

Genetic Algorithm Feature Selection for Enhanced Heart Disease Prediction: A Leakage Free, Stability Aware Multi Model Evaluation

Mohammad Subhi Al-Batah, Ihsan Alshraideh

Department of Computer Science, Faculty of Information Technology, Jadara University, Irbid, Jordan

Corresponding author: albatah@jadara.edu.jo | ORCID: 0000-0002-9341-1727

Abstract

Genetic-algorithm (GA) feature selection is frequently presented as a route to more accurate and efficient clinical machine-learning models, yet reported gains may be inflated when preprocessing and subset optimization are performed before validation. This study develops a stability-aware multiobjective genetic algorithm (SA-MOGA) and evaluates whether GA-selected variables genuinely enhance heart-disease prediction under a leakage-free protocol. The public Heart Failure Prediction Dataset contained 918 records, 11 demographic, clinical, and exercise-related predictors, and a binary HeartDisease endpoint. Physiologically implausible zero values in resting blood pressure and cholesterol were treated as missing and median-imputed within each training fold. SA-MOGA optimized bootstrap-stable mutual-information relevance while penalizing redundancy, relevance uncertainty, and subset size. Five-fold stratified outer validation generated out-of-fold predictions for logistic regression, support-vector machine, k-nearest neighbors, random forest, multilayer perceptron, XGBoost, and a mean-probability ensemble. The algorithm consistently retained eight variables (27.3% reduction; mean pairwise Jaccard stability 0.911). With all features, XGBoost achieved the best discrimination (AUROC 0.932, 95% CI 0.916–0.947), while the all-feature ensemble achieved 0.873 accuracy and a Brier score of 0.101. Contrary to a simple enhancement hypothesis, SA-MOGA reduced AUROC for every learner; the ensemble change was -0.0176 (95% bootstrap CI -0.0257 to -0.0095). The selected subset nevertheless preserved clinically interpretable variables and reduced fitting time for several models. These findings show that dimensionality reduction and predictive enhancement are distinct objectives, and that leakage-free, stability-aware evaluation is essential before claiming clinical benefit from evolutionary feature selection.

Keywords: heart disease prediction; genetic algorithm; feature selection; leakage-free validation; clinical machine learning

1. Introduction

Cardiovascular diseases remain a dominant global health burden and create continuing demand for reliable, accessible methods of early risk stratification [1]. Machine-learning models can combine demographic information, symptoms, resting measurements, exercise responses, and electrocardiographic descriptors to identify nonlinear patterns that are difficult to summarize through a single conventional score. However, clinical usefulness depends on more than a high accuracy value. A model should be evaluated with out-of-sample discrimination, calibration, sensitivity, specificity, uncertainty, subgroup behavior, and potential decision benefit.

Feature selection is often proposed as a means of improving machine-learning performance. Removing irrelevant or redundant variables can reduce measurement burden, simplify deployment, lower variance, and improve interpretability [3]. Genetic algorithms are attractive because they represent each candidate subset as a binary chromosome and can explore nonconvex combinations through selection, crossover, and mutation [4,5]. In medical prediction research, evolutionary feature selection has therefore been combined with support-vector machines, neural networks, random forests, and ensembles. Nevertheless, the phrase performance enhancement can conceal several distinct objectives: a smaller subset may be valuable even when discrimination is unchanged, whereas a nominal accuracy increase may be unreliable if selection was informed by the evaluation data.

The Heart Failure Prediction Dataset is widely used for binary heart-disease classification. It combines records originating from five established heart datasets and provides 918 unique observations with 11 common predictors [2,22]. Its modest dimensionality creates a particularly important test. When only 11 predictors are available, aggressive feature removal may discard complementary information, and the search procedure itself may overfit. Moreover, zero-coded resting blood pressure or cholesterol values are physiologically implausible and should not be treated as ordinary measurements. These properties make the dataset well suited to studying whether GA-based feature selection produces genuine, transferable gains under rigorous validation.

Previous heart-disease studies have reported promising results from feature ranking, wrapper selection, evolutionary optimization, ensembles, and low-code deployment [21–27]. Al-Batah and colleagues have also demonstrated the relevance of ranked features, multi-model clinical prediction, evolutionary feature selection, and careful data transformation in cardiac and other diagnostic tasks [24–29]. Yet much of the literature emphasizes the best point estimate, frequently after a single split, without simultaneously reporting feature-selection stability, paired uncertainty, probability calibration, decision-curve utility, and subgroup performance. This creates an evidence gap between algorithmic optimization and trustworthy clinical evaluation.

The present study addresses that gap by proposing SA-MOGA, a model-independent GA that combines bootstrap-stable relevance, low redundancy, low relevance uncertainty, and parsimony. Rather than optimizing a subset against one classifier and then evaluating the same classifier on data that influenced selection, SA-MOGA is executed entirely inside each outer training fold. The selected subset is transferred to six structurally different learners and a probability ensemble. This design permits a direct test of whether a compact, stable feature set generalizes across model families.

The study makes four contributions. First, it provides a leakage-free reconstruction of the complete preprocessing–selection–evaluation pipeline. Second, it introduces a stability-aware multiobjective

fitness function and reports selection reproducibility across outer folds. Third, it evaluates discrimination, accuracy, sensitivity, specificity, F1 score, Brier score, expected calibration error, decision curves, threshold trade-offs, computation time, and demographic subgroups using out-of-fold predictions. Fourth, it reports a scientifically important negative result: SA-MOGA improves parsimony and interpretability, but it does not improve predictive performance on this low-dimensional dataset. This distinction helps prevent optimistic claims that may arise from leakage or selection-induced overfitting.

2. Literature Review and Research Gap

2.1 Machine learning for heart-disease prediction

Classical learners remain competitive for structured clinical data. Logistic regression offers transparent additive effects; support-vector machines model nonlinear decision boundaries through kernels; k-nearest neighbors uses local similarity; random forests reduce tree variance; multilayer perceptrons represent nonlinear interactions; and gradient boosting sequentially corrects prediction errors [14–20]. Their relative performance depends on sample size, preprocessing, feature representation, class distribution, and validation design. On the 918-record dataset, Baghdadi et al. evaluated advanced machine-learning techniques for cardiovascular diagnosis [21], while Alzboon and Al-Batah reported that gradient boosting achieved 87.4% accuracy and AUROC 0.928 in a broad multi-model framework [24]. These findings establish a realistic benchmark near 0.92–0.93 AUROC for conventional tabular models.

Recent optimization studies report substantially higher estimates. Kumar et al. combined classical and quantum-inspired optimization with SVMs and reported 97.83% accuracy for a seven-feature quantum-inspired model and 94.02% for its classical counterpart [22]. Al-Mahdi et al. combined GA feature selection with a tunicate-swarm-optimized deep ensemble and reported 97.5% accuracy [23]. Such results demonstrate the potential of hybrid optimization, but comparison is complicated by differences in train-test splitting, tuning, feature-selection placement, and whether final probabilities are generated for observations untouched by all preprocessing and optimization decisions.

2.2 Feature selection, evolutionary search, and stability

Feature-selection methods are commonly divided into filters, wrappers, and embedded techniques [3]. Filters estimate feature–outcome association independently of a final classifier, wrappers search for subsets using predictive performance as the objective, and embedded methods perform selection during model fitting. Genetic algorithms are wrapper-like search procedures when their fitness is classifier performance. They can capture interactions but are computationally expensive and susceptible to selection-induced overfitting, especially when many candidate subsets are repeatedly scored on the same validation partitions [6–8].

Noroozi et al. compared 16 filter, wrapper, and evolutionary approaches with seven learners and found that the effect of selection depended on the model and metric [30]. Pathan et al. similarly emphasized that feature selection can improve efficiency and accuracy, but the benefit varies across datasets and algorithms [31]. Al-Batah et al. used a binary cuckoo-search approach to remove disturbance features in plant-image classification [27], demonstrating the broader potential of

evolutionary subset optimization. These studies motivate a model-independent fitness design that rewards stable information while avoiding direct optimization against the final test predictions.

Selection stability is a separate property from prediction performance. Two folds may achieve similar AUROC using very different subsets, reducing scientific interpretability. Stability can be quantified by selection frequency and pairwise Jaccard similarity [7]. Incorporating bootstrap variation of feature relevance into the fitness function can further discourage unstable variables whose apparent association changes under resampling.

2.3 Leakage, calibration, and clinical utility

Data leakage occurs when information from held-out observations influences preprocessing, oversampling, feature selection, or hyperparameter tuning. Scaling or imputing before cross-validation uses global distributional information; applying SMOTE before cross-validation can create synthetic observations whose neighborhoods contain information related to validation cases; and selecting features once on the complete dataset converts the validation set into part of the training process. Varma and Simon showed that non-nested selection and tuning can produce biased error estimates [8], while applied studies have documented severe optimism in small biomedical datasets [9].

Clinical risk models also require probability-oriented evaluation. AUROC measures ranking but does not establish that predicted probabilities are accurate. The Brier score and calibration plots quantify probabilistic agreement, and expected calibration error summarizes bin-wise deviations [11]. Decision-curve analysis translates predictions into net benefit across threshold probabilities [12]. These measures are especially important when a reduced feature set alters the confidence of predictions even if classification accuracy appears similar.

2.4 Research gap and hypotheses

Existing studies do not provide a complete answer to whether GA feature selection genuinely enhances this dataset under a training-fold-only search, transferred across multiple classifiers, with stability, uncertainty, calibration, and clinical utility reported together. The primary hypothesis was that SA-MOGA would reduce the number of predictors while maintaining or improving out-of-fold AUROC. Secondary hypotheses were that the selected subset would be stable across folds, improve computational efficiency, and preserve clinically plausible variables. The null result for the primary hypothesis is retained and interpreted rather than hidden.

