International Journal of Scientific Modern Research and Technology
(Volume: 24, Issue: 01, Number: 05) Paper ID: IJSMRT - 26070105
ISSN: 2582-8150

Early Detection of Cardiac Disease Using an Explainable, Sensitivity-Optimized Ensemble of Routine Clinical Variables

Gyaneshwar Kumar Dhoke[1], Dharmendra Sahu[2]
[1]Research Scholar, School of Computer Science, SAM Global University, Bhopal, Madhya Pradesh, India
[2]Professor, School of Computer Science, SAM Global University, Bhopal, Madhya Pradesh, India

Abstract: Cardiovascular diseases remain the leading global cause of death, and early risk identification can support timely clinical assessment. This study developed an explainable machine-learning framework for detecting angiographic heart disease from routinely collected clinical variables. Methods: The UCI Cleveland Heart Disease dataset (303 records; 13 predictors) was converted to a binary outcome indicating disease presence. Missing values were imputed within each training fold, numerical features were standardized, and categorical features were one-hot encoded. Logistic regression, linear support vector machine, random forest, and XGBoost models were evaluated using 5-fold stratified cross-validation repeated three times. Within every outer training set, three-fold out-of-fold predictions determined validation-AUC weights for a soft-voting ensemble and an F2-optimized threshold. Performance was assessed by accuracy, sensitivity, specificity, precision, F1, F2, ROC-AUC, precision-recall AUC, and Brier score. Results: Disease was present in 139 of 303 records (45.9%). The standard weighted ensemble achieved accuracy 85.6% (95% CI 82.9-88.2), sensitivity 80.3% (75.5-85.1), specificity 90.0% (87.5-92.6), ROC-AUC 0.919 (0.900-0.939), and Brier score 0.114. Its discrimination was not significantly higher than logistic regression (one-sided paired Wilcoxon p=0.793). A training-derived mean threshold of 0.217 increased sensitivity to 93.1% (90.2-95.9) while reducing specificity to 57.4% (50.0-64.7). Permutation analysis identified the number of major vessels, chest-pain type, thal status, and exercise-induced ST depression as the most influential variables. Conclusion: The ensemble provided strong discrimination and a transparent sensitivity-specificity trade-off. The method is suitable as a research prototype for pre-screening, but prospective multicentre validation and clinical-impact evaluation are required before clinical use.

Keywords: cardiac disease; coronary artery disease; early detection; ensemble learning; machine learning; clinical prediction; explainable artificial intelligence

How to cite this article: Ghanshyam Kumar Dhuware, Dharmendra Sahu. (2026). Early Detection of Cardiac Disease Using an Explainable, Sensitivity-Optimized Ensemble of Routine Clinical Variables, International Journal of Scientific Modern Research and Technology (IJSMRT), ISSN: 2582-8150, Volume-24, Issue-01, Number-05, July- 2026, pp. 34-41, URL: https://www.ijsmrt.com/wp-content/uploads/2026/07/IJSMRT-26070105.pdf

IJSMRT-26070105

I. INTRODUCTION

Cardiovascular diseases are a heterogeneous group of disorders affecting the heart and blood vessels. The World Health Organization estimated that 19.8 million people died from cardiovascular diseases in 2022, accounting for approximately 32% of global deaths, and emphasized the importance of early detection so that counselling and treatment can begin promptly [1]. Risk factors measurable in routine care include blood pressure, glucose, lipids, obesity, and behavioural exposures. Nevertheless, the relationships among symptoms, exercise-test findings, demographic variables, and angiographic disease are nonlinear and may be difficult to summarize with a single rule.

Clinical prediction models can combine multiple variables into an estimated probability of disease. The UCI Cleveland Heart Disease dataset was originally derived from patients evaluated for coronary artery disease and contains a commonly used subset of 13 predictors plus an angiographic outcome [2,3]. It remains a useful benchmark for testing methodological principles, although its age, sample size, and referral-centre population limit direct clinical generalization.

 ISSN: 2582-8150

Recent studies have reported high predictive performance using boosting, feature selection, synthetic oversampling, stacking, and explainable artificial intelligence [11,12]. However, performance estimates from small datasets can be optimistic when preprocessing, model selection, or threshold choice uses information from test folds. Furthermore, a single accuracy value does not capture the clinically important trade-off between missed cases and false-positive referrals. Transparent reporting guidelines such as TRIPOD+AI emphasize clear definition of the data, outcome, validation strategy, and performance measures [4].

