Afropolitan Journals

This open-access article is distributed under Creative Commons Attribution 4.0 license

ISSN: 3027-2882 www.afropolitanjournals.com

A Recall-Oriented Comparative Evaluation of Machine Learning Models for Heart Disease Detection on the Cleveland Heart Disease Dataset

Reginald E. Ogu [1]; Juliana C. Ihuefo [2] and Precious Iheanacho [3]

[1]Department of Computer Engineering, Federal University of Technology, Owerri, Nigeria. [2]Department of Senior Staff Establishment, Federal Teaching Hospital, Owerri, Nigeria. [3]IT Unit, Brightroot Empowerment Initiative, Owerri, Nigeria.

Corresponding author: reginald.ogu@futo.edu.ng

***DOI**: https://doi.org/10.62154/ajastr.2026.023.01023*

Abstract

Heart disease screening models are usually tuned for accuracy, although a missed diagnosis costs far more than a false alarm. This study compared six classifiers (Decision Tree, Random Forest, Support Vector Machine, XGBoost, Logistic Regression and a Multilayer Perceptron) on the 303-patient Cleveland Heart Disease dataset, in which 139 patients had angiographic disease. A widely shared copy of the dataset was found to carry a reversed outcome label, so all experiments used the original repository file. Hyperparameters were tuned by repeated nested cross-validation, and each model's decision threshold was fixed on training data to reach a recall of at least 0.90. Across 50 outer test folds, every model met the recall target (0.89 to 0.91). Logistic Regression achieved the highest mean specificity (0.68; 95% CI 0.60 to 0.77) and precision-recall AUC (0.91), although its advantage over Random Forest and Support Vector Machine was not statistically significant. The Decision Tree ruled out far fewer disease-free patients (specificity 0.43). Restricting the inputs to five pre-test variables reduced Logistic Regression specificity to 0.57. Logistic Regression is recommended as an interpretable, recall-oriented baseline, subject to external validation in unselected primary-care populations.

Keywords: Machine Learning, Heart Disease, Logistic Regression, Triage, Cost-Sensitive Learning.

Introduction

Cardiovascular diseases (CVDs) remain the leading cause of global mortality and morbidity, placing a heavy structural and economic burden on health systems (Banerjee & Paçal, 2025; Kumar et al., 2025; Suleiman et al., 2023; Wan et al., 2025). Coronary artery disease accounts for a large share of that burden, and its early course is often quiet. By the time chest pain or breathlessness brings a patient to a clinic, the definitive test is usually invasive coronary angiography, a procedure that carries procedural risk and cost and is not available everywhere (Chakraborty et al., 2026; Pedroso et al., 2026). In much of sub-Saharan Africa, including Nigeria, catheterisation laboratories are concentrated in a few tertiary centres, so

the decision about who should travel for further testing is made long before any imaging is done.

That decision is a triage problem. Front-line clinicians need a way to sort symptomatic patients into those who can safely wait and those who need prompt referral. Prediction tools can help here, and professional bodies increasingly expect preventive decisions to rest on an estimate of risk rather than on single risk factors (Khan et al., 2026). Machine learning (ML) models offer a flexible way to build such estimates from routinely recorded clinical data, and AI-assisted triage has already shortened the time to diagnosis in community programmes that used electrocardiograms to prioritise echocardiography (Pedroso et al., 2026; Rhee et al., 2026).

The difficulty is how such models are judged. Much of the heart disease ML literature leads with overall accuracy and treats the 0.5 probability cut-off as given (Gul et al., 2026; Ilyas et al., 2025; Mohan et al., 2019; Mohapatra et al., 2023). For triage, accuracy is the wrong summary. A false positive sends a patient for a further test that will usually clear them. A false negative sends a diseased patient home. Two models with the same accuracy can differ sharply in how many patients they miss, and a single accuracy figure hides that difference.

This paper reports a recall-oriented comparison of six commonly used classifiers on the Cleveland Heart Disease dataset. The recall target and the rule for choosing a model were fixed before the comparison, the models were evaluated (not externally validated) with repeated nested cross-validation and a held-out test set, and the analysis reports uncertainty, statistical tests, calibration and the effect of lower disease prevalence. In the course of the work, a commonly redistributed copy of the dataset was found to have its outcome label reversed; the consequences of that error are described because they bear on a substantial part of the published literature that uses the dataset.

Statement of the Problem

Classifiers for heart disease detection are commonly tuned and ranked by accuracy at a default decision threshold. In a triage setting this practice is unsafe, because the cost of a missed case is much higher than the cost of an unnecessary referral, and a model chosen for accuracy may still miss one in five diseased patients. Three further gaps compound the problem. First, few comparisons on the Cleveland data fix a sensitivity target in advance and then compare models on how many disease-free patients they can rule out at that target. Second, model rankings are often based on a single train-test split with no confidence intervals, so differences of one or two patients are reported as superiority. Third, published descriptions of the dataset frequently misstate what the variables are and how they are coded, which leads to claims that a model is non-invasive when its strongest inputs come from exercise testing, nuclear imaging and fluoroscopy.

Objectives and Research Questions

The aim of the study was to identify which of six classifiers is most suitable for recall-oriented heart disease triage on the Cleveland data. The specific objectives were:

- To verify the provenance and coding of the dataset before modelling.
- To compare Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, Logistic Regression and Multilayer Perceptron (MLP) models at a pre-specified recall target of 0.90, using repeated nested cross-validation with confidence intervals and corrected significance tests.
- To estimate how much discrimination is lost when the models are restricted to inputs available before any laboratory, exercise or imaging test.
- To identify which inputs, drive the predictions, and to test whether class weighting protects recall when the disease is rare in the training data.

These objectives correspond to four research questions:

- RQ1: Does the dataset used match the documented UCI Cleveland file in size, class balance and coding?
- RQ2: With recall held at or above 0.90, which classifier achieves the highest specificity, and are the differences between classifiers statistically meaningful?
- RQ3: How do recall, specificity and precision-recall performance change when only pre-test variables are used?
- RQ4: Which variables contribute most to the predictions, and does class weighting maintain recall at a disease prevalence of 15%?

The selection criterion was stated before the experiments were run. A model was regarded as meeting the screening requirement if its decision threshold, set on training data only, targeted a recall of at least 0.90. Among models meeting that requirement, the preferred model was the one with the highest mean specificity across the outer test folds. If two or more models could not be separated by a corrected resampled *t*-test at the 5% level, the simpler and more interpretable model was preferred.

