Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 01 May 2026
Revised 12 June 2026
Accepted 25 June 2026

Natural Resources for Human Health 2026, Vol. 6 (12s): 568–583

KEYWORDS:
cardiovascular disease; genetic algorithm; feature selection; LightGBM; clinical prediction; calibration.

Stability-Guided Genetic Feature Selection Produces a Compact Cardiovascular Disease Classifier with Preserved Discrimination

Mohammad Subhi Al-Batah[1]*, Abdulkarim Alanazi[1]

[1]Department of Computer Science, Faculty of Information Technology, Jadara University, Irbid, Jordan
*Correspondence: Mohammad Subhi Al-Batah (albatah@jadara.edu.jo)

ABSTRACT: Feature selection can reduce measurement burden in cardiovascular disease classification, but single-run evolutionary searches may be unstable and may overstate gains when feature selection and evaluation are not separated. We analyzed a public cardiovascular dataset containing 70,000 examination records. After removing duplicate profiles and applying prespecified physiological plausibility rules, 68,586 records were retained. Six clinically interpretable variables were engineered, giving 16 candidate predictors. A stability-guided, cost-aware genetic algorithm was run five times on training data, and features selected in at least two runs formed an eight-variable consensus panel. Logistic regression, Gaussian naive Bayes, random forest, and LightGBM were evaluated before and after selection; a validation-trained stacked ensemble was assessed secondarily. Final estimates were obtained from an untouched test set of 10,288 records with bootstrap confidence intervals, calibration, decision-curve analysis, and subgroup auditing. The consensus panel contained age, weight, systolic and diastolic blood pressure, cholesterol, physical activity, pulse pressure, and mean arterial pressure, reducing the feature set by 50%. Full-feature LightGBM achieved ROC-AUC 0.8027 (95% CI 0.7948-0.8109), while GA-LightGBM achieved 0.8003 (95% CI 0.7923-0.8088), accuracy 0.7341, and Brier score 0.1811, with an 18.6% reduction in measured training time. The AUC decrease was small but statistically detectable. The stacked model achieved AUC 0.7998 and did not improve on GA-LightGBM. Stability-guided consensus therefore produced a substantially smaller, clinically coherent panel while preserving nearly all discrimination, supporting parsimony with independent evaluation, calibration, and external validation.

1. INTRODUCTION

Cardiovascular diseases remain the leading global cause of death. The World Health Organization estimated that 19.8 million people died from cardiovascular diseases in 2022, accounting for approximately 32% of all deaths; elevated blood pressure, abnormal glucose and lipid levels, excess body weight, tobacco exposure, harmful alcohol use, and physical inactivity remain central modifiable pathways [1]. Earlier identification of individuals with an adverse cardiovascular phenotype can support timely counseling, diagnostic work-up, and treatment, but any data-driven tool intended to assist such decisions must be evaluated for discrimination, calibration, clinical usefulness, and transportability rather than accuracy alone.

Established cardiovascular prediction systems illustrate the value—and the data requirements—of rigorous modeling. The Framingham general cardiovascular risk profile combined routinely measured factors in sex-specific models [2], whereas SCORE2 was developed across European risk regions to estimate 10-year fatal and non-fatal cardiovascular events [3]. More recently, QR4 was derived and externally validated using more than 16 million UK primary-care records and achieved C-statistics of 0.835 in women and 0.814 in men while adding novel clinical risk groups [4]. These studies are prospective risk models with defined time horizons, in contrast to many public machine-learning datasets that encode the presence or absence

of disease at an examination. Distinguishing prognostic risk estimation from cross-sectional diagnostic classification is essential for valid interpretation.

Machine-learning research has nevertheless shown that compact panels of routine measurements can produce useful cardiovascular classification. Alqahtani et al. [5] reported an ensemble accuracy of 88.70% on a public 70,000-record dataset. Peng et al. [6] combined the same public source with hospital data and obtained an AUC of approximately 0.806 using an XGBH model; a three-feature version based on systolic blood pressure, cholesterol, and age retained an AUC near 0.799. These findings highlight the dominant signal carried by age and blood-pressure-related variables. Recent related work has also evaluated dataset-defined myocardial infarction risk, clinically accessible multi-model heart-disease prediction, and hospital-oriented cardiovascular decision support using machine-learning methods [7-9]. However, reported performance across studies is difficult to compare because preprocessing, outlier handling, resampling, threshold selection, and the separation of feature selection from final evaluation are often inconsistent. High point estimates can also conceal poor probability calibration or optimistic internal validation.

Feature selection is attractive because redundant or weak predictors can increase acquisition burden, reduce interpretability, and amplify variance. Ranked feature selection has likewise been applied to ECG-based heart-disorder classification, illustrating the diagnostic value of compact informative inputs [10]. Filter, wrapper, embedded, and evolutionary methods search the feature space using different assumptions [11-13]. Genetic algorithms encode candidate subsets as binary chromosomes and evolve them through selection, crossover, and mutation [14]. Their flexibility is valuable when interactions and nonlinearities are plausible, but their stochastic nature creates a reproducibility problem: a single run may return a locally favorable subset that is unstable under a new seed or resample. Stability selection principles suggest that repeated selection frequency can be used to distinguish persistent variables from chance discoveries [15]. A recent stability-aware genetic algorithm study used repeated runs and consensus retention to reduce single-run variability in another tabular classification setting [16], providing a methodological precedent for the present consensus strategy.

The empirical effect of feature selection is also model-dependent. Noroozi et al. [17] compared 16 selection methods on the Cleveland heart-disease dataset and found that selection improved some learners but reduced performance for others; their genetic algorithm mainly improved sensitivity, while the best accuracy arose from filter methods. Praveen et al. [18] combined hybrid metaheuristic selection with boosted decision fusion and reported an AUC of 0.89 and accuracy of 0.83. Collectively, the literature supports evolutionary selection as a potentially useful optimization layer, but not as a guarantee of superior discrimination. A stronger evaluation must quantify whether parsimony is purchased at a statistically and clinically meaningful performance cost.

Accordingly, this study develops and evaluates a stability-guided genetic feature-selection framework for cross-sectional cardiovascular disease classification. The contributions are fivefold: (i) prespecified physiological quality control and an untouched 15% test set; (ii) clinically interpretable feature engineering, including body mass index, pulse pressure, mean arterial pressure, metabolic burden, and lifestyle risk; (iii) five independent, cost-aware genetic searches summarized by consensus frequency; (iv) comparison of linear, probabilistic, bagging, boosting, and stacked models before and after selection; and (v) reporting aligned with TRIPOD+AI principles, including bootstrap uncertainty, calibration, decision-curve analysis, and subgroup performance [19]. The central hypothesis was that a stable GA panel would preserve most of the predictive information in the full engineered feature set while reducing model complexity.