The objective of this study was to develop and evaluate a leakage-controlled, explainable ensemble for early detection of heart disease using routine clinical variables. The main contributions are: (i) repeated outer-fold validation with all preprocessing confined to training data; (ii) validation-AUC weighting of diverse models; (iii) a screening-oriented threshold selected exclusively from inner training predictions; and (iv) model interpretation through permutation importance. We hypothesized that the ensemble would provide stable discrimination and that threshold optimization would substantially improve sensitivity, although potentially at the cost of specificity.

II. MATERIALS AND METHODS

Study design and reporting

This was a secondary analysis of a de-identified, publicly available benchmark dataset. The study was structured as a diagnostic prediction-model development and internal-validation experiment. Reporting was informed by the principles of TRIPOD+AI [4], and limitations were interpreted in light of PROBAST+AI concerns regarding small samples, representativeness, and absence of external validation [5]. No direct patient contact, intervention, or re-identification was performed.

Dataset and outcome

The Cleveland component of the UCI Heart Disease dataset contains 303 patient records and 13 routinely collected demographic, symptom, laboratory, and exercise-test predictors [2,3]. The original diagnosis variable ranges from 0 to 4. Consistent with the repository's conventional binary task, diagnosis 0 was classified as disease absent and values 1-4 as disease present. This produced 164 negative and 139 positive records, corresponding to a prevalence of 45.9%. The repository describes the outcome as angiographic disease status, with positive disease representing greater than 50% diameter narrowing in at least one major vessel [2].

Table 1. Predictor variables used in the analysis

Variable	Clinical description	Type	Coding/unit
Age	Age	Continuous	years
Sex	Sex	Binary categorical	0=female; 1=male
Cp	Chest-pain type	Categorical	four categories
Trestbps	Resting blood pressure	Continuous	mm Hg
Chol	Serum cholesterol	Continuous	mg/dL
Fbs	Fasting blood sugar >120 mg/dL	Binary categorical	0=no; 1=yes
Restecg	Resting ECG result	Categorical	three categories
Thalach	Maximum heart rate achieved	Continuous	beats/min
Exang	Exercise-induced angina	Binary categorical	0=no; 1=yes
Oldpeak	Exercise-induced ST depression	Continuous	relative to rest

 ISSN: 2582-8150

Variable	Clinical description	Type	Coding/unit
Slope	Slope of peak exercise ST segment	Categorical	three categories
Ca	Major vessels visualized by fluoroscopy	Categorical/ordinal	0-3
Thal	Thal test status	Categorical	normal/fixed/reversible defect

Preprocessing

The dataset contained 4 missing values for ca and 2 for thal. Missing numerical values were imputed with the median and categorical values with the most frequent category. Numerical predictors (age, resting blood pressure, cholesterol, maximum heart rate, and oldpeak) were standardized using parameters estimated only from the corresponding training partition. Categorical predictors were one-hot encoded with an unknown-category safeguard. Every preprocessing step was embedded in a model pipeline to prevent information leakage.

Candidate models

Four classifiers with complementary inductive biases were selected: class-weighted logistic regression, a class-weighted linear support vector machine, random forest, and XGBoost. Logistic regression provides a linear probabilistic baseline; support vector machines maximize the separating margin [6]; random forests aggregate randomized decision trees [7]; and XGBoost implements regularized gradient boosting with efficient handling of nonlinear interactions [8]. Hyperparameters were fixed before outer validation to avoid test-fold tuning.

Table 2. Prespecified classifier settings

Model	Principal settings	Purpose
Logistic regression	C=1.0; balanced class weights; liblinear solver	Interpretable linear baseline
Support vector machine	Linear kernel; C=1.0; balanced class weights; probability estimates enabled	Margin-based classifier
Random forest	200 trees; maximum depth 5; minimum leaf size 3; balanced subsampling	Nonlinear bagged trees
XGBoost	150 trees; depth 2; learning rate 0.05; subsample 0.9; column sample 0.8	Regularized boosted trees

Validation-weighted ensemble

The principal model was a weighted soft-voting ensemble. For each outer training partition, three-fold stratified out-of-fold predictions were generated separately for the four prespecified classifiers. Model weights were proportional to the amount by which the inner ROC-AUC exceeded random discrimination:

$$w_m = max\ (AUC_m - 0.5,\ 0.001) \ / \ \Sigma_k\ max(AUC_k - 0.5,\ 0.001)$$

The ensemble probability for patient i was then calculated as:

$$p_i,ensemble = \Sigma_m\ w_m\ p_i,m$$

This procedure allowed the relative contribution of each classifier to vary across training partitions while

 ISSN: 2582-8150

keeping every test fold completely unseen during weighting.