Literature Review

Risk Scores and Machine Learning

Conventional cardiovascular risk tools, such as the Framingham, SCORE2 and QRISK equations and the newer PREVENT equations, are regression models built for long-term risk estimation in the general population. A recent American Heart Association and American College of Cardiology scientific statement describes how predicted risk and expected treatment benefit should guide primary prevention decisions (Khan et al., 2026). Such tools answer a different question from the one addressed here. They estimate the probability of a future event in people without symptoms, whereas triage asks whether a symptomatic patient already has significant coronary narrowing. Machine learning has been proposed for the second task because it can model non-linear relationships and

interactions between clinical findings without specifying them in advance (Nagavelli et al., 2022; Safdar et al., 2018; Ullah et al., 2023).

Comparative Studies on the Cleveland Data and Related Datasets

The Cleveland data were collected at the Cleveland Clinic Foundation from patients referred for coronary angiography and were first used to test a probability algorithm for diagnosing coronary artery disease across four international centres (Detrano et al., 1989). The file has since become one of the most frequently used benchmarks in medical ML. Mohan et al. (2019) combined Random Forest with a linear method in a hybrid model and reported 88.7% accuracy. Latha and Jeeva (2019) showed that bagging, boosting and majority voting raised the accuracy of weak classifiers by up to 7%. Ahmad and Polat (2023) used the Jellyfish metaheuristic for feature selection before classification and reported the best results with an SVM, while Mohapatra et al. (2023) reported 92% accuracy for a stacked ensemble. Rehman et al. (2025) compared advanced classifiers for coronary heart disease risk assessment, and Ilyas et al. (2025) found Naive Bayes to be the strongest of several models on a 1,025-record Kaggle compilation of heart disease data.

Outside the Cleveland benchmark, Chakraborty et al. (2026) trained explainable models on 5,000 hospital records from Bangladesh and found Random Forest to perform best, with 98% accuracy and 96% recall, using SHapley Additive exPlanations (SHAP) to interpret the predictions. A 2026 systematic review concluded that Random Forest and boosting methods are the most practical choices for tabular clinical data, but cautioned that the reported gains rest largely on small, single-source datasets, that external validation is rare, and that calibration is seldom examined (Gul et al., 2026). Broader reviews reach similar conclusions about the dominance of accuracy as the headline metric (Banerjee & Paçal, 2025; Kumar et al., 2025; Wan et al., 2025). Taken together, these studies show no consistent winner. Rankings change with the dataset, its version and the metric, and most comparisons report a single split without uncertainty estimates.

Recall-Oriented and Cost-Sensitive Learning

When one type of error is more costly than the other, two families of remedy are available. Cost-sensitive learning changes the training objective, for example by weighting the minority or high-cost class more heavily (He & Garcia, 2009). Threshold moving keeps the trained model but shifts the probability cut-off so that a chosen sensitivity is reached. The second approach is attractive for triage, because the operating point can be set to match the clinical tolerance for missed cases, but the threshold must be chosen on training data, never on the test set, or the reported recall will be optimistic. Precision-recall curves are more informative than receiver operating characteristic (ROC) curves when the positive class is the focus and its prevalence may be low (Saito & Rehmsmeier, 2015). Calibration matters as well, since a model whose probabilities are systematically too high or too low can mislead clinicians even when its ranking of patients is good (Van Calster et al., 2019).

Finally, small datasets make model comparison fragile. Selecting hyperparameters and estimating performance on the same folds biases the estimate upwards (Varma & Simon, 2006), and the folds of repeated cross-validation are not independent, so naive paired *t*-tests overstate significance (Dietterich, 1998; Nadeau & Bengio, 2003).

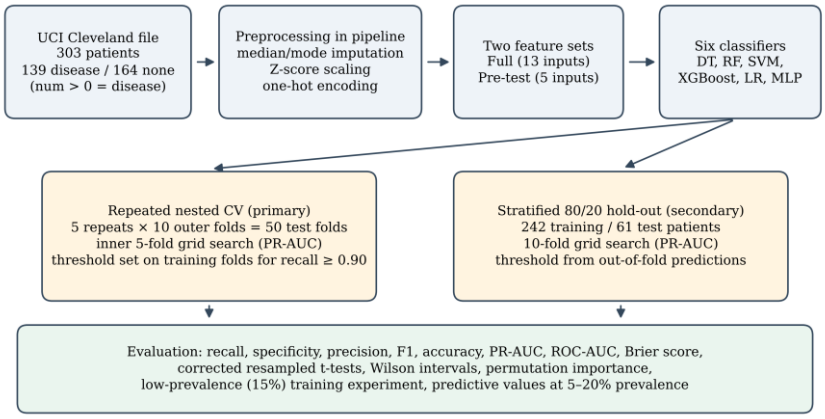
Research Gap

The literature reviewed above contains many accuracy-ranked comparisons on the Cleveland data, but few that fix a sensitivity target in advance, set the decision threshold without touching the test data, report specificity at that target with confidence intervals, and test whether the differences between models are larger than resampling noise. The effect of restricting the inputs to pre-test variables, and of a lower disease prevalence, is also rarely examined. The present study addresses these points.

Methodology

The study workflow is summarised in Figure 1.

Figure 1: Study workflow, from data verification to evaluation



Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

Dataset and Provenance Check

The study used the processed Cleveland file of the UCI Heart Disease repository (Janosi et al., 1989), which contains 303 patients evaluated at the Cleveland Clinic Foundation (Detrano et al., 1989). The original outcome, *num*, takes values from 0 to 4 and records angiographic disease status; following standard practice, values 1 to 4 were coded as disease present (diameter narrowing of more than 50% in at least one major vessel) and 0 as absent. This gives 139 patients with disease (45.9%) and 164 without (54.1%), so the classes are close to balanced. The cohort is a referral population of symptomatic patients

who were sent for angiography; it is not a screening population and contains no information on disease stage.

Before modelling, the file was compared row by row with a widely circulated derivative copy (303 rows, 165 positive and 138 negative) that is distributed through Kaggle and many code repositories. Matching on six continuous variables paired all 303 rows. The comparison showed that the derivative copy reverses the outcome: every patient labelled 1 in that copy has *num* = 0 in the UCI file, and every patient labelled 0 has disease. It also recodes chest pain type (0 in the copy corresponds to asymptomatic, UCI code 4), slope (0 corresponds to downsloping), and thal (1, 2 and 3 correspond to fixed defect, normal and reversible defect), stores the six missing values as 0 or 4 instead of leaving them missing, and duplicates one patient while omitting another. A model trained on that copy and evaluated on recall of class 1 is in fact measuring how well disease-free patients are identified. All results in this paper come from the UCI file.

