

Comparative Evaluation of Machine Learning Models for Heart Disease Prediction: Performance, Calibration, Threshold Selection, and Explainability

Stanley Kirunda Dhakaba¹

stanleykirunda11@gmail.com

Ernest Cook University

Patience Atukunda¹

Ernest Cook University

Anthony Bua²

Victoria University

Research Article

Keywords: heart disease prediction, machine learning, model calibration, classification threshold, permutation importance, Cleveland dataset

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11027885/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Cardiovascular disease remains the leading global cause of death, yet machine-learning studies report widely differing results on small benchmark datasets. This study compared five supervised classifiers using the processed Cleveland Heart Disease dataset, which contains 303 records: 139 with angiographic heart disease and 164 without disease. A stratified 70/30 split allocated 212 records to training and 91 to testing. Imputation, scaling, and categorical encoding were fitted only to training data. Logistic regression achieved the best overall test performance, with 85.71% accuracy, 85.37% precision, 83.33% sensitivity, 87.76% specificity, an F1 score of 84.34%, an area under the receiver operating characteristic curve (ROC-AUC) of 0.939, and a Brier score of 0.102. Raising the classification threshold from 0.50 to a training-derived value of 0.552 increased specificity to 89.80% but reduced sensitivity to 80.95% and did not improve accuracy. Permutation importance ranked the number of major vessels, chest-pain type, and thallium test result as the leading predictors. Cross-validation F1 increased as the training sample grew. These results support multi-metric evaluation but require external validation before generalization to populations not represented in the development data, including those in sub-Saharan Africa.

1. Introduction

Cardiovascular diseases are the leading cause of death globally. The World Health Organization estimated that 19.8 million people died from cardiovascular diseases in 2022, representing approximately 32% of all deaths; more than three quarters occurred in low- and middle-income countries [1]. Earlier identification of elevated risk can support timely clinical assessment and management. Machine learning can identify multivariable patterns in structured clinical data, but useful estimates depend on valid study design, reproducible preprocessing, appropriate performance measures, and evidence of generalizability [2].

The UCI Cleveland Heart Disease Dataset has become a benchmark for building heart disease prediction algorithms. The University of California, Irvine (UCI) Machine Learning Repository describes 303 records, 13 predictors, missing values in selected variables, and an angiographic outcome coded from zero to four [3]. Values from one to four are commonly recoded as disease present, while zero represents disease absent. Published performance varies with outcome coding, preprocessing, feature selection, data splitting, model tuning, and leakage control. Accuracy should therefore be interpreted together with class distribution, error-specific measures, discrimination, and calibration.

The main objective of this study was to build and evaluate machine learning algorithms for the prediction of heart disease.

This was achieved using Logistic Regression, Random Forest, Naive Bayes, Decision Trees, and K Nearest Neighbors. The models were evaluated using precision, sensitivity, specificity, F1-score, ROC-AUC, and calibration alongside accuracy..

2. Literature Review

Prior studies using the Cleveland data have reported markedly different results depending on preprocessing, model tuning, and validation. Chandrasekhar and Peddakrishna [4] compared six algorithms and found 90.16% accuracy for logistic regression (LR), while a soft-voting ensemble increased accuracy to 93.44%. Gao et al. [5] reported 98.6% for a bagged decision tree after principal component analysis (PCA). Noroozi et al. [6] obtained 85.5% with a support vector machine (SVM) after filter-based feature selection and found that feature selection did not improve every classifier. Bouqentar et al. [7] reported approximately 92% for an optimized SVM on Cleveland and 91.18% when validated on Statlog. Using 1,190 pooled records, Teja and Rayalu [8] reported 93% for XGBoost and bagged trees and 91% for random forest (RF) and k-nearest neighbours (KNN). These variations show why performance must be interpreted in relation to the data source and validation design. The present study compares five classifiers under the same split and preprocessing pipeline and reports class-specific performance, discrimination, and calibration rather than accuracy alone.

