ILKOM Jurnal Ilmiah Vol. 18, No. 2, August 2026, pp. 420-429
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Optimization of Naive Bayes Algorithm Using Genetic Algorithm for Heart Failure Prediction

Teguh Cahyono [a,1,*]; Yogiek Indra Kurniawan [a,2]; Bangun Wijayanto [a,3];
[a] *Universitas Jenderal Soedirman, Jl. Profesor DR. HR Boenyamin No.708, Jawa Tengah 53122, Indonesia*
[1] *teguh.cahyono@unsoed.ac.id;* [2] *yogiek@unsoed.ac.id;* [3] *bangun.wijayanto@unsoed.ac.id;*
** Corresponding author*

Article history: Received June 09, 2025; Revised June 12, 2026; Accepted July 24, 2026; Available online August 10, 2026.

Abstract

Feature selection is often reported to improve clinical prediction, yet optimization and evaluation on the same data can produce optimistic estimates. This study provides a leakage-controlled and reproducible reassessment of genetic algorithm (GA) feature selection for Gaussian Naive Bayes prediction of heart disease. The open-Heart Failure Prediction dataset contains 918 observations, 11 predictors, and a binary target. Clinically implausible zero values in resting blood pressure and cholesterol were recoded as missing; imputation, scaling, one-hot encoding, GA selection, and model fitting were confined to training data. Performance was estimated using repeated nested stratified cross-validation with five outer folds repeated five times and four inner folds for GA fitness. The GA used an 11-bit chromosome, population size 12, 10 generations, 0.80 uniform-crossover probability, 0.08 bit-flip mutation probability, tournament selection of size three, and two elites. Across 25 held-out outer folds, baseline Naive Bayes achieved 84.66% accuracy (SD 2.01%), whereas GA-selected Naive Bayes achieved 84.31% (SD 2.40%). The paired mean difference was −0.35 percentage points, with a bootstrap 95% confidence interval of −0.72 to 0.05 percentage points and a Wilcoxon p-value of 0.120. AUC values were 0.912 and 0.907, respectively. The GA retained 8.44 of 11 predictors on average; ExerciseAngina, FastingBS, ST_Slope, and Sex were selected in every outer fold. These results do not support a material accuracy gain from GA selection, but demonstrate modest dimensionality reduction and reveal predictors that are stable under resampling. The study emphasizes nested evaluation, uncertainty reporting, and full parameter disclosure for credible optimization claims.

Keywords: Heart Disease Prediction; Naive Bayes; Genetic Algorithm; Feature Selection; Nested Cross-Validation; Reproducibility.

Introduction

Cardiovascular disease remains a major clinical and public-health burden, motivating data-driven tools that can support risk stratification and earlier investigation. The dataset used here consolidates commonly measured demographic, symptomatic, electrocardiographic, exercise, and laboratory variables derived from established coronary-disease collections. Related work by Huzain Azis and colleagues at Universitas Muslim Indonesia evaluated K-NN with cross-validation on cardiovascular patient data [1], illustrating sustained regional interest in reproducible heart-disease classification.

Recent heart-disease prediction research has evaluated probabilistic classifiers, decision trees, support-vector machines, ensembles, neural architectures, and hybrid optimization pipelines [2]–[19]. Reported accuracies are often high, but direct comparison is difficult because studies differ in split strategy, preprocessing, feature selection, hyperparameter search, and reporting completeness. In particular, selecting variables before cross-validation or optimizing and evaluating on the same folds leaks outcome information and can overstate generalization.

