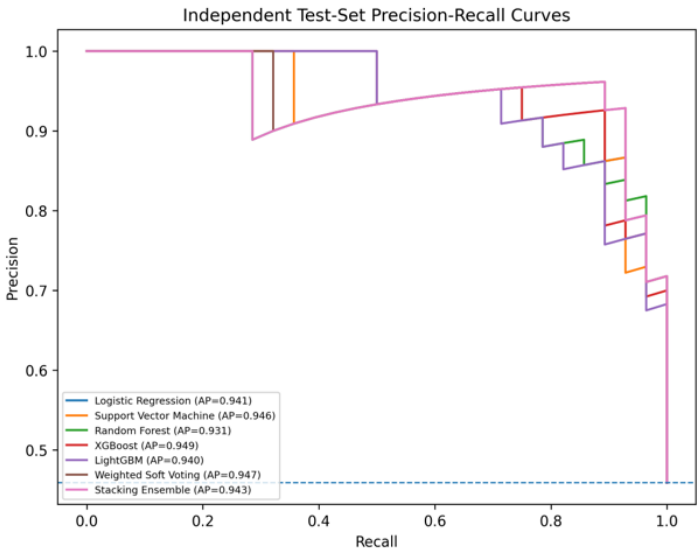

Figure 6. Test-set precision–recall curves

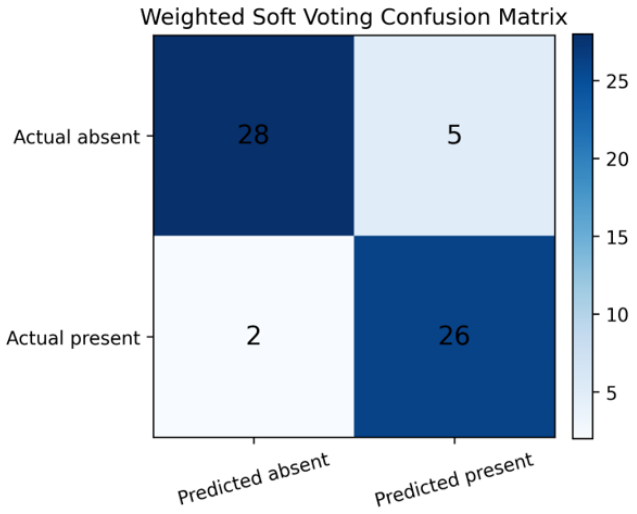

Figure 7a. Weighted-voting confusion matrix.

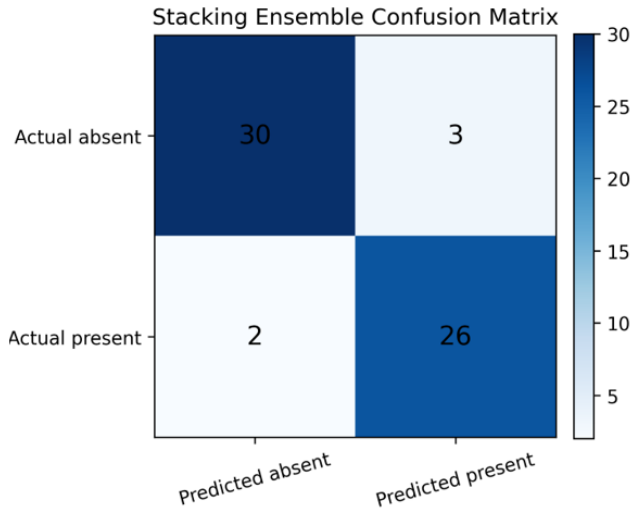

Figure 7b. Stacking confusion matrix.

3.4 Calibration and Statistical Comparison

Without changing the ROC-AUC, as predicted, there were no changes in ranking, so Sigmoid calibration decreased the weighted-voting Brier score from 0.0878 to 0.0849. Its bootstrap 95% confidence interval yielded a range of 0.904 to 1.000 for ROC-AUC and 0.802 to 0.958 for balanced accuracy. The global repeated fold comparison was not statistically significant (Friedman χ^2=4.364, p=0.628). All Holm adjusted pairwise p-values for weighted voting and the remaining models were 1.000. For weighted voting and logistic regression, the AUC difference was estimated to be in the bootstrap 95% confidence interval of about 0.000 to 0.011, two-sided bootstrap p=0.438. There was also no observed difference between paired test classifications in the McNemar (p=1.000). Favorable ensemble points estimate notwithstanding, H1 and H2 were not statistically confirmed.