2. MATERIALS AND METHODS

2.1 Study design and reporting framework

This was a secondary methodological analysis of a publicly available, de-identified tabular dataset. The task was defined as binary classification of the recorded cardiovascular disease label at the time represented by the dataset, not estimation of future event risk over a specified horizon. Model development and reporting were informed by TRIPOD+AI, with explicit separation of training, tuning, and final evaluation data [19]. Risk-of-bias considerations were interpreted using PROBAST concepts, particularly participant selection, predictor definition, outcome definition, and analysis [20]. Figure 1 summarizes the complete workflow.

Natural Resources for Human Health
eISSN: 2583-1194

Figure 1. Study workflow. Physiological quality control retained 68,586 records; 16 candidate features were reduced to an eight-feature consensus panel before evaluation on an untouched test set of 10,288 records.

2.2 Data source, variables, and outcome

The supplied file, cardio_train.csv, contained 70,000 rows and 13 columns separated by semicolons: an identifier, age in days, a recorded gender code, height, weight, systolic blood pressure, diastolic blood pressure, ordinal cholesterol, ordinal glucose, smoking, alcohol use, physical activity, and the binary cardio outcome. The target was coded 0 for no recorded cardiovascular disease and 1 for recorded cardiovascular disease. No missing values were present. The original public dataset page identifies these variables as objective, examination, and subjective features, but detailed information on recruitment setting, diagnostic adjudication, medications, comorbidities, geography, and timing is limited. The identifier was excluded from all modeling.

Table 1. Predictor dictionary and engineered features.

Variable	Source	Unit/coding	Transformation/QC	Interpretation
age_years	Derived	years	age / 365.25	Age at examination
gender	Original	code 1 or 2	as supplied	Recorded gender code; semantics require source verification
height	Original	cm	120–220 retained	Standing height
weight	Original	kg	30–250 retained	Body weight
ap_hi	Original	mmHg	70–250 retained	Systolic blood pressure
ap_lo	Original	mmHg	40–150 retained; < ap_hi	Diastolic blood pressure
cholesterol	Original	1–3	as supplied	Normal, above normal, well above normal
gluc	Original	1–3	as supplied	Normal, above normal, well above normal
smoke	Original	0/1	as supplied	Current smoking indicator
alco	Original	0/1	as supplied	Alcohol-use indicator
active	Original	0/1	as supplied	Physical-activity indicator
bmi	Derived	kg/m^2	weight / $height^2$	Body mass index
pulse_pressure	Derived	mmHg	ap_hi − ap_lo	Arterial pulse pressure
mean_arterial_pressure	Derived	mmHg	(ap_hi + 2ap_lo) / 3	Approximate mean arterial pressure

Natural Resources for Human Health
eISSN: 2583-1194

Variable	Source	Unit/coding	Transformation/QC	Interpretation
metabolic_burden	Derived	0–4	(cholesterol−1)+(gluc−1)	Combined ordinal metabolic abnormality
lifestyle_risk	Derived	0–3	smoke+alco+(1−active)	Count of adverse lifestyle indicators

Because the dataset is tabular and contains no radiological, histological, or other image files, representative data examples are presented as de-identified rows rather than image samples (Table 2). This prevents the introduction of synthetic visual material that is not part of the source dataset.

Table 2. Representative de-identified records from the cleaned dataset.

Age	Gender code	BMI	SBP	DBP	Chol.	Glucose	Active	CVD
53.5	1	32.0	120	90	1	1	1	1
49.3	1	19.8	100	70	2	1	0	0
47.4	1	28.0	150	90	1	1	1	0
60.5	1	24.5	120	80	1	1	1	1
42.1	1	27.8	150	90	2	2	1	1
60.0	1	37.4	115	70	3	1	0	1
53.8	1	23.9	120	80	1	1	0	0
59.6	2	22.8	120	80	1	1	1	0

2.3 Quality control and data partitioning

Exact duplicate profiles were removed after excluding the identifier. A prespecified physiological plausibility filter retained height from 120 to 220 cm, weight from 30 to 250 kg, systolic pressure from 70 to 250 mmHg, diastolic pressure from 40 to 150 mmHg, and systolic pressure greater than diastolic pressure. These broad limits were intended to remove obvious recording errors while avoiding aggressive trimming of clinically extreme observations. In total, 24 duplicates were removed and 1,390 additional rows failed one or more physiological rules, leaving 68,586 records (97.98% of the source data). No imputation or resampling was applied.

The cleaned cohort was split once using stratification by the outcome: 70% training (n=48,010), 15% validation (n=10,288), and 15% testing (n=10,288), with random seed 42. The training set was used for base-model fitting and GA development; the validation set was used for model stacking and the exploratory Youden threshold; and the test set remained untouched until final evaluation. This separation prevented the final test labels from influencing feature selection, hyperparameters, ensemble weights, or threshold choice.

2.4 Clinically informed feature engineering

Six derived variables were constructed before feature selection. Age was converted from days to years. Body mass index was calculated from weight and height. Pulse pressure quantified the systolic–diastolic difference, while mean arterial pressure

summarized average arterial load. Metabolic burden combined the ordinal cholesterol and glucose deviations from their normal category. Lifestyle risk counted smoking, alcohol use, and inactivity. These transformations were deterministic and were applied identically to all partitions. The resulting candidate space contained 16 variables: 10 retained source predictors and six derived features.

2.5 Stability-guided, cost-aware genetic algorithm

Each candidate subset was represented by a 16-bit chromosome, where 1 denoted inclusion. Five independent GA runs used seeds 100–104. Each run used a population of 18 chromosomes, 12 generations, tournament selection of size three, two elites, uniform crossover probability 0.80, bit-wise mutation probability 1/16, and a minimum subset size of three. To control runtime while preserving class balance, each run sampled at most 26,000 training observations and created an internal 70:30 stratified development split. A standardized logistic regression model served as a consistent surrogate evaluator.

The fitness function jointly optimized discrimination and parsimony: $\text{Fitness}(S) = \text{AUC\_validation}(S) - \lambda |S|/p$, where S is the selected subset, $p=16$, and $\lambda=0.015$. The size penalty was deliberately small so that a predictor was removed only when its validation contribution did not compensate for complexity. After five runs, each feature received a selection frequency from 0 to 1. The consensus panel included variables selected in at least two of five runs (frequency ≥0.40); if more than eight passed, the eight most frequent were retained. This strategy converts a stochastic optimizer into a stability summary rather than relying on one best chromosome.

2.6 Machine-learning and hybrid models

Four base learners represented complementary inductive biases. Logistic regression used z-score standardization, L2 regularization with C=1.0, and up to 1,500 iterations. Gaussian naive Bayes provided a low-complexity probabilistic benchmark. Random forest used 120 trees, square-root feature sampling, and a minimum leaf size of two [21]. LightGBM used 120 boosting iterations, learning rate 0.035, 25 leaves, row and column subsampling of 0.85, and L2 regularization of 1.0 [22]. Each base learner was trained separately on all 16 engineered predictors and on the eight GA consensus predictors.

