Original Paper

From Data to Decision: Interpretable and Reliable Heart Disease Risk Prediction

ANDREI-IONUȚ SPÎNU[1], RENATO CONSTANTIN IVĂNESCU[2], CHRISTIANA RALUCA DĂNCIULESCU[3], DANIEL-ROBERT STĂNESCU[3]

[1]University of Medicine and Pharmacy of Craiova, Romania
[2]University of Craiova, Romania
[3]Doctoral School, University of Medicine and Pharmacy of Craiova, Romania

ABSTRACT: Background: Heart disease is the leading cause of morbidity and mortality worldwide, highlighting the need for accurate and clinically meaningful risk prediction tool. Methodology: A retrospective analysis was performed on a dataset of 918 patients with complete clinical information. The relationship between variables and disease presence were explored using correlation and dependence measure. To prediction was made using logistic regression, linear discriminant analysis, and generalized additive models. The performances were evaluated using cross-validation, discrimination metrics, and calibration analysis. Results: Functional and exercise-related variables, particularly Oldpeak, maximum heart rate, chest pain type, and ST slope showed the strongest correlation with disease presence. All models achieved AUROC values higher than 0.91. Logistic regression and generalized additive models performed similarly, with no statistically significant difference between them. Calibration analysis showed a good agreement between predicted and observed probabilities. Conclusion: Heart disease prediction can be achieved with high accuracy using clinically accessible variables and interpretable models. Our findings suggest that simpler approaches remain suitable for practical clinical applications, while preserving transparency and usability.

KEYWORDS: *Heart disease, risk prediction, artificial intelligence, statistical analysis, interpretability.*

Introduction

Worldwide, the most important causes of death or long-term disabilities remain cardiovascular diseases, even if major progress has been made in terms of prevention, diagnosis, and treatment.

Among these diseases, a particular spot is occupied by heart failure, due to the fact that it is not only a disease entity, but more of a final destination of different cardiovascular disorders, such as ischemic heart disease, valvular disease, cardiomyopathy, and hypertension.

In 2022, AHA/ACC/HFSA published a guideline that emphasizes that heart failure must be diagnosed at an early stage, should account for phenotype-based classification, and different patient-centered treatment strategies should be developed.

Patients that suffer from heart failure require recurrent hospitalization and have functional decline and increased mortality [1].

Heart failure management has changed substantially in recent years, especially after discovering the evidence supporting sodium-glucose cotransporter 2 inhibitors across a large spectrum of heart failure phenotypes.

New therapeutic evidence for acute and chronic heart failure highlights the need for constant refinement of clinical decision-making in this field [2].

In parallel, an epidemiological study revealed that the burden of heart failure is risen because of population aging, improved survival after cardiac events, and the increasing prevalence of comorbidities such as diabetes, obesity, atrial fibrillation, and chronic kidney disease [3].

When diagnosing or prognosing heart failure we are dealing with multiple variables that interact between themselves.

Different symptoms are common, but not specific (e.g. dyspnea, fatigue, fluid retention, exercise intolerance, etc.), while underlying mechanism differ a lot from patient to patient.

This is the reason why current research focuses on early detection, risk stratification, and development of new predictive tool that can support medical decision-making without replacing clinical judgement.

Artificial Intelligence and machine learning algorithms are explored for this task.

For instance, Wang compared 18 machine learning techniques for heart failure prediction and reported that preprocessing choices, such as z-score and SMOTE, can improve the classification performance when dealing with class imbalance [4].

K-nearest neighbor, Support Vector Machines, Decision Trees, and XGBoost models were evaluated and the results show that these classical models can provide useful comparative baselines for predicting heart failure [5].

10.12865/CHSJ.52.02.06

A clinical metamodel for cardiac failure forecasting that combined several classifiers showed that ensemble-based approaches might be able to improve the prediction stability when compared to stand-alone algorithms [6].

In [7], the authors compared logistic regression, random forest, decision trees, multilayer perceptron, and k-nearest neighbor, and the results showed that random forest performed best, achieving a specificity of 0.93, AUC of 0.97, and Matthews correlation coefficient of 0.83.

Wang and Cao developed a weighted fusion LightGBM model that used bootstrap sampling and the results showed that the newly developed model outperform traditional prediction models, while reducing overfitting relative to the baseline model, [8].