Table 8. Statistical comparison and uncertainty analysis.

Analysis	Estimate	Interpretation
Friedman test, repeated CV ROC-AUC	χ^2=4.364; p=0.628	No global model difference
Weighted voting ROC-AUC bootstrap CI	0.904–1.000	Strong but imprecise discrimination
Weighted voting balanced-accuracy CI	0.802–0.958	Sampling uncertainty remains
AUC difference: voting − logistic regression	95% CI≈0.000–0.011; p=0.438	No significant superiority
McNemar: voting vs logistic regression	table [[53,1],[0,7]]; p=1.000	No paired error difference
Calibration effect	Brier 0.0878→0.0849	Small improvement after sigmoid calibration

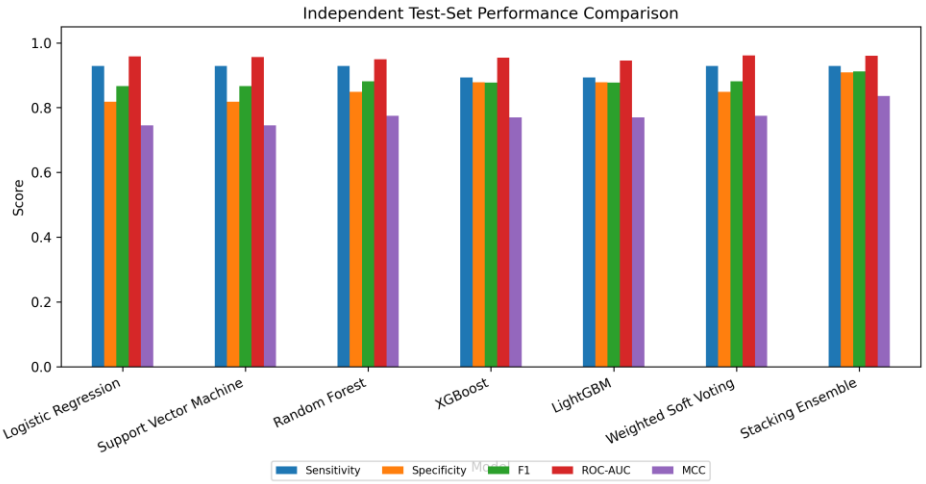

Figure 8. Test-set performance comparison

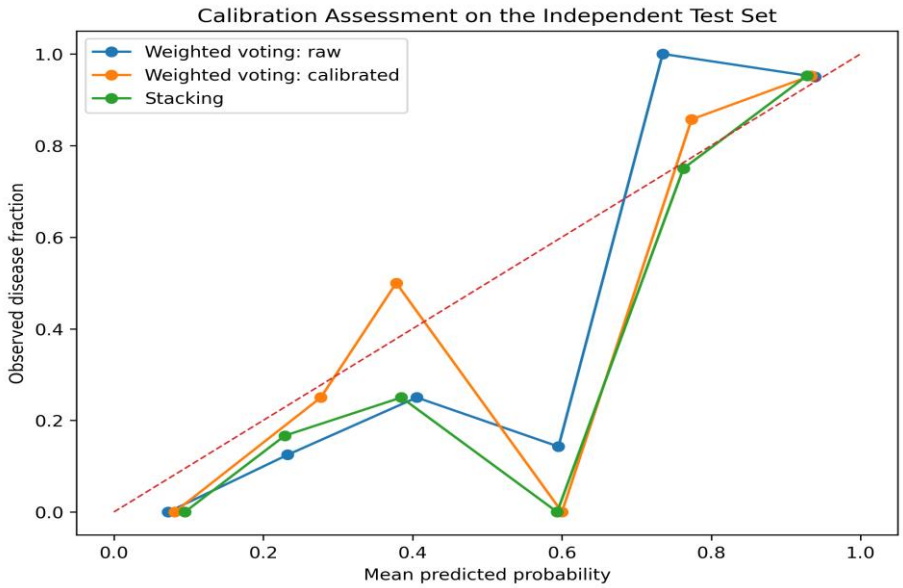

Figure 9. Calibration curves and Brier scores

3.5 Explainability Findings

According to SHAP analysis for the tuned XGBoost model, the number of major vessels (average absolute SHAP = 0.668) was the most important feature, followed by coding for asymptomatic chest pain (0.540), normal thallium test (0.439), female (0.265), reversible thallium defect (0.245), ST depression (0.217), age (0.177), flat ST slope (0.162), max heart rate (0.150), and cholesterol (0.128). The meaning of one-hot coded variables (as used in XGBoost) and the impact of these variables, are defined by the feature combinations and reference coding. Permutation analysis applied to the complete stacking pipeline provided a consistent and coarse ordering of the features, showing that characteristic a caused the highest mean ROC-AUC decrease at (0.074). Other notable features include type of chest pain (0.038), exercise angina (0.014), ST depression (0.013), thallium (0.013), and heart rate (0.011). These results are in favor of H3. The effects of age and fasting blood sugar were close to zero, and slightly negative in the test sample. These results should be interpreted as a lack of reliability, rather than evidence of no clinical relevance.

