International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

Evolutionary Dimensionality Reduction for Structured Heart-Disease Classification: Balancing Predictive Performance, Clinical Input Burden and Global Transparency

Research Scholar Rakesh Kumar Khillan, Associate Professor Dr. Abhinav Shukla
Department of IT & CS, Dr. C. V. Raman University, Bilaspur, India

Abstract — Background: In clinical machine learning, the task of feature selection is frequently stated as a step toward increased accuracy, but a smaller model can be just as useful as it can help to ease the burden of input and give a better global picture of the model. This study compared the performance-compactness trade-off between a full feature Random Forest and a Genetic Algorithm (GA) selected Random Forest in terms of their performance in binary classification of the recorded heart disease status. Data: A public structured dataset with 918 instances, 11 features and a binary target HeartDisease was used. The full-featured Random Forest employed all of the predictors. The binary chromosomes, population size of 20, number of generations of 10, tournament selection, two-point crossover, bit-flip mutation and fitness function of 20-fold Random Forest accuracy are used in a wrapper GA. A subset of 7 predictors was selected and compared to the full-feature model via 10 replications of 20-fold stratified cross-validation. Accuracy, precision, sensitivity, F1-score, ROC-AUC, predictor count and cross-validated permutation importance were measured. Results: The best repeated internal accuracy (87.11% ± 5.06%) and ROC-AUC (0.9285 ± 0.0396) was obtained by the full-feature Random Forest method. The GA-selected model reduced the predictor set from 11 to 7 (36.36%) and achieved accuracy of 83.67% ± 5.45% and ROC-AUC of 0.9075 ± 0.0439. The mean difference of accuracy between the two models in the paired accuracy was −3.44 percentage points in favor of the full-feature model. The largest mean decreases in validation ROC-AUC following permutation was from ST_Slope, followed by ChestPainType and Oldpeak. Conclusions: The evidence was not sufficient to support the assumption which led to the improvement of predictive accuracy through evolutionary features selection. On the contrary, GA has come up with a small, clinically identifiable prototype that had less intraclass discrimination. Thus, the full-featured versus the compact configuration are used in different ways: to maximize predictive performance versus to minimize both user input and global predictor transparency. Prior to clinical-use claims, the features should be fully nested and be externally validated.

Keywords— heart-disease classification; feature selection; genetic algorithm; random forest; model compactness; permutation importance; repeated cross-validation

I. INTRODUCTION

Cardiovascular disease still accounts for a high portion of deaths and disabilities worldwide and further research is still underway for computational techniques that can integrate demographic, physiological, electrocardiographic, symptom and exercise-response data [1]. In the context of this type of application, machine-learning models can be beneficial for capturing nonlinear relationships and interactions that may be hard to express with a few manually defined rules [18,19]. However, a clinical classification model that has an excellent head-line accuracy result is not the end-all, be-all of the model's value. Many factors impact the credibility of a reported result, including the target definition, data-processing boundary, validation design, uncertainty, error profile, probability reliability and intended use [16,17].

Reasons for feature selection include the "curse of dimensionality", noise reduction and interpretability [3-5]. Whereas these arguments are convincing when the source data is large, such as in genomic, imaging or signal processing applications, they need to be interpreted with care when the number of clinically identifiable variables in the source data is small. There may not be a significant computational savings if one takes an 11 predictor model and reduces it to seven predictors. Its usefulness may more readily be in the reduction of input required, the ease of data collection, limited exposure to missing data and global discussion of the variables retained. By contrast, omitting predictors may lose complementary information and lead to a decrease of discrimination. A smaller model and a more accurate model might therefore be two different configurations.

In wrapper-based feature selection, the search is flexible using evolutionary algorithms. A Genetic Algorithm encodes a candidate subset as a binary chromosome and applies selection, crossover and mutation to the population to evolve the new population [6,7]. The fitness function connects the search to the downstream classifier and has the ability to include interactions between predictors in its consideration of the subset to be chosen. But wrapper searches are random, feature dependent, expensive and prone to optimistic performance estimation due to the reuse of the same data for feature discovery and performance estimation [8,11]. Their value should instead be judged on a clear performance-simplicity basis and not taken for granted based on the use of an optimization algorithm.