3. Methodology

3.1 Dataset, Outcome, and Class Distribution

The study used the Heart Disease dataset from the UCI Machine Learning Repository [3]. It contains 303 records, 13 predictor variables, and one angiographic outcome variable (num). The outcome was recoded as disease absent when num equalled zero and disease present when num ranged from one to four. This produced 139 disease-positive and 164 disease-negative records. A stratified 70/30 split with random seed 42 produced 212 training records and 91 test records while preserving the outcome proportions.

Table 1 Outcome distribution in the full dataset and stratified split			
Partition	Total	Disease present	Disease absent
Full dataset	303	139	164
Training (70%)	212	97	115
Testing (30%)	91	42	49

3.2 Preprocessing and Feature Engineering

The predictors were age; resting blood pressure (trestbps); serum cholesterol (chol); maximum heart rate achieved during exercise (thalach); and exercise-induced ST-segment depression relative to rest (oldpeak). The categorical predictors were sex; chest-pain type (cp); fasting blood sugar indicator (fbs); resting electrocardiographic result (restecg); exercise-induced angina (exang); slope of the peak exercise ST segment (slope); number of major vessels visualized by fluoroscopy (ca); and thallium stress-test result (thal). Four ca values and two thal values were missing; no other predictor was incomplete.

All preprocessing parameters were estimated within the relevant training or cross-validation fold and then applied unchanged to validation or test data [9]. No synthetic records or test-set-derived transformations were used.

3.3 Algorithms and Performance Measures

Five classifiers were implemented in scikit-learn [10]: LR with a maximum of 2,000 iterations; RF with 500 trees, balanced class weights, and random seed 42; Gaussian naive Bayes (GNB); a decision tree (DT) with balanced class weights and random seed 42; and KNN with k equal to five. Each classifier used the same training and test partitions. Models were ranked primarily by test F1 score and secondarily by ROC-AUC.

Mathematical specification of the classifiers. Let $\mathbf{x}_i$ denote the preprocessed predictor vector for record i , y_i the binary outcome, n the number of training records, d the number of transformed predictors, τ the decision threshold, and $1\{\cdot\}$ the indicator function.

LR maps a linear predictor to the estimated probability of class 1. The coefficients were estimated with L2 regularization.

$$\eta_i = \beta_0 + \text{varvec}\beta^\top \mathbf{x}_i, \quad \hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp(-\eta_i)}.$$

$$\hat{y}_i = 1\{\hat{p}_i \geq \tau\}, \quad \tau = 0.50.$$

RF averages the class-1 probabilities from $B = 500$ bootstrap-trained decision trees. Balanced class weights are inversely proportional to the number of training records in each of the $K = 2$ classes.

$$\hat{p}_{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{p}_b(\mathbf{x}), \quad B = 500, \quad \hat{y} = 1\{\hat{p}_{\text{RF}}(\mathbf{x}) \geq \tau\}.$$

$$w_c = \frac{n}{Kn_c}, \quad c \in \{0, 1\}, \quad K = 2.$$

GNB assumes conditionally independent Gaussian predictors. In the equations below, π denotes the class prior, while μ and σ^2 denote the class-specific mean and variance.

$$P(y = c \mid \mathbf{x}) \propto \pi_c \prod_{j=1}^d \frac{1}{\sqrt{2\pi\sigma_{cj}^2}} \exp \left[-\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2} \right].$$

$$\hat{y} = \arg \max_{c \in \{0,1\}} \left\{ \log \pi_c - \frac{1}{2} \sum_{j=1}^d \left[\log(2\pi\sigma_{cj}^2) + \frac{(x_j - \mu_{cj})^2}{\sigma_{cj}^2} \right] \right\}.$$

For DT, $G(S)$ is the weighted Gini impurity of the records at node S and $p(c|S)$ is the weighted proportion of class c . A split on predictor j at value s creates left and right child sets; the selected split minimizes their weighted impurity.

$$G(S) = 1 - \sum_{c \in \{0,1\}} p(c|S)^2.$$

$$(j^*, s^*) = \arg \min_{j,s} \left[\frac{|S_L(j, s)|}{|S|} G(S_L(j, s)) + \frac{|S_R(j, s)|}{|S|} G(S_R(j, s)) \right].$$