Table 1 lists the 13 input variables, their UCI coding, and where each comes from in clinical practice. Five of the variables (age, sex, chest pain type, resting blood pressure and resting electrocardiogram) are available at a first consultation. The rest require a blood test (cholesterol, fasting blood sugar), a treadmill exercise test (maximum heart rate, exercise-induced angina, ST depression and ST slope), thallium myocardial perfusion scintigraphy (thal) or fluoroscopy (number of vessels coloured). The full model is therefore not non-invasive or laboratory-free, and is described in this paper as a model based on a complete non-invasive and semi-invasive cardiac work-up. The five first-consultation variables form a separate pre-test feature set.

Table 1: Input variables, UCI coding and clinical source

Variable	Description and UCI coding	Clinical source	Pre-test set
age	Age in years	History	Yes
sex	1 = male, 0 = female	History	Yes
cp	Chest pain: 1 typical angina, 2 atypical angina, 3 non-anginal pain, 4 asymptomatic	History	Yes
trestbps	Resting blood pressure (mm Hg) on admission	Examination	Yes
restecg	Resting ECG: 0 normal, 1 ST-T abnormality, 2 probable or definite left ventricular hypertrophy (Estes)	Resting ECG	Yes
chol	Serum cholesterol (mg/dL)	Laboratory	No
fbs	Fasting blood sugar > 120 mg/dL (1 = true)	Laboratory	No
thalach	Maximum heart rate achieved	Exercise test	No
exang	Exercise-induced angina (1 = yes)	Exercise test	No
oldpeak	ST depression induced by exercise relative to rest (mm)	Exercise test	No
slope	Slope of peak exercise ST segment: 1 upsloping, 2 flat, 3 downsloping	Exercise test	No

Variable	Description and UCI coding	Clinical source	Pre-test set
ca	Number of major vessels (0 to 3) coloured by fluoroscopy; 4 missing values	Fluoroscopy	No
thal	Thallium scintigraphy: 3 normal, 6 fixed defect, 7 reversible defect; 2 missing values	Nuclear imaging	No

Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

Preprocessing

All preprocessing steps were placed inside a scikit-learn pipeline so that they were learned from the training folds only and then applied to the corresponding test fold, which prevents information from the test data leaking into the model. The six missing values were imputed with the training-fold median (*ca*) or mode (*thal*). Chest pain type, resting ECG, slope and thal were one-hot encoded, since their numeric codes have no meaningful order. Continuous variables were standardised with the Z-score in Equation 1:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

where x is the observed value, and μ and σ are the mean and standard deviation of that variable in the training fold. Standardisation puts the inputs on a common scale, which matters for the distance-based SVM and for the gradient-based Logistic Regression and MLP; it makes no difference to the tree-based models.

Classifiers

Six classifiers were compared: a Decision Tree; a Random Forest of 200 trees (Breiman, 2001); an SVM with a radial basis function kernel; XGBoost (Chen & Guestrin, 2016); L2-regularised Logistic Regression; and an MLP with one or two hidden layers. All except XGBoost and the MLP were implemented in scikit-learn (Pedregosa et al., 2011). To counter the slight class imbalance, the Decision Tree, Random Forest, SVM and Logistic Regression used balanced class weights, the MLP received equivalent sample weights, and XGBoost used a positive-class weight equal to the ratio of negative to positive training cases. The balanced weight for class c is given by Equation 2:

$$w_c = \frac{N}{2N_c} \quad (2)$$

where N is the number of training patients and N_c the number in class c.

Evaluation Design and Hyperparameter Tuning

Two evaluation designs were used. The primary design was repeated nested cross-validation: stratified 10-fold cross-validation repeated five times with different shuffles, giving 50 outer test folds of about 30 patients each. Within every outer training fold, hyperparameters were chosen by an exhaustive grid search with inner stratified 5-fold cross-validation, and the outer test fold was used only once, to score the tuned model. This separation avoids the optimistic bias that arises when the same data are used to tune and

to evaluate (Varma & Simon, 2006). The secondary design was a single stratified 80/20 hold-out split, with 242 patients (111 with disease) for training and 61 (28 with disease) for testing; hyperparameters were tuned on the training part by grid search with stratified 10-fold cross-validation. The hold-out results are reported because they allow confusion matrices and precision-recall curves for individual patients to be shown, but they rest on a small test set and are treated as illustrative.

Hyperparameters were selected by maximising the cross-validated area under the precision-recall curve (PR-AUC). Recall was not used as the tuning criterion, because recall alone can be maximised trivially by predicting every patient as diseased. Instead, recall was controlled through the decision threshold. For each tuned model, out-of-fold predicted probabilities were obtained on the training data, and the threshold τ was set to the largest value that kept training recall at or above 0.90, as stated in Equation 3:

$$\tau = \max\{t : \mathrm{Recall}_{\mathrm{train}}(t) \geq 0.90\} \qquad (3)$$

The same threshold was then applied, unchanged, to the test patients. Each model was also evaluated at the default threshold of 0.5 for comparison. Table 2 gives the search grids and the values selected on the hold-out training data. In the nested design the selected values could differ from fold to fold; for Logistic Regression, for example, C = 0.1 was chosen in 26 of the 50 outer folds.

Table 2: Hyperparameter search grids and values selected on the hold-out training set (full feature set)

Model	Search grid	Selected values	Threshold τ
Decision Tree	max_depth {3, 4, 5, 6, none}; min_samples_leaf {1, 2, 4, 8}; criterion {gini, entropy}; class_weight = balanced	max_depth = 3; min_samples_leaf = 8; gini	0.23
Random Forest	200 trees; max_depth {3, 5, none}; min_samples_leaf {1, 5}; class_weight = balanced	max_depth = 3; min_samples_leaf = 5	0.33
SVM (RBF)	C {0.1, 1, 10}; gamma {scale, 0.01, 0.03, 0.1}; class_weight = balanced	C = 1; gamma = 0.01	0.28
XGBoost	n_estimators {100, 300}; max_depth {2, 3}; learning_rate {0.05, 0.1}; subsample = 0.8; scale_pos_weight = negative/positive ratio (1.18)	n_estimators = 100; max_depth = 2; learning_rate = 0.05	0.25
Logistic Regression	C {0.01, 0.03, 0.1, 0.3, 1, 3, 10}; L2 penalty; class_weight = balanced	C = 1	0.28
MLP	hidden layers {(16), (32, 16)}; alpha {0.01, 0.1, 1.0}; ReLU; L-BFGS solver; balanced sample weights	hidden layer (16); alpha = 1.0	0.05

Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

Performance metrics

Performance was summarised with the metrics in Equations 4 to 9, where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (6)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (7)$$

$$F_1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (8)$$

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \quad (9)$$

Recall (sensitivity) is the proportion of diseased patients who are flagged, and specificity is the proportion of disease-free patients who are correctly ruled out. In Equation 9, p_i is the predicted probability of disease for patient *i* and y_i is the observed outcome; the Brier score measures the accuracy of the probabilities themselves, with lower values indicating better calibrated and sharper predictions (Van Calster et al., 2019). Threshold-independent discrimination was summarised by PR-AUC, computed as average precision, and by the area under the ROC curve (ROC-AUC).