A secondary hybrid model used stacked generalization [23]. The four GA-panel base models generated validation probabilities, which were used to fit a logistic meta-learner. The fixed meta-learner was then applied to the corresponding test probabilities. This design ensured that the test set was not used to learn ensemble weights. The stack was evaluated as a proposed hybrid architecture, but superiority was not assumed; it was required to demonstrate a measurable improvement over the strongest component model.

2.7 Performance estimation and statistical analysis

The primary discrimination measure was area under the receiver-operating-characteristic curve (ROC-AUC). Secondary measures were area under the precision–recall curve (PR-AUC), accuracy, balanced accuracy, precision, sensitivity, specificity, F1-score, Matthews correlation coefficient, Brier score, and log loss. Classification metrics used a probability threshold of 0.50. For the stacked model, a validation-derived Youden threshold was reported only as a sensitivity analysis. Calibration was visualized in ten quantile bins; Brier score summarized overall probabilistic error [24,25]. Decision-curve analysis estimated net benefit across threshold probabilities and compared each model with treat-all and treat-none strategies [26].

Nonparametric bootstrap confidence intervals used 300 test-set resamples for the full-feature LightGBM, GA-LightGBM, and stacked model. Paired AUC differences used 500 bootstrap resamples because predictions were obtained on the same individuals. When no resample crossed zero, the empirical two-sided probability was reported as $p<0.004$ rather than zero. Prespecified subgroup summaries were produced for recorded gender codes and age bands (<50, 50–59, and ≥60 years). Subgroup results were considered diagnostic audits, not evidence of fairness or external validity.

2.8 Software and reproducibility

Analyses were implemented in Python 3.13.5 using pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0 [27], LightGBM 4.6.0, and Matplotlib 3.10.8. Random seeds, feature definitions, GA parameters, model settings, figure generation, and output tables are provided in Supplementary File S1. The source dataset itself is not redistributed. The reported training times are wall-clock measurements from the present computing environment and should be interpreted comparatively rather than as hardware-independent benchmarks.

Natural Resources for Human Health
eISSN: 2583-1194

3. RESULTS AND DISCUSSION

3.1 Cohort characteristics and data quality

The source data were approximately balanced. After quality control, participants with the cardiovascular label were older and had higher mean weight, systolic pressure, diastolic pressure, cholesterol category, body mass index, pulse pressure, mean arterial pressure, and metabolic burden than participants without the label (Table 3). Mean systolic pressure differed by 14.29 mmHg (133.89 versus 119.60 mmHg), and mean age differed by 3.24 years (54.93 versus 51.69 years). The recorded gender-code distribution was similar between outcome groups. Figure 2 shows separation in age, body mass index, systolic pressure, and cholesterol distributions while also illustrating substantial overlap, consistent with a multifactorial classification problem.

Table 3. Descriptive statistics in the cleaned cohort (mean ± standard deviation).

Characteristic	Overall	No CVD	CVD
Age (years)	53.29 ± 6.76	51.69 ± 6.77	54.93 ± 6.34
Recorded gender code	1.35 ± 0.48	1.35 ± 0.48	1.35 ± 0.48
Height (cm)	164.41 ± 7.91	164.51 ± 7.82	164.32 ± 8.00
Weight (kg)	74.12 ± 14.30	71.57 ± 13.26	76.72 ± 14.84
Systolic BP (mmHg)	126.67 ± 16.68	119.60 ± 12.56	133.89 ± 17.27
Diastolic BP (mmHg)	81.30 ± 9.42	78.12 ± 8.14	84.55 ± 9.54
Cholesterol category	1.36 ± 0.68	1.22 ± 0.53	1.52 ± 0.78
Glucose category	1.23 ± 0.57	1.18 ± 0.51	1.28 ± 0.62
Smoking indicator	0.09 ± 0.28	0.09 ± 0.29	0.08 ± 0.28
Alcohol indicator	0.05 ± 0.22	0.06 ± 0.23	0.05 ± 0.22
Physical activity	0.80 ± 0.40	0.82 ± 0.39	0.79 ± 0.41
Body mass index	27.46 ± 5.26	26.47 ± 4.81	28.47 ± 5.50
Pulse pressure (mmHg)	45.37 ± 11.67	41.48 ± 8.94	49.34 ± 12.73
Mean arterial pressure (mmHg)	96.43 ± 11.03	91.95 ± 8.89	101.00 ± 11.14
Metabolic burden	0.59 ± 1.07	0.39 ± 0.89	0.79 ± 1.19
Lifestyle risk	0.34 ± 0.57	0.33 ± 0.57	0.35 ± 0.57