Genetic algorithms are attractive for feature selection because they search non-differentiable combinations without exhaustively evaluating all subsets. Metaheuristic and wrapper-selection studies report that evolutionary search can remove redundant predictors and sometimes improve classification [20]–[30]. However, a gain observed inside the optimization sample is not equivalent to a gain on unseen data. Reproducible clinical artificial intelligence therefore requires transparent data handling, complete algorithm settings, uncertainty estimates, and leakage-controlled validation [31]–[41].

http://dx.doi.org/10.33096/ilkom.v18i2.3559.420-429

This work reconstructs and strengthens an earlier GA–Naive Bayes experiment in response to methodological review. The contributions are: (i) an explicit, executable preprocessing and optimization pipeline; (ii) repeated nested cross-validation that separates feature search from held-out evaluation; (iii) reporting of accuracy, precision, recall, specificity, F1-score, ROC AUC, confidence intervals, and paired tests; and (iv) stability analysis of selected predictors. The central research question is whether GA feature selection yields a reproducible out-of-sample improvement over Naive Bayes using all available predictors.

Method

A. Dataset and Data Quality

The Heart Failure Prediction dataset was downloaded from Kaggle and stored locally to support repeatable execution. It contains 918 records, 11 predictors, and the binary target HeartDisease. Five predictors are numeric (Age, RestingBP, Cholesterol, MaxHR, and Oldpeak), while six are categorical or binary (Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, and ST_Slope). The positive class comprises 508 records (55.34%), and **Table 1** summarizes the dataset and data-quality checks.

No conventional null values or duplicate rows were present. Nonetheless, one RestingBP value and 172 Cholesterol values were zero, which are physiologically implausible in this context. They were recoded as missing and imputed only from each training partition. This distinction between syntactic completeness and clinical plausibility is important because treating structural zeros as measurements can distort a Gaussian likelihood.

Table 1. Dataset and Quality Summary

Item	Value
Observations	918
Predictors / target	11 / 1
HeartDisease = 1	508 (55.34%)
Duplicate rows	0
RestingBP zeros recoded	1
Cholesterol zeros recoded	172

The distributions in **Figure 1** provide a descriptive view before model fitting. Heart-disease records are shifted toward older ages, lower maximum heart rates, and higher Oldpeak values. Cholesterol shows substantial overlap between classes, consistent with its lower and less stable selection frequency in the GA analysis. These patterns are descriptive associations and should not be interpreted as causal effects.

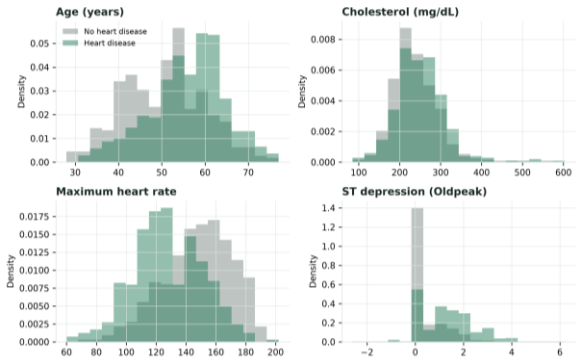

Figure 1. Distributions of four numeric predictors stratified by the HeartDisease target

B. Pre-processing Pipeline

All transformations were fitted within the relevant training fold. Numeric missing values were replaced by the training median and standardized to zero mean and unit variance. Categorical missing values were replaced by the training mode and encoded using one-hot indicators with unknown categories ignored. Feature selection operated on the 11 original clinical variables; when a categorical variable was selected, all one-hot columns generated from that variable entered the classifier. This design preserves interpretability at the source-variable level; the leakage-controlled sequence is shown in **Figure 2**.

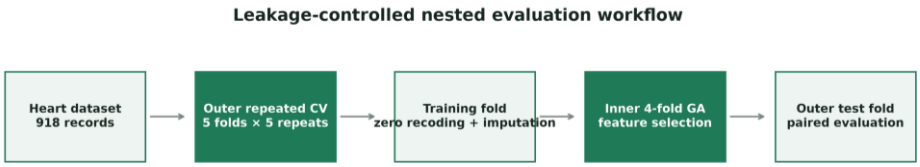

Figure 2. Leakage-controlled repeated nested evaluation workflow

C. Gaussian Naïve Bayes

Gaussian Naive Bayes estimates the posterior class probability by multiplying class-conditional feature likelihoods under conditional independence. For class c and transformed feature vector x, Equation (1) gives the decision quantity, where P(c) is the prior and p(xj|c) is modeled by a Gaussian density for continuous encoded inputs.

$$P(c \mid x) \propto P(c) \prod_j p(x_j \mid c) \quad (1)$$

The class with the largest posterior score was predicted. The variance-smoothing parameter was fixed at 10^{-9} for both the baseline and GA-selected models, ensuring that the comparison isolated feature selection rather than classifier tuning.

