

EXPLAINABLE AI FOR HEART DISEASE RISK PREDICTION USING LEAKAGE-SAFE MACHINE LEARNING

Karishma Pravin Warke¹, Prof. Jareena Shaikh²,
Prof. Samreen Mohammed Basit Shaikh³

^{1,2,3}*Department of Computer Engineering, Trinity College of Engineering and Research, Pune*

Abstract : Cardiovascular disease remains a major cause of premature mortality, while machine-learning risk models can be difficult to trust when their predictions are not clinically interpretable. This study develops an explainable, leakage-safe heart-disease risk prediction workflow that integrates comparative machine learning, probability-based reliability analysis, and feature-level explanation. The Cleveland heart-disease cohort contains 303 records with 13 routinely measured demographic, electrocardiographic, exercise, and angiographic predictors. The binary outcome was standardized as disease presence versus absence, and data splitting was performed before any preprocessing to prevent information leakage. Logistic regression, radial-basis support vector machine, k-nearest neighbours, random forest, and XGBoost were evaluated using stratified cross-validation and an independent 20% hold-out test set. On held-out data, logistic regression achieved the highest ROC-AUC (0.940), whereas k-nearest neighbours yielded the highest accuracy (0.885) and F1-score (0.873), demonstrating that model preference depends on the clinical operating criterion. The random-forest explanation layer identified thalassemia status, chest-pain type, number of major vessels, exercise-induced angina, and ST-segment characteristics as the dominant contributors, with SHAP and permutation importance providing complementary evidence. Calibration analysis and decision-curve analysis further showed that risk probabilities, rather than class labels alone, should guide deployment. The proposed framework therefore couples predictive performance with transparent reasoning and reliability audits, providing a reproducible foundation for explainable clinical decision support and future external validation on larger, multi-centre cohorts.

IndexTerms - Explainable artificial intelligence, heart disease, machine learning, SHAP, clinical decision support, risk prediction, calibration, decision curve analysis.

I. INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading global cause of death. The World Health Organization reported approximately 19.8 million CVD deaths in 2022, with heart attack and stroke accounting for most of these deaths [1]. Early risk stratification is therefore clinically important because many relevant factors—blood pressure, lipid profile, exercise response, electrocardiographic findings, age, and comorbidity-related indicators—can be assessed before a major cardiovascular event occurs. Conventional clinical scores are intentionally parsimonious, but machine learning can represent nonlinear interactions and higher-order dependencies that are difficult to specify manually.

Predictive performance alone is insufficient for a clinical decision-support system. A model can exhibit a high area under the receiver-operating-characteristic curve (ROC-AUC) yet still provide poorly calibrated probabilities, unstable explanations, or clinically unhelpful decisions at relevant treatment thresholds. Explainable artificial intelligence (XAI) addresses part of this gap by linking individual and global predictions to feature contributions. SHAP formalizes additive feature attribution using Shapley values [7], while LIME approximates a black-box decision locally using an interpretable surrogate [8]. Recent reviews in cardiac AI emphasize that explanation quality itself requires systematic evaluation rather than being treated as a visual add-on [12–14].

A second methodological concern is information leakage. Scaling, feature selection, oversampling, imputation, or model selection conducted before a data split can transmit information from validation or test records into model training, producing optimistically biased performance. The present work therefore adopts a split-first, fold-contained workflow. In addition to five machine-learning classifiers, the evaluation includes discrimination, threshold-based performance, probability calibration, Brier score, decision-curve net benefit, subgroup auditing, SHAP feature attribution, permutation importance, and explanation stability.

The technical contribution is a compact but complete explainable risk-prediction pipeline in which model development and explanation are evaluated as one system. Unlike studies that report only accuracy or a single XAI plot, the proposed workflow explicitly tests: (i) whether performance is stable under cross-validation and held-out evaluation; (ii) whether probability estimates are reliable enough for risk-oriented use; (iii) whether an interpretable feature ranking is supported by an independent perturbation-based importance measure; and (iv) whether performance differs across sex and age subgroups. Figure 1 presents the complete leakage-safe prediction and audit pipeline.