Random Forest is chosen as the common family of classifiers because it is a model that can capture nonlinear predictor relationships, has an ensemble aggregation to improve the stability of individual decision tree and it can be applied to the mixed structured data [2]. Importantly, the work presented in the current investigation is not to state that Random Forest is always better than Support vector Machine and Logistic Regression or boosting models. It's a narrower research question: In one random forest-based system, what information is discarded and retained when a subset of the predictors is deleted by an evolutionary wrapper?

The data source is publicly released under the name “Heart Failure Prediction Dataset”, but the target variable is HeartDisease and it indicates whether the patient has heart disease or not [21]. The task is thus a binary cross-sectional one for the classification of heart disease. It is not a risk model for future, it does not predict the time of onset, it is not a heart failure phenotype classifier and it is not a causal analysis. This separation is critical for the classification of the diagnostic status and the prediction of future risk, which has different outcomes and validation designs.

This study has three contributions. Firstly, it compares the full-feature and GA-selected Random Forest configuration, not by examining a single partition, but by repeatedly examining one internal validation. The first is to compare the full-feature and GA-selected Random Forest configuration by repeatedly examining one internal validation, instead of examining a single partition. Secondly, it provides a quantification of the predictor reduction and the corresponding accuracy and ROC-AUC changes. Third, it investigates the compact model with cross-validated permutation importance, where importance is not causal, but rather as global predictive relevance. Don't say that the central hypothesis is that GA doesn't need to be more accurate, it's maybe that, if they can find a sensible deal that's less obtrusive in terms of input and more transparent globally, then they're in a position to accept this one.

II. MATERIALS AND METHODS

1. Study design, dataset and outcome

In this study, a public Heart Failure Prediction Dataset provided by Soriano [21] was used, in a retrospective secondary-data study. There are 918 records, 11 input variables and a binary target HeartDisease. Class 0 contains 410 records (44.66%) and class 1 contains 508 records (55.34%). Class 1 was considered as the positive class. Stratified validation was possible with the near-balanced distribution without synthetic oversampling.

Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak and ST_Slope were the predictors. These variables range from demographic context to resting measurement, metabolic measurements, resting electrocardiography, symptoms and exercise response, among other variables. The importance of clinically identifiable variables aids interpretation, but no variable importance nor permutation importance provides biological causation.

Table 1. Predictor domains in the structured heart-disease dataset.

Domain	Predictors	Analytical role
Demographic	Age; Sex	Baseline patient context
Symptoms	ChestPainType; ExerciseAngina	Reported and exercise-induced symptoms
Resting physiology / metabolism	RestingBP; Cholesterol; FastingBS	Resting and metabolic information
Electrocardiographic / exercise response	RestingECG; MaxHR; Oldpeak; ST_Slope	Electrical and stress-response information

2. Information boundaries and preprocessing

The target was pre-removed from the predictor matrix. The target is removed prior to transformation. Categorical variables were translated into binary variables with two numbers (binary encoding) and nominal variables were encoded using a fixed representation of their categories (nominal encoding). The learned preprocessing operations were only applied to the training data and its validation data, which mitigated information leakage caused by the preprocessing [11]. Tree-

based models did not need to be numerically standardized for their construction as the splits.

A better separation is related to feature-selection leakage. The internal cross-validation portion of the GA was used to look for a candidate subset from the available set, which mask was then fixed and then the extraction process was repeated with the full-feature model for comparison. The design is made to preprocess leakage while it is not a complete nested estimate of pipeline with feature selection. In a fully nested design, the GA would be run again in each outer-training split and tested once on the outer-validation observations that were not touched in the first run. The reported GA model is therefore not an unbiased external model of a deployed selection procedure, but merely an internal model.

3. Full-feature Random Forest

The reference model was a Random Forest with 100 trees, Gini impurity, square-root feature sampling at candidate splits, bootstrap aggregation and random state 42. Used all 11 source predictors. Random forest was selected as it can capture nonlinear interactions and can lower the variance of a single tree [2]. The same classifiers family was adopted as fitness evaluator for the GA, so the full and reduced set-up could be compared in the same modelling context.

4. Genetic Algorithm wrapper selection

There were 11 bits in each candidate subset to represent whether the corresponding predictor was included or not. The mean accuracy across an internal 20-fold stratified cross-validation procedure using the random forest model was used as the fitness for candidates. The GA was run on 20 chromosomes, 10 generations, tournament selection, two-point crossover, crossover probability of 0.80 and the bit-flip mutation probability of 0.05.