D. Genetic Algorithm Feature Selection

Each candidate subset was encoded as an 11-bit chromosome. A bit value of one retained the corresponding original predictor and zero removed it; empty chromosomes were repaired by activating one randomly chosen predictor. The initial population was sampled with a fixed random seed. Candidate fitness was the mean inner-fold validation accuracy defined in Equation (2), calculated after fitting the complete preprocessing and Naive Bayes pipeline on each inner-training fold.

$$Fitness(s) = (1/K) \Sigma_k Accuracy_k(s), \; K = 4 \quad (2)$$

Parents were chosen by tournament selection. Uniform crossover independently exchanged genes between two parents with probability 0.80, and each offspring bit mutated with probability 0.08. The two highest-fitness chromosomes survived unchanged through elitism. The search terminated after 10 generations. The population size and generation budget were deliberately modest because the search space contains only 2^11 possible subsets; crossover and mutation provide exploration while tournament selection and elitism retain strong candidates. No feature-count penalty was used, so a smaller subset emerged only when it preserved or improved validation accuracy. Duplicate subsets were cached to avoid repeated evaluation. **Table 2** lists all fixed GA settings.

Table 2. Genetic Algorithm Parameters

Parameter	Setting
Chromosome	11 binary genes
Population	12
Generations	10
Selection	Tournament, k = 3
Crossover	Uniform, p = 0.80
Mutation	Bit-flip, p = 0.08/gene
Elitism	2 chromosomes
Fitness	4-fold CV accuracy
Feature penalty	0
Base seed	20260714

E. Validation and Statistical Analysis

Generalization was estimated by five-fold stratified outer cross-validation repeated five times, producing 25 held-out evaluations. Within every outer-training partition, a separate four-fold stratified cross-validation optimized the GA. The baseline used all 11 predictors, whereas the optimized model used only the subset returned by the inner search. Thus, no outer-test label influenced imputation, encoding, scaling, selection, or fitting. Nested validation follows established guidance for avoiding selection bias [40], [41].

Reported metrics were accuracy, positive predictive value (precision), sensitivity (recall), specificity, F1-score, and ROC AUC. A recent evaluation by Azis used five-fold cross-validation and the same core classification metrics for logistic-regression heart-disease detection [42]. Means, sample standard deviations, and normal-approximation 95% confidence intervals were computed across outer folds. The paired accuracy difference was additionally assessed with a 20,000-resample bootstrap confidence interval and a two-sided Wilcoxon signed-rank test. Exact McNemar tests compared paired predictions within each complete repeat; metric and test selection followed recent machine-learning evaluation guidance [36]. A two-sided $p < 0.05$ was considered statistically significant.

Results and Discussion

A. Predictive Performance

Baseline Naive Bayes achieved a mean accuracy of 84.66%, compared with 84.31% after GA selection. The paired difference was −0.35 percentage points (GA minus baseline), and its bootstrap 95% confidence interval ranged from −0.72 to 0.05 percentage points. The interval includes zero and the Wilcoxon test was not significant ($p = 0.120$). Consequently, the experiment provides no evidence that GA selection materially improves out-of-sample accuracy. **Table 3** reports the complete metric comparison, and **Table 4** reports the paired statistical and search summary.