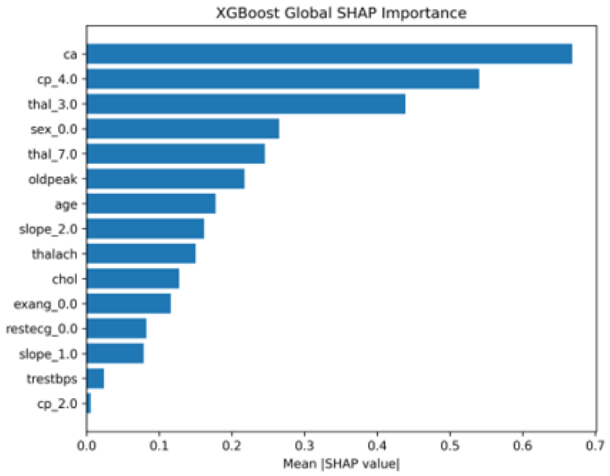

Figure 10. XGBoost global SHAP importance

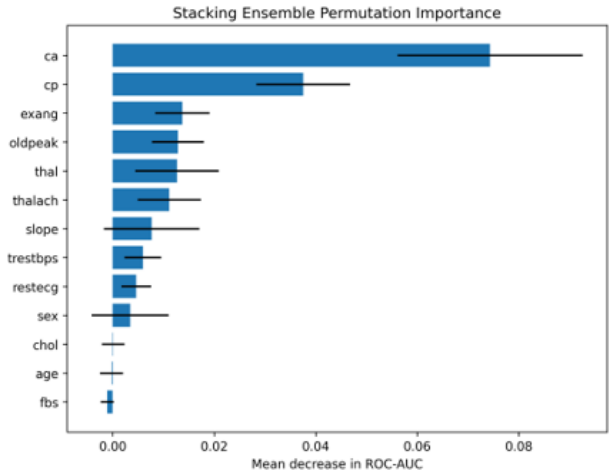

Figure 11. Stacking permutation importance.

4. Discussion

This study evaluated heterogeneous supervised learners using an internal-validation design. The best thresholder holdout result was achieved through stacking, which classified 56 out of 61 cases, with a sensitivity of 92.9%. Comparing methods using weighted voting achieved the highest ROC-AUC and Brier score. These results address the first of the posed questions. On the Cleveland benchmark, the subjected supervised methods demonstrated the ability to classify the repository's angiographic disease subcategories with a high degree of internal discrimination. These results, however, do not address the ability of the same model to predict a future acute coronary event, to work reliably on the current population, or to affect the clinical work result in a positive way.

The second question considered whether the superiority of ensembles could be demonstrated. The evidence was insufficient, despite the favorability of point estimates. Weighted voting only marginally improved holdout ROC-AUC compared to logistic regression, and repeated-fold rankings were inconsistent. The Friedman test, Holm-adjusted comparisons, the bootstrap AUC difference, and the McNemar test showed no significance. Thus, the evidence available does not support the claim that the hybrid ensemble is superior to all individually tuned base learners. This is in line with findings that, while machine learning has the potential to substantially improve discrimination in large, complex data-rich cohorts, its application in low-dimensional clinical datasets may have little or no impact over well-defined regression. In this regard, ensembles are particularly important because they yield different operating points, and while stacking addressed specificity and improved the MCC, calibrated voting improved probability accuracy.