KNN used standardized and encoded predictors. Euclidean distance identifies the set $N_k(\mathbf{x})$ of $k = 5$ nearest training records, and the estimated class-1 probability is the proportion of positive outcomes among those neighbours.

$$d(\mathbf{x}, \mathbf{x}_i) = \left[\sum_{j=1}^d (x_j - x_{ij})^2 \right]^{1/2}.$$

$$\hat{p}_{\text{KNN}}(\mathbf{x}) = \frac{1}{k} \sum_{i \in N_k(\mathbf{x})} 1\{y_i = 1\}, \quad \hat{y} = 1\{\hat{p}_{\text{KNN}}(\mathbf{x}) \geq \tau\}, \quad k = 5.$$

Performance was summarized using accuracy, precision, sensitivity (recall), specificity, F1 score, ROC-AUC, and Brier score. Accuracy, precision, sensitivity, specificity, and F1 were calculated from true-positive, true-negative, false-positive, and false-negative counts. ROC-AUC measures discrimination across decision thresholds [11], while the Brier score is the mean squared difference between predicted probability and observed outcome [12]. Lower Brier scores indicate better probabilistic accuracy.

3.4 Calibration, Threshold, Explainability, and Learning-Curve Analysis

Five-fold stratified cross-validation generated out-of-fold training probabilities for LR. The threshold that maximized Youden's J (sensitivity + specificity - 1) was fixed and evaluated once on the held-out test set [13]. The conventional 0.50 threshold remained the reference operating point. Calibration was assessed on the test set with six quantile-based bins and the Brier score [14].

Predictor importance was assessed with 50-repeat permutation importance and ROC-AUC as the scoring function [15]. Larger decreases in ROC-AUC after shuffling indicated greater model reliance on a predictor. Learning behaviour was examined with five-fold cross-validation F1 scores at five increasing training-set sizes.

3.5 Exploratory Subgroup Evaluation

LR performance at the 0.50 threshold was summarized by sex and by age groups below 50 years, 50–65 years, and above 65 years. Each subgroup summary included sample size, class counts, accuracy, precision, sensitivity, specificity, and F1 score. Because the groups were small, the analysis was descriptive and was not treated as a definitive fairness assessment [16].

4. Results

4.1 Comparative Model Performance

On the 91-record test set, LR recorded the highest F1 score (84.34%) and ROC-AUC (0.939). Its accuracy was 85.71%, with precision of 85.37%, sensitivity of 83.33%, and specificity of 87.76%. RF and KNN produced F1 scores of 80.46% and 80.00%, respectively. GNB detected 90.48% of disease-positive cases but correctly classified only 51.02% of disease-negative cases, resulting in 69.23% accuracy.

Table 2
Comparative test performance of the five machine-learning models

Algorithm	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 (%)	ROC-AUC	Brier
LR	85.71	85.37	83.33	87.76	84.34	0.939	0.102
RF	81.32	77.78	83.33	79.59	80.46	0.933	0.113
KNN	82.42	84.21	76.19	87.76	80.00	0.907	0.121
GNB	69.23	61.29	90.48	51.02	73.08	0.846	0.266
DT	69.23	65.91	69.05	69.39	67.44	0.692	0.308

4.2 Calibration Assessment

LR had the lowest Brier score (0.102), followed by RF (0.113) and KNN (0.121). GNB and DT recorded larger scores of 0.266 and 0.308, respectively. The calibration points in Fig. 1 vary between bins because each bin contains relatively few observations from the 91-record test set.

4.3 Classification Threshold Analysis

The conventional 0.50 threshold and the training-derived 0.552 threshold produced the same accuracy (85.71%) on the test set. At 0.552, false positives declined from six to five and specificity increased from 87.76% to 89.80%. False negatives increased from seven to eight, sensitivity declined from 83.33% to 80.95%, and F1 decreased from 84.34% to 83.95%.

Table 3
LR performance at the conventional and training-derived thresholds

Operating point	Threshold	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1 (%)
Conventional 0.50 threshold	0.500	85.71	85.37	83.33	87.76	84.34
Training-derived Youden J threshold	0.552	85.71	87.18	80.95	89.80	83.95

4.4 Model Explainability

Shuffling ca produced the largest mean decrease in ROC-AUC (0.067), followed by cp (0.047), thal (0.029), slope (0.023), and sex (0.021). The decreases indicate how strongly test discrimination depended on each predictor in the fitted LR model.