Natural Resources for Human Health
eISSN: 2583-1194

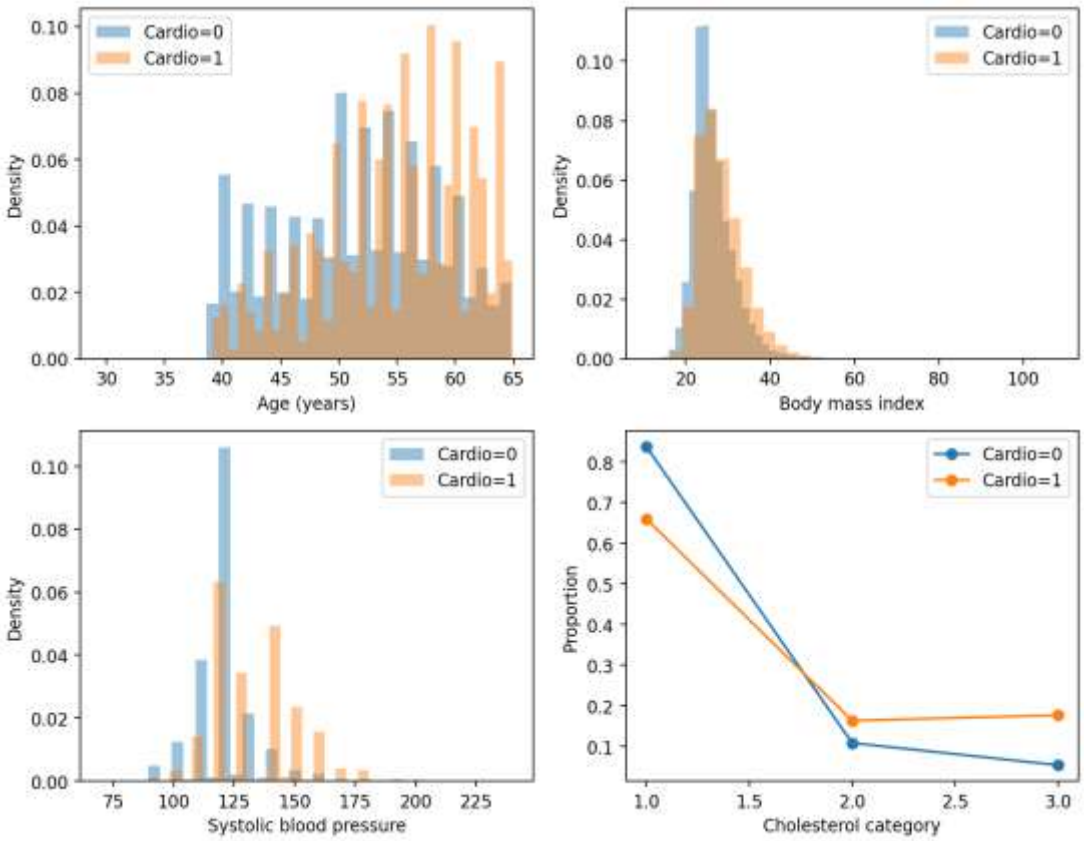

Figure 2. Phenotype distributions by cardiovascular disease label. Continuous panels show density histograms; the cholesterol panel shows category proportions.

3.2 Genetic algorithm convergence and feature stability

The five GA runs converged to subsets containing five to seven variables, with best internal validation AUCs ranging from 0.7882 to 0.7958 during the search. Age was selected in every run; weight, cholesterol, and diastolic pressure appeared in four runs; physical activity appeared in three; and systolic pressure, pulse pressure, and mean arterial pressure appeared in two. The 0.40 consensus threshold therefore produced an eight-variable panel: age, weight, systolic pressure, diastolic pressure, cholesterol, physical activity, pulse pressure, and mean arterial pressure. Gender code, height, alcohol use, and the composite lifestyle-risk variable were never selected. Table 4 details run-specific subsets and frequencies, and Figure 3 demonstrates both convergence and between-run stability.

Table 4. Genetic algorithm run summaries and consensus selection frequencies.

Run (seed)	Selected variables	Best penalized fitness
1 (100)	age, weight, SBP, cholesterol, activity, mean arterial pressure	0.7845
2 (101)	age, weight, SBP, DBP, cholesterol	0.7899
3 (102)	age, weight, DBP, cholesterol, smoking, activity, mean arterial pressure	0.7816
4 (103)	**age, weight, DBP, glucose, pulse pressure, metabolic burden**	**0.7868**
5 (104)	age, DBP, cholesterol, activity, BMI, pulse pressure	0.7887

Run (seed)	Selected variables	Best penalized fitness
Consensus features (selected in ≥2 of 5 runs)		
age	5/5	1.0
weight	4/5	0.8
cholesterol	4/5	0.8
DBP	4/5	0.8
activity	3/5	0.6
SBP	2/5	0.4
pulse pressure	2/5	0.4
mean arterial pressure	2/5	0.4

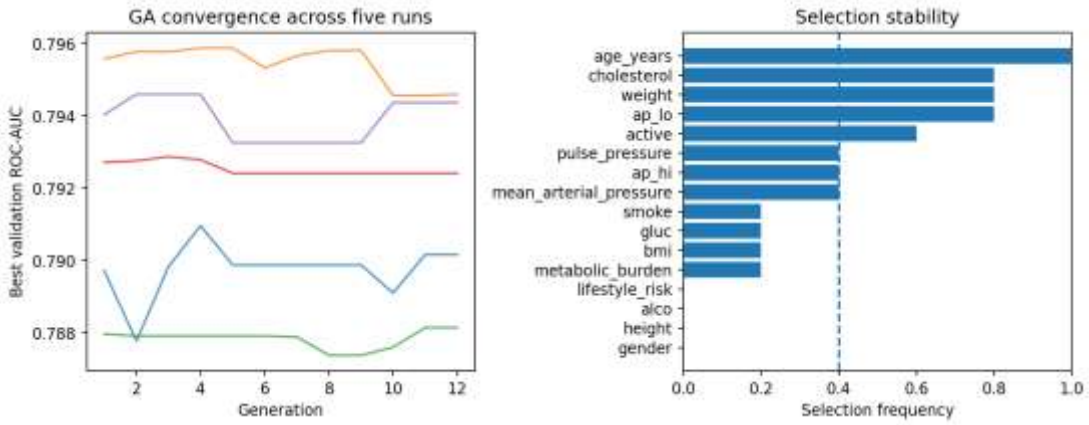

Figure 3. Genetic algorithm behavior. Left: best validation ROC-AUC in each independent run. Right: selection frequency; the dashed line marks the 0.40 consensus threshold.

3.3 Test-set predictive performance

Among models using all 16 engineered features, LightGBM achieved the strongest discrimination (ROC-AUC 0.8027; PR-AUC 0.7904), accuracy 0.7334, F1-score 0.7214, and the lowest Brier score (0.1800). Logistic regression followed with AUC 0.7916. Random forest and Gaussian naive Bayes achieved AUCs of 0.7853 and 0.7801, respectively. The relative ordering is consistent with moderate nonlinear structure but limited incremental information beyond age, blood pressure, and metabolic variables.

After GA selection, LightGBM achieved AUC 0.8003, accuracy 0.7341, F1-score 0.7217, and Brier score 0.1811 using half as many variables. Relative to its full-feature counterpart, measured training time decreased from 0.598 to 0.487 seconds (18.6%), while accuracy increased by 0.0007. The paired AUC difference was −0.00243 (bootstrap 95% CI −0.00377 to −0.00113; $p<0.004$), indicating a small but statistically detectable discrimination loss. Gaussian naive Bayes improved markedly after selection (AUC 0.7904 versus 0.7801), whereas random forest declined (0.7781 versus 0.7853), reinforcing that feature selection effects depend on the learner.

The proposed stacked model achieved AUC 0.7998, accuracy 0.7343, F1-score 0.7188, and Brier score 0.1813. Its paired AUC difference from GA-LightGBM was −0.00051 (95% CI −0.00148 to 0.00036; $p=0.312$). Thus, the stack did not improve

discrimination over its strongest component and should not be described as superior. At the default threshold, its sensitivity was 0.6866 and specificity 0.7809. The validation-derived Youden threshold of 0.5326 increased specificity to 0.8032 but reduced sensitivity to 0.6617.

Table 5. Held-out test performance before and after GA feature selection.