Not only does test sensitivity matter because of false negatives, classifying false negatives for coronary-disease may lead to dangerous delays to further testing. Stacking and voting both resulted in two false negatives, but stacking also decreased the false positive rate from five to three. However, maximizing balanced accuracy when setting a classification threshold does not imply that a clinical utility threshold has been reached. Sensitivity would be prioritized in a screening application, and a long and costly pathway to testing would shift the balance in favor of a clinical utility threshold for specificity. Operational thresholds would require decision-curve analysis, outcome-specific costs, a reasonable impact on the expected workflow, and the override of clinicians to balance competing interests. Also, after transport, the implicit cost of measuring disease and the shifting patterns of referral, measurement, and coding, would require recalibration [38], [55].

The results of the explanation were diagnostic in nature and were therefore very contextually relevant. A lack of normative data was shown, along with Chest Pains, Exercise Angina, ST Depression, Thallium, and Max Heart Rate (provisional) for these assessments. Therefore, none of these would be contextually relevant as they were part of specialized assessments. Without these at the intended point of decision, the model would not be a low-cost community risk assessment. SHAP and Permutations show the signals that were used, and reveal weak/unstable contributions. They answered the 3rd and 5th research questions, respectively. However, neither transforms the model into a true causal system, nor does stability testing of contributions rationalize showing these to clinicians or patients.

The fourth question dealt with class management, tuning, and calibration. The 54:46 split fell short of being critically unbalanced. Training-fold SMOTE provided a small logistic-regression MCC boost, with no consistent benefit across models; support-vector alternatives had a minor negative impact. Therefore, H4 was only partially

confirmed. Hyperparameter tuning and fold-contained processing were crucial, while voting Brier score improvement due to calibration confirmed that resampling was not required for the final model. This shows that the rationale behind the evaluation of imbalance techniques is far more critical than the implementation, as in small tabular datasets, the synthetic neighbor's methodology can prove to be noisy, and distribute new values across the feature space. Recent studies have built off or compared to the framework provided here, but the comparison is problematic as studies present different versions of the 'heart disease' data, different types of target encodings, and validation that is not truly independent. For example, Tiwari and others have a version that reports 92.34% accuracy for a stacked ensemble on a larger dataset [11]. Other versions of the study present accuracy improvements stemming from feature engineering, hybrid learning, and explainable systems [15]–[19], [24]–[27]. Such accuracy is not comparable across disparate implementations, dataset splits, preprocessing steps, feature sets, endpoints, and confidence intervals. Systematic reviews continue to show inconsistent outcomes and a high risk of bias for clinical machine-learning prediction models [13], [14], [37]. For this reason, the current study prioritizes traceability of outcomes over stating a benchmark.

There are three aspects that need to be distinguished in clinical translation. The first is the statistical performance of a sample when a judgment is made. The second is the predictive validity or the degree to which the behavior is expected to remain similar in the target population. The last is the clinical utility of the model where value is given to a model through the expected improvement in outcomes and efficiencies or the promotion of equity through the decisions aided by the model. Also, external generalizability means that performance must be shown across different institutions, times, sub-populations, and measures, and is expected beyond a high internal AUC. TRIPOD+AI, PROBAST+AI, CONSORT-AI, SPIRIT-AI, and early DECIDE-AI stage guidance provide steps for design, reporting, and evaluation [33]-[37]. A proactive model is needed to monitor data drift, calibration, subgroup error rates, fidelity of offered explanations, and the risk of automation bias.

Table 9. Contextual comparison with selected recent literature.

Study	Data/design	Approach	Comparison with present study
Krittanawong et al., 2020 [9]	Meta-analysis; 103 cohorts	Multiple ML families	Promising AUCs but marked heterogeneity
Zhao et al., 2021 [10]	Systematic review	ML with social determinants	Representation and fairness often incomplete
Tiwari et al., 2022 [11]	Combined heart-disease dataset	Stacked ensemble	Reported 92.34% accuracy; dataset differs
Ishaq et al., 2021 [16]	Heart-failure survival	Tree ensemble + SHAP	Illustrates explainable ensemble workflow
Subramani et al., 2023 [17]	Cardiovascular classification	ML comparison	Supports optimization and multiple metrics
Mahajan et al., 2023 [12]	Disease-prediction review	Bagging, boosting, stacking	Ensembles can improve robustness but require diversity
Teshale et al., 2024 [13]	Systematic review	AI CVD risk models	External validation and reporting remain limited
Present study	UCI Cleveland; n=303	Voting + stacking	Leakage-controlled; test AUC 0.961; no significant CV superiority