Table 3. Complete Outer-Fold Metric Comparison

Model	All Features	GA Subset	Δ
Accuracy	0.8466 ± 0.0201	0.8431 ± 0.0240	-0.0035
Precision	0.8692 ± 0.0244	0.8656 ± 0.0286	-0.0036
Recall	0.8519 ± 0.0337	0.8495 ± 0.0349	-0.0024
Specificity	0.8400 ± 0.0359	0.8351 ± 0.0424	-0.0049
F1-score	0.8599 ± 0.0195	0.8569 ± 0.0224	-0.0030
ROC AUC	0.9115 ± 0.0254	0.9069 ± 0.0278	-0.0046

Table 4. Paired Comparison And Search Summary

Statistic	Results
Accuracy Δ (GA − all)	-0.0035 (-0.35 pp)
Bootstrap 95% CI	-0.0072 to +0.0005
Wilcoxon signed-rank	p = 0.120
Exact McNemar range	p = 0.078–1.000
Selected predictors	8.44 ± 1.36 of 11
Unique subsets evaluated	Mean 59.72 per fold

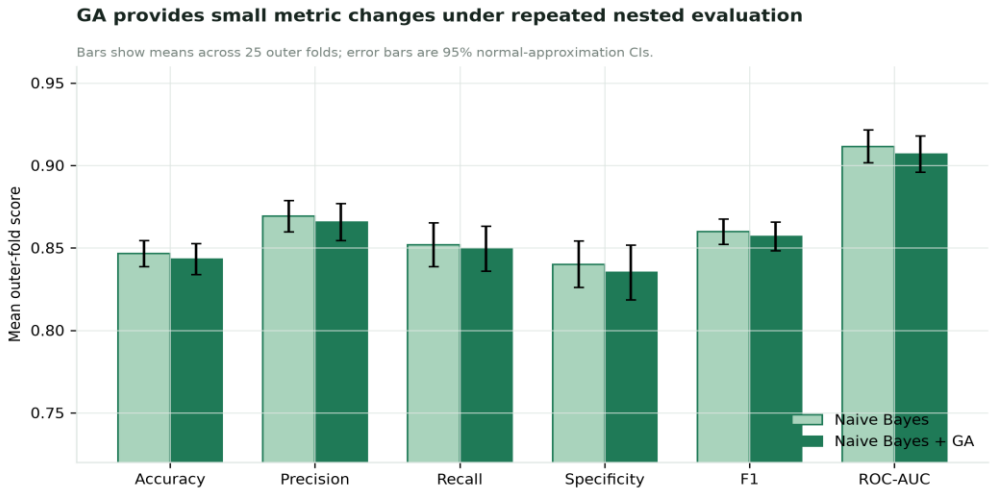

Figure 3. Mean held-out performance across 25 outer folds; error bars show 95% confidence intervals

As shown in **Figure 3**, the same direction was observed for F1-score and AUC, although differences were small. Repeat-level exact McNemar p-values ranged from 0.078 to 1.000, and none crossed the significance threshold. These

findings illustrate why a single train–test split or the best optimization fold can be misleading: a seemingly favorable subset may reflect sampling variation rather than a stable improvement.

The aggregated confusion matrices in **Figure 4** contain 4,590 out-of-fold prediction instances from five complete repeats. The all-feature model produced 1,722 true negatives, 328 false positives, 376 false negatives, and 2,164 true positives. The GA-selected model produced 1,712, 338, 382, and 2,158, respectively. Relative to the baseline, GA selection therefore added 10 false positives and six false negatives, matching the small negative accuracy difference.

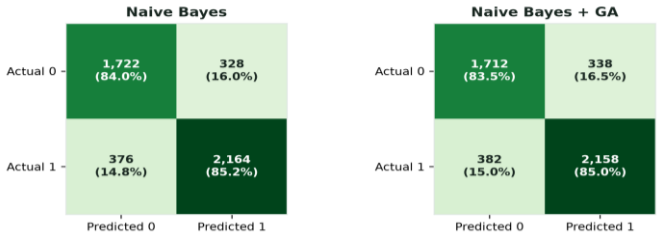

Figure 4. Aggregated held-out confusion matrices; cell percentages are normalized by actual class

B. Feature Reduction and Stability

Table 5 and **Figure 5** show that the GA selected 8.44 ± 1.36 of 11 variables on average and evaluated 59.72 unique subsets per outer fold. ExerciseAngina, FastingBS, ST_Slope, and Sex were retained in all 25 searches. ChestPainType and Oldpeak were selected in 88% of folds, whereas Cholesterol was selected in only 40%. Stable inclusion suggests consistent incremental value for the classifier; lower rates indicate redundancy, weak conditional contribution, or sensitivity to the training sample.