Model	Panel	Accuracy	Sens.	Spec.	F1	ROC-AUC	PR-AUC	Brier	Time (s)
Logistic regression	All 16	0.7275	0.6682	0.7857	0.7082	0.7916	0.7754	0.1866	4.934
Gaussian naive Bayes	All 16	0.7197	0.6124	0.8247	0.6837	0.7801	0.7549	0.2363	0.024
Random forest	All 16	0.7259	0.7039	0.7474	0.7176	0.7853	0.7710	0.1887	1.807
LightGBM	All 16	0.7334	0.6978	0.7682	0.7214	0.8027	0.7904	0.1800	0.598
Logistic regression	GA 8	0.7262	0.6686	0.7826	0.7073	0.7905	0.7739	0.1872	0.015
Gaussian naive Bayes	GA 8	0.7276	0.6171	0.8359	0.6915	0.7904	0.7747	0.2206	0.007
Random forest	GA 8	0.7150	0.6998	0.7299	0.7084	0.7781	0.7652	0.1934	1.183
LightGBM	GA 8	0.7341	0.6969	0.7705	0.7217	0.8003	0.7872	0.1811	0.487
Proposed GA-stacked ensemble	GA 8	0.7343	0.6866	0.7809	0.7188	0.7998	0.7847	0.1813	1.692

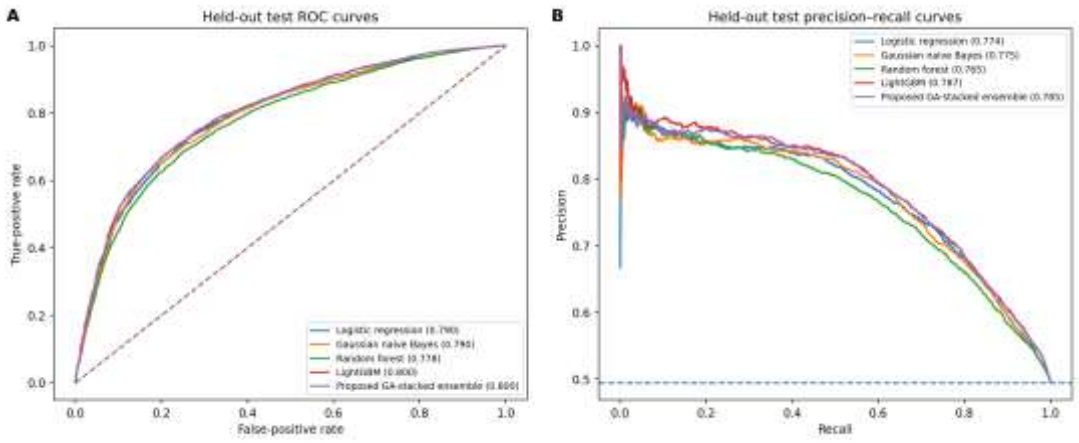

Figure 4. Held-out discrimination using the GA consensus panel. (A) ROC curves with AUC in parentheses. (B) Precision–recall curves with average precision in parentheses; the dashed horizontal line represents outcome prevalence.

Natural Resources for Human Health
eISSN: 2583-1194

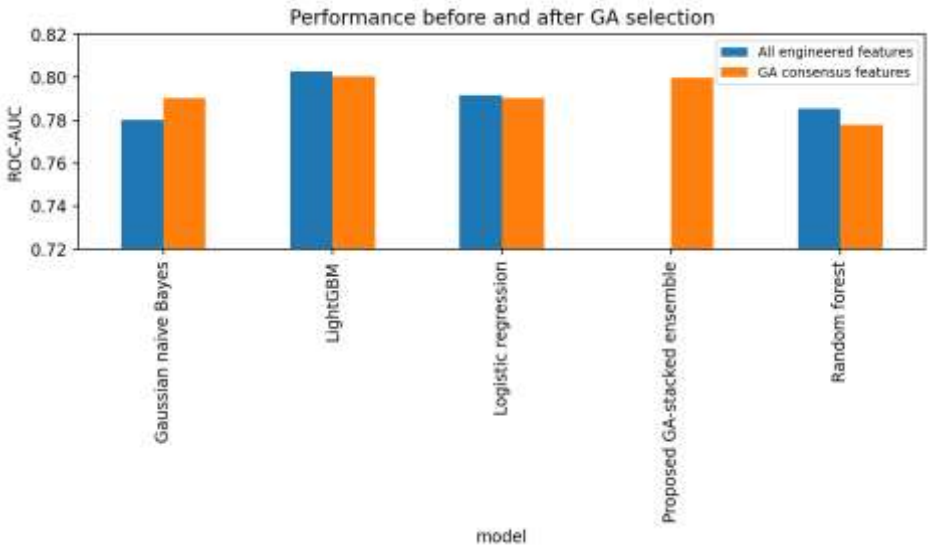

Figure 5. ROC-AUC before and after genetic feature selection. The stacked ensemble is defined only for the GA panel.

3.4 Uncertainty, calibration, and decision utility

Bootstrap intervals for the three load-bearing models were narrow because the test set contained more than 10,000 records (Table 6A). Full-feature LightGBM retained the best point estimates for ROC-AUC and Brier score, while GA-LightGBM and the stack were closely grouped. Calibration plots showed reasonable agreement between predicted and observed proportions across deciles, although the models slightly underpredicted in some upper-middle bins and did not span the full 0–1 probability range. Decision curves showed positive net benefit relative to treat-none over most of the evaluated threshold range and were substantially better than treat-all at moderate and high thresholds (Figure 6). Differences between GA-LightGBM and the stack were small, again indicating that added ensemble complexity did not translate into a clear decision-analytic advantage.

Table 6. Bootstrap uncertainty and subgroup audit for the proposed models.

Model	**ROC-AUC (95% CI)**	**PR-AUC (95% CI)**	**Accuracy (95% CI)**	**F1 (95% CI)**	**Brier (95% CI)**
Proposed GA-stacked ensemble	0.7998 (0.7916–0.8081)	0.7847 (0.7708–0.7962)	0.7343 (0.7261–0.7429)	0.7188 (0.7078–0.7278)	0.1813 (0.1774–0.1851)
LightGBM with GA features	0.8003 (0.7923–0.8088)	0.7872 (0.7753–0.8002)	0.7341 (0.7260–0.7431)	0.7217 (0.7119–0.7311)	0.1811 (0.1772–0.1847)
LightGBM with all features	0.8027 (0.7948–0.8109)	0.7904 (0.7779–0.8032)	0.7334 (0.7249–0.7423)	0.7214 (0.7125–0.7305)	0.1800 (0.1763–0.1835)
B. Proposed stacked model subgroup audit at probability threshold 0.50					
Subgroup	n	Prevalence	ROC-AUC	Sensitivity	Specificity
Gender code 1	**6723**	**0.495**	**0.7993**	**0.6682**	**0.7934**
Gender code 2	3565	0.494	0.8024	0.7216	0.7573
Age <50 years	3194	0.367	0.8246	0.5828	0.9110
Age 50–59 years	5248	0.516	0.7636	0.6531	0.7671
Age ≥60 years	1846	0.654	0.7399	0.8626	0.4232