3. Materials and Methods

3.1 Dataset and endpoint

The analysis used the publicly available Heart Failure Prediction Dataset [2]. Despite its common name, the target is a binary indicator of the presence or absence of heart disease, not longitudinal heart-failure incidence, hospitalization, or mortality. The dataset contains 918 unique records, 508 positive cases (55.3%) and 410 negative cases (44.7%), and 11 predictors. The source dataset combines Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog heart records using common variables [2,22].

Table 1. Dataset variables and operational definitions.

Variable	Type	Description	Values/units
Age	Numeric	Age in years	Continuous
Sex	Categorical	Biological sex	M/F
ChestPainType	Categorical	Chest-pain category	TA, ATA, NAP, ASY
RestingBP	Numeric	Resting systolic blood pressure	mmHg
Cholesterol	Numeric	Serum cholesterol	mg/dL
FastingBS	Binary	Fasting blood sugar >120 mg/dL	0/1
RestingECG	Categorical	Resting ECG result	Normal, ST, LVH
MaxHR	Numeric	Maximum achieved heart rate	60–202 bpm
ExerciseAngina	Binary categorical	Exercise-induced angina	N/Y
Oldpeak	Numeric	Exercise-related ST depression	Numeric
ST_Slope	Categorical	Slope of peak exercise ST segment	Up, Flat, Down
HeartDisease	Target	Presence of heart disease	0/1

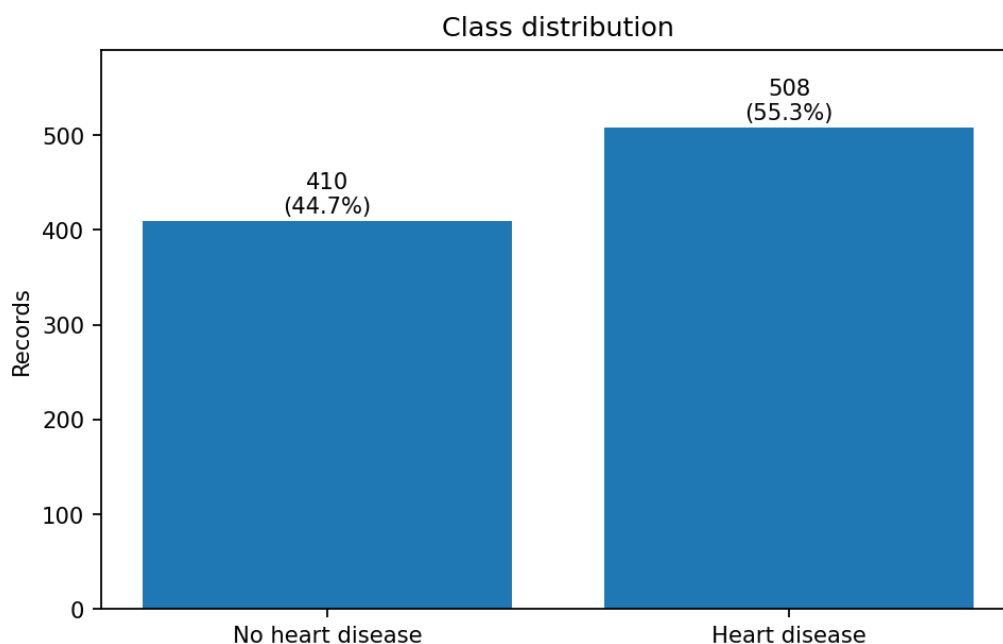


Figure 1. Distribution of the binary HeartDisease endpoint.

3.2 Data-quality audit and preprocessing

The raw data contained one zero value for RestingBP and 172 zero values for Cholesterol. A resting systolic pressure of zero and a serum cholesterol concentration of zero are not plausible clinical measurements; these values were therefore recoded as missing. To prevent leakage, medians were learned only from the current outer training fold and applied to both its training and held-out observations. Numeric predictors were standardized using training-fold means and standard deviations. Categorical predictors were one-hot encoded with fixed category definitions and unknown-category handling. No oversampling was applied because the class distribution was moderately balanced and because synthetic oversampling can complicate probability calibration.

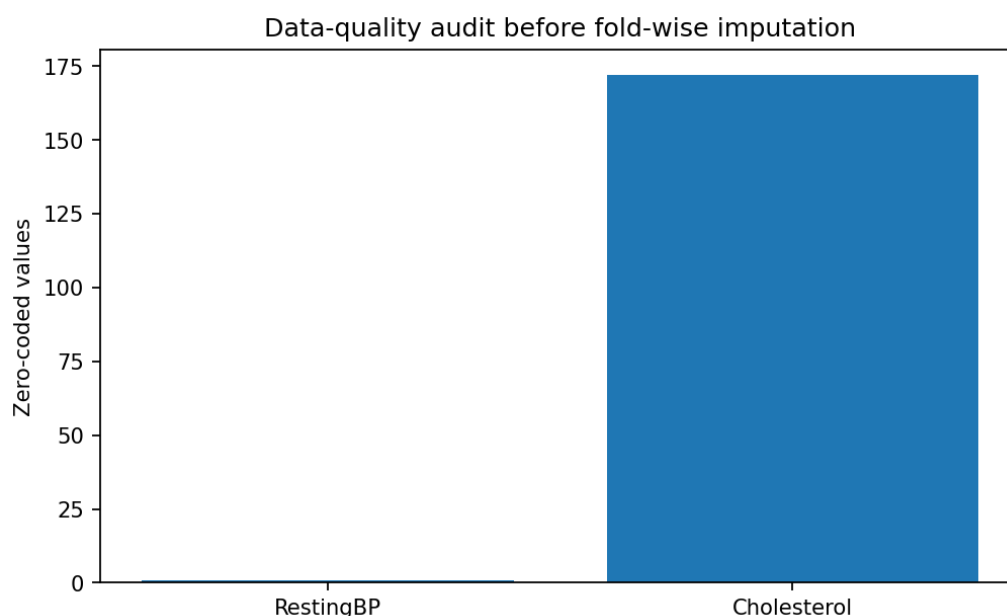


Figure 2. Data-quality audit showing zero-coded values treated as missing.

3.3 Proposed stability-aware multiobjective genetic algorithm

Each chromosome was an 11-bit vector in which 1 indicated inclusion of the corresponding original predictor. Feasible chromosomes contained four to eight variables. A population of 50 chromosomes evolved for 40 generations with tournament selection, 0.85 crossover probability, and mutation applied to 30% of offspring, with bit-level probability 1/11. Elitism retained the two best chromosomes. All random processes used the prespecified seed 20260729 plus the fold number.

For each outer training fold, predictor relevance was estimated by mutual information in 30 bootstrap resamples. Let r_j denote normalized mean mutual information for predictor j , u_j its normalized bootstrap standard deviation, and ρ_{jk} the absolute Spearman correlation between predictors j and k after training-fold ordinal representation. For subset S , the fitness was:

$$Fitness(S) = \sum_{j \in S} r_j - 0.18 |S| \text{mean}_{j < k \in S}(|\rho_{jk}|) - 0.04 \text{mean}_{j \in S}(u_j) - 0.08 |S|.$$

This formulation differs from a conventional accuracy-only wrapper in three ways. It is classifier-independent, so it can be transferred fairly across learners; it rewards cumulative stable relevance; and it penalizes redundancy, relevance uncertainty, and measurement burden. The GA never accessed the outer test fold.

3.4 Machine-learning models and ensemble

Table 2. Prespecified model configurations.

Model	Configuration	Role
Logistic regression	C=1; liblinear; class weighting; 2,000 iterations	Linear probabilistic baseline
RBF-SVM	C=2; probability estimates; class weighting	Nonlinear margin classifier
KNN	k=15; distance weighting	Local similarity model
Random forest	120 trees; sqrt features; min leaf=2	Bagged nonlinear ensemble
Neural network	Hidden layers 24 and 12; $\alpha=0.001$; early stopping	Nonlinear multilayer perceptron
XGBoost	100 trees; depth=3; learning rate=0.04; row/column sampling=0.85	Regularized boosted trees

Model	Configuration	Role
Mean-probability ensemble	Unweighted mean of six predicted probabilities	Diversity-based soft voting

For each outer fold, every model was trained twice: once with all 11 predictors and once with the SA-MOGA subset selected from that fold's training data. The mean-probability ensemble averaged the six model probabilities without using the test labels. This produced a complete out-of-fold probability for every record under both conditions.

3.5 Validation design and performance measures

A five-fold stratified outer cross-validation with shuffling preserved the outcome ratio. The complete sequence—zero recoding, imputation, scaling, one-hot encoding, GA search, and model fitting—was repeated independently inside each training fold. The held-out fold was used once for prediction. This design estimates the performance of the entire modeling process rather than the performance of a subset selected in advance.