Screening-oriented threshold selection

The default operating point classified probabilities of at least 0.50 as positive. A second operating point was designed for early screening, where missed disease may be more consequential than additional referrals. Within each outer training partition, the threshold between 0.10 and 0.80 that maximized the F2 score was selected using inner out-of-fold ensemble probabilities. F2 weights recall four times as strongly as precision:

$$F2 = 5 \times (\text{precision} \times \text{sensitivity}) / (4 \times \text{precision} + \text{sensitivity})$$

The selected threshold was then applied once to the corresponding untouched outer test fold. Thus, threshold optimization did not use outer-fold outcomes.

Performance measures and statistical analysis

Outer validation used 5-fold stratified cross-validation repeated three times, yielding 15 test-fold evaluations. The primary discrimination measure was ROC-AUC. Additional measures were accuracy, sensitivity, specificity, precision, F1, F2, precision-recall AUC, and the Brier score. The Brier score was calculated as the mean squared difference between observed outcomes and predicted probabilities [9]. Mean values and normal-approximation 95% confidence intervals were calculated across outer folds. The weighted ensemble was compared with the best mean-AUC individual model using a one-sided paired Wilcoxon signed-rank test. Because repeated cross-validation folds are not statistically independent, the p value was interpreted descriptively rather than as definitive confirmatory evidence.

Explainability and software

Global feature importance was estimated for the final XGBoost component by repeatedly permuting each original predictor and measuring the mean decrease in ROC-AUC. Permutation importance is model-agnostic in principle and captures the loss of predictive information after disrupting a feature, although correlated predictors may share importance. Analyses were conducted in Python using pandas, NumPy, scikit-learn, XGBoost, SciPy, and Matplotlib. The complete analysis script and processed dataset are supplied with the publication package.

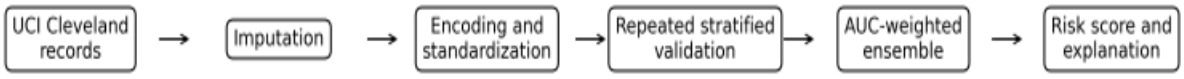

Figure 1. Leakage-controlled workflow for preprocessing, repeated validation, ensemble weighting, threshold selection, and explanation.

III. RESULTS

Dataset characteristics

The dataset included 303 records with 139 disease-positive and 164 disease-negative outcomes. The positive-class proportion was 45.9%. Missingness was limited to six observations across handled within training folds. Figure 2 shows the class distribution.

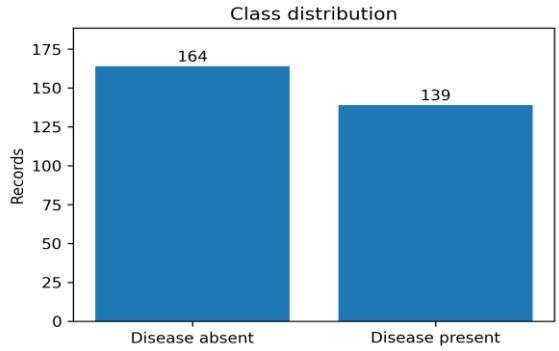

Figure 2. Distribution of disease-absent and disease-present records in the Cleveland dataset.

Predictive performance

Table 3 summarizes repeated outer-fold performance. Logistic regression provided the highest mean ROC-AUC among individual models (0.920), while the standard weighted ensemble achieved the highest mean accuracy (85.6%), specificity (90.0%), precision (87.4%), and F1 score (83.4%). The ensemble ROC-AUC was 0.919 and its Brier score was 0.114. The one-sided paired comparison did not show an AUC advantage over logistic regression (Wilcoxon statistic 39.5, p=0.793), indicating that the main benefit of the ensemble was operating-point balance rather than a statistically detectable improvement in discrimination.