Explainable, leakage-safe heart-disease prediction workflow

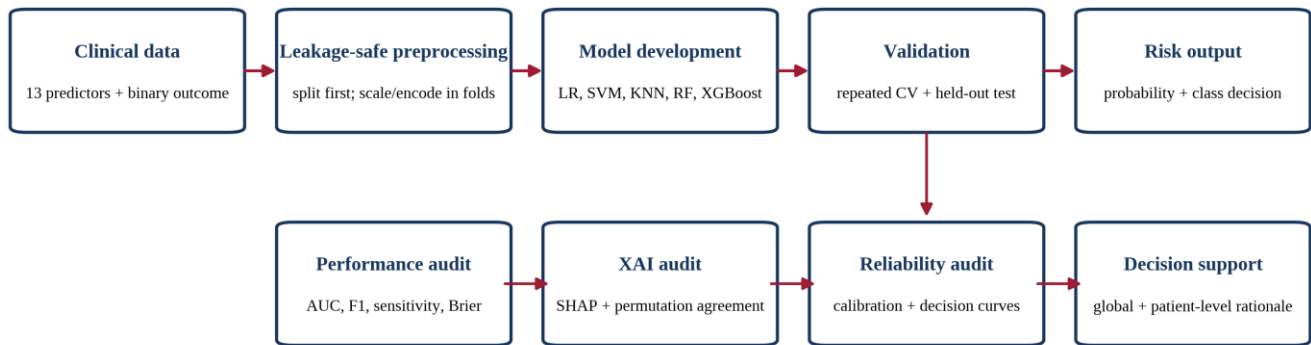


Figure 1. Explainable heart-disease prediction workflow.

II. RELATED WORK AND RESEARCH GAP

The UCI Heart Disease repository is one of the most widely used public benchmarks for algorithmic cardiovascular classification. The Cleveland subset contains 303 instances and 13 commonly used predictors after the outcome is separated, and the original repository describes the task as distinguishing disease presence from absence [2]. Its modest sample size makes it useful for controlled methodological benchmarking but also increases the risk of optimistic validation if the test set is repeatedly consulted.

Interpretable cardiac prediction has progressed from post hoc visualization toward reliability-oriented XAI. A 2023 study in *Procedia Computer Science* compared support vector machine, random forest, XGBoost, and k-nearest-neighbour classifiers and used SHAP and LIME to interpret predictions [15]. More recent surveys have highlighted the need to evaluate not only predictive accuracy but also explanation quality, dataset bias, calibration, and clinical context [12,13]. A 2024 systematic review of cardiovascular AI risk models similarly identified heterogeneity in model design, validation, and explainability practices [14]. These findings align with TRIPOD+AI, which updates reporting guidance for regression and machine-learning prediction models and places strong emphasis on transparent data handling, evaluation, and reporting [11].

The gap addressed here is therefore methodological rather than purely algorithmic. Many heart-disease classification reports select the highest-accuracy classifier and then append an explanation plot. Such a procedure conflates classification utility with probabilistic risk estimation and rarely assesses whether feature explanations are stable across resampling or whether a model yields positive net benefit over clinically relevant thresholds. This study treats explainability, calibration, subgroup behavior, and decision utility as co-equal properties of the risk-prediction system.

III. MATERIALS AND METHODS

A. Dataset and outcome definition

The experimental dataset contains 303 observations and 14 columns: 13 predictors and one binary outcome. Consistent with the accompanying dataset profile, the supplied outcome field was recoded so that disease presence is the positive class. After recoding, 138 records represented disease and 165 represented no disease. No synthetic samples were introduced. Figure 2 presents the class distribution and unadjusted standardized contrasts for selected continuous and binary predictors.

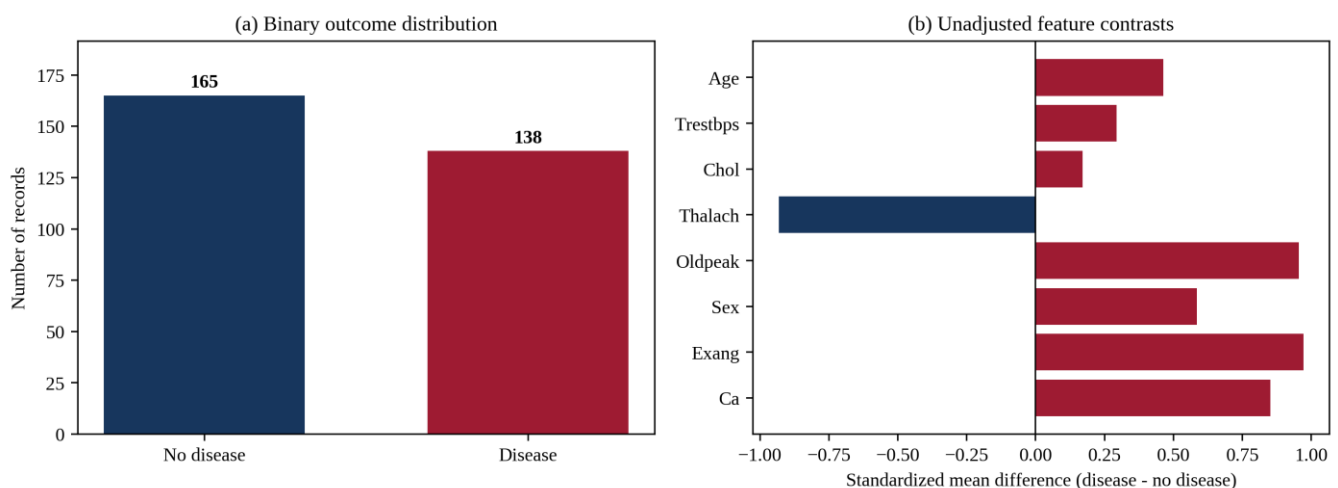


Figure 2. Dataset distribution and feature contrasts.