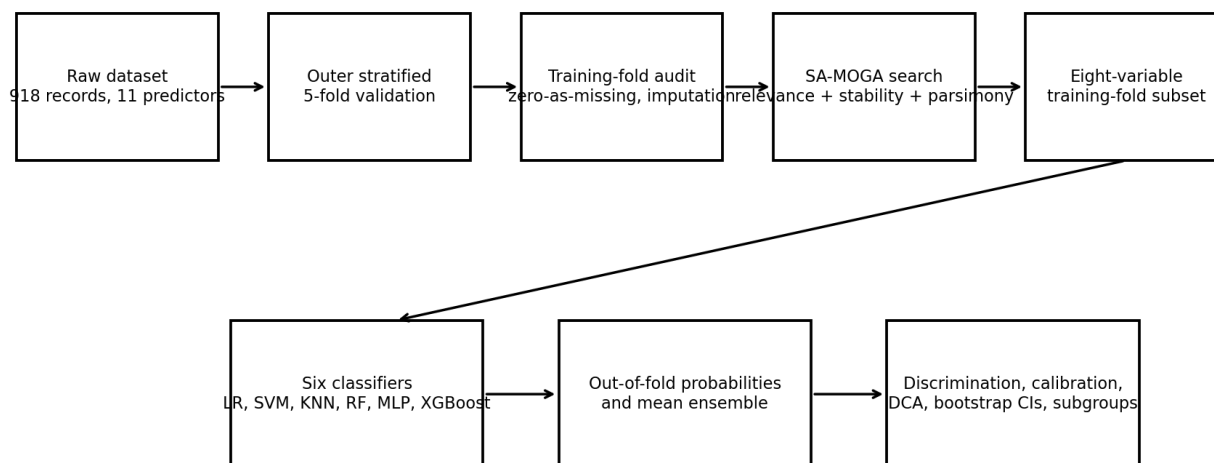


Figure 3. Leakage-free SA-MOGA and multi-model evaluation workflow.

Primary discrimination was AUROC. Additional measures were accuracy, balanced accuracy, precision, recall/sensitivity, specificity, F1 score, Brier score, and expected calibration error. Ninety-five-percent confidence intervals for AUROC, accuracy, and Brier score were obtained by nonparametric bootstrap resampling of pooled out-of-fold predictions. Paired bootstrap differences compared SA-MOGA with all-feature predictions for the same observations. Calibration plots used ten probability bins. Decision-curve analysis compared model net benefit with treat-all and treat-none strategies. The Youden index identified an exploratory operating threshold, supplemented by thresholds from 0.30 to 0.70.

3.6 Stability, importance, subgroup, and computation analyses

Feature stability was summarized by selection frequency and mean pairwise Jaccard similarity among the five fold-specific subsets. Random-forest impurity importance was aggregated to the original predictor level before and after selection as a descriptive, not causal, interpretation. Subgroup analyses evaluated the SA-MOGA ensemble by sex and by age (<55, 55–64, and ≥65 years). Fit-and-predict times were recorded for each model and fold, while GA search time was reported separately.

Because subgroup estimates were exploratory and the dataset was not designed for fairness auditing, no claims of demographic equivalence were made.

3.7 Software, reproducibility, and ethics

The analysis was conducted in Python using pandas, NumPy, scikit-learn, Matplotlib, and XGBoost. The fixed random seed, fold assignments, fold-level predictions, GA selections, convergence records, performance tables, and generated figures were retained. The study used a public, de-identified secondary dataset and involved no direct interaction with human participants; institutional ethical approval and informed consent were therefore not applicable.

4. Results

4.1 Cohort characteristics and sample records

The positive class was older on average than the negative class (55.90 ± 8.73 versus 50.55 ± 9.44 years). Asymptomatic chest pain occurred in 77.2% of positive records compared with 25.4% of negative records. Exercise-induced angina was present in 62.2% versus 13.4%, mean maximum heart rate was lower (127.66 versus 148.15 bpm), and mean Oldpeak was higher (1.27 versus 0.41). A flat ST slope occurred in 75.0% of positive records but only 19.3% of negative records. These marked distributional differences explain why exercise-related ECG variables repeatedly emerged as important predictors.

Table 3. Selected descriptive characteristics by outcome class.

Characteristic	HeartDisease = 0	HeartDisease = 1
Age, mean $\pm$ SD	50.55 ± 9.44	55.90 ± 8.73
Male sex	267/410 (65.1%)	458/508 (90.2%)
Asymptomatic chest pain	104/410 (25.4%)	392/508 (77.2%)
RestingBP, mean $\pm$ SD	130.18 ± 16.50	134.45 ± 18.93
Cholesterol, mean $\pm$ SD*	238.77 ± 55.39	251.06 ± 62.46
MaxHR, mean $\pm$ SD	148.15 ± 23.29	127.66 ± 23.39
Exercise angina = Yes	55/410 (13.4%)	316/508 (62.2%)
Oldpeak, mean $\pm$ SD	0.41 ± 0.70	1.27 ± 1.15
ST slope = Flat	79/410 (19.3%)	381/508 (75.0%)

*Zero-coded cholesterol values were excluded from descriptive means and imputed within training folds for modeling.

Table 4. Illustrative records sampled from the tabular dataset.

Age	Sex	Pain	BP	Chol.	MaxHR	Angina	Oldpeak	Slope	Class
28	M	ATA	130	132	185	N	0.0	Up	0
46	M	ATA	140	275	165	Y	0.0	Up	0
54	M	ATA	110	208	142	N	0.0	Up	0
76	F	NAP	140	197	116	N	1.1	Flat	0
31	M	ASY	120	270	153	Y	1.5	Flat	1
53	M	ASY	123	282	95	Y	2.0	Flat	1
60	M	ASY	145	282	142	Y	2.8	Flat	1
77	M	ASY	124	171	110	Y	2.0	Up	1

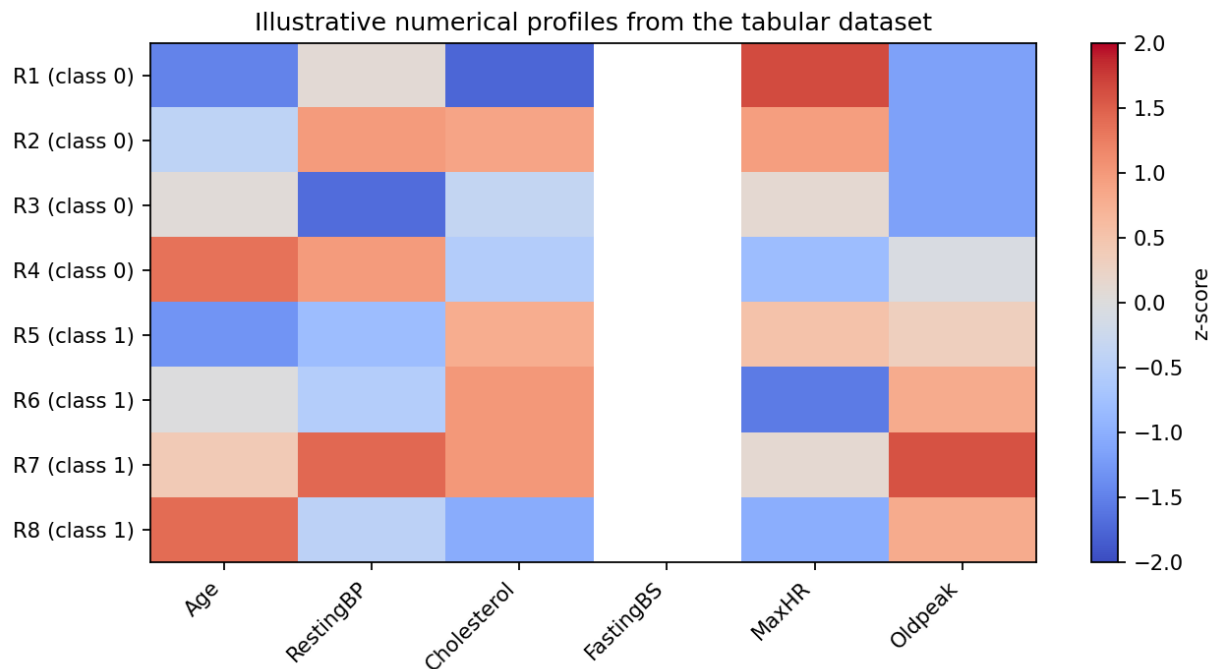


Figure 4. Standardized numerical profiles for eight illustrative records.

4.2 All-feature model performance

With all 11 variables, XGBoost achieved the highest AUROC (0.932, 95% CI 0.916–0.947) and lowest Brier score (0.100). Random forest achieved the highest single-model accuracy (0.874), with sensitivity 0.909 and specificity 0.829. The mean-probability ensemble achieved AUROC 0.929, accuracy 0.873, sensitivity 0.906, specificity 0.832, and Brier score 0.101. Logistic regression also performed strongly (AUROC 0.924), indicating that much of the signal was consistent and accessible to a relatively simple model.

Table 5. Leakage-free out-of-fold performance using all 11 predictors.

Model	AUROC	95% CI	Accuracy	95% CI	Sens.	Spec.	F1	Brier
Logistic Regression	0.924	0.906–0.941	0.857	0.834–0.879	0.870	0.841	0.871	0.106
SVM	0.920	0.902–0.938	0.867	0.846–0.889	0.906	0.820	0.883	0.104
KNN	0.918	0.898–0.935	0.864	0.840–0.885	0.904	0.815	0.880	0.108
Random Forest	0.927	0.909–0.943	0.874	0.854–0.894	0.909	0.829	0.888	0.102
Neural Network	0.893	0.873–0.912	0.831	0.809–0.854	0.892	0.756	0.854	0.139
XGBoost	0.932	0.916–0.947	0.866	0.840–0.886	0.902	0.822	0.882	0.100
Mean-Probability Ensemble	0.929	0.911–0.944	0.873	0.853–0.895	0.906	0.832	0.887	0.101