The fitness function used was a maximum of classification accuracy with no explicit penalty for subset size. The seven-predictor solution is therefore termed the compact subset obtained with the generated search, which is not necessarily mathematically minimal nor necessarily the best solution for the different values of seed. Age, Sex, ChestPainType, MaxHR, ExerciseAngina, Oldpeak and ST_Slope were the only columns left in the final mask.

5. Repeated internal validation and performance measures

Full feature and fixed GA selected Random Forest were tested by 10 repetitions of 20 stratified cross validation. The same split schedule was used for both models. This design was able to produce 200 matched fold-level results per configuration and minimize reliance on the one partitioning sequence. The 200 observations were not considered independent experimental replications as each training set is correlated by overlap between folds and repetitions [23]. The main summaries used were mean and standard deviation.

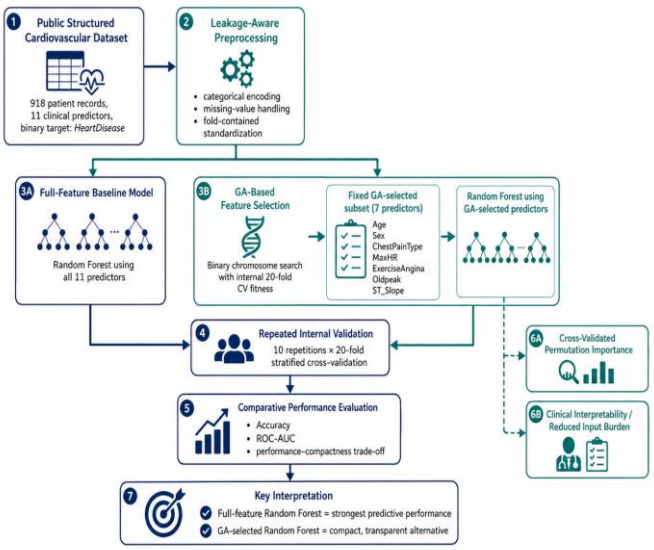

Figure 1. Analytical workflow used to compare the full-feature Random Forest with the fixed GA-selected configuration.

Accuracy, precision, sensitivity, F1 score and ROC AUC were calculated per fold. The main compactness measure reported was predictor count. The difference between the accuracy of GA-selected RF and full-feature RF for pairs was calculated and averaged. Results of a nominal paired t-test and Wilcoxon tests were available from the original analysis but the test results are used here as exploratory as conventional fold-level tests can underestimate uncertainty in repeated cross-validation scores where repeated scores are correlated. The scientific interpretation was thus based on the size of the difference in performance and its direction between repeated performances and not on the single p-value.

6. Cross-validated permutation importance

Permutation importance was used on validation data to measure the relevance of the different predictors in the compact model. Values for each predictor that was retained were randomly permuted and the resulting drop in ROC-AUC was noted. The larger the mean decrease, the more important this predictor is for the fitted model. The procedure does not specify effect direction, patient specific explanations and causal inferences [12,13,30].

III. RESULTS

1. Repeated predictive performance

The full feature Random Forest had the highest repeated internal results. Its mean accuracy was 87.11% (SD 5.06%), compared with 83.67% (SD 5.45%) for the GA-selected model. Mean ROC-AUC was 0.9285 (SD 0.0396) for the full-feature model and 0.9075 (SD 0.0439) for the compact model. The full-feature also performed better with higher mean precision, sensitivity and F1 score. The results show that the four variables that were not included in GA had complementary predictive information.

Table 2. Repeated 10 × 20-fold internal-validation performance.

Configuration	Predictors	Accuracy, %	Precision, %	Sensitivity, %	F1-score, %	ROC-AUC
Full-feature RF	11	87.11 ± 5.06	86.99 ± 5.57	90.58 ± 6.05	88.59 ± 4.52	0.9285 ± 0.0396
GA-selected RF	7	83.67 ± 5.45	84.96 ± 5.79	86.04 ± 6.71	85.33 ± 4.94	0.9075 ± 0.0439

The mean difference in accuracy between the two was −3.44 percentage points (GA-selected RF minus full-feature RF) with the reported nominal interval ranging from −4.00 to −2.88 percentage points. The direction of the tendency was always in favor of the full-featured model and the intensity of the tendency was always high. This interval is a descriptive internal comparison since it was based on correlated values of the fold-level.