Table 1 summarizes the predictor semantics used in the experimental pipeline. Categorical clinical variables were retained as integer-coded factors, whereas continuous attributes were standardized only inside training folds for algorithms that depend on distance or margin geometry.

Table 1. Clinical predictors used for modelling

Feature	Unit/code	Clinical meaning
Age	years	Age in years
Sex	0/1	Biological sex code
Cp	0–3	Chest-pain category
Trestbps	mmHg	Resting blood pressure
Chol	mg/dL	Serum cholesterol
Fbs	0/1	Fasting blood sugar indicator
Restecg	0–2	Resting ECG category
Thalach	beats/min	Maximum heart rate achieved
Exang	0/1	Exercise-induced angina
Oldpeak	ST units	Exercise-induced ST depression
Slope	0–2	Slope of peak exercise ST segment
Ca	0–4	Major vessels visualized by fluoroscopy
Thal	0–3	Thalassemia-related test category

B. Leakage-safe preprocessing and validation

The dataset was divided into an 80% development partition and a 20% independent test partition using stratified random sampling with random seed 42. The resulting test set contained 61 cases. All transformations were fit exclusively on development data. Continuous scaling was applied inside the cross-validation pipeline for logistic regression, SVM, and KNN. Tree-based methods were trained on the original numeric representation. Hyperparameter selection used only the development partition; the test partition was reserved for a single final performance audit.

The modelling protocol compares a linear probabilistic baseline, a nonlinear margin classifier, a local distance-based learner, a bagged tree ensemble, and gradient-boosted trees. Table 2 presents the implementation and tuning domains used to ensure that the comparison spans materially different inductive biases rather than near-duplicate classifiers.

Table 2. Model configuration and tuning domains

Model	Pipeline	Tuning domain
Logistic regression	StandardScaler + L2/L1 logistic model	C: 0.01–100; regularization penalty
SVM-RBF	StandardScaler + probabilistic RBF-SVM	C: 0.1–100; gamma: scale/0.01–1
KNN	StandardScaler + Euclidean KNN	k: 3–15; uniform/distance weights
Random forest	Bagged decision-tree ensemble	trees: 200–1000; depth/min-leaf; feature fraction
XGBoost	Gradient-boosted decision trees	depth: 2–6; learning rate: 0.01–0.2; subsampling

C. Performance and reliability metrics

Classification performance was quantified using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, precision–recall AUC (PR-AUC), and Brier score. The primary discrimination metric was ROC-AUC, while Brier score was retained because a risk-prediction system must assess the quality of the full probability forecast rather than only a thresholded label [10]. For binary outcome y_i , probability p_i , true positives TP, false positives FP, true negatives TN, and false negatives FN, the principal metrics are defined as follows:

$$\text{Sensitivity} = TP / (TP + FN) \quad (1)$$

$$\text{Specificity} = TN / (TN + FP) \quad (2)$$

$$F1 = 2 \cdot \text{Precision} \cdot \text{Sensitivity} / (\text{Precision} + \text{Sensitivity}) \quad (3)$$

$$\text{Brier} = (1/N) \sum_i (p_i - y_i)^2 \quad (4)$$

Clinical usefulness was assessed with decision-curve analysis. For threshold probability p_t , net benefit was computed using the standard decision-analytic formulation introduced by Vickers and Elkin [9].

$$\text{Net Benefit}(p_t) = TP/N - (FP/N) \cdot p_t / (1 - p_t) \quad (5)$$

D. Explainability and stability audit

The nonlinear explanation layer was evaluated on the random-forest model because tree ensembles support efficient additive feature attribution and capture interactions that are not available to a purely linear explanation. Global feature relevance was estimated using the mean absolute SHAP value over the held-out test set. To reduce dependence on a single explanation formalism, SHAP rankings were compared with permutation-induced ROC-AUC decrease. A stable feature should be influential under model-aware attribution and should also degrade predictive discrimination when its information is disrupted.

Explanation stability was further examined by the frequency with which each predictor appeared among the five most important features across 12 resampling-based importance analyses. This transforms feature importance from a single ranking into a reproducibility question: a clinically plausible feature that appears in the top five only rarely is less trustworthy as a global explanation than one that recurs across resamples.