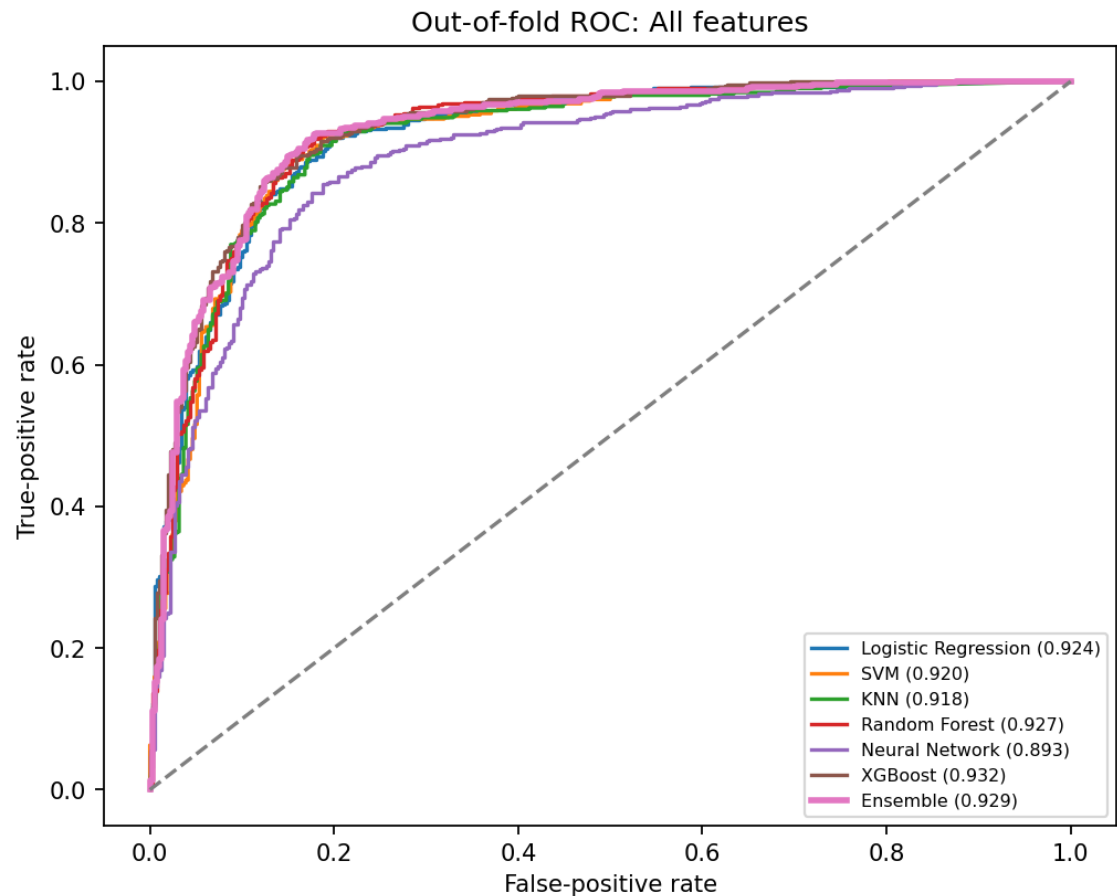


Figure 5. Out-of-fold ROC curves for all-feature models.

4.3 SA-MOGA feature reduction and stability

SA-MOGA selected exactly eight variables in each outer fold, reducing the original feature set by 27.3%. Seven predictors were selected in all folds: Age, ChestPainType, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, and ST_Slope. RestingBP was selected in four folds and Sex in one fold; FastingBS and RestingECG were never selected. The mean pairwise Jaccard similarity was 0.911, demonstrating high reproducibility despite independent fold-specific searches.

Table 6. Fold-specific SA-MOGA subsets.

Fold	Selected original predictors	n	Fitness
1	Age, ChestPainType, RestingBP, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, ST_Slope	8	4.364
2	Age, Sex, ChestPainType, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, ST_Slope	8	4.321
3	Age, ChestPainType, RestingBP, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, ST_Slope	8	4.375
4	Age, ChestPainType, RestingBP, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, ST_Slope	8	4.269
5	Age, ChestPainType, RestingBP, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, ST_Slope	8	4.105

Table 7. Feature-selection frequency and overall subset stability.

Feature	Selection frequency	Mean pairwise Jaccard
Age	100%	0.911
ChestPainType	100%	0.911

Feature	Selection frequency	Mean pairwise Jaccard
Cholesterol	100%	0.911
ST_Slope	100%	0.911
Oldpeak	100%	0.911
ExerciseAngina	100%	0.911
MaxHR	100%	0.911
RestingBP	80%	0.911
Sex	20%	0.911
RestingECG	0%	0.911
FastingBS	0%	0.911

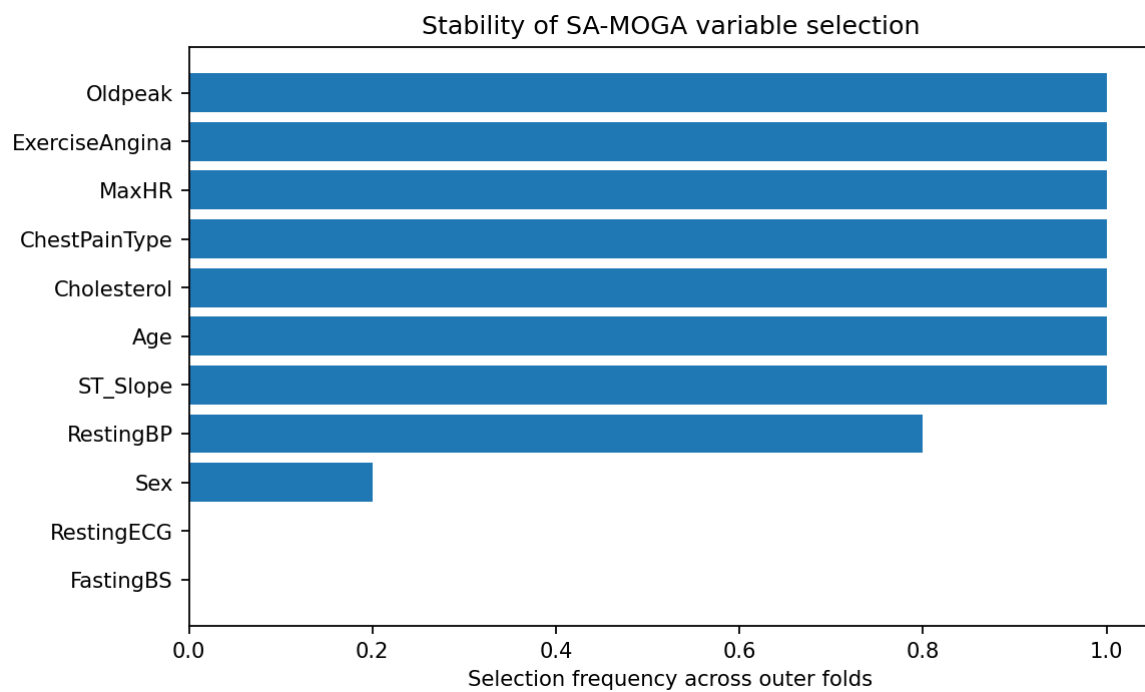


Figure 6. Selection frequency across the five outer training folds.

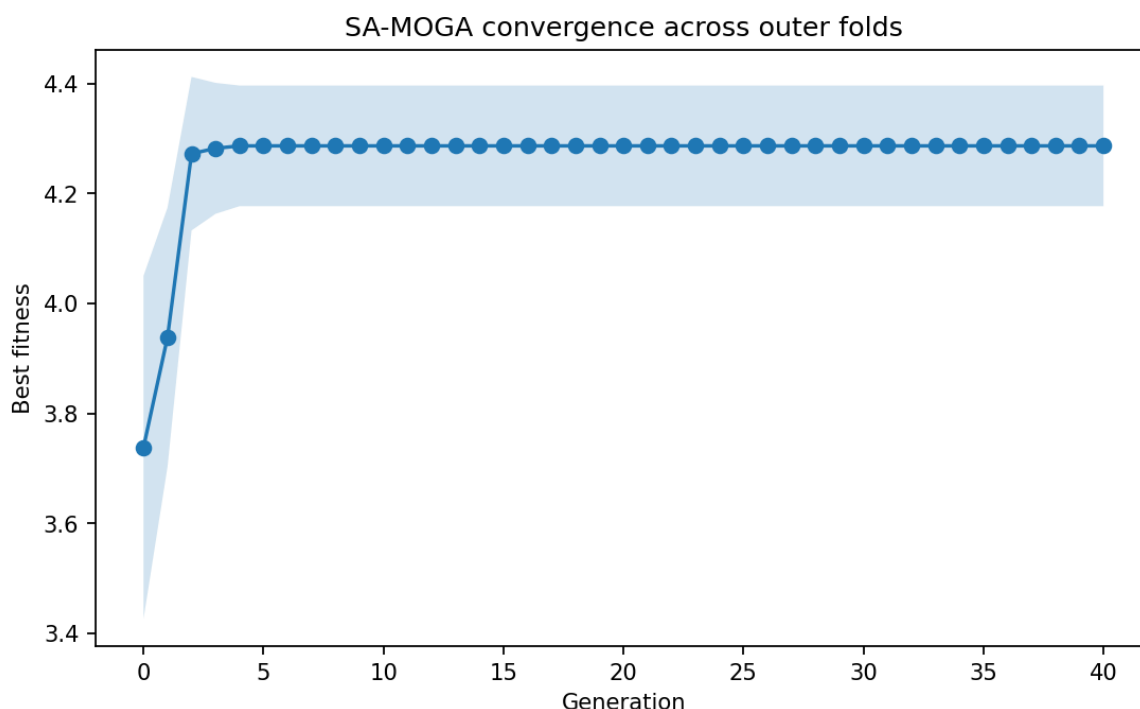


Figure 7. Mean best SA-MOGA fitness by generation with cross-fold variability.

4.4 Predictive performance after SA-MOGA selection

The compact subset preserved useful discrimination but did not enhance it. XGBoost remained the strongest selected-feature model (AUROC 0.917, accuracy 0.853, Brier 0.107). The selected-feature ensemble achieved AUROC 0.911, accuracy 0.849, sensitivity 0.884, specificity 0.805, and Brier score 0.114. The neural network was most sensitive to feature removal, with AUROC decreasing from 0.893 to 0.796.