Table 4
Leading predictors by repeated permutation importance for LR

Rank	Predictor	Mean ROC-AUC decrease	Definition
1	ca	0.067	Number of major vessels visualized by fluoroscopy
2	cp	0.047	Chest-pain type
3	thal	0.029	Thallium test result category
4	slope	0.023	Slope of the peak exercise ST segment
5	sex	0.021	Recorded sex

4.5 Learning-Curve Assessment

As the training size increased from 33 to 169 records, mean training F1 declined from 0.899 to 0.841, while mean cross-validation F1 rose from 0.725 to 0.799 (Fig. 3). The narrowing gap between the curves indicates more stable generalization as additional training observations were included.

4.6 Exploratory Subgroup Performance

Female records had 93.55% accuracy and 100% specificity, compared with 81.67% accuracy and 79.31% specificity among male records. Sensitivity was similar (81.82% and 83.87%). The 50-65-year group contained 56 records and achieved 85.71% accuracy. The group above 65 years contained only nine records; its 100% sensitivity was based on four disease-positive cases.

Table 5
Exploratory LR performance by sex and age group (% except counts)

Group	n	Positive	Negative	Accuracy	Precision	Sensitivity	Specificity	F1
Sex: Female	31	11	20	93.55	100.00	81.82	100.00	90.00
Sex: Male	60	31	29	81.67	81.25	83.87	79.31	82.54
Age: <50	26	9	17	88.46	100.00	66.67	100.00	80.00
Age: 50–65	56	29	27	85.71	86.21	86.21	85.19	86.21
Age: >65	9	4	5	77.78	66.67	100.00	60.00	80.00

5. Discussion

LR was the best-performing model on the main selection criteria, with an F1 score of 84.34%, ROC-AUC of 0.939, and Brier score of 0.102. RF had a slightly lower ROC-AUC (0.933) and lower precision and specificity. KNN achieved the same specificity as LR (87.76%) but lower sensitivity (76.19%). GNB produced the highest sensitivity (90.48%) but the lowest specificity (51.02%), whereas DT recorded the lowest ROC-AUC (0.692). Thus, the preferred model changes when attention shifts from detecting positive cases to avoiding false positives or estimating probabilities.

Changing the LR threshold from 0.50 to 0.552 did not improve accuracy and reduced F1 from 84.34% to 83.95%. The reduction of one false positive was offset by one additional false negative. The value 0.552 should therefore be treated as a training-derived analytical comparison rather than a clinical decision threshold.

The learning curve suggests that sample size constrained performance. Cross-validation F1 improved as more records were added, while the training-validation gap narrowed. Independent data are therefore more likely to strengthen the evidence than further tuning on the same 303 records.

5.1 Comparison with Previous Studies

LR achieved 85.71% accuracy in the present study. Chandrasekhar and Peddakrishna [4] reported a higher accuracy of 90.16% for LR when comparing six algorithms on the Cleveland dataset, and their soft-voting ensemble reached 93.44%. Noroozi et al. [6] obtained 85.5% with an SVM after comparing 16 feature-selection methods. Additionally, Bouqentar et al. [7] reported approximately 92% for an optimized SVM on Cleveland and 91.18% after validation on the Statlog dataset. Gao et al. [5] reported 98.6% for bagged DT after PCA, while Teja and Rayalu [8] achieved 93% with XGBoost and bagged trees on a

pooled 1,190-record dataset. The variation is consistent with differences in sample composition, feature selection, model tuning, and validation design.

5.2 Interpretation and Research Implications

Permutation importance identified ca, cp, and thal as the leading predictors for LR, with mean ROC-AUC decreases of 0.067, 0.047, and 0.029, respectively. In this fitted model, discrimination therefore depended most on angiographic information, chest-pain presentation, and thallium stress-test findings.