In [9], the authors evaluated Adaptive Boosting, Naïve Bayes, XGBoost, Bagging, and Logistic Regression, and demonstrated that XGboost reached 94.34% accuracy and 95.19% F1-score.

Alzboon et al. proposed a prediction framework which combined Python-based analysis with low-code deployment and reported that Gradient Boosting achieved the highest accuracy 87.4% and AUC 0.928, while emphasizing interpretability [10].

This research was further extended towards federated learning for cardiovascular heart analysis, indicating a growing interest in privacy-preserving AI frameworks for medical prediction tasks [11].

Objective

In this study we aim to provide a comprehensive and clinically meaningful evaluation of heart disease prediction using a merger between classical statistical modeling with more flexible, data-driven approaches, with a particular interest in interpretability, stability and practical applicability.

In contrast to other existing studies that focus primarily on the models' performances, this study is designed to examine how different modeling strategies behave across multiple analytical dimensions, such as effect size estimation, discrimination, association structure, and calibration.

Methods

Study design and data

The study was conducted on a publicly available dataset, https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data.

The data represents a curated aggregation of five previously independent heart disease cohorts that include data from Cleveland, Hungary, Switzerland, Long Beach VA, and Statlog.

After different preprocessing steps which included harmonization and cleaning, the final dataset comprises 918 patients, described by a set of clinical variables [12].

There is no longitudinal follow-up of patients, the dataset design being a typical cross-sectional clinical structure.

The variables that describe each patient cover demographic information, cardiovascular risk factors, and exercise-related parameters.

These include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic findings, maximum heart rate achieved, exercise-induced angina, ST-segment depression (Oldpeak), and ST slope.

Since the output is binary, signaling the presence (1) or absence (0) of heart disease, the problem can be formulated as a supervised classification one.

From a clinical point of view, the variables reflect standard diagnostic and functional parameters, that are commonly used in cardiovascular evaluation.

Therefore, this dataset is a realistic representation of patient-level information which is available in routine clinical practice, without relying on advanced imaging, biomarkers, or longitudinal follow-up data.

The combination of multiple original cohorts increases the variability in patient characteristics, thus improving the robustness of the statistical analyses.

Our study followed a stepwise analytical approach, starting with exploratory data analysis and continuing with regression modeling, non-linear modeling, and model validation.

Exploratory data analysis

At first, we were interested in exploring the relationships between variables.

For continuous variables the linear associations were quantified using Pearson correlation coefficients, while for the monotonic relationships were analyzed using Spearman correlation [13].

In the case of categorical variables, the dependence was assessed using chi-square test, and the strength of the association was quantified using Cramer's V [14].

The relationship between continuous predictors and the dependent binary output was assessed using point-biserial correlation, providing a direct measure of association with disease presence.

Logistic regression analysis

For estimating the contribution of each variable to the risk of a disease, we have used the logistic regression method.

This model describes the relationship between the predictors and the probability of a person to have that disease using the following formula:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_j \beta_j x_j ,$$

where p is the probability of having heart disease, [15]. In order to allow direct clinical interpretation, the estimated coefficients were expressed as odds ratios $OR = e^{\beta_j}$.

We have assessed the model's performance using pseudo-R^2 measures and likelihood based statistics.

Because we wanted to allow comparison between variables measured on different scales, we have standardized the coefficients.

Generalized additive modeling

In medicine, there is rarely a linear effect between predictive and dependent variables, hence to account for a potential non-linear effect, a generalized additive model (GAM) was used for fitting.

Hence, we have replaced the linear predictor with a sum of smooth functions:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_j f_i(x_j),$$

where each function f_i represents a data-driven smooth effect. In order to ensure a stable estimation and to avoid overfitting, we have used penalized splines.

Each effect's complexity was evaluated using the effective degrees of freedom.

A value that is close to 1 indicates approximately linear relationships [16].

We have assessed the uncertainty in the estimated effects by using bootstrap resampling, generating confidence intervals for the smooth functions.

The dataset was divided into training, validation, and testing subsets, while maintaining the proportion of cases and controls.

We have also used the 10-fold cross-validation for the evaluation of the model's performance [17].

The discriminative ability was quantified using the area under the ROC curve (AUROC), while the differences between models was assessed using the DeLong test [18].