Table 5. Feature-Selection Stability

Predictor	Folds	Rate
ExerciseAngina	25/25	100%
FastingBS	25/25	100%
ST_Slope	25/25	100%
Sex	25/25	100%
ChestPainType	22/25	88%
Oldpeak	22/25	88%
Age	19/25	76%
MaxHR	13/25	52%
RestingECG	13/25	52%
RestingBP	12/25	48%
Cholesterol	10/25	40%

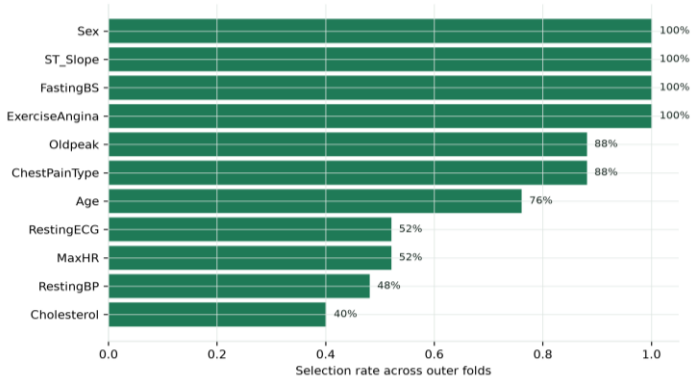

Figure 5. Predictor selection frequency across 25 independent outer-fold searches

Feature stability is more informative than presenting one ‘optimal’ chromosome. The variation across folds shows that several alternative subsets provide similar inner-CV accuracy. This is expected in correlated clinical data and

cautions against interpreting GA inclusion as causal importance. The GA nevertheless reduced the source-variable count by approximately 23% without a large loss of discrimination, which may be useful where acquisition cost or questionnaire burden matters.

C. Search Behavior and Discrimination

Figure 6 shows that mean best fitness improved rapidly during the first generations and then plateaued, indicating that the chosen population and generation budget generally reached a stable region of the small 2^{11}-subset search space. The fixed termination rule makes runtime predictable, but does not prove global optimality. Because fitness contained no sparsity penalty, the algorithm appropriately prioritized predictive validation performance rather than minimizing variables at any cost.