5. Conclusion

A reproducible hybrid supervised framework was created using logistic regression, SVM, random forest, XGBoost, LightGBM, stacking, and calibrated weighted voting for the UCI Cleveland coronary-heart-disease

endpoint. To avoid major data leakage, out-of-loop ensemble construction and fold-inclusive preprocessing were utilized. Having an unseen sample size of 61, the stacking method yielded, 91.8% accuracy; 92.9% sensitivity; 90.9% specificity; F1=0.912, and, MCC=0.836. Calibrated weighted voting resulted in an optimal ROC-AUC of 0.961 and the best Brier score. Major vessels, pain type, thallium, exercise-induced angina, ST depression, and maximum heart rate were the most influential variables. Although repeated cross-validation did not indicate a major variation between the methods, the statistical difference between the ensemble and logistic regression was not provided. Thus, the study confirmed the applicability of heterogeneous ensemble classification, but brought to question the need for heterogeneous ensemble classification in this case. The endpoint is a measure of the current state of angiographic disease and should not be seen as an indicator of an impending myocardial infarction. Before being put to use, it must undergo evaluation across multiple centers, prospective labeling of events, recent measurements, subgroup analysis, recalibration, clinical impact, and applicability assessment.

6. Limitations

This dataset is small and derived mainly from one Cleveland clinical source and is also historical. We expect some ref erral and spectrum bias. Important current predictors (e.g. longitudinal biomarkers, medications, renal function, imaging, socioeconomic factors, and repeated measures) are also missing. The endpoint is cross-sectional coronary disease instead of incident infarction. One holdout has a large degree of uncertainty. Cross-validation, however, is internal. A hyperparameter search and threshold selection will likely still overfit the training sample, even with an out-of-sample build. The thallium and vessel variables may limit generalizability prior to specialized testing. Race, ethnicity, and social determinants are absent, thus a fairness audit is not possible. Implementation of SHAP and permutation importance is sample-based and suggests an association rather than a cause. We also lack temporal, external, prospective, or clinical-impact validation.

7. Future Research Directions

In the future, multicenter longitudinal studies with prediction horizons and adjudicated myocardial-infarction or major-adverse-cardiovascular-event outcomes must be employed. Nested validation needs to be followed by geographic and temporal external validation. Federated learning may allow cross-institutional studies without centralizing records, but still would require harmonization and governance. Where possible, studies at the point of care need to integrate data from ECGs, medical imaging, cardiac biomarkers, renal function, medication histories, and wearables along with longitudinal electronic health records. Prior to undertaking prospective studies, there must be cost-sensitive approaches, decision-curve analyses, calibration of subgroups, fairness audits, uncertainty quantification, stability of explanation, and clinician-in-the-loop simulations. With an emphasis on patient-perceived outcomes and safety, post-marketing research needs to go beyond discrimination.

8. Reproducible Implementation Logic

1. Load the validated UCI Cleveland table; map Diagnosis 0→0 and 1–4→1.
2. Split dataset into 80% training and 20% untouched testing set (random_state = 42).
3. For numeric features, use Column Transformer with median + StandardScaler; for categorical features use
4. mode + OneHotEncoder.
5. Use Pipeline for all sklearn/imblearn classifiers; insert SMOTE after preprocessing only for the comparison
6. Pipeline.
7. Use StratifiedKFold (5-fold) inside the training set with RandomizedSearchCV and scoring='roc_auc'.
8. Compute out-of-fold probabilities with cross_val_predict (…, method='predict_proba') for each tuned base
9. learner.
10. Fit sigmoid calibration and tune only the threshold with training data; optimize non-negative soft-voting
11. weights on the out-of-fold (OOF) probabilities.
12. Build OOF meta-features and fit the logistic L2 model with stacking; choose the model's threshold from OOF
13. balanced accuracy.
14. Refit all base learners on the complete training set and compute independent test set probabilities.
15. Compute various metrics, bootstrap confidence intervals, and perform Friedman/Wilcoxon tests and the
16. McNemar test.
17. Compute SHAP with Tree Explainer for XGBoost and repeated permutation importance for the stacking
18. pipeline.