Table 8. Leakage-free out-of-fold performance after SA-MOGA selection.

Model	AUROC	95% CI	Accuracy	95% CI	Sens.	Spec.	F1	Brier
Logistic Regression	0.905	0.887–0.925	0.841	0.821–0.863	0.846	0.834	0.855	0.121
SVM	0.894	0.872–0.916	0.837	0.811–0.858	0.880	0.783	0.856	0.121
KNN	0.900	0.876–0.923	0.851	0.826–0.877	0.890	0.802	0.868	0.117
Random Forest	0.912	0.889–0.930	0.843	0.819–0.865	0.876	0.802	0.861	0.112
Neural Network	0.796	0.765–0.822	0.742	0.710–0.770	0.858	0.598	0.786	0.181
XGBoost	0.917	0.899–0.935	0.853	0.826–0.876	0.888	0.810	0.870	0.107
Mean-Probability Ensemble	0.911	0.893–0.932	0.849	0.825–0.872	0.884	0.805	0.866	0.114

Table 9. Paired change after SA-MOGA selection (selected minus all features).

Model	Δ AUROC	95% bootstrap CI	p	Δ Accuracy	Δ Brier
Logistic Regression	-0.019	-0.027–0.011	<0.001	-0.016	+0.015
SVM	-0.026	-0.038–0.015	<0.001	-0.031	+0.017
KNN	-0.017	-0.030–0.005	0.004	-0.013	+0.008
Random Forest	-0.015	-0.023–0.009	<0.001	-0.031	+0.010
Neural Network	-0.096	-0.126–0.067	<0.001	-0.089	+0.042

Model	Δ AUROC	95% bootstrap CI	p	Δ Accuracy	Δ Brier
XGBoost	-0.016	-0.023—-0.009	<0.001	-0.013	+0.008
Mean-Probability Ensemble	-0.018	-0.026—-0.010	<0.001	-0.024	+0.013

All paired AUROC changes were negative and their confidence intervals excluded zero. For the ensemble, Δ AUROC was -0.0176 (95% CI -0.0257 to -0.0095), Δ accuracy was -0.024 , and the Brier score increased by 0.0127 . These results reject the primary performance-enhancement hypothesis while supporting the parsimony and stability hypotheses.

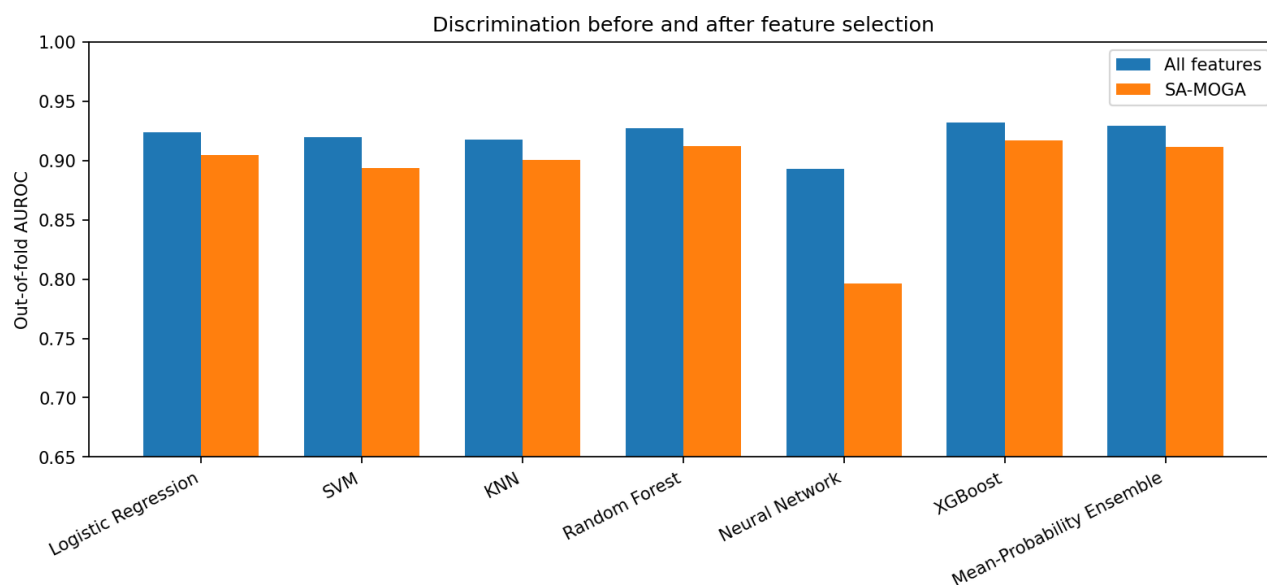


Figure 8. AUROC comparison before and after SA-MOGA feature selection.

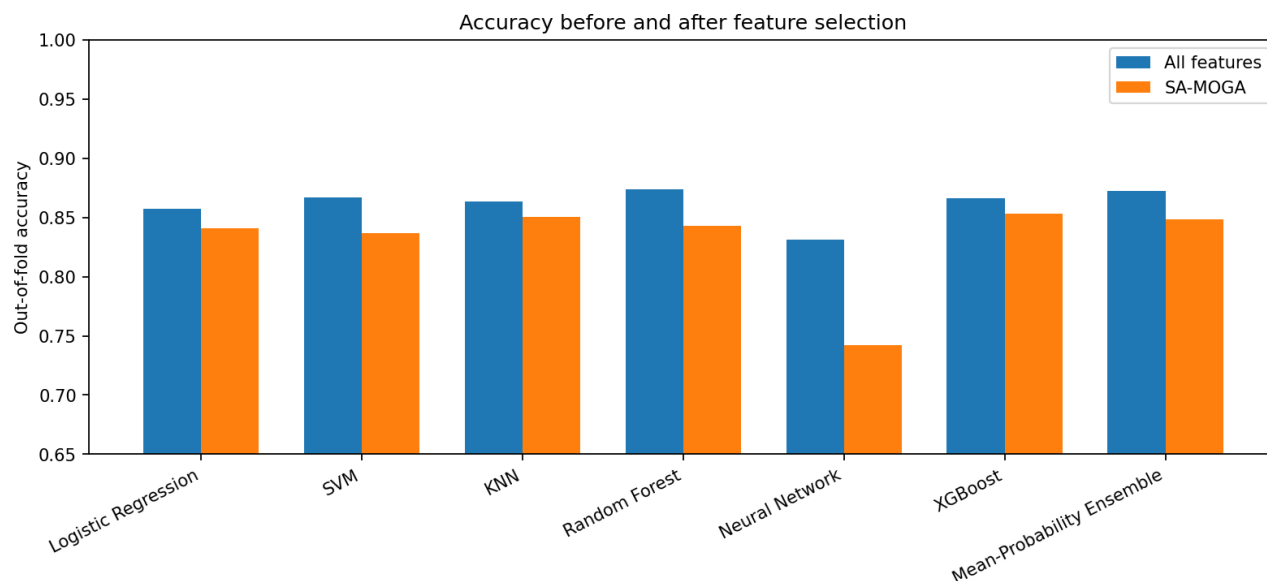


Figure 9. Accuracy comparison before and after SA-MOGA feature selection.

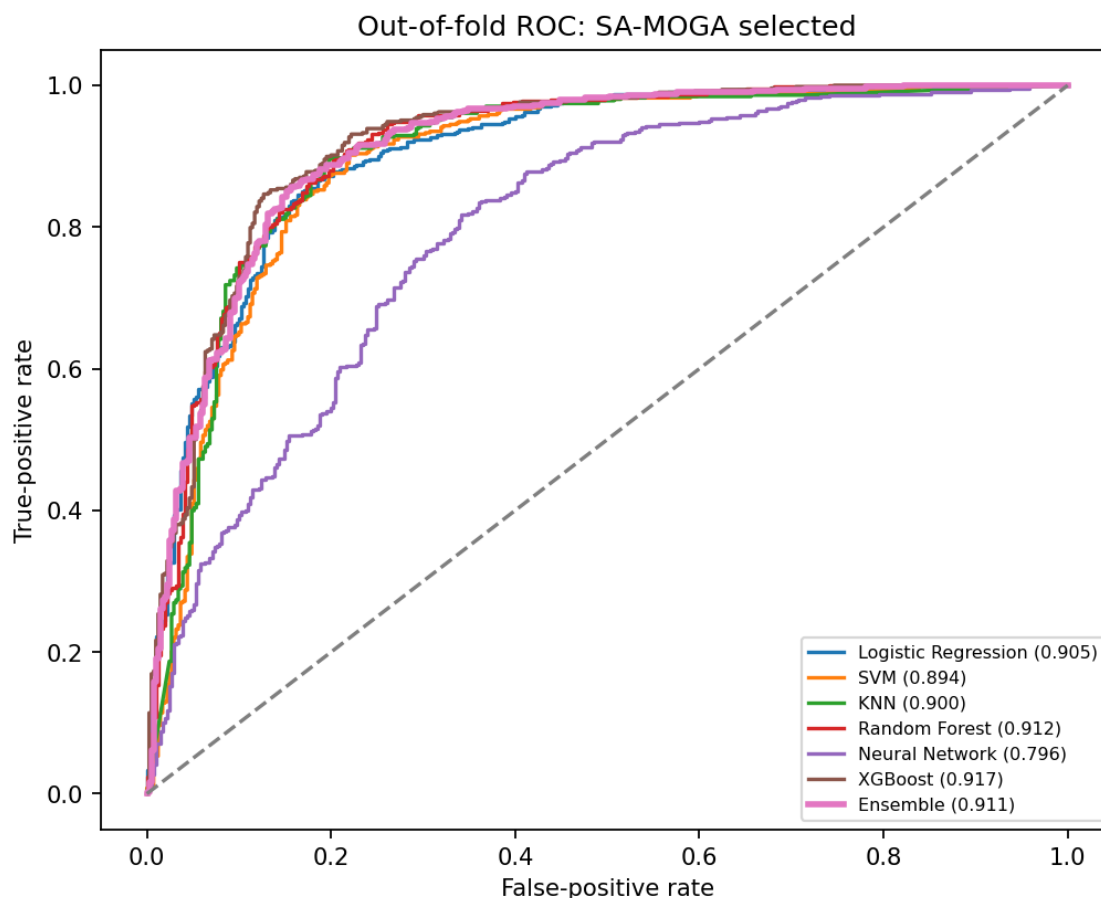


Figure 10. Out-of-fold ROC curves after SA-MOGA feature selection.