Subgroup performance varied by sex and age. Female records had 93.55% accuracy compared with 81.67% for male records, while the group above 65 years had 100% sensitivity but 60% specificity. Because the oldest group contained nine records and only four disease-positive cases, these figures are descriptive estimates rather than population-level evidence. Larger studies are needed to determine whether the differences persist [16].

5.3 Limitations

This study used 303 records from the Cleveland dataset and a single 70/30 stratified split. The small test set limits the precision of overall, calibration, and subgroup estimates; model selection and exploratory interpretation also relied on the same held-out sample. The dataset was not collected in sub-Saharan Africa and does not represent the region's patient mix, diagnostic pathways, or data-collection practices. The findings therefore cannot be generalized to sub-Saharan African populations without external validation using data from the region. Permutation importance may also be affected by correlated predictors.

5.4 Future Research

Future studies should evaluate these models on larger, independently collected datasets from sub-Saharan Africa; use repeated or nested validation; pre-specify clinical outcomes and threshold criteria; and assess calibration and subgroup performance prospectively. Coefficient intervals and local explanation methods can be added after stable external performance has been established.

6. Conclusion

This study found that LR provided the strongest overall performance among the five models, achieving 85.71% accuracy, an F1 score of 84.34%, ROC-AUC of 0.939, and a Brier score of 0.102 on the held-out test set. RF achieved comparable discrimination (ROC-AUC 0.933), while GNB gave the highest sensitivity (90.48%) at the expense of specificity. The most influential LR predictors were ca, cp, and thal. Raising the decision threshold to 0.552 increased specificity but reduced sensitivity and F1, so it did not improve the model overall. Cross-validation F1 reached 0.799 at the largest assessed training size of 169 records. LR is therefore the preferred benchmark model in this analysis, but validation on larger independent datasets, including data from sub-Saharan Africa, is required before clinical use.

Declarations

Clinical trial number: not applicable.

Authors' declaration

The authors declare that this manuscript presents original work, has not been published previously, and is not under consideration for publication elsewhere. All authors reviewed and approved the manuscript and agree to its submission to Discover Artificial Intelligence.

Ethics approval and consent to participate

Not applicable. This study used a publicly available, de-identified benchmark dataset and involved no direct recruitment of human participants or collection of new human or animal data.

Consent for publication

All authors approved the manuscript and consent to its publication.

Competing interests

The authors declare no competing interests.

Funding

The authors declare that this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Built a supervised machine-learning model for binary prediction of heart disease. Compare the models using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and Brier score. identify the predictors that most influence the discrimination of the best-performing model. To determine an evidence-based classification threshold from training data and evaluate it on held-out test data.

Data availability

The dataset analyzed during the current study is publicly available from the UCI Machine Learning Repository Heart Disease dataset.

References

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. Geneva: World Health Organization. 2025 Jul 31 [cited 2026 Aug 22]. Available from: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29>
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28:31–8. <https://doi.org/10.1038/s41591-021-01614-0>.
3. Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease [dataset]. Irvine (CA): University of California, Irvine, Machine Learning Repository; 1989. <https://doi.org/10.24432/C52P4X>
4. Chandrasekhar N, Peddakrishna S. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*. 2023;11(4):1210. <https://doi.org/10.3390/pr11041210>.
5. Gao XY, Ali AA, Hassan HS, Anwar EM, Khalil AM. Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity*. 2021;2021:6663455. <https://doi.org/10.1155/2021/6663455>.
6. Noroozi Z, Orooji A, Erfannia L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci Rep*. 2023;13:22588. <https://doi.org/10.1038/s41598-023-49962-w>.
7. Bouqentar MA, Terrada O, Hamida S, Saleh S, Lamrani D, Cherradi B, et al. Early heart disease prediction using feature engineering and machine learning algorithms. *Heliyon*. 2024;10(19):e38731. <https://doi.org/10.1016/j.heliyon.2024.e38731>.
8. Teja MD, Rayalu GM. Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. *BMC Cardiovasc Disord*. 2025;25:212. <https://doi.org/10.1186/s12872-025-04627-6>.
9. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. eGEMs (Wash DC). 2013;1(3):1035. <https://doi.org/10.13063/2327-9214.1035>. Strategies for handling missing data in electronic health record derived data.
10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30. <https://jmlr.org/papers/v12/pedregosa11a.html>.
11. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861–74. <https://doi.org/10.1016/j.patrec.2005.10.010>.
12. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1–3.
13. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–5. [https://doi.org/10.1002/1097-0142.\(1950\)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3](https://doi.org/10.1002/1097-0142.(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3).
14. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>.
15. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340–7. <https://doi.org/10.1093/bioinformatics/btq134>.