References

1. S. S. Martin et al., “2024 heart disease and Stroke Statistics: A Report of US and Global Data from the American Heart Association,” Circulation, vol. 149, pp. e347–e913, 2024, doi: 10.1161/CIR.0000000000001209.
2. F. L. J. Visseren et al., “2021 ESC Guidelines on cardiovascular disease prevention in clinical practice,” European Heart Journal, vol. 42, no. 34, pp. 3227–3337, 2021, doi: 10.1093/eurheartj/ehab484.
3. M. Gulati et al., “2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain,” Circulation, vol. 144, pp. e368–e454, 2021, doi: 10.1161/CIR.0000000000001029.
4. SCORE2 Working Group and ESC Cardiovascular Risk Collaboration, “SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe,” European Heart Journal, vol. 42, no. 25, pp. 2439–2454, 2021, doi: 10.1093/eurheartj/ehab309.
5. S. S. Khan et al., “Development and Validation of the American Heart Association’s PREVENT Equations,” Circulation, vol. 149, pp. 430–449, 2024, doi: 10.1161/CIRCULATIONAHA.123.067626.
6. K. Thygesen et al., “Fourth universal definition of myocardial infarction (2018),” European Heart Journal, vol. 40, no. 3, pp. 237–269, 2019, doi: 10.1093/eurheartj/ehy462.
7. A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, Heart Disease [Dataset], UCI Machine Learning Repository, 1989, doi: 10.24432/C52P4X.
8. R. Detrano et al., “International application of a new probability algorithm for the diagnosis of coronary artery disease,” American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989, doi: 10.1016/0002-9149(89)90524-9.
9. C. Krittanawong et al., “Machine learning prediction in cardiovascular diseases: a meta-analysis,” Scientific Reports, vol. 10, art. 16057, 2020, doi: 10.1038/s41598-020-72685-1.
10. Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, “Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review,” American Journal of Preventive Medicine, vol. 61, no. 4, pp. 596–605, 2021, doi: 10.1016/j.amepre.2021.04.016.
11. A. Tiwari, A. Chugh, and A. Sharma, “Ensemble framework for cardiovascular disease prediction,” Computers in Biology and Medicine, vol. 146, art. 105624, 2022, doi: 10.1016/j.compbiomed.2022.105624.
12. P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, “Ensemble Learning for Disease Prediction: A Review,” Healthcare, vol. 11, no. 12, art. 1808, 2023, doi: 10.3390/healthcare11121808.
13. A. B. Teshale et al., “A Systematic Review of Artificial Intelligence Models for Cardiovascular Disease Risk Prediction,” Journal of Medical Systems, vol. 48, art. 82, 2024.
14. T. Liu et al., “Machine learning based prediction models for cardiovascular disease risk using electronic health records: a systematic review for primary prevention,” European Heart Journal – Digital Health, vol. 6, no. 1, pp. 7–20, 2025.
15. V. Doppala, D. Bhattacharyya, M. J. Vidyarthi, and T.-H. Kim, “A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques,” Journal of Healthcare Engineering, vol. 2022, art. 2585235, 2022, doi: 10.1155/2022/2585235.
16. A. Ishaq et al., “Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques,” IEEE Access, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.
17. S. Subramani et al., “Cardiovascular disease prediction using machine learning algorithms,” Frontiers in Medicine, vol. 10, art. 1150933, 2023, doi: 10.3389/fmed.2023.1150933.
18. D. G. Honi and L. Szathmary, “A review of machine learning’s role in cardiovascular disease prediction: recent advances and future challenges,” Informatics in Medicine Unlocked, art. 101535, 2024, doi: 10.1016/j.imu.2024.101535.
19. D.-I. Kasartzian and T. Tsiampalis, “Transforming Cardiovascular Risk Prediction: A Review of Machine Learning and Artificial Intelligence Innovations,” Life, vol. 15, no. 1, art. 94, 2025, doi: 10.3390/life15010094.
20. S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?” PLoS ONE, vol. 12, no. 4, e0174944, 2017, doi: 10.1371/journal.pone.0174944.
21. A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, “cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants,” PLoS ONE, vol. 14, no. 5, e0213653, 2019, doi: 10.1371/journal.pone.0213653.
22. I. Kakadiaris et al., “Machine Learning Outperforms ACC/AHA CVD Risk Calculator in MESA,” Journal of the American Heart Association, vol. 7, e009476, 2018, doi: 10.1161/JAHA.118.009476.
23. M. Abdar et al., “A new machine learning technique for an accurate diagnosis of coronary artery disease,” Computer Methods and Programs in Biomedicine, vol. 179, art. 104992, 2019, doi: 10.1016/j.cmpb.2019.104992.
24. M. Tarawneh and O. Saboor, “heart disease identification method using machine learning classification in e-healthcare,” IEEE Access, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
25. D. Shah, S. Patel, and S. K. Bharti, “Heart Disease Prediction using Machine Learning Techniques,” SN Computer Science, vol. 1, art. 345, 2020, doi: 10.1007/s42979-020-00365-y.
26. M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, “Decision tree-based diagnosis of coronary artery disease: CART model,” Computer Methods and Programs in Biomedicine, vol. 192, art. 105400, 2020, doi: 10.1016/j.cmpb.2020.105400.
27. L. Verma, S. Srivastava, and P. C. Mate, “Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques,” Health and Technology, vol. 11, pp. 63–73, 2021, doi: 10.1007/s12553-020-00508-4.