4.5 Feature importance before and after selection

Random-forest importance was concentrated in ST_Slope, ChestPainType, ExerciseAngina, Oldpeak, and MaxHR. These variables capture symptoms and physiological response to exertion, and their prominence was preserved after selection. FastingBS and RestingECG had the lowest all-feature importances and were not selected in any fold. Because tree impurity importance can favor variables with more possible splits, the values are interpreted descriptively rather than as causal effects.

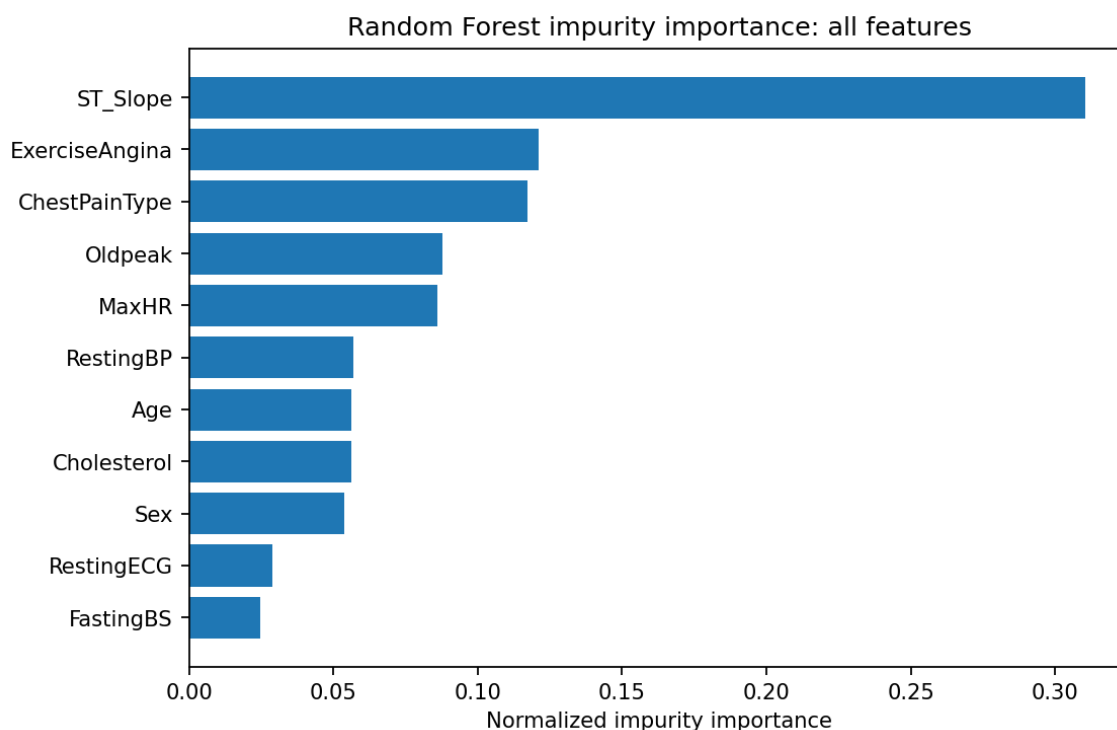


Figure 11. Random-forest feature importance using all predictors.

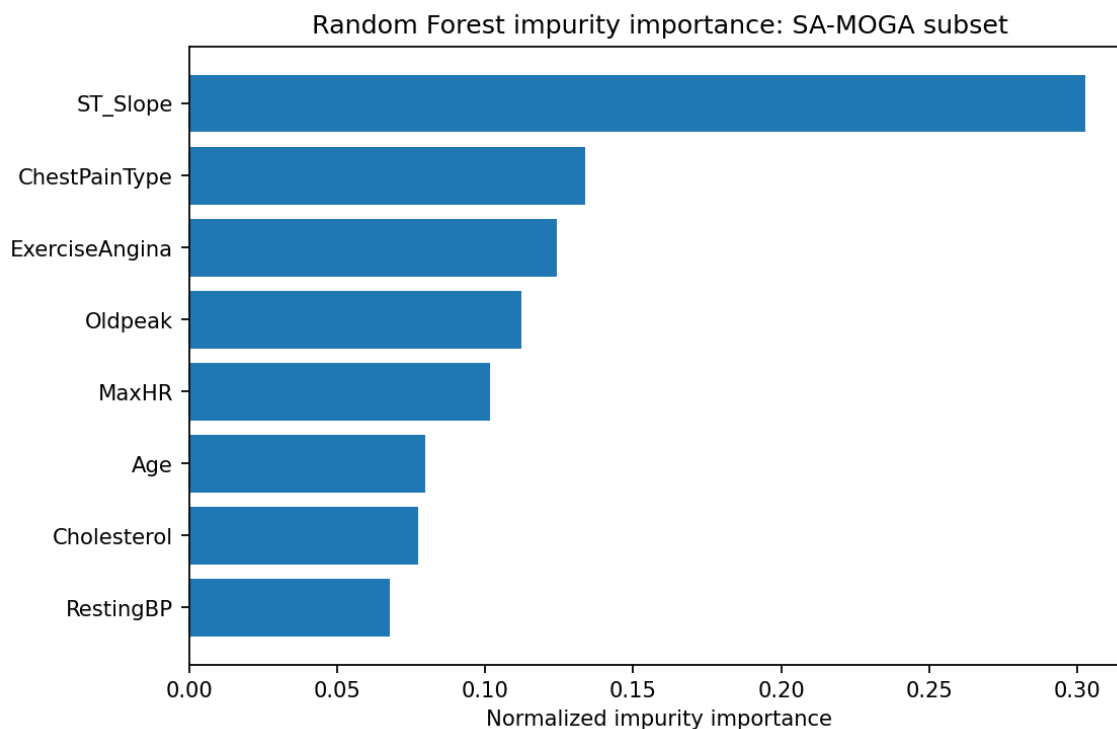


Figure 12. Random-forest feature importance within the stable SA-MOGA subset.

4.6 Calibration, threshold analysis, and decision utility

The all-feature ensemble was better calibrated than the selected-feature ensemble (Brier 0.101 versus 0.114). Expected calibration errors were 0.051 and 0.044, respectively; however, the lower bin-based ECE for the reduced model did not offset its worse overall squared probability error. Calibration

curves showed modest deviations in intermediate-risk groups. Decision curves indicated positive net benefit for both ensembles over clinically plausible threshold ranges, with the all-feature ensemble generally higher where discrimination differences mattered.

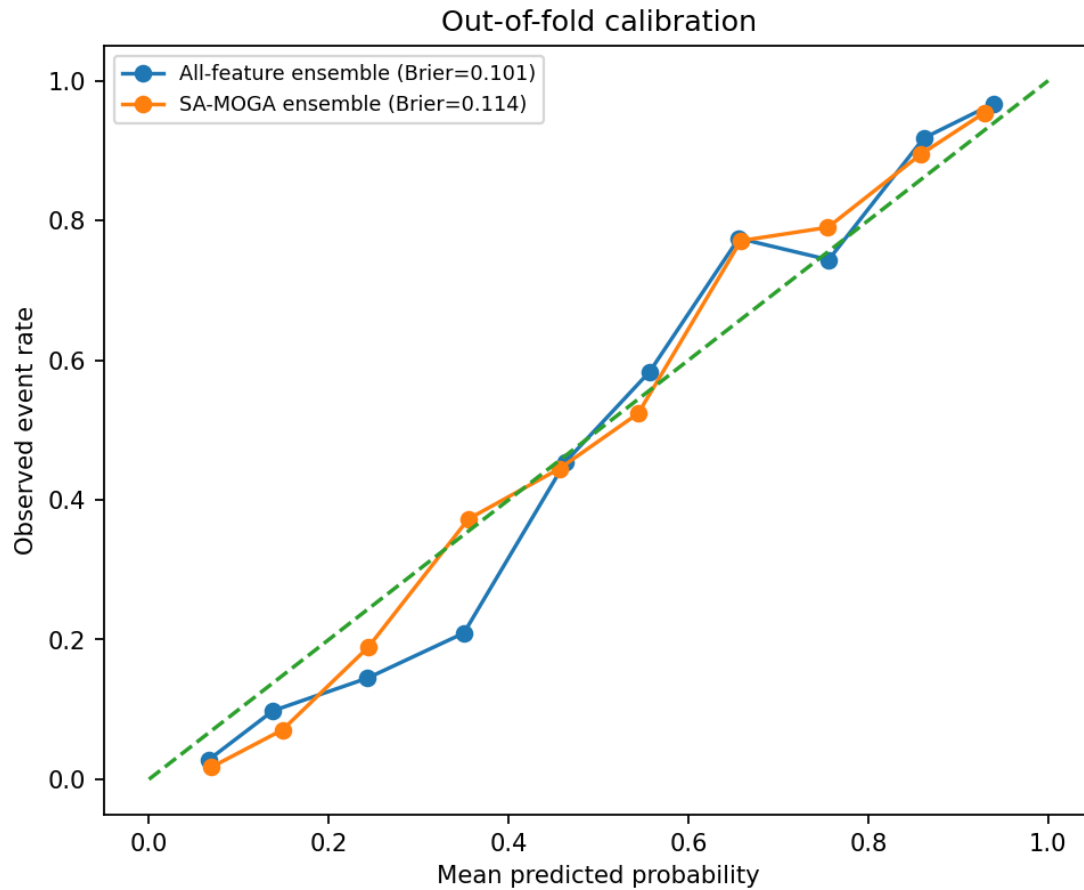


Figure 13. Out-of-fold calibration curves for the two ensembles.