Because the Cleveland prevalence of 45.9% is far higher than would be seen in primary care, the positive and negative predictive values that a model would achieve at a lower prevalence π were estimated from its sensitivity (Se) and specificity (Sp) with Bayes' theorem (Altman & Bland, 1994), as shown in Equations 10 and 11:

$$\text{PPV}(\pi) = \frac{\text{Se}\cdot\pi}{\text{Se}\cdot\pi+(1-\text{Sp})(1-\pi)} \quad (10)$$

$$\text{NPV}(\pi) = \frac{\text{Sp}(1-\pi)}{\text{Sp}(1-\pi)+(1-\text{Se})\pi} \quad (11)$$

Uncertainty and Statistical Comparison

For the nested design, each metric is reported as the mean over the 50 outer folds with a 95% confidence interval. Because the training sets of different folds overlap, the ordinary standard error understates the true variability. The interval and the paired comparisons between models therefore used the corrected resampled *t*-statistic of Nadeau and Bengio (2003), given in Equation 12:

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k}+\frac{n_{\text{test}}}{n_{\text{train}}}\right)\hat{\sigma}^2{}_d}} \quad (12)$$

where $\bar{d}$ and $\hat{\sigma}_d^2$ are the mean and variance of the fold-wise differences between two models, $k = 50$ is the number of outer folds, and $n_{\text{test}}/n_{\text{train}} \approx 30/273$ is the ratio of test to training patients in each fold. The statistic was referred to a *t*-distribution with $k - 1$ degrees of freedom. On the hold-out test set, 95% confidence intervals for recall and specificity were computed with the Wilson score method (Wilson, 1927), and paired differences in correct classification were tested with the exact McNemar test (Dietterich, 1998).

Feature Attribution

The contribution of each input was measured by permutation importance (Breiman, 2001). In every outer test fold of the nested design, the values of one original variable were shuffled ten times, and the mean drop in ROC-AUC relative to the unshuffled data was recorded. Shuffling the raw variable, before encoding, means that the importance of a categorical variable such as chest pain type is measured as a whole rather than split across its one-hot columns. The fold-wise importances were averaged over the 50 folds for each model.

Low-Prevalence Experiment

To test whether class weighting protects recall when the disease is rare, the training part of each outer fold (five repeats of 10-fold cross-validation) was subsampled so that only 15% of the training patients had disease, while all disease-free patients were kept. Each model was trained on this reduced set with the hyperparameters selected on the hold-out training data, once with class weighting and once without, and was then evaluated at the default 0.5 threshold on the untouched outer test fold.

Software and Reproducibility

The analysis was carried out in Python 3.13.18 with scikit-learn 1.9.1, XGBoost 3.4.1, NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1, statsmodels 0.15.0 and Matplotlib 3.11.2. A single random seed (42) controlled the data splits, model initialisation and permutations, and the seed for each inner cross-validation was offset by the outer fold index. The analysis scripts are available from the corresponding author on reasonable request.

Results

Dataset verification (RQ1)

The UCI file contained 303 patients, of whom 139 (45.9%) had angiographic disease, with four missing values in *ca* and two in *thal*. These figures match the repository documentation. The derivative copy, which reports 165 patients with disease and 138 without, was confirmed to carry a reversed outcome label and a different coding of chest pain type, slope and thal, as described in Section 5.1.

Repeated Nested Cross-Validation (RQ2)

Table 3 summarises performance across the 50 outer test folds with the full feature set. At the screening threshold all six models reached their recall target, with mean recall between 0.89 and 0.91, and none of the pairwise differences in recall was statistically significant (all $p > 0.70$ for comparisons with Logistic Regression). The models differed markedly in what that recall cost. Logistic Regression ruled out the largest share of disease-free patients (specificity 0.68; 95% CI 0.60 to 0.77), closely followed by SVM (0.68) and Random Forest (0.68). XGBoost was somewhat lower (0.63), the MLP lower again (0.59), and the Decision

Tree reached the recall target only by flagging most patients, with a specificity of 0.43 (Figure 2a).

Table 3: Repeated nested cross-validation results at the screening threshold, full feature set (mean over 50 outer folds, 95% CI in brackets)

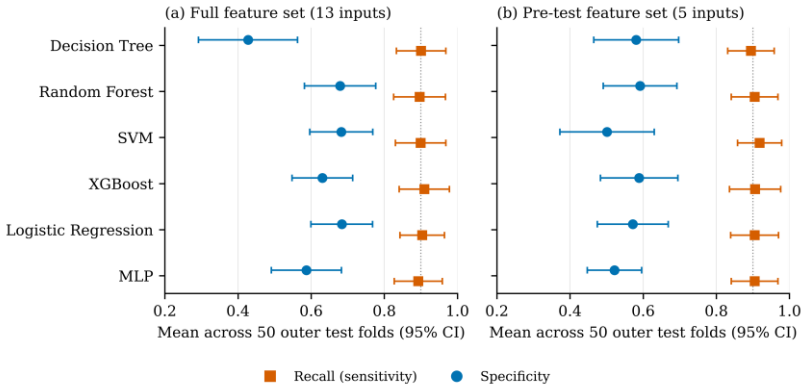
Model	Recall	Specificity	Prec.	F1	Acc.	PR-AUC	ROC-AUC	Brier
Decision Tree	0.90 [0.83, 0.97]	0.43 [0.29, 0.56]	0.58	0.70	0.65	0.81	0.83	0.168
Random Forest	0.90 [0.83, 0.97]	0.68 [0.58, 0.78]	0.71	0.79	0.78	0.91	0.91	0.128
SVM	0.90 [0.83, 0.97]	0.68 [0.60, 0.77]	0.71	0.79	0.78	0.91	0.91	0.118
XGBoost	0.91 [0.84, 0.98]	0.63 [0.55, 0.71]	0.68	0.78	0.76	0.90	0.90	0.127
Logistic Regression	0.90 [0.84, 0.97]	0.68 [0.60, 0.77]	0.72	0.80	0.79	0.91	0.91	0.120
MLP	0.89 [0.83, 0.96]	0.59 [0.49, 0.68]	0.65	0.75	0.73	0.88	0.87	0.155

Note. Prec. = precision; Acc. = accuracy. Intervals use the corrected resampled variance (Nadeau & Bengio, 2003). *Source:* Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

The corrected *t*-tests confirmed that the Decision Tree had significantly lower specificity than every other model ($p \leq 0.04$) and that the MLP was significantly less specific than Logistic Regression ($p = 0.009$), SVM ($p = 0.020$) and Random Forest ($p = 0.044$). The differences between Logistic Regression and SVM ($p = 0.94$), Random Forest ($p = 0.88$) and XGBoost ($p = 0.10$) were not significant. The same pattern held for PR-AUC: Logistic Regression had the highest mean value (0.913), but it could not be separated from Random Forest (0.911), SVM (0.909) or XGBoost (0.903), whereas the Decision Tree (0.814) and the MLP (0.876) were significantly lower. SVM produced the best-calibrated probabilities (Brier score 0.118), with Logistic Regression a close second (0.120).