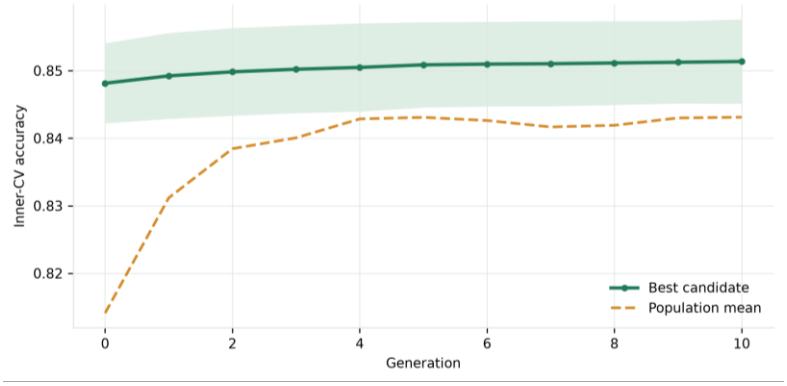

Figure 6. Mean inner-validation fitness trajectory across outer-fold GA runs

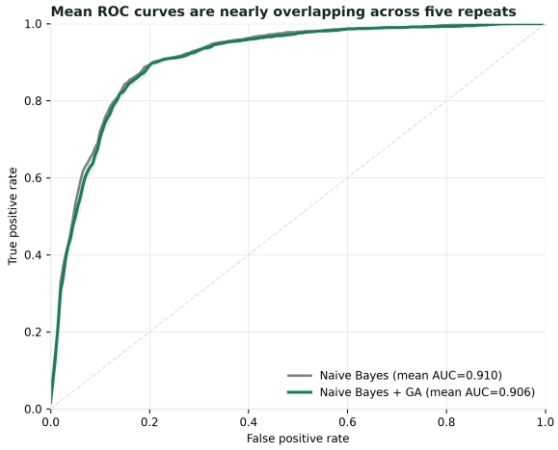

Figure 7. Aggregated held-out ROC curves for baseline and GA-selected Naive Bayes

The aggregated held-out ROC curves in **Figure 7** were close, consistent with mean AUC values of 0.912 for the baseline and 0.907 for the GA-selected model. The GA therefore did not reveal a clinically meaningful separation advantage at alternative thresholds. Any deployment decision would still require threshold calibration to explicit clinical costs and prospective validation.

D. Comparison with Prior Work and Implications

Many recent heart-disease studies report improvements after feature selection or hybrid optimization [2]–[30]. Azis, Abdullah, and Ismail also showed that weighting and combination rules can materially affect heterogeneous ensemble performance across five health datasets, including heart disease and heart failure [43]. The present result is deliberately more conservative because every subset decision is made inside the outer-training data and uncertainty is estimated across repeated held-out folds. It does not imply that evolutionary selection is ineffective in general; rather, for this dataset, classifier, and parameterization, the evidence supports dimensionality reduction but not an accuracy improvement.

This distinction addresses a central reproducibility concern in medical machine learning. Transparent reporting frameworks emphasize the intended use, data provenance, preprocessing, validation boundaries, uncertainty, and limitations [31]–[39]. Incorrectly performing feature selection before validation can generate substantial optimism [40], [41]. By reporting the random seed, complete GA settings, fold-level results, and an executable workflow, this reconstruction converts an optimization claim into a testable result.

The study has limitations. The dataset is retrospective, relatively small, and assembled from multiple source cohorts; institution-specific sampling and coding may limit transportability. Class labels are treated as ground truth without adjudication data. Gaussian Naive Bayes assumes conditional independence and simple likelihood forms. The GA budget is intentionally modest and only one classifier and fitness objective were tested. Finally, repeated cross-validation quantifies internal variability but cannot replace temporal or external validation. Future work should preregister clinically meaningful margins, evaluate calibration and decision utility, test external cohorts, and compare GA against simpler filters and embedded regularization.

Conclusion

A reproducible, leakage-controlled reassessment found that GA feature selection did not significantly improve Gaussian Naive Bayes heart-disease prediction. Mean accuracy changed from 84.66% to 84.31%, with a paired 95% confidence interval compatible with a small loss or negligible gain. The GA retained 8.44 of 11 predictors on average and consistently selected ExerciseAngina, FastingBS, ST_Slope, and Sex. The primary value of the optimization was therefore modest variable reduction and stability information, not superior predictive performance. Nested validation and uncertainty reporting should accompany future claims that metaheuristic feature selection improves clinical prediction.

Acknowledgement

This work was supported by Universitas Jenderal Soedirman through the Riset Dasar Scheme, contract no. 1218/UN23/PT.01.02/2023. The authors thank the reviewers whose methodological comments motivated the reproducible reconstruction and expanded validation.