2. Predictor reduction and the performance-compactness trade-off

GA cut the number of source predictors from 11 to 7 - a 36.36% reduction. The reduction in accuracy cost was 3.44% and the ROC-AUC cost was 0.0210. Thus, the compact model was able to maintain a high degree of discrimination in fewer recorded variables but had poor preservation of the maximum performance of the full model. While the two configurations are relevant to different priorities, the one preferred is the all-feature configuration when the need for internal discrimination is important, the other is the GA-selected configuration, when the choice of input reduction and global transparency is important.

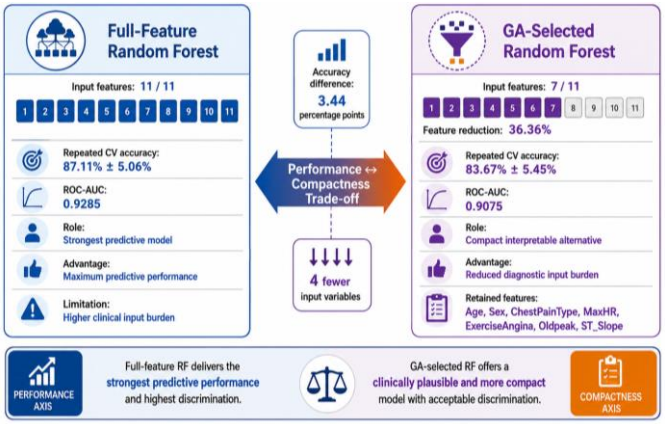

Figure 2. Performance-compactness trade-off between the full-feature and GA-selected Random Forest configurations.

Table 3. Compactness and performance differences.

Quantity	Full-feature RF	GA-selected RF	Difference / reduction
Predictor count	11	7	−4 predictors (−36.36%)
Mean accuracy	87.11%	83.67%	−3.44 percentage points
Mean ROC-AUC	0.9285	0.9075	−0.0210

3. Global predictor relevance in the compact model

After permutation the most important variable in the compact model was ST_Slope with a mean ROC-AUC decrease of 0.1368. ChestPainType (0.0399), Oldpeak (0.0278) and Sex (0.0210) formed the next tier. The mean values of MaxHR, ExerciseAngina and Age were smaller and their standard deviations were greater than or close to the mean, suggesting that these variables' marginal contribution varied among folds. These variables can still make a contribution via interactions or specific portions of the predictor space.

Table 4. Cross-validated permutation importance of GA-selected predictors.

Predictor	Mean ROC-AUC decrease	SD	Interpretation
ST_Slope	0.1368	0.0346	Dominant global predictor

ChestPainType	0.0399	0.0259	Substantial but variable contribution
Oldpeak	0.0278	0.0167	Moderate contribution
Sex	0.0210	0.0166	Moderate, variable contribution
MaxHR	0.0079	0.0175	Low average; unstable across folds
ExerciseAngina	0.0078	0.0193	Low average; unstable across folds
Age	0.0062	0.0129	Low average; unstable across folds

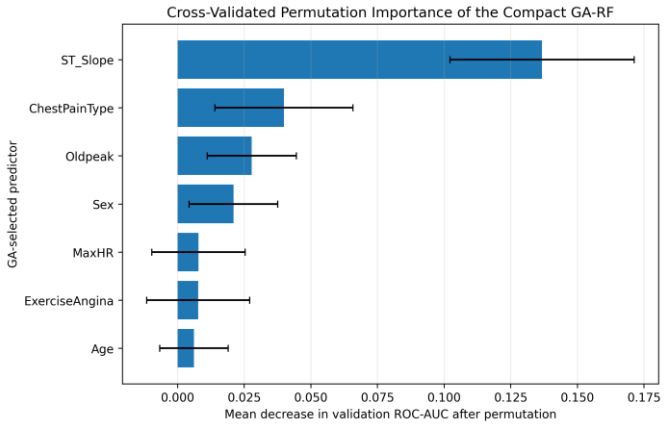

Figure 3. Cross-validated permutation importance of the seven predictors retained by the GA-selected Random Forest.

IV. DISCUSSION

1. Principal findings

The key result is that there was no common configuration for predictive performance and compactness of the model. The highest repeated accuracy, sensitivity, F1-score and ROC-AUC were obtained by the full-feature Random Forest. After removing more than one third of the predictors, GA achieved a small decrease in repeated discrimination, however. The discrete model had a lower repeated discrimination after GA reduced the predictor set by more than one third. This evidence challenges the generally accepted notion that an optimization based selector will lead to an increase in accuracy.