Figures

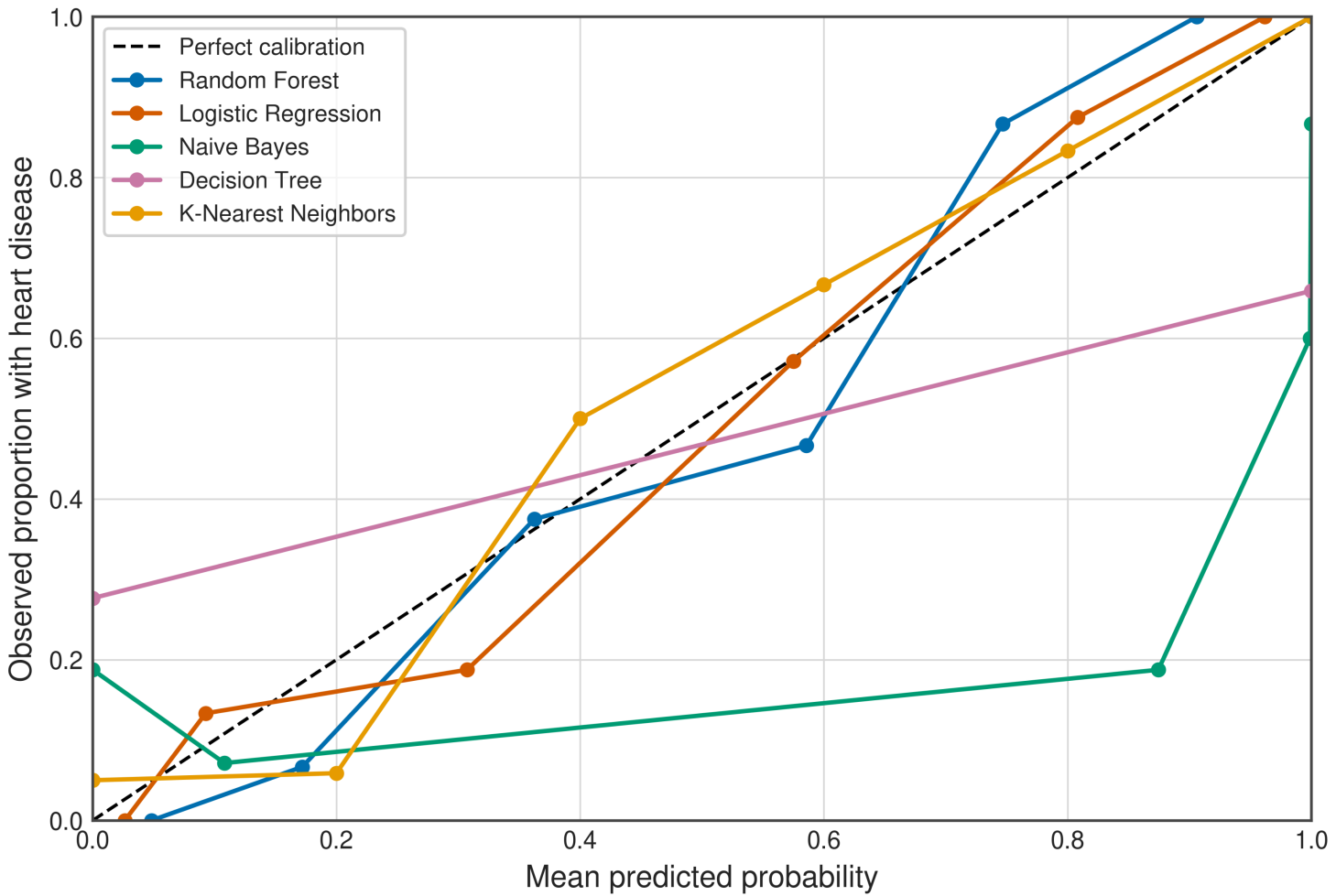


Figure 1
Calibration curves for all models on the held-out test set