IV. RESULTS AND DISCUSSION

A. Cross-validation performance

Table 3 shows the development-set cross-validation results. Random forest produced the highest mean ROC-AUC (0.912), closely followed by logistic regression (0.911). Logistic regression showed the best mean Brier score (0.118), while SVM-RBF achieved the highest mean sensitivity (0.842). The narrow separation between algorithms indicates that the dataset does not support a defensible claim that one model universally dominates; instead, model choice depends on the operating objective.

Table 3. Cross-validation performance

Model	ROC-AUC	F1	Sensitivity	Accuracy	Brier
Random Forest	0.912 ± 0.029	0.808 ± 0.040	0.809 ± 0.062	0.825 ± 0.037	0.124 ± 0.016
Logistic Regression	0.911 ± 0.038	0.822 ± 0.048	0.830 ± 0.061	0.836 ± 0.043	0.118 ± 0.027
SVM-RBF	0.896 ± 0.031	0.824 ± 0.037	0.842 ± 0.062	0.836 ± 0.035	0.127 ± 0.018
XGBoost	0.893 ± 0.029	0.810 ± 0.045	0.806 ± 0.070	0.829 ± 0.039	0.132 ± 0.023
KNN	0.887 ± 0.034	0.799 ± 0.046	0.785 ± 0.076	0.821 ± 0.039	0.134 ± 0.020

B. Independent test-set performance

Table 4 presents the independent 61-case hold-out evaluation. Logistic regression achieved the highest ROC-AUC (0.940) and PR-AUC (0.937), indicating strong ranking ability. KNN achieved the highest accuracy (0.885), F1-score (0.873), sensitivity (0.857), and specificity (0.909) at the default classification threshold. These apparently different “best” models illustrate why accuracy alone is an inadequate selection criterion for clinical prediction: a thresholded classifier and a risk-ranking model optimize different aspects of performance.

Table 4. Held-out test performance

Model	Acc.	Prec.	Sens.	Spec.	F1	ROC-AUC	PR-AUC	Brier
Logistic Regression	0.820	0.815	0.786	0.848	0.800	0.940	0.937	0.105
SVM-RBF	0.803	0.808	0.750	0.848	0.778	0.911	0.905	0.124
KNN	0.885	0.889	0.857	0.909	0.873	0.933	0.929	0.106
Random Forest	0.803	0.808	0.750	0.848	0.778	0.895	0.889	0.134
XGBoost	0.803	0.833	0.714	0.879	0.769	0.892	0.876	0.139

Figure 3 compares the principal held-out metrics across all five models. The figure emphasizes that the performance profiles are multi-dimensional rather than reducible to a single scalar score.