Table 3. Repeated outer-fold performance, mean (95% CI)

Model	Accuracy %	Sensitivity %	Specificity %	Precision %	F1 %	ROC-AUC	Brier
Logistic regression	85.0 (82.3-87.8)	81.3 (76.1-86.4)	88.2 (85.5-90.9)	85.5	83.0	0.920 (0.901-0.940)	0.112
Support vector machine	84.3 (81.2-87.3)	78.1 (74.1-82.2)	89.4 (86.0-92.9)	86.6	81.9	0.911 (0.892-0.931)	0.117
Random forest	84.0 (81.5-86.6)	81.0 (76.7-85.4)	86.6 (84.0-89.2)	83.8	82.2	0.920 (0.902-0.937)	0.121
XGBoost	84.4 (81.9-86.9)	80.3 (75.1-85.5)	87.8 (85.3-90.3)	85.0	82.2	0.909 (0.890-0.929)	0.122
Weighted ensemble	85.6 (82.9-88.2)	80.3 (75.5-85.1)	90.0 (87.5-92.6)	87.4	83.4	0.919 (0.900-0.939)	0.114
Sensitivity-optimized ensemble	73.7 (70.6-76.8)	93.1 (90.2-95.9)	57.4 (50.0-64.7)	65.9	76.7	0.919 (0.900-0.939)	0.114

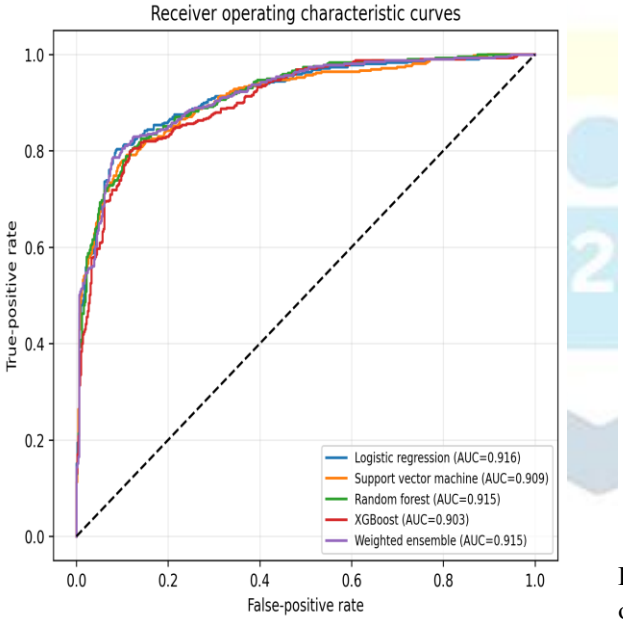

Figure 3. Receiver operating characteristic curves based on pooled repeated outer-fold predictions. Repeated predictions are shown for visualization; fold-level metrics in Table 3 are the primary estimates.

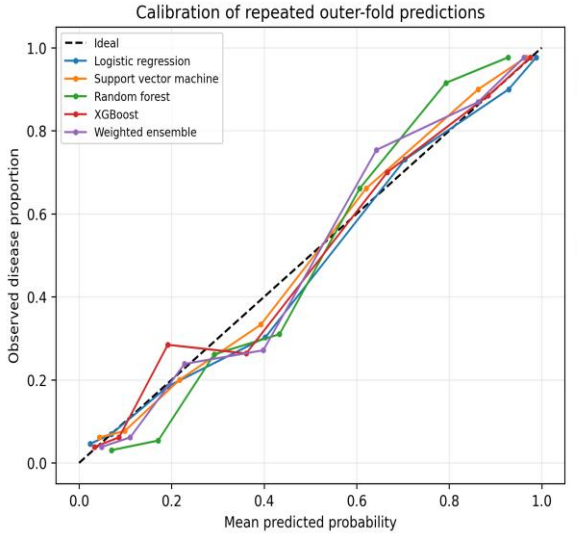

Figure 4. Calibration curves from pooled repeated outer-fold probabilities. The diagonal indicates perfect agreement between predicted and observed event rates.