Natural Resources for Human Health
eISSN: 2583-1194

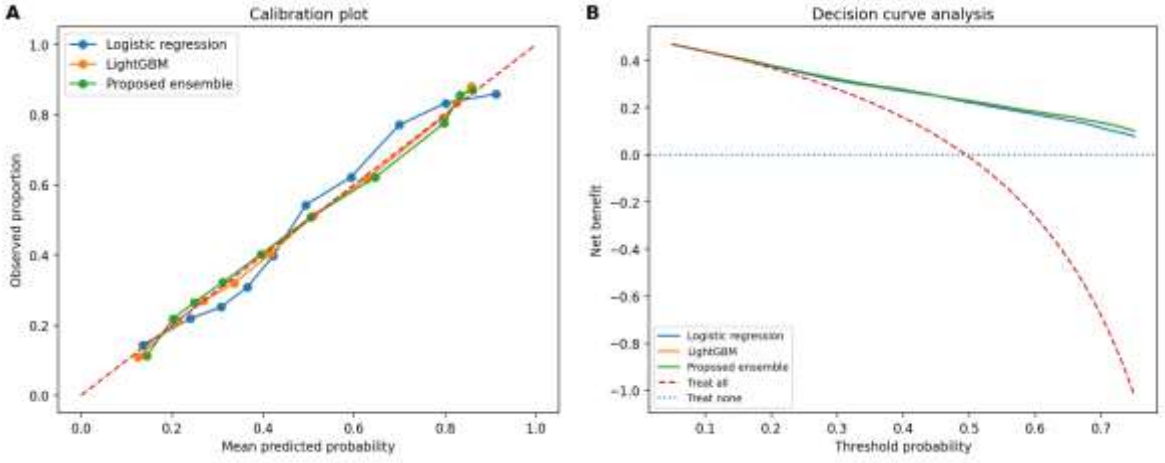

Figure 6. Probability quality and potential clinical utility. (A) Calibration by probability deciles. (B) Decision-curve analysis; higher net benefit is preferred.

3.5 Feature importance and subgroup behavior

Gain-based importance from GA-LightGBM was concentrated in systolic pressure (63.9%), followed by age (14.7%), mean arterial pressure (8.6%), and cholesterol (7.7%). Weight, activity, pulse pressure, and diastolic pressure together accounted for the remaining 5.1% (Figure 7). This ranking was clinically coherent and closely aligned with the top variables reported by Peng et al. [6]. Nevertheless, gain importance is model-specific and can distribute credit unevenly among correlated blood-pressure variables; it should not be interpreted as causal or as evidence that low-ranked variables are biologically unimportant.

Recorded gender-code subgroups had similar AUCs (0.7993 and 0.8024), although sensitivity and specificity differed. Age stratification revealed a more important gradient: AUC was 0.8246 below age 50, 0.7636 at age 50–59, and 0.7399 at age 60 or older. Among older participants, sensitivity reached 0.8626 but specificity fell to 0.4232, consistent with high outcome prevalence and upward-shifted predicted probabilities. The test-set confusion matrix for the stacked model contained 4,059 true negatives, 1,139 false positives, 1,595 false negatives, and 3,495 true positives (Figure 8). These findings caution against using one universal threshold across age groups without prospective clinical validation.

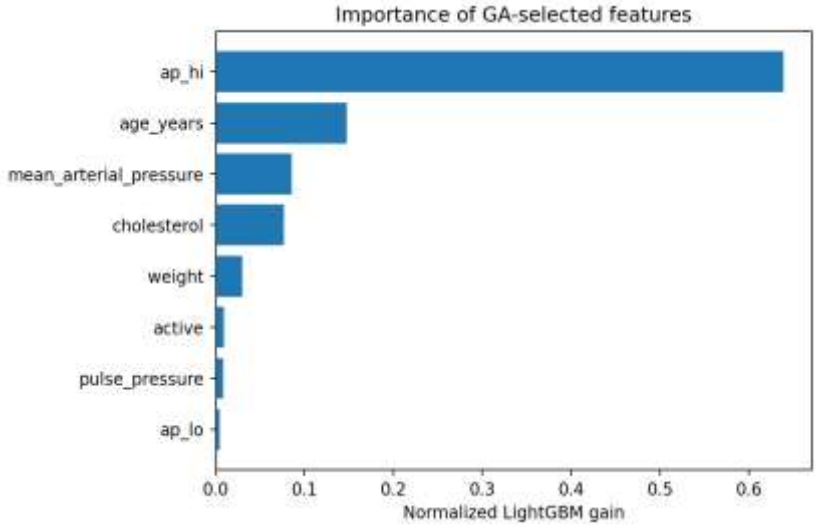

Figure 7. Normalized LightGBM gain importance within the GA consensus panel.

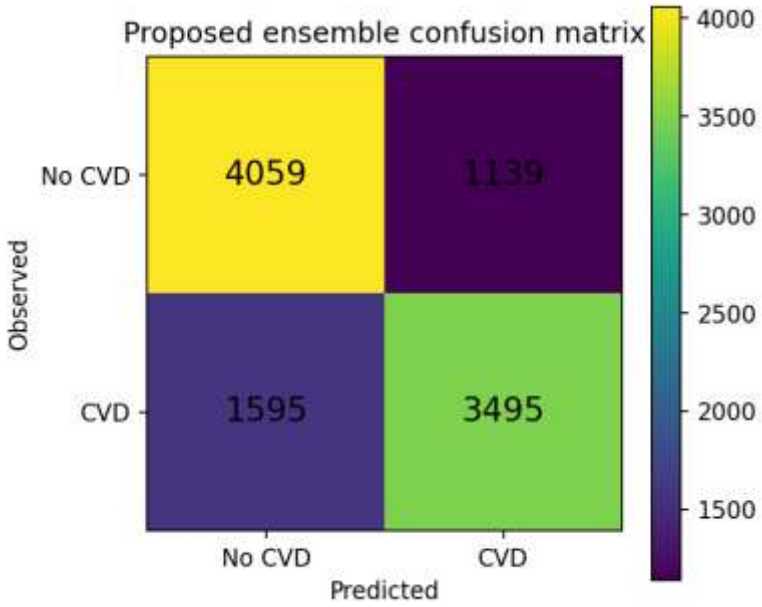

Figure 8. Held-out confusion matrix for the proposed GA-stacked ensemble at threshold 0.50.