The benchmarking process was further expanded by using Net Reclassification Improvement (NRI) and Integrated Discrimination Improvement (IDI) [19].

Calibration analysis

Besides discrimination, we have also evaluated calibration, in order to assess the agreement between the observed outcomes and the predicted probabilities.

For this task, we have used reliability diagrams and computed the expected calibration error (ECE) [20]:

$$ECE = \sum_n \frac{|B_m|}{n} |observed_m - predicted_m|.$$

We have quantified the prediction error using Brier score:

$$BS = \frac{1}{n}\sum (y_i - \hat{p}_i)^2.$$

To ensure a monotonic relationship between predicted and observed probabilities, we have used isotonic regression as post-processing step.

In this manner, we have preserved calibration [21, 22].

Results

We have included in our statistical analysis the 918 patients with their complete data.

All the variables have been used in their original clinical form.

The first step in our statistical analysis involved examining the relationship between the presence or absence of the disease and the predictive variables.

The relationships between continuous variables were modest, with a low correlation score, which ultimately suggests redundancy.

We can see that there is a positive association between age and resting blood pressure and Oldpeak, and a negative correlation between age and maximum heart rate.

Interestingly enough, cholesterol did not have any strong associations with other variables.

After computing and comparing the Pearson and Spearman coefficients, we can see that most relationships can be reasonable approximated as linear.

In what regards the categorical variables, the results show that there are moderate associations between chest pain type, ST slope, and the

 10.12865/CHSJ.52.02.06

presence of heart disease, while sex and fasting blood sugar showed weaker associations.

If we were to focus directly on the outcome, then age and Oldpeak showed the strongest positive correlates, whereas maximum heart rate showed a negative correlation.

Again, no correlation with the cholesterol value (Figure 1).

Figure 1. Correlation and association structure of clinical variables.

The results of the multivariable analysis pointed out the following conclusions.

When using the logistic regression method, we have discovered that asymptomatic chest pain, male sex, and flat ST slope are the most statistically significant influential predictors, each one of them increasing the odds of the disease.

Two other variables, fasting blood sugar and exercise-induced angina, also had strong positive correlations.

In contrast, cholesterol and maximum heart rate were associated with a lower probability of disease when we took into consideration all the variables.

Age and resting blood pressure had a limited contribution.

We can see that there is a clear separation between healthy and diseased patients, and that most observations are clustered at the extremes of the risk spectrum (Figure 2).