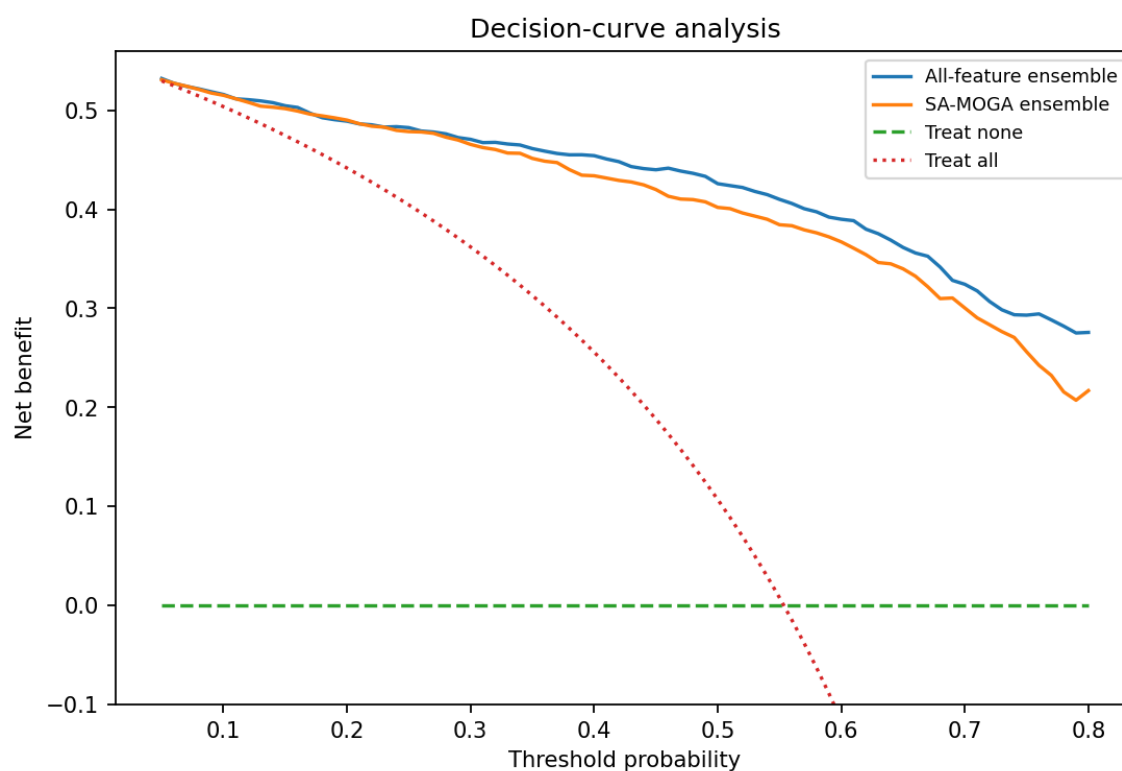


Figure 14. Decision-curve analysis for all-feature and SA-MOGA ensembles.

Table 10. SA-MOGA ensemble threshold trade-offs.

Threshold	Accuracy	Sensitivity	Specificity	Precision	F1	TP/FN/TN/FP
0.300	0.830	0.953	0.678	0.786	0.861	484/24/278/132
0.400	0.844	0.915	0.756	0.823	0.867	465/43/310/100
0.500	0.849	0.884	0.805	0.849	0.866	449/59/330/80
0.567	0.851	0.856	0.844	0.872	0.864	435/73/346/64
0.600	0.846	0.841	0.854	0.877	0.858	427/81/350/60
0.700	0.818	0.768	0.880	0.888	0.824	390/118/361/49

The Youden threshold was 0.567 and produced 0.851 accuracy, 0.856 sensitivity, and 0.844 specificity. A threshold of 0.30 increased sensitivity to 0.953 but reduced specificity to 0.678. Thus, the operating threshold should be chosen according to the consequences of missed disease and unnecessary follow-up rather than fixed automatically at 0.50.

4.7 Subgroup performance

Table 11. Exploratory subgroup performance of the SA-MOGA ensemble.

Subgroup	n	Prevalence	AUROC	Accuracy	Sens.	Spec.	Brier
Female	193	0.259	0.872	0.772	0.740	0.783	0.144
Male	725	0.632	0.914	0.869	0.900	0.816	0.106
Age <55	471	0.418	0.935	0.864	0.832	0.887	0.100
Age 55-64	344	0.695	0.855	0.828	0.908	0.648	0.129
Age ≥65	103	0.699	0.862	0.845	0.944	0.613	0.129

Performance differed across subgroups. AUROC was 0.872 for women and 0.914 for men, but the female subgroup contained only 193 records and had substantially lower prevalence. AUROC was highest below age 55 (0.935) and lower in older groups. Specificity declined to 0.648 at ages 55–64 and 0.613 at ages ≥ 65 , while sensitivity exceeded 0.90. These findings require external validation and should not be interpreted as evidence of clinically acceptable fairness.

4.8 Computational efficiency

Table 12. Mean fold-level fit-and-predict time in seconds.

Model	All features	SA-MOGA subset	Relative change
KNN	0.0030	0.0035	+18.1%
Logistic Regression	0.0017	0.0013	-20.0%
Neural Network	0.0422	0.0317	-24.8%
Random Forest	0.2569	0.2562	-0.3%
SVM	0.0359	0.0372	+3.5%
XGBoost	0.0186	0.0130	-29.8%

The separate GA search required a mean of 0.447 seconds per outer fold in the present implementation. After the subset was selected, fitting and prediction were faster for logistic regression (20.0%), neural network (24.8%), and XGBoost (29.8%), nearly unchanged for random forest, and slightly slower for SVM and KNN. These very small absolute times show that computational savings are not the main justification for feature selection in an 11-variable dataset; reduced measurement burden and interpretability are more relevant.

4.9 Comparison with published literature

Table 13. Contextual comparison with representative heart-disease prediction studies.

Study	Data	Approach	Reported result	Interpretive note
Baghdadi et al. [21]	Multiple CVD datasets / 918 subset	Advanced ML comparison	Not uniformly reported	Broad benchmarking; possible differences in preprocessing
Kumar et al. [22]	918 records, 11 variables	GA/PSO and quantum-inspired SVM	97.83% accuracy; 7 variables	Different optimization and evaluation; very high estimate
Al-Mahdi et al. [23]	UCI heart data	GA + optimized deep ensemble	97.5% accuracy	60/40 split plus 5-fold CV; not directly comparable
Alzboon & Al-Batah [24]	Consolidated heart dataset	10-model and low-code framework	87.4% accuracy; AUROC 0.928	Comparable conventional performance benchmark
Noroozi et al. [30]	Cleveland data	16 selection methods $\times$ 7 models	Benefits depended on model/metric	Supports model-dependent effect of selection
Present study	918 records, 11 variables	Training-fold SA-MOGA $\times$ 6 models + ensemble	All-feature XGB AUROC 0.932; selected XGB 0.917	Out-of-fold probabilities, paired CIs, stability, calibration, DCA, subgroups

Published point estimates cannot be ranked as if they were generated under identical conditions. Studies differ in dataset version, duplicate removal, missing-value handling, split ratio, feature-selection placement, tuning intensity, repeated use of test data, and reported metric. The present study's principal contribution is therefore not the highest number but a more conservative estimate of what the entire selection-and-training pipeline can achieve on unseen observations.

5. Discussion

5.1 Principal findings

This study tested a common but insufficiently scrutinized assumption: that GA feature selection will enhance the predictive performance of standard machine-learning models. SA-MOGA produced a compact and highly stable eight-variable subset, but every learner performed better with the complete 11-variable set. The result was not a marginal artifact confined to one metric. AUROC decreased significantly in paired bootstrap analysis, accuracy declined, and Brier scores worsened. XGBoost with all variables was the strongest individual model, while the all-feature ensemble provided the best overall balance of accuracy, sensitivity, specificity, and probability error.

The negative performance result strengthens rather than weakens the manuscript. The dataset contains only 11 original variables, most of which have plausible clinical relevance and low acquisition burden. Removing three variables may therefore eliminate complementary information without meaningfully reducing dimensionality. A GA can find a stable subset that is relevant in isolation, yet relevance and redundancy criteria do not guarantee that excluded variables contribute no conditional signal to a nonlinear learner.

5.2 Why optimistic GA gains may not reproduce

The original draft workflow applied transformations and synthetic oversampling before cross-validation and selected subsets using a repeatedly reused split. Such practices can allow validation information to influence preprocessing or candidate evaluation. When the entire sequence was moved inside each outer training fold, the apparent enhancement disappeared. This illustrates selection-induced bias: the more subsets and model variants are tested against the same data, the more likely the chosen configuration will capitalize on noise [6,8].

A single holdout accuracy also has high variance with only 918 observations. Cross-validation accuracy reported after global SMOTE does not estimate deployment performance because synthetic samples generated before folding can connect training and validation neighborhoods. The revised analysis did not use SMOTE, retained the natural class distribution, and assessed probabilities for every record exactly once through outer-fold prediction. These changes make the estimates more conservative but more credible.