The default threshold of 0.5 tells a different story. At that cut-off, the mean recall of every model fell to between 0.75 and 0.80, while specificity rose to between 0.82 and 0.89 and accuracy to between 0.79 and 0.85. In other words, a model chosen and operated for accuracy would have missed about one diseased patient in five. Across the five repetitions, Logistic Regression at the 0.5 threshold missed 136 of 695 positive cases, compared with 67 at the screening threshold. The thresholds that delivered the 0.90 recall target were consistently below 0.5; for Logistic Regression they ranged from 0.24 to 0.39 across the 50 folds.

Figure 2: Mean recall and specificity at the screening threshold across 50 outer test folds, with corrected 95% confidence intervals, for (a) the full and (b) the pre-test feature set



Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989). The dotted line marks the recall target of 0.90.

Hold-Out Test Set

Table 4 reports the hold-out results at the screening threshold, and Figures 3 and 4 show the precision-recall curves and confusion matrices. On these 61 patients, Logistic Regression, Random Forest, XGBoost and the MLP identified all 28 diseased patients, while the SVM and the Decision Tree each missed one. Logistic Regression again ruled out the most disease-free patients (23 of 33; specificity 0.70). The small size of the test set is evident in the intervals: a recall of 28 out of 28 has a Wilson 95% interval of 0.88 to 1.00, and the specificity intervals of the four leading models overlap widely. The exact McNemar test found Logistic Regression to classify significantly more patients correctly than the Decision Tree (p = 0.012) and the MLP (p = 0.008), but not more than Random Forest, SVM or XGBoost ($p \geq 0.13$).

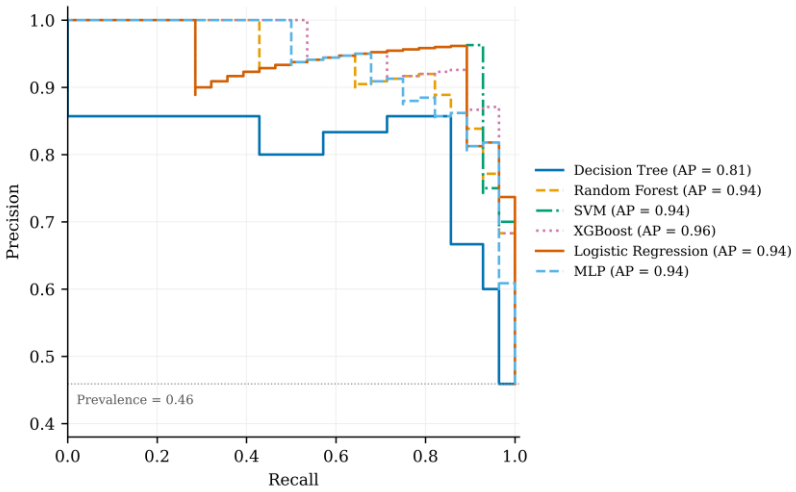
Table 4: Hold-out test set results at the screening threshold, full feature set (61 patients, 28 with disease)

Model	TP	FN	TN	FP	Recall [95% CI]	Specificity [95% CI]	Prec.	F1	Acc.	PR-AUC
Decision Tree	27	1	15	18	0.96 [0.82, 0.99]	0.45 [0.30, 0.62]	0.60	0.74	0.69	0.81
Random Forest	28	0	20	13	1.00 [0.88, 1.00]	0.61 [0.44, 0.75]	0.68	0.81	0.79	0.94
SVM	27	1	22	11	0.96 [0.82, 0.99]	0.67 [0.50, 0.80]	0.71	0.82	0.80	0.94
XGBoost	28	0	19	14	1.00 [0.88, 1.00]	0.58 [0.41, 0.73]	0.67	0.80	0.77	0.96
Logistic Regression	28	0	23	10	1.00 [0.88, 1.00]	0.70 [0.53, 0.83]	0.74	0.85	0.84	0.94

Model	TP	FN	TN	FP	Recall [95% CI]	Specificity [95% CI]	Prec.	F1	Acc.	PR-AUC
MLP	28	0	15	18	1.00 [0.88, 1.00]	0.45 [0.30, 0.62]	0.61	0.76	0.70	0.94

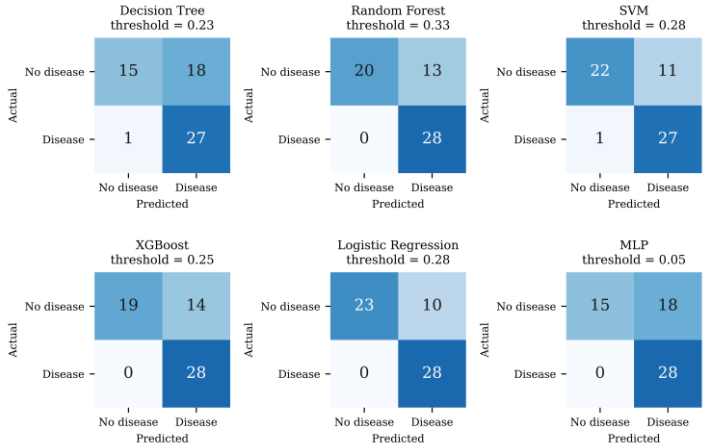
Note. TP = true positives; FN = false negatives; TN = true negatives; FP = false positives; Prec. = precision; Acc. = accuracy. Recall and specificity intervals are Wilson score intervals.
Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

Figure 3: Precision-recall curves of the six models on the hold-out test set (full feature set)



Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989). AP = average precision.