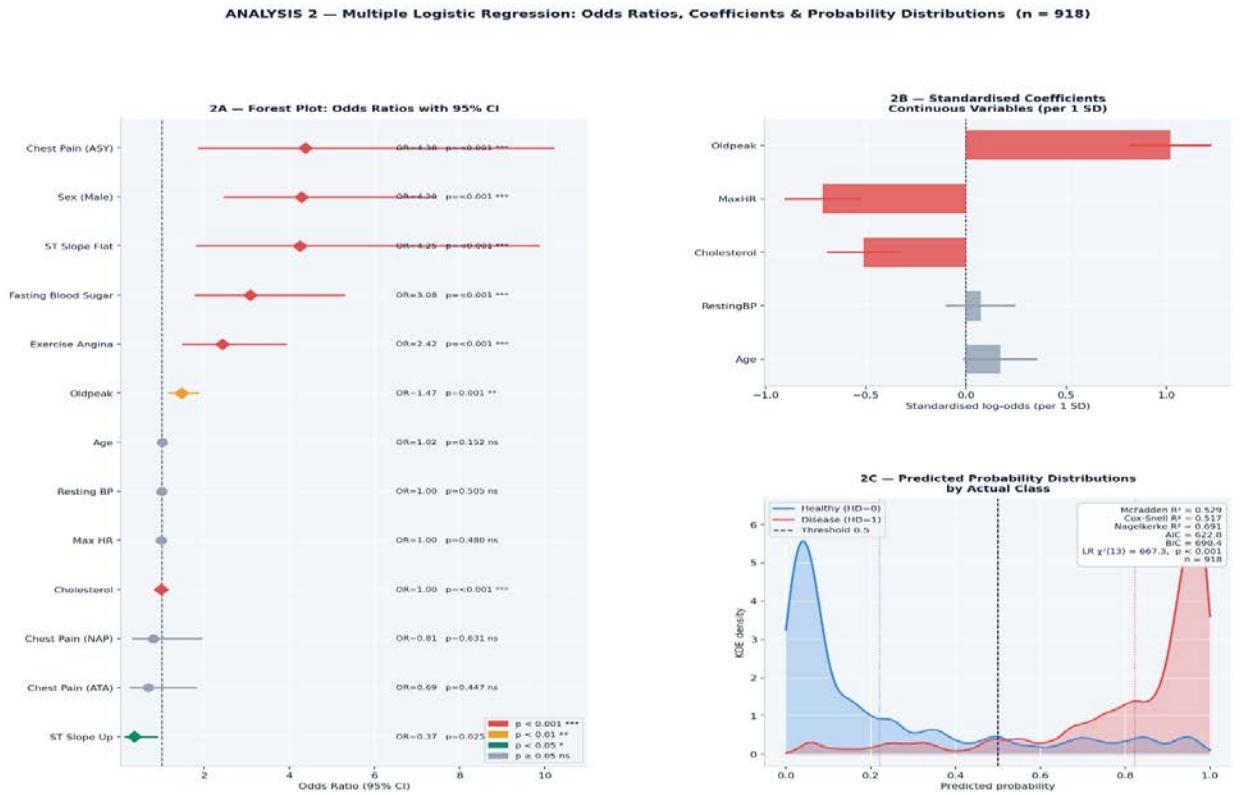

Figure 2. Logistic regression results and predicted probability distribution.

In order for us to explore the linearity of the variables' effects, we used generalized additive model.

Most of the predictive variables behaved approximately linearly, except Oldpeak, which showed a clear non-linear pattern, with an increasing effect on risk at higher values.

The stability of these findings was also supported by the confidence intervals obtained through bootstrap resampling.

We can see that indeed certain variables carry additional information that is better captured through flexible modeling (Figure 3).

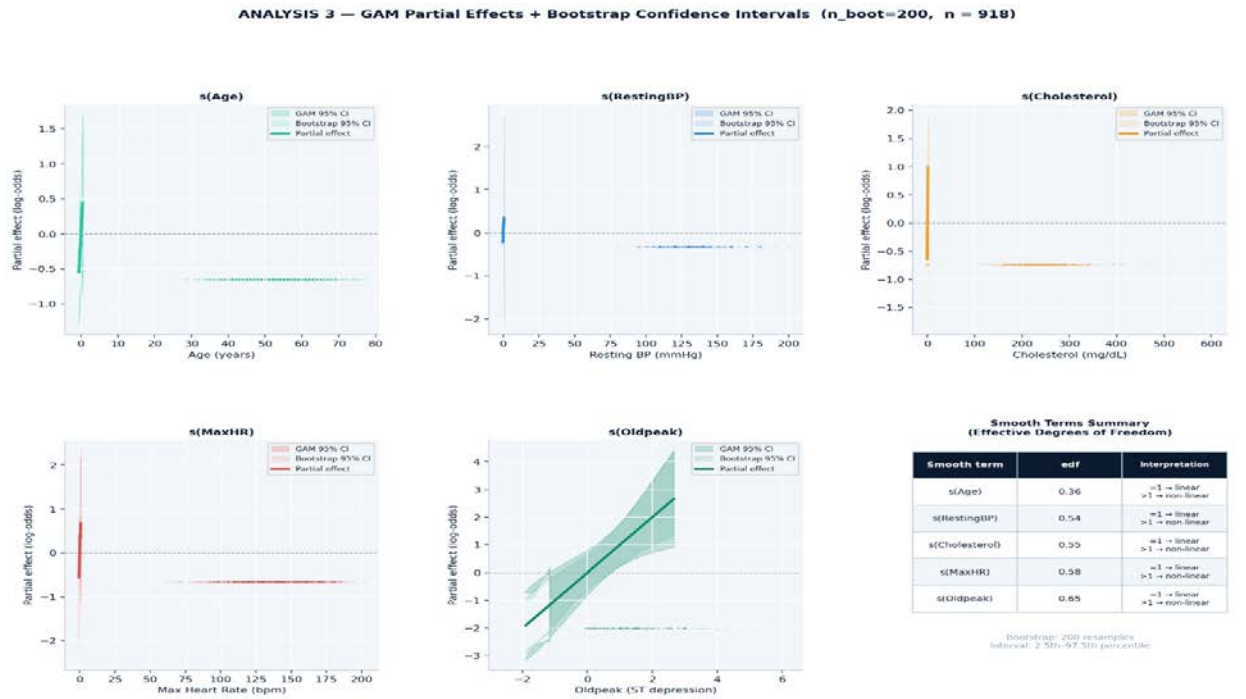

Figure 3. Partial effects estimated by generalized additive model (GAM).