What results are obtained are methodologically probable. Random Forest may take the combination of conditionally informative, weak and redundant variables. If a predictor is not very important on its own, it may be quite useful to help partition a subset of trees and/or may balance another variable in specific observations. The input space was simplified by removing RestingBP, Cholesterol, FastingBS and RestingECG, but complementary signals were also omitted. The resulting loss in accuracy is therefore not indicative of failure of GA and it measures the price of simplification in the search implemented.

2. Clinical input burden and global transparency

When data completeness or the burden of data acquisition or simplicity of the interface is more significant than optimizing discrimination, a seven-variable model may be operationally appealing. The variables that are included in the retained subset are: demographic variables, chest-pain type, maximum heart rate, exercise-induced angina, ST depression and ST-segment slope. These are clinically identifiable variables and enable the model's dependence on the world to be discussed more directly than a larger list of variables.

The use of the term reduced input burden should be done with great care, however. This doesn't necessarily mean the lower the number of variables, the lower the cost; the cost of the variables varies. There are some variables that require exercise testing while other ones may be routinely available and excluded from the list. A formal burden analysis would take a lot of time, cost and data on missingness, equipment and flow. The present study does not provide a validated health-economic benefit, but rather it establishes predictor count reduction.

3. Interpretation of the selected predictors

The top three features, ST_Slope, ChestPainType and Oldpeak, all have high predictive value in the source data set, consistent with the scores of these features. The interpretation of the permutation importance is however dependent on the fitted model, coding scheme, correlated predictors and analyzed population. These variables don't signify separate clinical causes, nor do they establish a diagnostic rule of the use of the model.

In this particular case it may be noted that the three factors Age, MaxHR and ExerciseAngina are relatively unimportant and that this was not a clear-cut hierarchy of medical importance, but rather a binary selection. Even if the average permutation effect of a predictor is small, it can be useful because of interactions. The seven-variable mask can only be considered a clinical signature that can be reproduced in multiple GA seeds and independent cohorts.

4. Relation to prior work

Many works in the feature-selection literature have highlighted benefits of reduced subsets to include better generalization, computation and interpretation and at the same time have pointed out that wrapper methods are data- and model-dependent [3-5,8]. The same issues external validation, full reporting and overinterpretation are highlighted in recent cardiovascular AI reviews [25,26]. The present results uphold a less radical stance: subset reduction is an objective of the design under question that is measurable and does not have to be defended on specious grounds of numerical superiority.

The study also adheres to current reporting and risk of bias reporting guidelines, which recommend that outcomes be clearly defined, methods of validation, calibration and claims of applicability be transparent and that the claims be made with caution [16,17]. The analysis does not choose a smaller model just because it achieved a good outcome in one realization of the cross validation, but rather separates out the predictive reference from the compact alternative.

5. Implications for decision-support research

When using research prototypes, a good idea is to keep both setups. The full-featured model can be used as the predictive model; the compact model can be used for sensitivity analysis, simplicity of user interfaces and analysis of cases where some inputs are missing. In a future clinical study, it would be interesting to determine if the small decrease in discrimination is more than compensated for by the increased data completeness or increased acquisition or user acceptance. It is not enough to risk count to make such a decision; prospective workflow evidence and decision-curve analysis are needed [20,27-29].

Limitations

- A moderate-sized public dataset was used and there was no independent geographical, temporal or prospective validation.
- The target is not a diagnosis of heart failure, probability of future events, severity or survival, but is a recorded HeartDisease value.
- The feature discovery as part of the GA was not reprobed in all outer folds and selection induced optimism cannot be ruled out for the compact model.
- One mask was configured to search and the stability of the mask with the 7 predictors over random seeds was not tested.
- Neither did the fitness function explicitly optimize the number of predictors, nor the acquisition cost, nor the fact that predictors are missing, nor the clinical burden of the predictor.
- Repeated fold scores in cross validation are correlated and nominal significance tests at the fold level are therefore exploratory.
- The permutation importance is a measure of significance for the prediction process, but it is not a measure of direction of effect, explanation in the local domain, fairness or causality.

V. CONCLUSION

The full-feature Random Forest resulted in the best repeated internal performance in structured heart-disease status classification, while GA generated a seven-predictor model with less accuracy and ROC-AUC measures, but reduced input dimensionality and transparent global predictor analysis. The study thus lends weight to a performance-compactness interpretation instead of an accuracy improvement through optimization. The full-feature model is the more suitable predictive reference for the analyzed data set, while the GA selected model is a suitable research alternative in case of decreased burden of input and in the context of increased transparency in the global application of the analysis. The compact subset must be validated from independent data sets, tested for stability, perform selection and optimization and be burden sensitive before it can be used clinically.