Sensitivity-optimized operating point

The mean F2-optimized threshold was 0.217 (SD 0.073), substantially below 0.50. At this threshold, mean sensitivity increased from 80.3% to 93.1%, and

F2 increased from 81.5% to 85.6%. The trade-off was a reduction in specificity from 90.0% to 57.4% and in accuracy from 85.6% to 73.7%. Figure 5 displays the mean row-normalized confusion matrix for this screening-oriented operating point. The result demonstrates that the same probability model can support different clinical workflows depending on the threshold and the relative costs of false negatives and false positives.

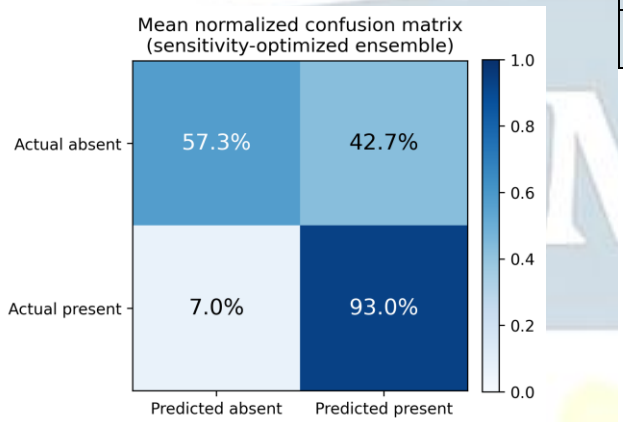

Figure 5. Mean row-normalized confusion matrix for the sensitivity-optimized ensemble. Percentages are averaged across the 15 outer test folds.

Model weighting and feature importance

When fitted to the complete dataset for descriptive interpretation, cross-validated AUC weighting assigned similar contributions to logistic regression (25.3%), support vector machine (24.8%), random forest (25.4%), and XGBoost (24.6%). This near-uniform distribution indicates that no single learner dominated the ensemble.

Permutation analysis of the XGBoost component ranked ca, chest-pain type, thal status, oldpeak, age, and sex as the most influential predictors. The ranking is clinically plausible for a referral population assessed for coronary artery disease, but it should not be interpreted causally. In particular, ca and thal are test-derived variables and may be unavailable in low-resource primary screening.

Table 4. Leading permutation-importance values for the final XGBoost component

Rank	Predictor	Mean decrease in ROC-AUC
1	Ca	0.0554
2	Cp	0.0324
3	Thal	0.0264
4	oldpeak	0.0191
5	Age	0.0128
6	sex	0.0127
7	chol	0.0104
8	slope	0.0069

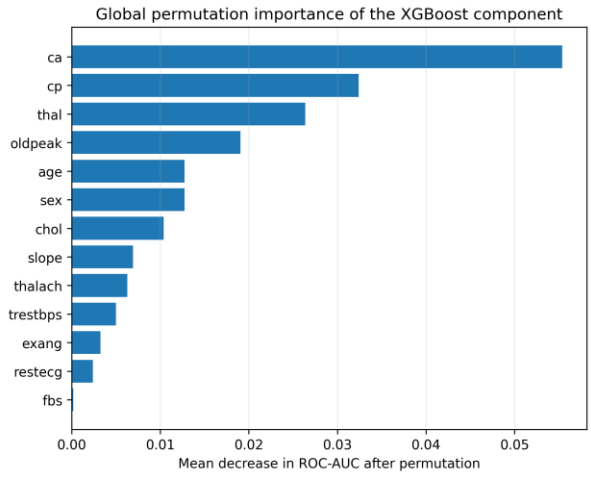

Figure 6. Global permutation importance for the final XGBoost component. Values represent the reduction in ROC-AUC after permutation of each original predictor.

IV. DISCUSSION

Principal findings

This study developed an internally validated ensemble for detecting angiographic heart disease from 13 routine clinical variables. The standard ensemble achieved strong discrimination and favoured specificity, whereas the F2-derived operating point shifted the model toward screening sensitivity. The results highlight three practical lessons. First, a diverse ensemble may improve the balance of classification metrics even when its AUC is not significantly greater than a strong linear model. Second, reported performance depends materially on the chosen decision threshold. Third, careful separation of preprocessing, weighting, and threshold selection from the outer test folds is essential in small datasets.