3.6 Principal findings

This study shows that a multi-run genetic algorithm can identify a compact cardiovascular feature panel that retains nearly all of the discriminative information available in a broader engineered set. The eight-variable GA-LightGBM model reduced the candidate space by 50% and achieved AUC 0.8003 compared with 0.8027 for full-feature LightGBM. The absolute difference was only 0.0024, but the confidence interval excluded zero, so the correct interpretation is not exact equivalence. Rather, GA selection provided a favorable complexity–performance trade-off: half as many variables, modestly shorter training time, essentially unchanged accuracy, and a small loss in ranking performance.

The hybrid stack did not improve on GA-LightGBM. This negative comparative result is important because ensemble complexity is frequently presented as an innovation without testing whether it adds independent signal. The learned meta-model placed its largest weight on LightGBM, indicating that the other learners contributed little complementary information. When base learners share the same small, strongly predictive variables, their probability errors may be too correlated for stacking to improve AUC. The proposed stack remains useful as a reproducible benchmark, but the parsimonious boosted model is preferable under the present data.

3.7 Biological and clinical interpretation of the selected panel

The selected variables reflect established cardiovascular pathways. Systolic and diastolic pressure, pulse pressure, and mean arterial pressure characterize different aspects of arterial load and vascular stiffness. Age aggregates cumulative exposure and vascular aging; cholesterol captures atherogenic metabolic burden; weight approximates adiposity; and physical activity represents a modifiable behavioral factor. The concentration of LightGBM gain in systolic pressure is consistent with the global importance of elevated blood pressure and with prior analyses of this dataset [1,6]. The repeated selection of both original and derived blood-pressure features suggests that nonlinear models used distinct thresholds or interactions, although their collinearity means that importance should not be interpreted as independent mechanistic effects.

Notably, body mass index was selected in only one GA run, while raw weight appeared in four. Height contributed little once weight and other variables were present. This may reflect the narrow adult height distribution and the ability of tree-based models to combine weight with age and gender code. Glucose, smoking, and alcohol indicators were also unstable or absent. Their low selection frequency may result from coarse binary or ordinal encoding, underreporting, or mediation through other factors; it does not imply that these exposures are unimportant for cardiovascular biology. Feature selection optimizes predictive utility in a particular dataset, not biological causality.

3.8 Comparison with previous studies

Our results closely match the AUC range reported by Peng et al. [6], who obtained approximately 0.806 with a larger XGBH model and approximately 0.799 with three variables. The convergence is reassuring because both studies identify age, systolic pressure, cholesterol, and other blood-pressure measures as dominant. The present work extends that line by using repeated evolutionary selection, a strictly untouched test set, paired uncertainty for the performance cost of reduction, and probability-focused evaluation. In contrast, Alqahtani et al. [5] reported much higher accuracy on the public dataset; differences in preprocessing, validation, class thresholds, and potential leakage make direct ranking inappropriate. Table 7 places these studies alongside larger prospective risk models and emphasizes the distinction between prevalent-disease classification and future-event prediction.

Noroozi et al. [17] found that feature selection can improve some models while degrading others. Our results reproduce that heterogeneity at much larger sample size: Gaussian naive Bayes improved after selection, logistic regression was nearly unchanged, and random forest declined. This pattern supports the view that feature selection should be evaluated as a model-specific intervention. Praveen et al. [18] reported stronger performance with a more elaborate hybrid optimizer and boosted fusion, but the dataset and evaluation design differed. Without common external cohorts, point estimates should not be interpreted as a league table. Transparent partitioning, calibration, and independent validation are more informative than maximizing a single accuracy value.

Table 7. Comparison with representative cardiovascular prediction and feature-selection studies.

Study	Data	Outcome	Method	Reported performance	Interpretive note
D'Agostino et al. [2]	Framingham cohort; 8,491	10-year incident CVD	Sex-specific Cox models	Widely used general risk profile	Prospective outcome; not directly comparable
Alqahtani et al. [5]	Public CVD; 70,000	Recorded CVD label	ML/DL ensemble	Accuracy 0.887	Higher estimate; validation details differ
Peng et al. [6]	Public + hospital; 70,000	Recorded CVD label	XGBH + importance ranking	AUC 0.806; 3 features 0.799	Closest dataset-level comparator
Noroozi et al. [17]	Cleveland; 297 after QC	Heart-disease diagnosis	16 feature-selection methods	Best accuracy 0.855	Selection helped some models, hurt others
Hippisley-Cox et al. [4]	UK EHR; >16 million	10-year incident CVD	QR4 cause-specific Cox	C 0.835 women; 0.814 men	Large external validation; prospective
Praveen et al. [18]	Heart-disease data	Heart-disease diagnosis	Hybrid GOL2-2T + ABDF	AUC 0.89; accuracy 0.83	Different data and resampling workflow
Current study	Public CVD; 68,586 after QC	Recorded CVD label	Five-run GA + LightGBM/stack	AUC 0.803 full; 0.800 GA	Untouched test, CIs, calibration, DCA, subgroups

3.9 Strengths and methodological implications

The primary strength is the evaluation architecture. Feature engineering and GA search were confined to development data, stacking weights were learned on a separate validation partition, and the final test set was not consulted until all choices were fixed. Repeated GA runs exposed selection instability rather than hiding it behind one chromosome. Confidence intervals and paired differences quantified the uncertainty around small AUC changes. Calibration and decision curves addressed whether predicted probabilities could support threshold-based use, while subgroup results revealed performance variation that an overall AUC would miss. The complete code package records seeds, rules, model settings, and outputs, supporting computational reproducibility.

The findings also suggest a practical standard for evolutionary feature-selection papers. A proposed optimizer should be compared against the same learner with all features; the performance cost of reduction should be reported with paired uncertainty; selection frequencies should be shown across independent runs; and any hybrid ensemble should demonstrate incremental benefit over its best component. Under this standard, a compact model may be valuable even when its AUC is marginally lower, provided the reduction has a clear operational rationale.

3.10 Limitations and future work

Several limitations restrict clinical interpretation. First, the public dataset provides limited documentation of participant recruitment, disease definition, measurement protocols, and ethical governance. The label is cross-sectional, so the models must not be described as estimating 5- or 10-year cardiovascular risk. Second, this analysis used an internal random split rather than geographic, temporal, or institutional external validation. Random splitting can preserve unmeasured source-specific patterns and usually yields more optimistic transportability than evaluation in a new population.

Third, physiological filtering was rule-based. Although the limits were broad and removed only 2.02% of records, some valid extreme measurements may have been excluded, and repeated measurements from the same person cannot be identified. Fourth, several predictors have coarse or ambiguous coding. The meaning of gender codes was not documented in the supplied file, and sex-specific biological conclusions are therefore inappropriate. Smoking, alcohol, activity, cholesterol, and glucose are simplified categories that cannot capture dose, duration, medication, or laboratory values. Race or ethnicity, socioeconomic status, comorbidities, and treatment information were unavailable.