Declarations

Ethics statement: This study did not involve secondary data from a public de-identified data set or the solicitation of new health information from participants, nor the recruitment of participants for the study. The analysis is made not as a standalone diagnostic system, but as the study of research.

Availability of data: The Heart Failure Prediction Dataset is open sourced on Kaggle [21].

Availability of source code: The Python source code file used to build the workflow of the Random Forest and evolutionary feature-selection will be made available from the authors on reasonable request and should be published in an open repository after acceptance.

REFERENCES

1. World Health Organization. Cardiovascular diseases (CVDs). 2025. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29.
2. L. Breiman, Random forests, Machine Learning 45 (2001) 5-32. https://doi.org/10.1023/A:1010933404324.

3. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157-1182.
4. Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507-2517. https://doi.org/10.1093/bioinformatics/btm344.
5. J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Computing Surveys 50 (2017) 94. https://doi.org/10.1145/3136625.
6. J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, 1975.
7. D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.
8. B. Xue, M. Zhang, W.N. Browne, X. Yao, A survey on evolutionary computation approaches to feature selection, IEEE Transactions on Evolutionary Computation 20 (2016) 606-626. https://doi.org/10.1109/TEVC.2015.2504420.
9. T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.
10. D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6. https://doi.org/10.1186/s12864-019-6413-7.
11. S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: Formulation, detection and avoidance, ACM Transactions on Knowledge Discovery from Data 6 (2012) 1-21. https://doi.org/10.1145/2382577.2382579.
12. A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.
13. E. Tjoa, C. Guan, A survey on explainable artificial intelligence: Toward medical XAI, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314.
14. E.W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation and Updating, 2nd ed., Springer, 2019. https://doi.org/10.1007/978-3-030-16399-0.
15. B. Van Calster, D.J. McLernon, M. van Smeden, L. Wynants, E.W. Steyerberg, Calibration: The Achilles heel of predictive analytics, BMC Medicine 17 (2019) 230. https://doi.org/10.1186/s12916-019-1466-7.
16. G.S. Collins, K.G.M. Moons, P. Dhiman, et al., TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378. https://doi.org/10.1136/bmj-2023-078378.
17. K.G.M. Moons, J.A.A. Damen, T. Kaul, et al., PROBAST+AI: An updated quality, risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods, BMJ 388 (2025) e082505. https://doi.org/10.1136/bmj-2024-082505.
18. A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, New England Journal of Medicine 380 (2019) 1347-1358. https://doi.org/10.1056/NEJMra1814259.
19. E.J. Topol, High-performance medicine: The convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44-56. https://doi.org/10.1038/s41591-018-0300-7.
20. E.H. Shortliffe, M.J. Sepúlveda, Clinical decision support in the era of artificial intelligence, JAMA 320 (2018) 2199-2200. https://doi.org/10.1001/jama.2018.17163.
21. F. Soriano, Heart Failure Prediction Dataset, Kaggle, 2021. Retrieved June 5, 2026, from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.
22. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825-2830.
23. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1145.
24. D.G. Altman, P. Royston, What do we mean by validating a prognostic model?, Statistics in Medicine 19 (2000) 453-473. https://doi.org/10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5.
25. Y. Cai, Y.-Q. Cai, L.-Y. Tang, et al., Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review, BMC Medicine 22 (2024) 56. https://doi.org/10.1186/s12916-024-03273-7.
26. P. Shah, M. Shukla, N.H. Dholakia, H. Gupta, Predicting cardiovascular risk with hybrid ensemble learning and explainable AI, Scientific Reports 15 (2025) 17927. https://doi.org/10.1038/s41598-025-01650-7.
27. J. Wiens, S. Saria, M. Sendak, et al., Do no harm: A roadmap for responsible machine learning for health care, Nature Medicine 25 (2019) 1337-1340. https://doi.org/10.1038/s41591-019-0548-6.
28. C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine 17 (2019) 195. https://doi.org/10.1186/s12916-019-1426-2.

29. A.J. Vickers, E.B. Elkin, Decision curve analysis: A novel method for evaluating prediction models, Medical Decision Making 26 (2006) 565-574. https://doi.org/10.1177/0272989X06295361.
30. C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022. https://christophm.github.io/interpretable-ml-book/.