References

[1] H. Azis, P. Purnawansyah, F. Fattah, I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 81-86, 2020. doi: 10.33096/ilkom.v12i2.507.81-86

[2] M. A. Islam, M. Z. H. Majumder, M. S. Miah, S. Jannaty, "Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction," *Computers in Biology and Medicine*, vol. 176, pp. 108432, 2024. doi: 10.1016/j.compbiomed.2024.108432

[3] Z. Noroozi, A. Orooji, L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Scientific Reports*, vol. 13, no. 1, pp. 22588, 2023. doi: 10.1038/s41598-023-49962-w

[4] C. M. Bhatt, P. Patel, T. Ghetia, P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, pp. 88, 2023. doi: 10.3390/a16020088

[5] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi, et al., "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151-80173, 2022. doi: 10.1109/access.2022.3165792

[6] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, et al., "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2, pp. 144, 2024. doi: 10.3390/diagnostics14020144

[7] N. Chandrasekhar, S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, pp. 1210, 2023. doi: 10.3390/pr11041210

[8] A. Saboor, M. Usman, S. Ali, A. Samad, et al., "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems*, vol. 2022, pp. 1-9, 2022. doi: 10.1155/2022/1410169

[9] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," *Sensors*, vol. 22, no. 3, pp. 1184, 2022. doi: 10.3390/s22031184

[10] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, et al., "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, pp. 104672, 2021. doi: 10.1016/j.compbiomed.2021.104672

[11] D. Bertsimas, L. Mingardi, B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3627-3637, 2021. doi: 10.1109/jbhi.2021.3066347

[12] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, et al., "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," *Applied Sciences*, vol. 11, no. 18, pp. 8352, 2021. doi: 10.3390/app11188352

[13] R. Katarya, S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," *Health and Technology*, vol. 11, no. 1, pp. 87-97, 2021. doi: 10.1007/s12553-020-00505-7

[14] D. Shah, S. Patel, S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Computer Science*, vol. 1, no. 6, pp. 345, 2020. doi: 10.1007/s42979-020-00365-y

[15] A. K. Gárate-Escamila, A. Hajjam El Hassani, E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, pp. 100330, 2020. doi: 10.1016/j.imu.2020.100330

[16] P. Ghosh, S. Azam, M. Jonkman, A. Karim, et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," *IEEE Access*, vol. 9, pp. 19304-19326, 2021. doi: 10.1109/access.2021.3053759

[17] J. P. Li, A. U. Haq, S. U. Din, J. Khan, et al., "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562-107582, 2020. doi: 10.1109/access.2020.3001149

[18] F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, et al., "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Information Fusion*, vol. 63, pp. 208-222, 2020. doi: 10.1016/j.inffus.2020.06.008

[19] W. DeGroat, H. Abdelhalim, K. Patel, D. Mendhe, et al., "Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine," *Scientific Reports*, vol. 14, no. 1, pp. 1, 2024. doi: 10.1038/s41598-023-50600-8

[20] P. Agrawal, H. F. Abutarboush, T. Ganesh, A. W. Mohamed, "Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019)," *IEEE Access*, vol. 9, pp. 26766-26791, 2021. doi: 10.1109/access.2021.3056407

[21] M. Rostami, K. Berahmand, S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection," *Journal of Big Data*, vol. 8, no. 1, pp. 2, 2021. doi: 10.1186/s40537-020-00398-3

[22] E. Ileberi, Y. Sun, Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *Journal of Big Data*, vol. 9, no. 1, pp. 24, 2022. doi: 10.1186/s40537-022-00573-8

[23] K. Rajwar, K. Deep, S. Das, "An exhaustive review of the metaheuristic algorithms for search and optimization: taxonomy, applications, and open challenges," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13187-13257, 2023. doi: 10.1007/s10462-023-10470-y

[24] B. F. Azevedo, A. M. A. C. Rocha, A. I. Pereira, "Hybrid approaches to optimization and machine learning methods: a systematic literature review," *Machine Learning*, vol. 113, no. 7, pp. 4055-4097, 2024. doi: 10.1007/s10994-023-06467-x

[25] Z. M. Elgamal, N. B. M. Yasin, M. Tubishat, M. Alswaitti, et al., "An Improved Harris Hawks Optimization Algorithm With Simulated Annealing for Feature Selection in the Medical Field," *IEEE Access*, vol. 8, pp. 186638-186652, 2020. doi: 10.1109/access.2020.3029728