Figure 4: Confusion matrices on the hold-out test set at each model's screening threshold (full feature set)



Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

PR-AUC on the hold-out set ranked XGBoost highest (0.96), with SVM, Logistic Regression, Random Forest and the MLP all at 0.94. This ordering differs from the nested cross-validation ranking, in which Logistic Regression led narrowly. On a test set of 61 patients, a shift of one or two patients in the ranking of predicted probabilities is enough to reorder the models, which is why the nested results are treated as the main evidence.

Pre-test Feature Set (RQ3)

When the models were restricted to age, sex, chest pain type, resting blood pressure and resting ECG, they could still be tuned to a recall of about 0.90, but they paid for it in specificity (Figure 2b). Logistic Regression specificity fell from 0.68 to 0.57 ($p = 0.002$), its PR-AUC from 0.91 to 0.82 ($p < 0.001$) and its ROC-AUC from 0.91 to 0.84 ($p < 0.001$). Table 5 shows that the loss was general: PR-AUC fell significantly for every model except the Decision Tree, and no pre-test model ruled out more than 59% of disease-free patients at the 0.90 recall target. Logistic Regression retained the highest PR-AUC (0.82) and ROC-AUC (0.84) among the pre-test models.

Table 5: Repeated nested cross-validation results at the screening threshold, pre-test feature set (age, sex, chest pain type, resting blood pressure, resting ECG)

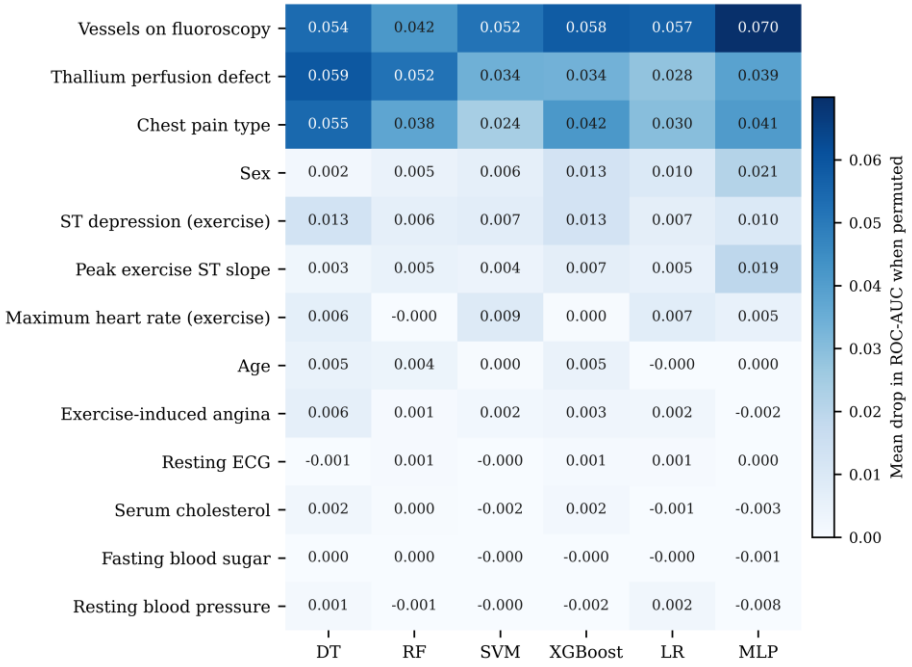
Model	Recall	Specificity	Precision	PR-AUC	ROC-AUC	Brier
Decision Tree	0.90	0.58	0.65	0.75	0.82	0.178
Random Forest	0.91	0.59	0.66	0.81	0.83	0.171
SVM	0.92	0.50	0.62	0.81	0.83	0.174
XGBoost	0.91	0.59	0.66	0.79	0.82	0.170
Logistic Regression	0.91	0.57	0.65	0.82	0.84	0.164
MLP	0.91	0.52	0.62	0.76	0.81	0.180

Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

Feature Attribution (RQ4)

Figure 5 shows the permutation importance of each variable. Three variables dominated every model: the number of vessels coloured by fluoroscopy, the thallium perfusion result and chest pain type. For Logistic Regression, permuting the vessel count reduced ROC-AUC by 0.057 on average, chest pain type by 0.030 and the thallium result by 0.028. Sex, exercise ST depression and the ST slope made smaller contributions, and serum cholesterol, fasting blood sugar, resting blood pressure and the resting ECG contributed almost nothing once the other variables were present. The two strongest predictors therefore come from imaging and fluoroscopy, which explains why the pre-test models lost so much specificity.

Figure 5: Permutation importance of each input variable, averaged over 50 outer test folds (full feature set)

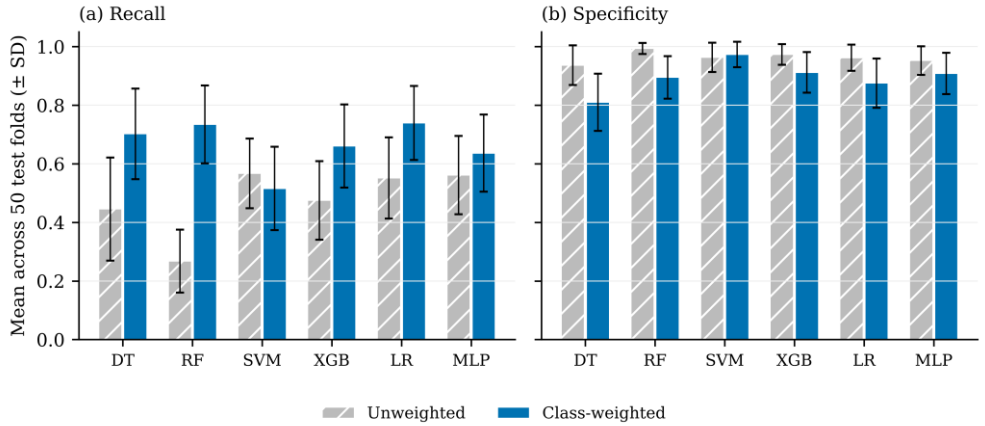


Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989). DT = Decision Tree; RF = Random Forest; LR = Logistic Regression.

Low Prevalence and Predictive Values (RQ4)

When the training data were subsampled to a disease prevalence of 15%, the unweighted models became heavily biased towards the disease-free class at the 0.5 threshold (Figure 6). Mean recall fell to 0.27 for Random Forest, 0.45 for the Decision Tree, 0.48 for XGBoost and 0.55 for Logistic Regression. Class weighting restored much of the lost recall: Logistic Regression rose to 0.74, Random Forest to 0.73, the Decision Tree to 0.70 and XGBoost to 0.66, at the cost of a modest drop in specificity (for Logistic Regression, from 0.96 to 0.88). The SVM was the exception, with recall of 0.57 unweighted and 0.52 weighted. Its probability outputs come from an internal Platt-scaling step fitted on the imbalanced training data, which appears to cancel the effect of the class weights at the 0.5 cut-off. None of the weighted models reached a recall of 0.90 at the default threshold, which confirms that weighting alone is not a substitute for setting the threshold explicitly.