Fifth, the GA used logistic regression as a surrogate fitness model and a fixed parsimony penalty. A fully nested multi-objective optimization could model a Pareto frontier between AUC, feature count, calibration, and measurement cost, but would require substantially more computation. Sixth, gain importance is not causal and may be biased among correlated variables. Seventh, measured training times depend on hardware and software conditions. Finally, the dataset contains phenotypic and clinical variables rather than genomic, transcriptomic, proteomic, or metabolomic measurements. The current work therefore evaluates cardiovascular phenotype classification, not molecular mechanisms. Future studies should integrate molecular biomarkers, perform repeated nested validation, recalibrate in external cohorts, assess threshold-specific harms, and compare GA consensus with simpler embedded selection and clinically prespecified panels.

4. CONCLUSION

A stability-guided genetic algorithm reduced 16 engineered cardiovascular predictors to an eight-variable panel centered on age, blood pressure, cholesterol, weight, and physical activity. GA-LightGBM preserved nearly all of the full model's discrimination while halving the feature set, but the small AUC decrease was statistically detectable. A stacked hybrid model did not improve on GA-LightGBM, demonstrating that additional architectural complexity is not automatically beneficial. The main contribution is therefore a transparent, reproducible framework for quantifying the trade-off between parsimony and predictive performance. External validation—and, for molecular-bioscience applications, integration of molecular biomarkers—is required before clinical or mechanistic claims are justified.

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

Mohammad Subhi Al-Batah: Conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing-original draft, and writing-review and editing.

Abdulkarim Alanazi: Dataset preparation, Orange workflow execution, result extraction, and preliminary draft preparation. Both authors approved the final manuscript and accept responsibility for the integrity of the work.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study was a secondary analysis of a publicly available, de-identified dataset and involved no direct interaction with human participants. No new informed consent was obtained. The source files supplied for this analysis did not document the original recruitment ethics, consent procedure, or institutional review board determination.

DATA AVAILABILITY STATEMENT

The source Cardiovascular Disease Dataset is publicly available from Kaggle: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. The analysis used the supplied cardio_train.csv file. Cleaned derivative data are not redistributed; representative rows and aggregate outputs are reported in the manuscript. The complete Python analysis, environment requirements, parameter settings, generated result tables, and final figures are provided in Supplementary File S1.

REFERENCES

[1] World Health Organization, 2025. Cardiovascular diseases (CVDs). Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed 21 July 2026).

[2] D'Agostino, R.B. Sr., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., et al., 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 117, 743-753. doi: 10.1161/CIRCULATIONAHA.107.699579.

[3] SCORE2 Working Group and ESC Cardiovascular Risk Collaboration, 2021. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal 42, 2439-2454. doi: 10.1093/eurheartj/ehab309.

[4] Hippisley-Cox, J., Coupland, C.A.C., Bafadhel, M., et al., 2024. Development and validation of a new algorithm for improved cardiovascular risk prediction. Nature Medicine 30, 1440-1447. doi: 10.1038/s41591-024-02905-y.

[5] Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L., Javed, T., 2022. Cardiovascular disease detection using ensemble learning. Computational Intelligence and Neuroscience 2022, 5267498. doi: 10.1155/2022/5267498.

[6] Peng, M., Hou, F., Cheng, Z., Shen, T., Liu, K., Zhao, C., et al., 2023. Prediction of cardiovascular disease risk based on major contributing features. Scientific Reports 13, 4778. doi: 10.1038/s41598-023-31870-8.

[7] Al-Batah, M.S., Alourani, A., 2026. Machine learning-based prediction of dataset-defined myocardial infarction risk: a retrospective computational study for precision cardiovascular risk assessment. Digital Health 12, 20552076261464747. doi: 10.1177/20552076261464747.

[8] Alzboon, M.S., Al-Batah, M.S., 2026. A clinically accessible heart disease prediction framework: multi-model evaluation with ensemble learning and low-code deployment. Discover Applied Sciences 8, 339. doi: 10.1007/s42452-026-08353-2.

[9] Al-Batah, M.S., Alzboon, M.S., Alazaidah, R., 2023. Intelligent Heart Disease Prediction System with Applications in Jordanian Hospitals. International Journal of Advanced Computer Science and Applications 14(9), 508-517. doi: 10.14569/IJACSA.2023.0140954.

[10] Al-Batah, M.S., 2019. Automatic Diagnosis System for Heart Disorder using ESG Peak Recognition with Ranked Features Selection. International Journal of Circuits, Systems and Signal Processing 13, 391-398.

[11] Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157-1182.

[12] Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273-324. doi: 10.1016/S0004-3702(97)00043-X.

[13] Xue, B., Zhang, M., Browne, W.N., Yao, X., 2016. A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation 20, 606-626. doi: 10.1109/TEVC.2015.2504420.
[14] Siedlecki, W., Sklansky, J., 1989. A note on genetic algorithms for large-scale feature selection. Pattern Recognition Letters 10, 335-347. doi: 10.1016/0167-8655(89)90037-8.
[15] Meinshausen, N., Buhlmann, P., 2010. Stability selection. Journal of the Royal Statistical Society: Series B 72, 417-473. doi: 10.1111/j.1467-9868.2010.00740.x.
[16] Al-Batah, M.S., Alabbas, M.S., 2026. Stability-Aware Genetic Algorithm Feature Selection and a Calibrated Hybrid Ensemble for Imbalanced Smart-Meter Loss Detection. Journal of Intelligent Decision Making and Information Science 3(2), 833-854. doi: 10.59543/jidmis.v3.1736.
[17] Noroozi, Z., Orooji, A., Erfannia, L., 2023. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Scientific Reports 13, 22588. doi: 10.1038/s41598-023-49962-w.
[18] Praveen, S.P., Hasan, M.K., Abdullah, S.N.H.S., Sirisha, U., Tirumanadham, N.S.K.M.K., Islam, S., et al., 2024. Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2T and adaptive boosted decision fusion with babysitting refinement. Frontiers in Medicine 11, 1407376. doi: 10.3389/fmed.2024.1407376.
[19] Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., et al., 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385, e078378. doi: 10.1136/bmj-2023-078378.
[20] Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., et al., 2019. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine 170, 51-58. doi: 10.7326/M18-1376.
[21] Breiman, L., 2001. Random forests. Machine Learning 45, 5-32. doi: 10.1023/A:1010933404324.
[22] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al., 2017. LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30, 3146-3154.
[23] Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5, 241-259. doi: 10.1016/S0893-6080(05)80023-1.
[24] Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1-3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
[25] Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., 2019. Calibration: the Achilles heel of predictive analytics. BMC Medicine 17, 230. doi: 10.1186/s12916-019-1466-7.
[26] Vickers, A.J., Elkin, E.B., 2006. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making 26, 565-574. doi: 10.1177/0272989X06295361.
[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825-2830.