10.12865/CHSJ.52.02.06

Further on in our statistical analysis, we have used model calibration to understand how well predicted probability matched the observed outcomes.

From Figure 4 we can see that both the logistic regression and GAM showed good agreement with the ideal calibration line, having just small deviations in the mid-probability range.

In both models we can see that the calibration error is low, and that when applying isotonic regression, we further improve the alignment, by effectively removing residual bias.

The Brier score confirmed our findings, that there is a similar overall prediction across the used models, while the distribution of the predicted probabilities showed an appropriate split, without signs of overconfidence.

Figure 4. Calibration performance of predictive models.

Our statistical analysis ended with a comparison of the predictive performance of the models.

All models achieved high discrimination, the logistic regression having the highest AUC, closely followed by GAM and LDA.

There were no statistically significant differences between the logistic regression and GAM, but there were statistically significant differences between the logistic regression and LDA.

The reclassification metrics showed a small advantage of GAM in assigning patients into the correct risk categories when compared to LDA (Figure 5).

The obtained results how that heart disease prediction in this dataset is driven by a merger between functional and demographic variables.

While traditional models like logistic regression perform well, we can use more flexible methods such as GAM to provide additional insight into how risk evolves across clinical measurements.

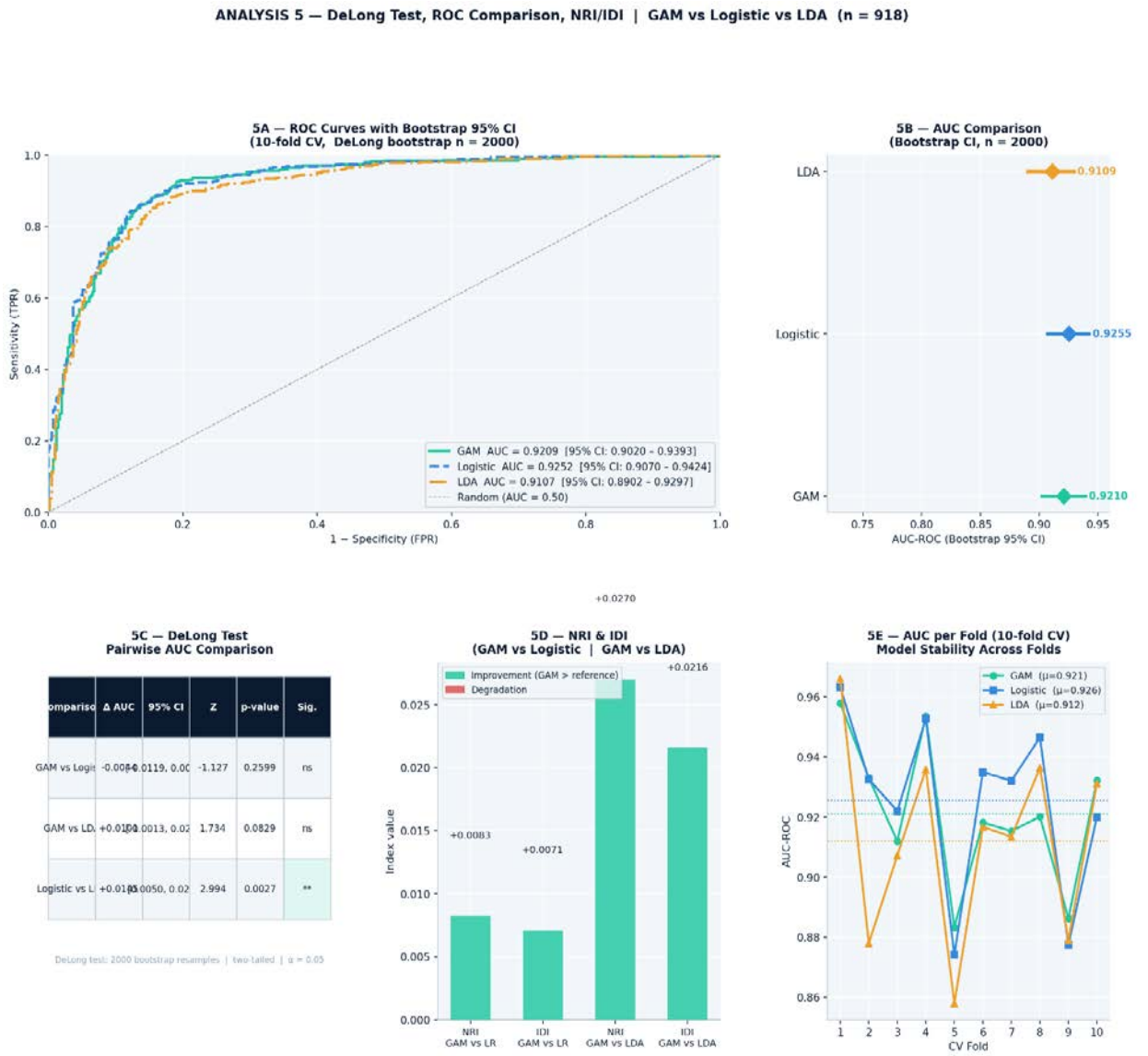

Figure 5. Model comparison and validation results.

Discussion

In this study we present a comprehensive statistical analysis that shows that heart disease prediction is driven by a combination of functional and clinical variables.

Exercise-related parameters proved to be informative also. Oldpeak, chest pain type, ST slope, and maximum heart rate had the strongest correlation, hence influence, on disease classification, while classical biochemical markers such as cholesterol had a lower influence in this multivariable setting.

This pattern is in line with current clinical understanding, where functional impairment and ischemic response proved direct evidence of cardiovascular pathology, rather than isolated laboratory values [1,5].

From an AI modeling perspective, we can have discovered that both the logistic regression and GAM have achieved a high discriminative performance, having non-significant statistical differences between them.