Figure 6: Recall and specificity at the 0.5 threshold after training at 15% disease prevalence, with and without class weighting



Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989). Bars show the mean and standard deviation across 50 test folds; XGB = XGBoost.

Table 6 translates the nested cross-validation sensitivity and specificity of Logistic Regression into the predictive values expected at lower prevalence. At a prevalence of 10%, about one in four referred patients would have disease (PPV 0.24), while 98.5% of those not referred would be disease-free. Per 1,000 patients assessed, the model would refer about 375 and miss about 10.

Table 6: Expected predictive values of Logistic Regression at the screening threshold for different disease prevalences

Prevalence	PPV	NPV	Referred per 1,000	Missed per 1,000
5%	0.13	0.993	346	5
10%	0.24	0.985	375	10
20%	0.42	0.966	434	19
45.9% (Cleveland)	0.71	0.893	586	44

Note. Values derived from the nested cross-validation recall (0.904) and specificity (0.684) using Equations 10 and 11.

Source: Authors' analysis of the UCI Heart Disease dataset, Cleveland file (Janosi et al., 1989).

Discussion

The central finding is that, once recall is held at 0.90, the choice between Logistic Regression, SVM and Random Forest makes little difference on these data. All three ruled out about two-thirds of disease-free patients, and their specificities, PR-AUCs and Brier scores were within resampling noise of one another. Logistic Regression was marginally ahead on specificity and PR-AUC in the nested design and on specificity in the hold-out set, but the evidence does not support calling it superior. Under the selection rule fixed in

Section 3, it is preferred because it cannot be separated from the best of the other models and is the simplest of them: its coefficients can be read directly, its probabilities were well calibrated, and it can be deployed as a scoring sheet or a few lines of code on a mobile device.

The results also show why single-split, default-threshold comparisons can mislead. A Decision Tree can appear to lead on recall when each model is run at a default threshold on a single split, because a shallow tree with balanced class weights tends to label more patients as diseased. Once every model is tuned to the same recall target, the Decision Tree is shown to be the least useful of the six: it reaches that recall only by referring most disease-free patients as well. The same reasoning applies to accuracy. At the 0.5 threshold, Logistic Regression and SVM both reached about 85% accuracy, close to the 88.7% reported by Mohan et al. (2019), yet at that operating point they missed about one diseased patient in five. Figures reported on derivative versions of the data, such as the 1,025-record Kaggle compilation used by Ilyas et al. (2025), are not directly comparable, because those versions differ in size, coding and, as shown here, sometimes in the meaning of the outcome label itself.

The label reversal deserves attention beyond this paper. In the reversed copy, the class reported as positive is the disease-free class, so a model described as highly sensitive is in fact good at confirming health. Variable importances computed on that copy remain valid as measures of association, which is why such errors can go unnoticed, but any statement about missed diagnoses, recall or triage safety is inverted. Studies that use the Cleveland data should state which file they used and how the outcome was coded.

The feature analysis agrees with clinical knowledge. The number of diseased vessels seen on fluoroscopy and the thallium perfusion result are close to the reference standard of angiography, and chest pain character has long been a cornerstone of pre-test probability estimation (Detrano et al., 1989). The flip side is that the full model depends on tests that are themselves part of a cardiology work-up. When the model is restricted to first-consultation variables, specificity drops to about 0.57, so a front-line clinic using such a model would refer roughly two disease-free patients for every three it clears. That may still be useful in a setting with no structured referral criteria, but it is a much more modest tool than the full model, and the two should not be confused.

The low-prevalence experiment speaks to how the models would behave outside a referral cohort. Without weighting, a Random Forest trained at 15% prevalence identified only about a quarter of diseased patients at the default threshold. Weighting the classes recovered most of that recall for every model except the probability-calibrated SVM, but not enough to meet the 0.90 target, so explicit threshold selection remains necessary. The predictive values in Table 6 make the trade-off concrete: at 10% prevalence the model would miss about ten patients per thousand while referring a little over a third of those assessed. Whether that balance is acceptable is a clinical and economic judgement that depends on the cost and availability of further testing.

Limitations

Several limitations qualify these findings. The study used a single, small dataset of 303 patients collected more than three decades ago at one centre, so the results describe internal performance only; no external or temporal validation was possible. The patients were a referral cohort who had been selected for angiography, which inflates disease prevalence and probably makes the classes easier to separate than in an unselected population. The strongest predictors come from tests that are not available at a first contact, so the full model is not suitable for community screening as it stands. The thresholds were chosen to meet a recall target on training data, and recall in new data may fall below that target if the case mix differs. Calibration was assessed with the Brier score only, without calibration curves or recalibration. Finally, the grids searched were modest, and more extensive tuning or other model families might shift the results slightly, although the overlap between the leading models suggests that the data, rather than the algorithm, set the ceiling.

Conclusion

On the Cleveland Heart Disease data, and with the decision threshold set to deliver a recall of 0.90, Logistic Regression, SVM and Random Forest performed equivalently, each ruling out about two-thirds of disease-free patients. Logistic Regression is recommended among them because it matched the best performance while being the simplest and most transparent model. The Decision Tree and the MLP were clearly weaker at the same recall. The full model relies on exercise testing, thallium imaging and fluoroscopy, and its performance falls considerably when only first-consultation variables are used. These conclusions apply to a symptomatic referral population and to internal validation on a single historical dataset.

Recommendations

For practitioners and developers of clinical decision support tools, the following recommendations arise from the study:

- Fix the acceptable rate of missed cases before choosing a model, and set the decision threshold on training or validation data to meet it; do not rely on the default 0.5 cut-off or on accuracy.
- Report specificity, predictive values at a realistic prevalence, and confidence intervals alongside recall, so that the referral burden implied by a recall target is visible.
- Verify the source file and outcome coding of public datasets before modelling; for the Cleveland data, use the UCI file and state how *num* was dichotomised.
- Describe the inputs honestly: a model that uses exercise, imaging or laboratory results is not a pre-test or laboratory-free tool.

For future research, the following are recommended:

- External and temporal validation in contemporary Nigerian and West African cohorts, including unselected primary-care patients.
- Development of pre-test models from variables that can be collected at a first consultation, with additional low-cost inputs such as a resting ECG interpreted by AI (Pedroso et al., 2026).
- Calibration analysis, recalibration and decision-curve analysis to weigh missed cases against unnecessary referrals.
- Prospective evaluation of a recall-oriented triage tool, measuring time to diagnosis, referral volume and patient outcomes.