28. P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, “A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets,” Health and Technology, vol. 14, pp. 597–613, 2024, doi: 10.1007/s12553-024-00835-w.
29. A. Suliman et al., “Predictive performance of machine learning compared to conventional risk models for secondary cardiovascular disease prevention,” BMJ Open, vol. 14, e082654, 2024, doi: 10.1136/bmjopen-2023-082654.
30. A. Rajkomar, J. Dean, and I. Kohane, “Machine Learning in Medicine,” New England Journal of Medicine, vol. 380, pp. 1347–1358, 2019, doi: 10.1056/NEJMra1814259.
31. E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nature Medicine, vol. 25, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
32. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, “Key challenges for delivering clinical impact with artificial intelligence,” BMC Medicine, vol. 17, art. 195, 2019, doi: 10.1186/s12916-019-1426-2.
33. X. Liu et al., “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension,” Nature Medicine, vol. 26, pp. 1364–1374, 2020, doi: 10.1038/s41591-020-1034-x.
34. S. Cruz Rivera et al., “Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension,” Nature Medicine, vol. 26, pp. 1351–1363, 2020, doi: 10.1038/s41591-020-1037-7.
35. G. S. Collins et al., “TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods,” BMJ, vol. 385, e078378, 2024, doi: 10.1136/bmj-2023-078378.
36. K. G. M. Moons et al., “PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods,” BMJ, vol. 388, e082505, 2025, doi: 10.1136/bmj-2024-082505.
37. C. L. Andaur Navarro et al., “Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review,” BMJ, vol. 375, n2281, 2021, doi: 10.1136/bmj.n2281.
38. B. Van Calster et al., “Calibration: the Achilles heel of predictive analytics,” BMC Medicine, vol. 17, art. 230, 2019, doi: 10.1186/s12916-019-1466-7.
39. E. Christodoulou et al., “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” Journal of Clinical Epidemiology, vol. 110, pp. 12–22, 2019, doi: 10.1016/j.jclinepi.2019.02.004.
40. R. D. Riley et al., “Calculating the sample size required for developing a clinical prediction model,” BMJ, vol. 368, m441, 2020, doi: 10.1136/bmj.m441.
41. L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
42. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
43. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
44. G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in Advances in Neural Information Processing Systems 30, 2017, pp. 3146–3154.
45. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
46. D. H. Wolpert, “Stacked generalization,” Neural Networks, vol. 5, no. 2, pp. 241–259, 1992, doi: 10.1016/S0893-6080(05)80023-1.
47. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.
48. S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nature Machine Intelligence, vol. 2, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
49. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
50. G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning,” Journal of Machine Learning Research, vol. 18, no. 17, pp. 1–5, 2017.
51. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves,” Biometrics, vol. 44, no. 3, pp. 837–845, 1988, doi: 10.2307/2531595.
52. T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLoS ONE, vol. 10, e0118432, 2015, doi: 10.1371/journal.pone.0118432.
53. D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, art. 6, 2020, doi: 10.1186/s12864-019-6413-7.
54. A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in Proc. 22nd International Conference on Machine Learning, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.
55. E. W. Steyerberg, Clinical Prediction Models, 2nd ed. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-16399-0.