At the same time, GAM analysis added an important layer of interpretation showing that even if most of the predictors are linear, Oldpeak showed a non-linear effect, which suggests a progressive increase in risk rather than a proportional relationship.

The calibration analysis shows that these models can be clinical usable, because they produce well-calibrated probability estimate, with minor deviations that were corrected by the isotonic regression.

This is very relevant in the medical context, because accurate risk estimation is more important than classification [2].

When compared to other machine learning techniques applied on the similar datasets, the

10.12865/CHSJ.52.02.06

performance obtained in this study is consistent with the upper range reported in literature [4,7,9].

Even if other complex hybrid models might achieve superior improvements, these gains are modest and come at the cost of reduced interpretability [10].

In this context, our findings support the idea that simpler models offer a favorable balance between performance, clinical interpretability, and transparency.

Nevertheless, we should acknowledge several limitations.

The dataset size is relatively small when compared to large-scale clinical registries and thus may limit the generalizability.

Secondly, the study is based on retrospective structured data, with no imagining, longitudinal follow-up, treatment variables, which are known to influence the disease's progression [3].

Future work will focus on validating these findings on larger cohorts on diverse populations, and also on integrating other data sources, such as electronic heart records.

The use of explainable AI frameworks may enhance the interpretability of complex models, and thus allow clinicians to understand individual risk profiles.

Conclusions

In this study, we have estimated the heart disease risk using traditional AI models.

The results indicate that classical models capture most of the predictive structure in data, while flexible approaches refine how the risk is expressed, but do not improve the overall performance.

Overall, our findings suggest that clinically meaningful prediction can be achieved without increasing model complexity beyond what is necessary.

Acknowledgement

None to declare.

Author Contributions

Conceptualization, A.S. and C.R.D.; Methodology, A.S. C.R.D. and R.C.I.; Investigation, Methodology, A.S. C.R.D. and R.C.I.; Data analysis, R.C.I.; Manuscript writing and initial draft preparation, Methodology, A.S. C.R.D. and R.C.I.; Manuscript review and editing Methodology, A.S. C.R.D. and R.C.I.; Supervision, R.C.I.

All authors read and approved the final manuscript.