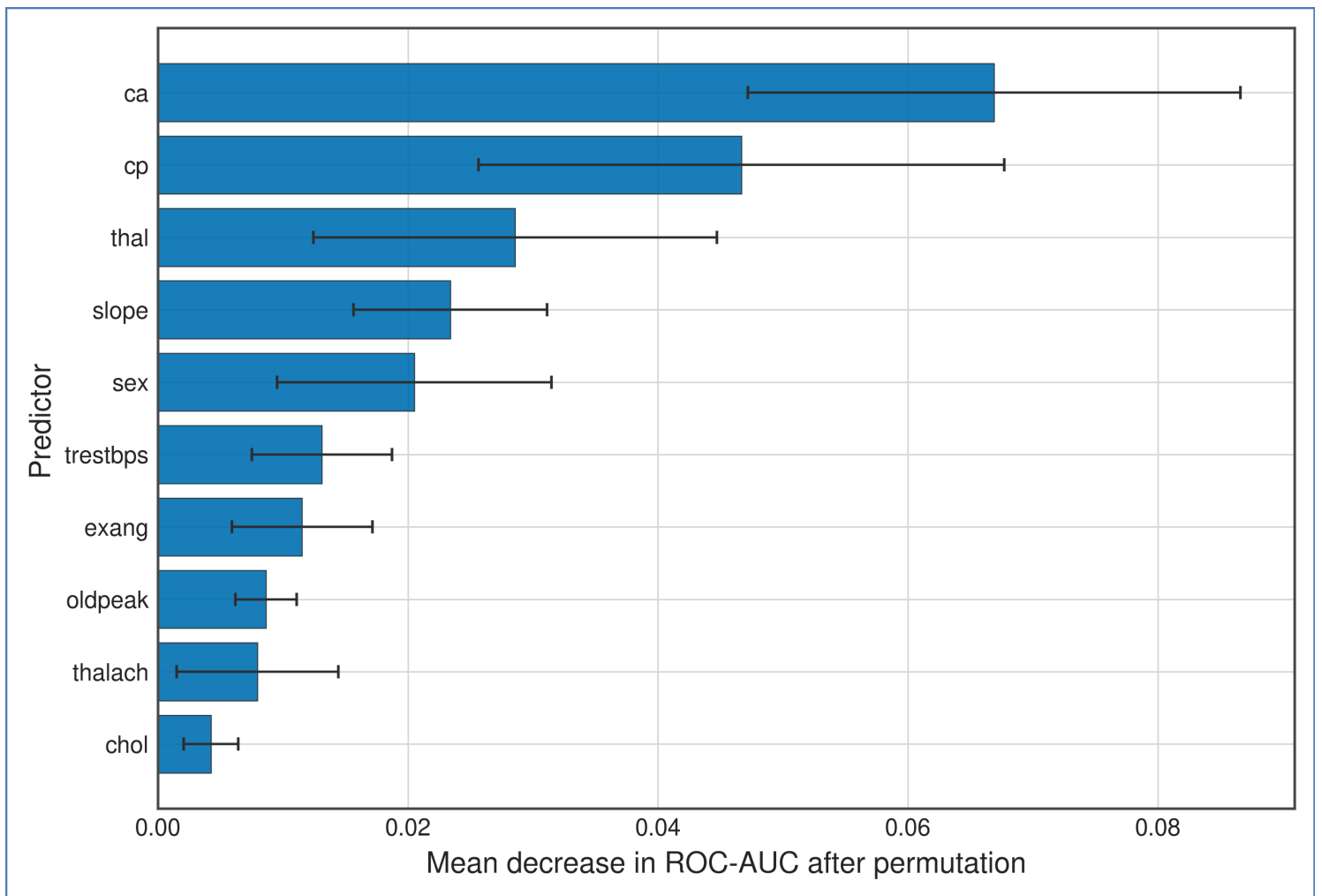


Figure 2

Permutation importance for the ten highest-ranked predictors