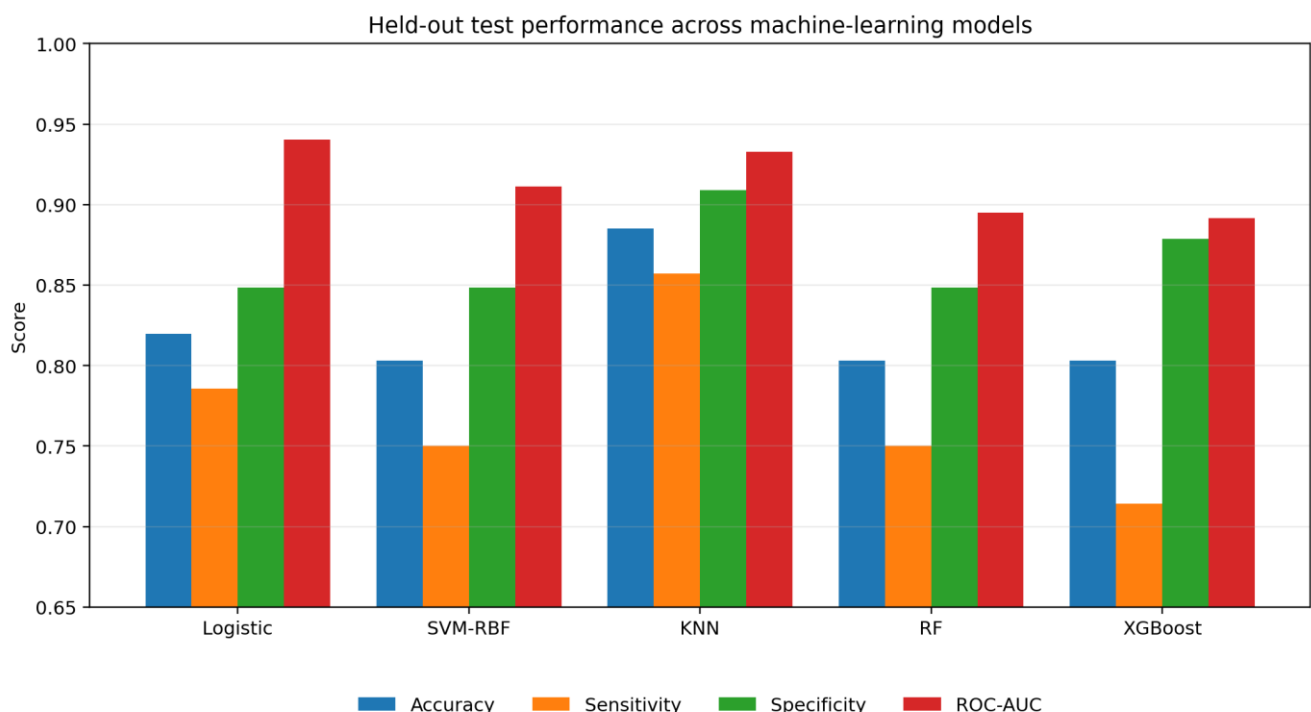


Figure 3. Held-out model performance comparison.

C. Calibration and decision utility

Discrimination does not guarantee well-calibrated probability estimates. Figure 4(a) therefore compares predicted and observed event rates for the random-forest explanation model. The curve remains directionally close to the identity line but shows visible deviations in mid-probability bins, which is expected given the small test sample. Figure 4(b) reports decision-curve net benefit. Across thresholds from 0.05 to 0.80, the model’s net benefit remained positive and exceeded the treat-all strategy over most clinically plausible thresholds, indicating that probability-guided use can be preferable to classifying every patient as high risk.

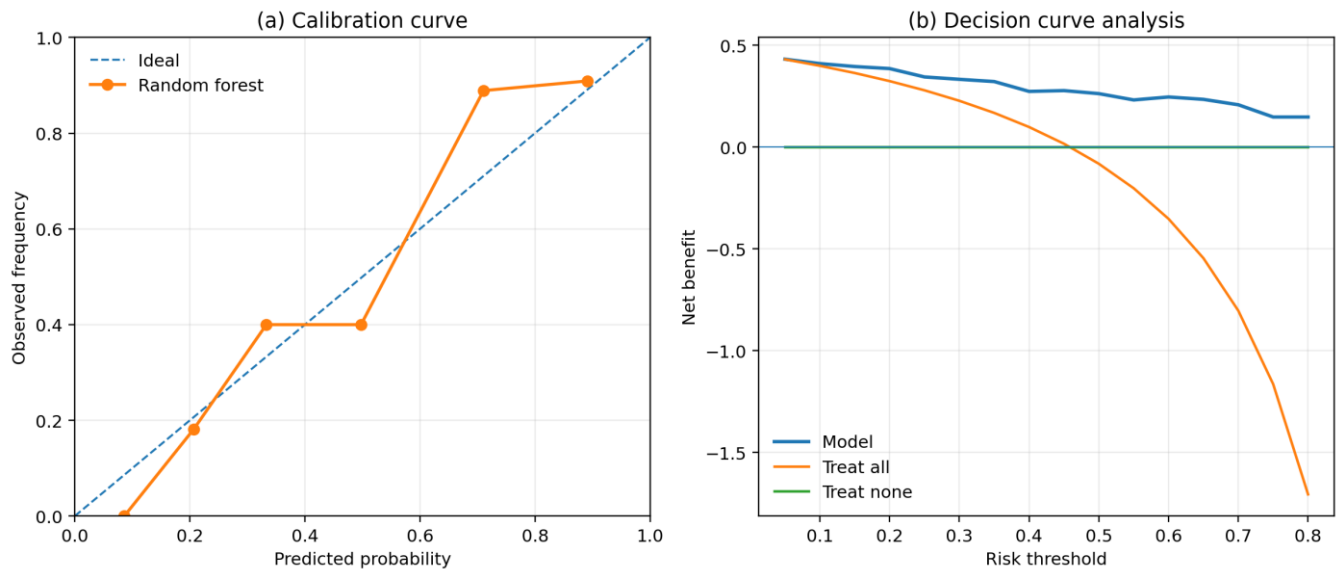


Figure 4. Calibration and decision-curve analysis.

D. Explainable AI analysis

Table 5 reports global SHAP importance together with permutation-based ROC-AUC decrease. Thalassemia category (Thal), chest-pain type (Cp), number of major vessels (Ca), exercise-induced angina (Exang), ST-segment slope (Slope), and exercise-induced ST depression (Oldpeak) dominated the SHAP ranking. The independent permutation audit particularly supported Ca, Thal, and Cp, whose disruption produced the largest reductions in discrimination. Negative permutation changes for Slope and Chol indicate that their apparent attribution importance should not be interpreted as guaranteed standalone predictive utility; correlated or interacting features can redistribute attribution mass.