Funding

This work was supported by the Regional Program South-West Oltenia 2021-2027, under Financing Contract No. 797/19.02.2026 (SMIS 347141), funded by the European Regional Development Fund (ERDF), Priority 1 "Competitiveness through innovation and dynamic enterprises".

Conflicts of Interest

The authors declare no competing interests.

Institutional Review Board

The study was conducted according to the guidelines of the Declaration of Helsinki.

Consent Statement

Not applicable.

Data Availability

The anonymized data are available at the following link:

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data.

References

1. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, Deswal A, Drazner MH, Dunlay SM, Evers LR, Fang JC, Fedson SE, Fonarow GC, Hayek SS, Hernandez AF, Khazanie P, Kittleson MM, Lee CS, Link MS, Milano CA, Nnacheta LC, Sandhu AT, Stevenson LW, Vardeny O, Vest AR, Yancy CW. 2022 AHA/ACC/HFSA guideline for the management of heart failure. Circulation, 2022, 145(18):e895-e1032.
2. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, Burri H, Butler J, Čelutkienė J, Chioncel O, Cleland JGF, Coats AJS, Crespo-Leiro MG, Farmakis D, Gilard M, Heymans S, Hoes AW, Jaarsma T, Jankowska EA, Lainscak M, Lam CSP, Lyon AR, McMurray JJV, Mebazaa A, Mindham R, Muneretto C, Piepoli MF, Price S, Rosano GMC, Ruschitzka F, Skibelund AK. 2023 focused update of the 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J, 2023, 44(37):3627-3639.
3. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. Eur J Heart Fail, 2020, 22(8):1342-1356.
4. Wang J. Heart failure prediction with machine learning: a comparative study. J Phys Conf Ser, 2021, 2031(1):012068.
5. Zeng M. The prediction of heart failure based on four machine learning algorithms. Highlights Sci Eng Technol, 2023, 39:1377-1382.
6. Mahmud I, Hasan M, Rahman MA, Islam MR, Hossain MS, Uddin S. Cardiac failure forecasting based on clinical data using a metamodel. Diagnostics, 2023, 13(15):2540.
7. Chulde-Fernández B, Enríquez-Ortega D, Guevara C, Sánchez-García A, López-González JL, Rodríguez-Mazahua L, Hernández-Suárez CM. Classification of heart failure using machine learning: a comparative study. Life, 2025, 15(3):496.

8. Wang Y, Cao H. Heart failure prediction based on bootstrap sampling and weighted fusion LightGBM model. Appl Sci, 2025, 15(8):4360.
9. Tabassum S, Muhammad F, Khan MA, Alqahtani AS, Alshammari SA, Alsubhi K. A machine learning-based framework for heart disease diagnosis using a comprehensive patient cohort. Comput Mater Contin, 2025, 84(1):1253-1278.
10. Alzboon MS, Alzboon NM, Al-Hadhrami T, Al-Smadi M, Alshurideh M. A clinically accessible heart disease prediction framework: multi-model evaluation with ensemble learning and low-code deployment. Discov Appl Sci, 2026, 8:339.
11. Sarwar F, Farooq MS, Samee NA, Jamjoom MM, Ashraf I. A federated learning approach for cardiovascular health analysis and detection. Comput Mater Contin, 2025, 84(3):5897-5914.
12. Kaggle Inc. Heart Failure Prediction Dataset [online]. Available at https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data [Accessed 08.05.2026].
13. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J, 2012, 24(3):69-71.
14. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed, Lawrence Erlbaum Associates, 1988, Hillsdale, 216-220.
15. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley, 2013, New York, 35-70.
16. Hastie T, Tibshirani R. Generalized additive models. Stat Sci, 1986, 1(3):297-310.
17. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Joint Conf Artif Intell, 1995, 14(2):1137-1145.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated ROC curves. Biometrics, 1988, 44(3):837-845.
19. Pencina MJ, D'Agostino RB, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker. Stat Med, 2008, 27(2):157-172.
20. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proc Int Conf Mach Learn, 2017, 70:1321-1330.
21. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev, 1950, 78(1):1-3.
22. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. Proc ACM SIGKDD Int Conf, 2002, 694-699.

Corresponding Author: Christiana Raluca Dănciulescu, Doctoral School, University of Medicine and Pharmacy of Craiova, Romania, e-mail: raluca.danciulescu@umfcv.ro

 10.12865/CHSJ.52.02.06