Data Availability Statement

The data analysed in this study are openly available in the UCI Machine Learning Repository at https://doi.org/10.24432/C52P4X (Heart Disease dataset, processed Cleveland file).

Ethics Approval

Not applicable. The study used a publicly available, de-identified dataset.

Consent to Participate

Not applicable.

Author Contributions

Reginald E. Ogu: conceptualisation and research administration; Juliana C. Ihuefo: writing of the original draft; Precious Iheanacho: programming and editing. All authors reviewed and approved the revised manuscript.

Conflict of Interests

The authors declare that they have no conflict of interest related to this study or its publication.

Funding

The authors received no financial support or funding for the research, authorship, or publication of this manuscript.

Declaration of Generative AI

Generative AI tools were used during the preparation and revision of this manuscript. During the revision, an AI assistant (Claude, Anthropic) helped to verify the dataset, write and run the analysis code, prepare the figures and edit the text. The authors checked the code, the outputs and the text, and remain fully responsible for the content, accuracy and integrity of the manuscript.

References

Ahmad, A., & Polat, H. (2023). Prediction of heart disease based on machine learning using jellyfish optimization algorithm. *Diagnostics, 13*(14), 2392. https://doi.org/10.3390/diagnostics13142392

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *BMJ, 309*(6947), 102. https://doi.org/10.1136/bmj.309.6947.102

Banerjee, T., & Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. *Turkish Journal of Biology, 49*(5), 600–634. https://doi.org/10.55730/1300-0152.2766

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chakraborty, A., Das, U. K., Sazzad, S., Das, P., & Khan, M. M. H. (2026). Explainable machine learning for early heart disease risk prediction: Insights from a clinical dataset in Bangladesh. *Intelligence-Based Medicine, 13*, 100352. https://doi.org/10.1016/j.ibmed.2026.100352

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology, 64*(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Gul, G., Korejo, I. A., Hakro, D. N., Alqahtani, H., Abbasi, A., Babar, M., Al Rahbi, O., & Ali, N. I. (2026). Machine learning and ensemble methods for cardiovascular disease prediction: A systematic review of approaches, performance trends, and research challenges. *Computers, 15*(1), 25. https://doi.org/10.3390/computers15010025

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Ilyas, A. R., Javaid, S., & Kharisma, I. L. (2025). Heart disease prediction using ML. *Engineering Proceedings, 107*(1), 124. https://doi.org/10.3390/engproc2025107124

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). *Heart disease* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X

Khan, S. S., Bhave, N., Blumenthal, R. S., Coresh, J., Huang, X., Joseph, J. J., Khera, A., Ho, J. E., Lloyd-Jones, D. M., Low Wang, C. C., Lu, Y., Morris, P. B., Nasir, K., Natarajan, P., Rangaswami, J., Rodriguez, F., Sperling, L. S., Virani, S. S., Zhang, S., & Ndumele, C. E. (2026). Use of predicted risk and expected benefit to guide decision-making in cardiovascular-kidney-metabolic syndrome for the primary prevention of cardiovascular disease: A scientific statement from the American Heart Association and American College of Cardiology. *Circulation, 154*(4), e159–e180. https://doi.org/10.1161/CIR.0000000000001447

Kumar, R., Garg, S., Kaur, R., Johar, M. G. M., Singh, S., Menon, S. V., Kumar, P., Hadi, A. M., Hasson, S. A., & Lozanović, J. (2025). A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions. *Frontiers in Artificial Intelligence, 8*, 1583459. https://doi.org/10.3389/frai.2025.1583459

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked, 16*, 100203. https://doi.org/10.1016/j.imu.2019.100203

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Mohapatra, S., Maneesha, S., Mohanty, S., Patra, P. K., Bhoi, S. K., Sahoo, K. S., & Gandomi, A. H. (2023). A stacking classifiers model for detecting heart irregularities and predicting cardiovascular disease. *Healthcare Analytics, 3*, 100133. https://doi.org/10.1016/j.health.2022.100133

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning, 52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering, 2022*, 7351061. https://doi.org/10.1155/2022/7351061

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pedroso, A. F., Nascimento, B. R., Dhingra, L. S., Shankar, S. V., Vinhal, W. C., e Reges, R. B., Cardoso, C. S., Sable, C., Ribeiro, A. L. P., & Khera, R. (2026). Artificial intelligence-enabled electrocardiography to triage echocardiography for structural heart disease diagnosis in a low-resource setting. *American Journal of Preventive Cardiology, 27*, 101539. https://doi.org/10.1016/j.ajpc.2026.101539

Rehman, M. U., Naseem, S., Butt, A., Mahmood, T., Khan, A. R., Khan, I., Khan, J., & Jung, Y. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports, 15*, 13361. https://doi.org/10.1038/s41598-025-96437-1

Rhee, T.-M., Kang, S., Lee, M. S., Han, G. I., Yoo, A.-H., Jang, J.-H., Jo, Y.-Y., Son, J. M., Kwon, J., Choi, S.-Y., Lee, H. S., & Lee, H. (2026). Artificial intelligence–driven electrocardiogram screening for asymptomatic left ventricular systolic dysfunction in the general population. *JACC: Advances, 5*(4), 102660. https://doi.org/10.1016/j.jacadv.2026.102660

Safdar, S., Zafar, S., Zafar, N., & Khan, N. F. (2018). Machine learning based decision support systems (DSS) for heart disease diagnosis: A review. *Artificial Intelligence Review, 50*(4), 597–623. https://doi.org/10.1007/s10462-017-9552-8

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

Suleiman, A. B., Luka, S., & Ibrahim, M. (2023). Cardiovascular disease prediction using random forest machine learning algorithm. *FUDMA Journal of Sciences, 7*(6). https://doi.org/10.33003/fjs-2023-0706-2128

Ullah, I., Inayat, T., Ullah, N., Alzahrani, F., & Khan, M. I. (2023). Clinical decision support system (CDSS) for heart disease diagnosis and prediction by machine learning algorithms: A systematic literature review. *Journal of Mechanics in Medicine and Biology, 23*(10), 2330001. https://doi.org/10.1142/S0219519423300016

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., Macaskill, P., Moons, K. G. M., & Vickers, A. J. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91. https://doi.org/10.1186/1471-2105-7-91

Wan, S., Wan, F., & Dai, X. (2025). Machine learning approaches for cardiovascular disease prediction: A review. *Archives of Cardiovascular Diseases, 118*(10), 554–562. https://doi.org/10.1016/j.acvd.2025.04.055

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953