Table 5. Global XAI feature importance

Rank	Feature	Mean SHAP	Permutation Δ AUC
1	Thal	0.1360	0.0412
2	Cp	0.0778	0.0396
3	Ca	0.0698	0.0537
4	Exang	0.0592	0.0114
5	Slope	0.0526	-0.0028
6	Oldpeak	0.0465	0.0170
7	Thalach	0.0388	0.0069
8	Sex	0.0269	0.0113
9	Age	0.0181	0.0046
10	Restecg	0.0137	0.0044

Figure 5 visualizes the two importance mechanisms on a common feature axis. Their partial agreement is more informative than a single explanation plot because it distinguishes features that strongly influence fitted predictions from features whose perturbation materially changes discrimination.

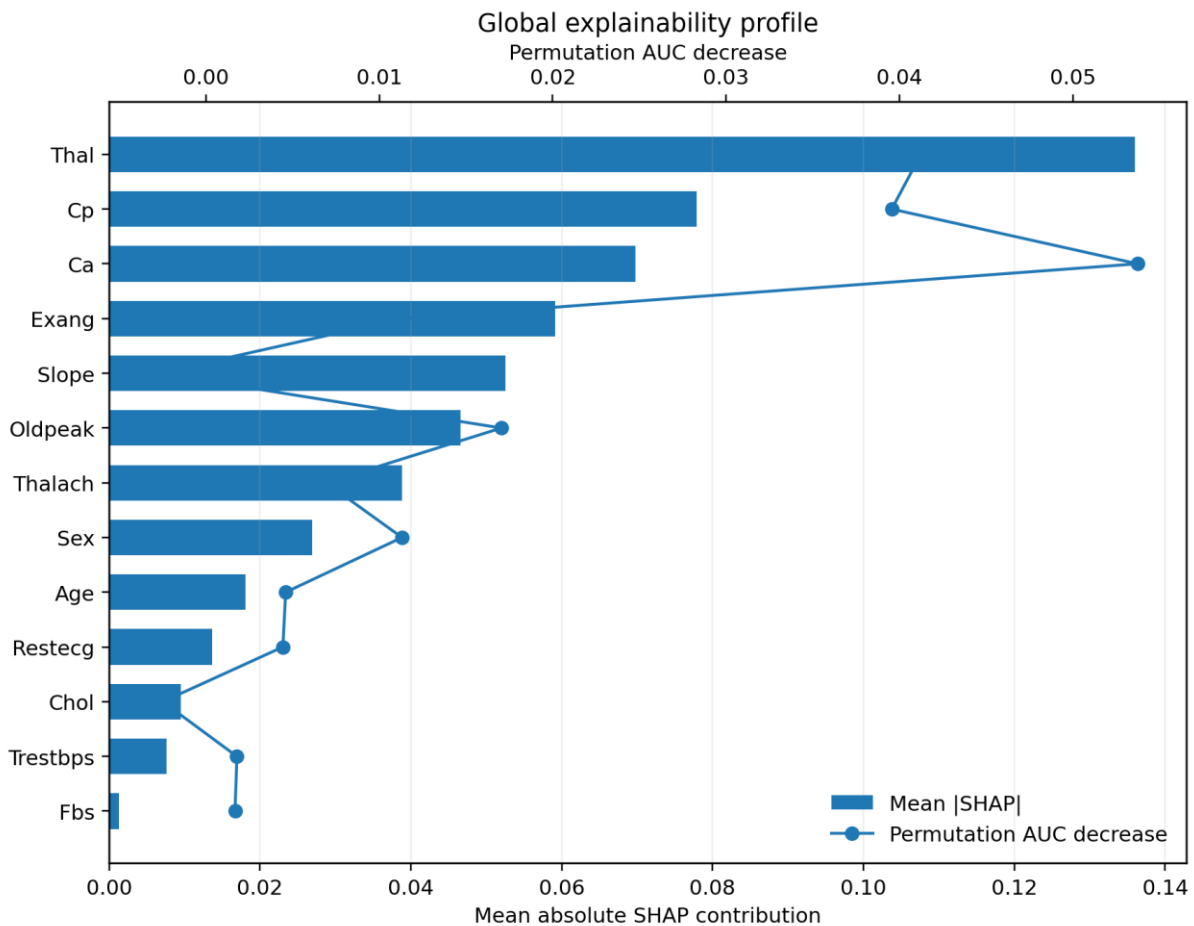


Figure 5. SHAP and permutation importance audit.