[26] D. S. A. Elminaam, A. Nabil, S. A. Ibraheem, E. H. Houssein, "An Efficient Marine Predators Algorithm for Feature Selection," *IEEE Access*, vol. 9, pp. 60136-60153, 2021. doi: 10.1109/access.2021.3073261

[27] M. Tubishat, M. Alswaitti, S. Mirjalili, M. A. Al-Garadi, et al., "Dynamic Butterfly Optimization Algorithm for Feature Selection," *IEEE Access*, vol. 8, pp. 194303-194314, 2020. doi: 10.1109/access.2020.3033757

[28] T. Bhattacharyya, B. Chatterjee, P. K. Singh, J. H. Yoon, et al., "Mayfly in Harmony: A New Hybrid Meta-Heuristic Feature Selection Algorithm," *IEEE Access*, vol. 8, pp. 195929-195945, 2020. doi: 10.1109/access.2020.3031718

[29] H. Elmannai, N. El-Rashidy, I. Mashal, M. A. Alohali, et al., "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," *Diagnostics*, vol. 13, no. 8, pp. 1506, 2023. doi: 10.3390/diagnostics13081506

[30] A. A. Alhussan, A. A. Abdelhamid, E. S. M. El-Kenawy, A. Ibrahim, et al., "A Binary Waterwheel Plant Optimization Algorithm for Feature Selection," *IEEE Access*, vol. 11, pp. 94227-94251, 2023. doi: 10.1109/access.2023.3312022

[31] J. Amann, A. Blasimme, E. Vayena, et al., "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 310, 2020. doi: 10.1186/s12911-020-01332-6

[32] E. Tjoa, C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, 2021. doi: 10.1109/tnnls.2020.3027314

[33] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199-217, 2021. doi: 10.1038/s42256-021-00307-0

[34] F. Cabitza, A. Campagner, F. Soares, L. García de Guadiana-Romualdo, et al., "The importance of being external. methodological insights for the external validation of machine learning models in medicine," *Computer Methods and Programs in Biomedicine*, vol. 208, pp. 106288, 2021. doi: 10.1016/j.cmpb.2021.106288

[35] A. Bailly, C. Blanc, É. Francis, T. Guillotin, et al., "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Computer Methods and Programs in Biomedicine*, vol. 213, pp. 106504, 2022. doi: 10.1016/j.cmpb.2021.106504

[36] O. Rainio, J. Teuho, R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, pp. 6086, 2024. doi: 10.1038/s41598-024-56706-x

[37] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, pp. e078378, 2024. doi: 10.1136/bmj-2023-078378

[38] K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, et al., "FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, vol. 388, pp. e081554, 2025. doi: 10.1136/bmj-2024-081554

[39] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, et al., "Data leakage inflates prediction performance in connectome-based machine learning models," *Nature Communications*, vol. 15, no. 1, pp. 1829, 2024. doi: 10.1038/s41467-024-46150-w

[40] D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of Cheminformatics*, vol. 6, no. 1, pp. 10, 2014. doi: 10.1186/1758-2946-6-10

[41] A. Demircioğlu, "Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics," *Insights into Imaging*, vol. 12, no. 1, pp. 172, 2021. doi: 10.1186/s13244-021-01115-1

[42] H. Azis, "Assessing the Performance of Logistic Regression in Heart Disease Detection through 5-Fold Cross-Validation," *International Journal of Artificial Intelligence in Medical Issues*, vol. 2, no. 1, pp. 1-11, 2024. doi: 10.56705/ijaimi.v2i1.137

[43] H. Azis, M. Abdullah, S. Ismail, "Performance Comparison of Various Weight and Combination Function on Heterogeneous Parallel Ensemble Machine Learning Model," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 56, no. 4, pp. 342-353, 2025. doi: 10.37934/araset.56.4.342353