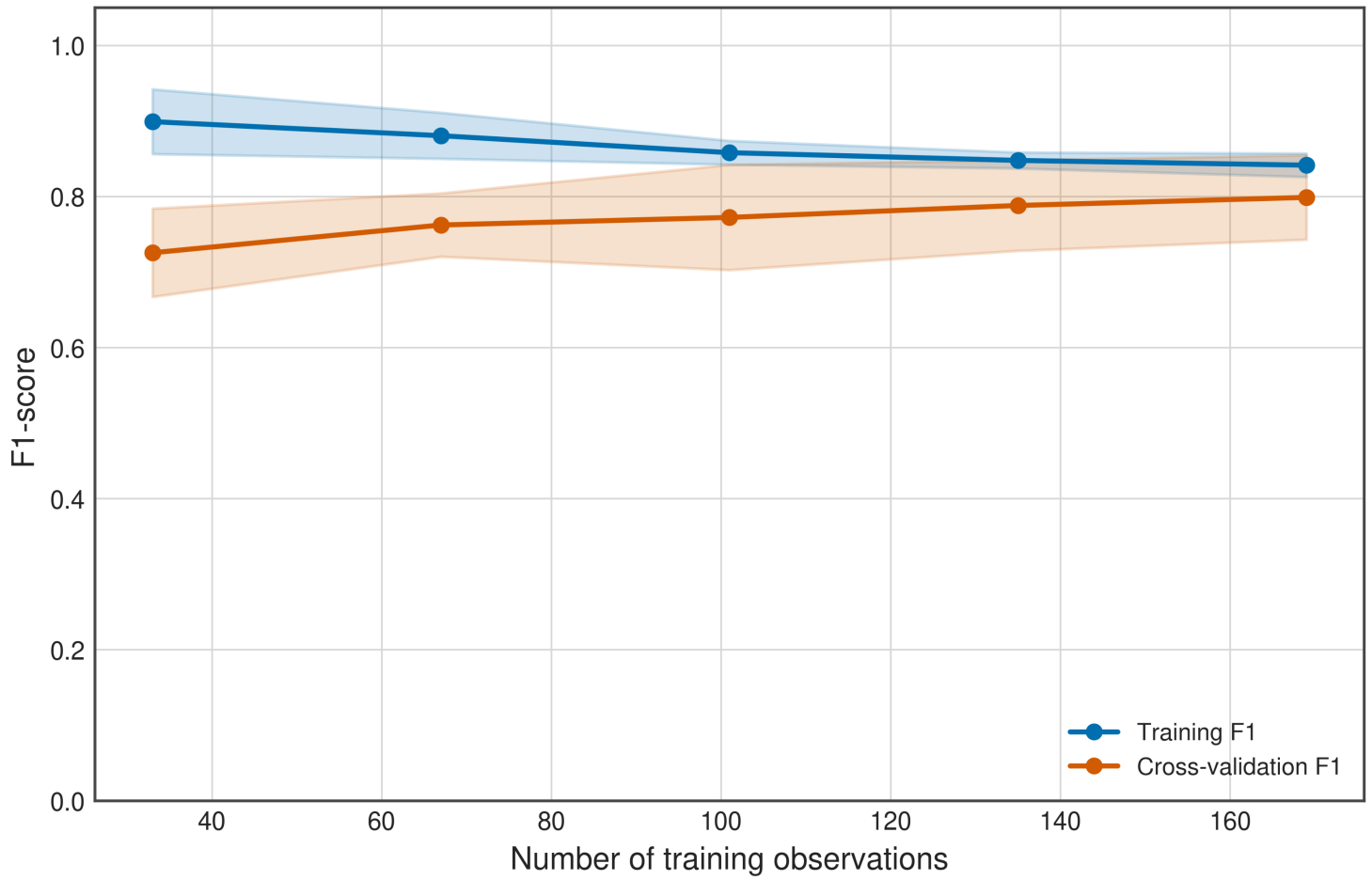


Figure 3

Logistic regression learning curve using five-fold cross-validation F1-score