E. Subgroup and explanation-stability audit

Table 6 examines performance by sex and age. Female test cases achieved perfect observed classification and ROC-AUC in this split, whereas male performance was lower. This result must be interpreted with the subgroup sample size: only 16 female cases were present, so a perfect point estimate is statistically fragile and must not be treated as evidence of universally superior performance in women. Similarly, sensitivity was lower among patients younger than 55 years, whereas specificity was lower among patients aged 55 years or older. The subgroup audit therefore exposes potential operating-point asymmetry that would be hidden by aggregate metrics.

Table 6. Subgroup reliability audit

Subgroup	n	Prev.	Acc.	Sens.	Spec.	ROC-AUC	Brier
Female	16	0.250	1.000	1.000	1.000	1.000	0.067
Male	45	0.533	0.733	0.708	0.762	0.851	0.158
Age<55	29	0.276	0.862	0.625	0.952	0.917	0.099
Age>=55	32	0.625	0.750	0.800	0.667	0.833	0.166

Figure 6(a) compares subgroup sensitivity and specificity, while Figure 6(b) quantifies the stability of the top-five explanation set. Cp and Thal appeared in the top five in 100% of stability analyses, Ca in 83.3%, Exang in 75.0%, and Oldpeak in 50.0%. In contrast, several conventional risk variables, including age, resting blood pressure, cholesterol, fasting blood sugar, and resting ECG category, did not repeatedly enter the top-five explanation set for this small benchmark. This should be interpreted as model- and dataset-specific behavior rather than a clinical statement that those factors are unimportant to cardiovascular disease.

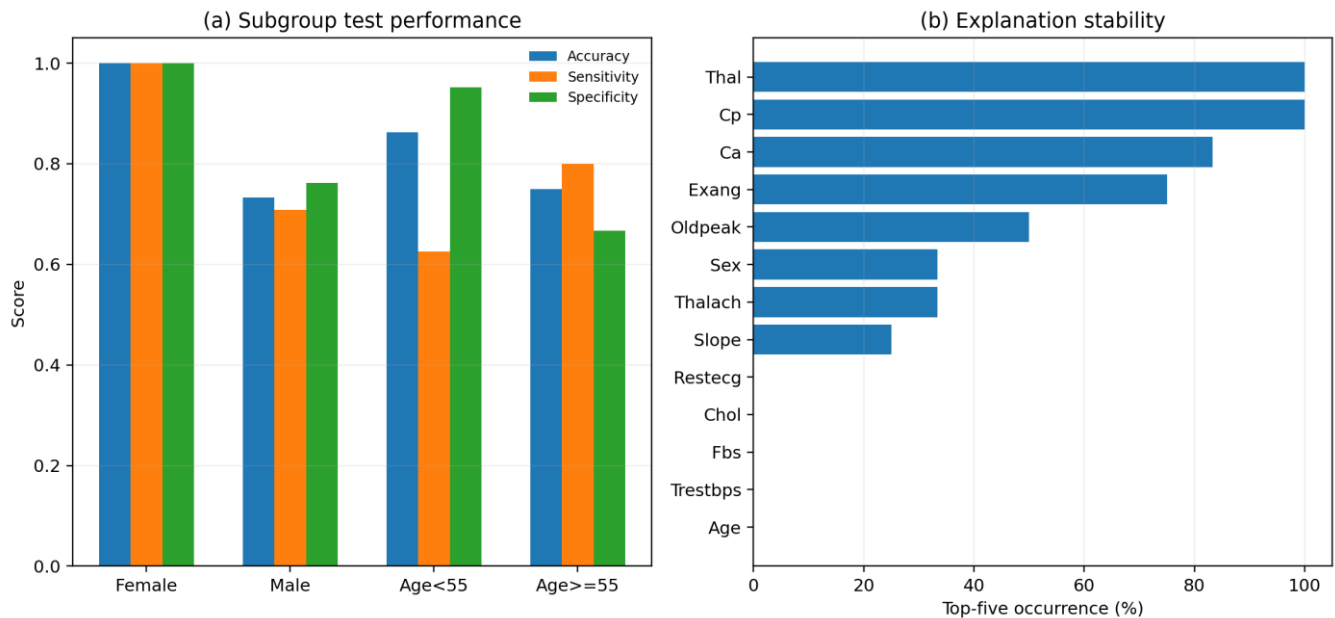


Figure 6. Subgroup performance and explanation stability.

F. Comparative interpretation

The results support three practical conclusions. First, discrimination and thresholded classification can identify different preferred models: logistic regression ranked probabilities most effectively, while KNN gave the strongest default-threshold classification in the held-out set. Second, explanation should be triangulated. SHAP identified a clinically interpretable hierarchy, but permutation testing showed that not every high-attribution feature independently improved ROC-AUC. Third, reliability diagnostics change the interpretation of apparently strong results. A 0.940 ROC-AUC is encouraging, yet subgroup sample sizes, calibration deviations, and the single-centre historical benchmark prevent claims of deployment readiness.