The ensemble AUC of 0.919 is consistent with the broad range reported in recent heart-disease prediction research, although direct comparisons are inappropriate because datasets, feature definitions,

validation schemes, and resampling methods differ. El-Sofany and colleagues reported high performance using XGBoost, feature selection, SMOTE, and combined public/private data [11]. Ganie and colleagues found that stacking and voting ensembles improved prediction across multiple datasets and used explainability methods to support interpretation [12]. The present work differs by focusing on leakage-controlled threshold selection and by explicitly reporting the sensitivity-specificity consequences of a screening objective.

Clinical interpretation

At the default threshold, the ensemble may be viewed as a relatively specific decision-support model: approximately nine of ten disease-negative cases were correctly identified, while about four of five disease-positive cases were detected. For early screening, the lower threshold increased sensitivity above 93%, but false-positive referrals increased substantially. Neither operating point is intrinsically superior. Selection should depend on the clinical setting, downstream test burden, prevalence, patient preferences, and consequences of missed disease.

The feature ranking emphasizes variables related to coronary anatomy, symptoms, and exercise-induced ischemia. However, the strongest predictor, ca, is the number of major vessels visualized by fluoroscopy and is not a simple pre-consultation measurement. A deployable community-screening system would need to exclude invasive or post-test features and be retrained using accessible predictors such as history, blood pressure, laboratory values, and noninvasive ECG measurements. Therefore, the present model is best interpreted as a methodological benchmark rather than a ready-to-use screening device.

Strengths

The study has several methodological strengths: preprocessing was performed within model pipelines; test folds were isolated from ensemble weighting and threshold selection; repeated stratified validation reduced dependence on a single split; probability quality was assessed with calibration and Brier score in addition to classification metrics; and the threshold trade-off was reported explicitly. The analysis code, data file, fold-level metrics, and figures are provided for reproducibility.

Limitations

The primary limitation is the small, historical, single-centre Cleveland dataset. It may not represent current diagnostic pathways, contemporary therapies, geographic diversity, or the prevalence encountered in primary care. The endpoint was angiographic disease rather than future cardiovascular events. Repeated cross-validation is internal validation and cannot establish transportability. Confidence intervals based on fold variation are approximate because repeated folds are correlated. Hyperparameters were prespecified rather than optimized, which reduces adaptive overfitting but may not maximize each learner. Probability calibration was assessed descriptively and was not externally recalibrated. Finally, permutation importance does not establish causality and can underestimate correlated predictors.

Future work

Future research should validate the framework prospectively across multiple hospitals, report subgroup performance by sex, age, ethnicity, and comorbidity, and evaluate models restricted to variables available before advanced testing. External calibration, decision-curve analysis, net-benefit assessment, missingness robustness, and comparison with established clinical risk scores are required. Federated learning may support multicentre development without centralizing identifiable patient data. Most importantly, clinical-impact studies should determine whether model-assisted screening improves referral appropriateness, time to diagnosis, patient outcomes, and resource use.

V. CONCLUSION

A validation-AUC-weighted ensemble of logistic regression, support vector machine, random forest, and XGBoost achieved strong internal discrimination for heart-disease detection on the UCI Cleveland dataset. The standard threshold produced 85.6% accuracy and 90.0% specificity, while a training-derived threshold increased sensitivity to 93.1% at the expense of specificity. These findings demonstrate the importance of leakage-controlled validation and explicit threshold selection. The model should not be used for clinical diagnosis without contemporary multicentre external validation and formal assessment of clinical utility.

REFERENCES

[1] World Health Organization. Cardiovascular diseases (CVDs). WHO Fact Sheet. Updated 31 July 2025.

[2] Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease [Dataset]. UCI Machine Learning Repository; 1989. doi:10.24432/C52P4X.

[3] Detrano R, Janosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989;64(5):304-310. doi:10.1016/0002-9149(89)90524-9.

[4] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

[5] Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505.

[6] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273-297. doi:10.1007/BF00994018.

[7] Breiman L. Random forests. Mach Learn. 2001;45:5-32. doi:10.1023/A:1010933404324.

[8] 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. 2016:785-794. doi:10.1145/2939672.2939785.

[9] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1-3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[10] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.

[11] El-Sofany H, Bouallegue B, Abd El-Latif YM. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Sci Rep. 2024;14:23277. doi:10.1038/s41598-024-74656-2.

[12] Ganie SM, Pramanik PKD, Zhao Z. Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. Sci Rep. 2025;15:13912. doi:10.1038/s41598-025-97547-6.