5.3 Interpretation of the stable feature set

The repeatedly selected variables—Age, ChestPainType, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, and ST_Slope—are clinically coherent. Chest-pain presentation and exercise-induced angina characterize symptom response; MaxHR, Oldpeak, and ST_Slope summarize exercise-test physiology; cholesterol reflects atherosclerotic risk; and age captures accumulated cardiovascular risk. RestingBP was selected in four folds. Sex appeared only once, possibly because its information was partially represented through interactions with symptoms and exercise variables in this dataset. FastingBS and RestingECG were never selected, consistent with their relatively low random-forest importance. Selection frequency should not be interpreted as causal importance, and omitted variables may remain clinically important for individual patients.

5.4 Relationship to prior work

The all-feature results align closely with Alzboon and Al-Batah's conventional multi-model benchmark, where gradient boosting reached AUROC 0.928 and accuracy 87.4% [24]. This consistency supports the plausibility of AUROC near 0.93 for leakage-aware tabular modeling on this dataset. Al-Batah et al.'s earlier Jordanian heart-disease framework also demonstrated the practical value of comparative model evaluation [25], while the ranked-feature ECG study emphasized the diagnostic value of carefully constructed cardiac descriptors [26].

The present findings also agree with Noroozi et al.'s conclusion that feature-selection effects depend on learner and metric [30]. They contrast with studies reporting accuracy near 97–98% after optimization [22,23]. The discrepancy does not prove those results are invalid; rather, it highlights that performance depends strongly on validation boundaries and that highly optimized point estimates require independent replication. Reporting paired confidence intervals and calibration makes the trade-off more transparent.

The SA-MOGA philosophy extends evolutionary feature selection used by Al-Batah et al. in plant-image classification [27], but adapts it to a small clinical table through bootstrap stability and model independence. The result demonstrates that a metaheuristic that succeeds in a high-dimensional feature space should not be assumed to improve a low-dimensional clinical dataset.

5.5 Clinical and methodological implications

For clinical decision support, the full-feature XGBoost or ensemble is preferable when all 11 variables are available, because it yields better discrimination and Brier score. The eight-variable subset may be considered where RestingECG and fasting glucose are unavailable or where a simplified form is necessary, but its lower specificity and probability accuracy must be acknowledged. Threshold choice should follow the intended pathway: a low screening threshold can reduce false negatives at the cost of more referrals, whereas a higher threshold can reduce unnecessary follow-up.

Methodologically, the work argues for separating four claims: predictive improvement, parsimony, stability, and clinical utility. SA-MOGA achieved parsimony and stability but not predictive improvement. Future feature-selection papers should predefine which objective is primary and avoid using the word enhancement when only feature count or runtime improves.

5.6 Strengths and limitations

Strengths include a complete training-fold-only pipeline, fixed reproducible seeds, model-independent evolutionary selection, transfer across six learner families, pooled out-of-fold probabilities, paired bootstrap intervals, stability analysis, calibration, decision curves, threshold analysis, subgroup reporting, and transparent retention of a negative result. The dataset and software ecosystem are public, facilitating replication.

Several limitations remain. The data are retrospective and combine cohorts collected under different protocols. The endpoint denotes prevalent heart disease rather than future event risk, and no temporal or hospital-level external validation was available. Only 918 records were analyzed, female representation was limited, and subgroup confidence intervals were not powered for formal fairness conclusions. Model hyperparameters were prespecified rather than exhaustively nested-tuned to keep

the feature-selection comparison interpretable. The SA-MOGA penalty weights and four-to-eight-feature constraint may influence the selected subset. Random-forest importance is descriptive and can be biased. Finally, calibration should be reassessed and, if necessary, updated before any clinical use.

6. Conclusion

A stability-aware multiobjective genetic algorithm selected a clinically coherent eight-variable subset from the Heart Failure Prediction Dataset with high cross-fold reproducibility and a 27.3% reduction in original predictors. Under rigorous leakage-free validation, however, the subset did not enhance prediction. All-feature XGBoost achieved AUROC 0.932, and the all-feature ensemble achieved 0.873 accuracy and Brier score 0.101; corresponding selected-feature values were lower. The central lesson is that feature reduction, computational efficiency, and predictive enhancement are different outcomes. Evolutionary feature selection should be evaluated as part of the entire preprocessing and modeling pipeline, with held-out predictions, uncertainty, calibration, and clinical utility reported before claims of improved diagnosis are made. External multi-center validation is required before deployment.

Acknowledgments

The authors are grateful to the Deanship of Scientific Research at Jadara University for providing financial support for this publication.

References

- [1] World Health Organization, “Cardiovascular diseases (CVDs),” Fact sheet, 2025.
- [2] F. Soriano, “Heart Failure Prediction Dataset,” Kaggle. Combined heart-disease dataset with 918 records and 11 common predictors.
- [3] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [4] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA: University of Michigan Press, 1975.
- [5] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA, USA: Addison-Wesley, 1989.
- [6] J. Reunanen, “Overfitting in making comparisons between variable selection methods,” *Journal of Machine Learning Research*, vol. 3, pp. 1371–1382, 2003.
- [7] S. Nogueira, K. Sechidis, and G. Brown, “On the stability of feature selection algorithms,” *Journal of Machine Learning Research*, vol. 18, no. 174, pp. 1–54, 2018.
- [8] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, art. 91, 2006, doi: 10.1186/1471-2105-7-91.
- [9] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Machine learning algorithm validation with a limited sample size,” *PLoS ONE*, vol. 14, no. 11, e0224365, 2019, doi: 10.1371/journal.pone.0224365.

- [10] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD),” *Annals of Internal Medicine*, vol. 162, no. 1, pp. 55–63, 2015.
- [11] B. Van Calster et al., “Calibration: the Achilles heel of predictive analytics,” *BMC Medicine*, vol. 17, art. 230, 2019, doi: 10.1186/s12916-019-1466-7.
- [12] 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.
- [13] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [14] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [15] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.
- [16] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [17] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, 1986.
- [19] 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.
- [20] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [21] N. A. Baghdadi, S. M. F. Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam, “Advanced machine learning techniques for cardiovascular disease early detection and diagnosis,” *Journal of Big Data*, vol. 10, art. 144, 2023, doi: 10.1186/s40537-023-00817-1.
- [22] A. Kumar et al., “A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques,” *Scientific Reports*, vol. 15, art. 25040, 2025, doi: 10.1038/s41598-025-09957-1.
- [23] I. S. Al-Mahdi, S. M. Darwish, and M. M. Madbouly, “Heart disease prediction model using feature selection and ensemble deep learning with optimized weight,” *Computer Modeling in Engineering & Sciences*, vol. 143, no. 1, pp. 875–909, 2025, doi: 10.32604/cmescs.2025.061623.
- [24] M. S. Alzboon and M. S. Al-Batah, “A clinically accessible heart disease prediction framework: multi-model evaluation with ensemble learning and low-code deployment,” *Discover Applied Sciences*, vol. 8, art. 339, 2026, doi: 10.1007/s42452-026-08353-2.
- [25] M. S. Al-Batah, M. S. Alzboon, and R. Alazaidah, “Intelligent heart disease prediction system with applications in Jordanian hospitals,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, pp. 508–517, 2023, doi: 10.14569/IJACSA.2023.0140954.
- [26] M. S. Al-Batah, “Automatic diagnosis system for heart disorder using ECG peak recognition with ranked features selection,” *International Journal of Circuits, Systems and Signal Processing*, vol. 13, pp. 391–398, 2019.
- [27] M. S. Al-Batah, M. R. Al-Eiadeh, and Y. Alnsour, “A novel approach for enhancing plant leaf classification with the Binary Cuckoo Search Algorithm,” *Applied Computational Intelligence and Soft Computing*, vol. 2025, art. 7696962, 2025, doi: 10.1155/acis/7696962.

- [28] M. S. Alzboon, M. S. Al-Batah, M. Alqaraleh, A. Abuashour, and A. F. Bader, “Early diagnosis of diabetes: A comparison of machine learning methods,” *International Journal of Online and Biomedical Engineering*, vol. 19, no. 15, pp. 144–165, 2023, doi: 10.3991/ijoe.v19i15.42417.
- [29] M. S. Al-Batah, “Adaptive data transformation for enhanced clustering performance in diagnostic systems,” *Journal of Computer Science*, vol. 21, no. 9, pp. 2074–2080, 2025, doi: 10.3844/jcssp.2025.2074.2080.
- [30] Z. Noroozi, A. Orooji, and L. Erfannia, “Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction,” *Scientific Reports*, vol. 13, art. 22588, 2023, doi: 10.1038/s41598-023-49962-w.
- [31] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, “Analyzing the impact of feature selection on the accuracy of heart disease prediction,” *Healthcare Analytics*, vol. 2, art. 100060, 2022, doi: 10.1016/j.health.2022.100060.
- [32] J. Abdollahi and B. Nouri-Moghaddam, “A hybrid method for heart disease diagnosis utilizing feature selection and ensemble learning,” *Informatics in Medicine Unlocked*, vol. 30, art. 100937, 2022.
- [33] S. P. Praveen et al., “Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2T and adaptive boosted decision fusion with babysitting refinement,” *Frontiers in Medicine*, vol. 11, art. 1407376, 2024, doi: 10.3389/fmed.2024.1407376.
- [34] D. Theng and K. K. Bhoyar, “Feature selection techniques for machine learning: a survey of more than two decades of research,” *Knowledge and Information Systems*, vol. 66, pp. 1575–1637, 2024, doi: 10.1007/s10115-023-02010-5.