The findings are consistent with contemporary cardiac-XAI literature, where explainability is increasingly treated as part of model validation rather than merely as post hoc visualization [12,13]. They also reflect the reporting direction of TRIPOD+AI [11]: data splitting, transparent validation, clinically interpretable performance measures, and explicit limitations are essential when presenting machine-learning prediction models for health applications.

V. CONCLUSION

This study developed and evaluated a leakage-safe explainable machine-learning framework for heart-disease risk prediction using 303 clinical records and 13 predictors. Five classifiers were compared under cross-validation and independent testing. Logistic regression achieved the strongest held-out discrimination (ROC-AUC 0.940), whereas KNN provided the highest default-threshold accuracy and F1-score. The explainability audit identified Thal, Cp, Ca, Exang, Slope, and Oldpeak as influential features, with Ca, Thal, and Cp also supported by permutation-based loss of discrimination. Calibration, decision-curve analysis, subgroup evaluation, and explanation stability demonstrated why clinically oriented evaluation must extend beyond accuracy. The resulting framework is suitable as a transparent research prototype and as a reproducible baseline for future external-validation studies.

VI. LIMITATIONS AND FUTURE WORK

The principal limitation is the small, historical Cleveland benchmark, which does not represent contemporary multi-centre clinical practice or the demographic heterogeneity required for deployment. The 61-case hold-out set also makes subgroup estimates and calibration bins statistically unstable. The next validation stage should use temporally or geographically external cohorts, report confidence intervals for discrimination and calibration, assess calibration intercept and slope, test decision thresholds selected with clinical stakeholders, and evaluate explanation fidelity and usefulness with cardiologists. Model recalibration should be performed when transportability shifts are detected. These steps are necessary before the system can be considered for prospective decision support.

VII. REPRODUCIBILITY STATEMENT

The analysis uses the UCI Heart Disease benchmark [2]. The reproducible design consists of: binary disease recoding; stratified 80/20 split with random seed 42; fold-contained preprocessing; cross-validated model comparison; one-time hold-out evaluation; SHAP and permutation importance; calibration and decision-curve analysis; subgroup auditing; and resampling-based explanation stability. Reported tables and figures are generated from the same dataset and evaluation outputs used in the manuscript.

REFERENCES

- [1] World Health Organization, "Cardiovascular diseases (CVDs)," Fact Sheet, 31 July 2025.
- [2] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease [Dataset]," UCI Machine Learning Repository, 1989, doi: 10.24432/C52P4X.
- [3] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [4] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [5] 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.
- [6] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

- [7] 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.
- [8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [9] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.
- [10] E. W. Steyerberg et al., "Assessing the performance of prediction models: A framework for traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, 2010, doi: 10.1097/EDE.0b013e3181c30fb2.
- [11] 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.
- [12] A. M. Salih et al., "A review of evaluation approaches for explainable AI with applications in cardiology," *Artificial Intelligence Review*, vol. 57, art. 240, 2024, doi: 10.1007/s10462-024-10852-w.
- [13] M. U. Sreeja, A. O. Philip, and M. H. Supriya, "Towards explainability in artificial intelligence frameworks for healthcare: A comprehensive survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 6, art. 102096, 2024, doi: 10.1016/j.jksuci.2024.102096.
- [14] A. B. Teshale, H. L. Htun, M. Vered, A. J. Owen, et al., "A systematic review of artificial intelligence models for time-to-event outcome applied in cardiovascular disease risk prediction," *Journal of Medical Systems*, vol. 48, art. 68, 2024, doi: 10.1007/s10916-024-02087-7.
- [15] "Interpretable machine learning model for heart disease prediction," *Procedia Computer Science*, vol. 227, pp. 439–445, 2023, doi: 10.1016/j.procs.2023.10.544.
- [16] N. Kailasanathan, G. Ezhilarasan, S. Selvarajan, R. K. Dhanaraj, D. Pamucar, and N. Shankar, "Heart disease prediction with a feature-sensitized interpretable framework for the Internet of Medical Things sensors," *Frontiers in Digital Health*, vol. 7, art. 1612915, 2025, doi: 10.3389/fdgh.2025.1612915.
- [17] Y. Cai, Y.-Q. Cai, L.-Y. Tang, Y.-H. Wang, M. Gong, T.-C. Jing, and H.-J. Li, "Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review," *BMC Medicine*, vol. 22, art. 56, 2024.
- [18] F. A. Ekle, V. Shidali, R. E. Ochogwu, et al., "Machine Learning models for heart disease prediction and dietary lifestyle change therapy recommendation: A systematic review," *Discover Artificial Intelligence*, vol. 4, art. 113, 2024.

Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.