ISSN: 2832-7708

Research Article

Insights of Cardiovascular Pharmacology Research

Machine Learning Models for Predicting Heart Failure: Unveiling Patterns and Enhancing Precision in Cardiac Risk Assessment

Mahdi Navaei[1*] and Zohreh Doogchi[2]

[1]*Bachelor's in Information Technology, University of Applied Science and Technology Informatics of Iran, Iran*

[2]*Education manager and research officer, University of Applied Science and Technology Informatics of Iran, Iran*

[*]Corresponding Author
Mahdi Navaei, Bachelor's in Information Technology, University of Applied Science and Technology Informatics of Iran, Iran.

Submitted: 2024, Jul 16; **Accepted**:2024, Aug 09: **Published**: 2024, Aug 19

Citation: Navaei, M., Doogchi, Z. (2024). Machine Learning Models for Predicting Heart Failure: Unveiling Patterns and Enhancing Precision in Cardiac Risk Assessment. *Insig Cardio Vasc Phar Res, 4*(1),1-12.

Abstract

Purpose

This study aims to evaluate the efficacy of various machine learning models in predicting heart failure incidence using medical data, focusing on the innovative aspect of a novel dataset. Unlike previous studies that predominantly examined features such as smoking status or age, this research explores novel features. The primary challenge addressed is the utilization of these new features, coupled with machine learning techniques, to accurately diagnose heart failure.

Methods

Five machine learning models, including logistic regression, support vector classifier, decision tree classifier, random forest classifier, and K-nearest neighbors, were applied to analyze medical data from a dataset comprising over 900 individuals. The dataset encompasses diverse parameters such as age, sex, chest pain severity, blood pressure, cholesterol levels, blood sugar levels, and electrocardiogram results, introducing a novel approach to feature selection.

Results

The evaluation of machine learning models unveiled varying performances in predicting heart failure. Logistic regression and support vector classifier exhibited the highest accuracy of 88%, followed by the decision tree classifier (below 85%), random forest classifier (84%), and K-nearest neighbors (82%). Additionally, the analysis revealed a balanced dataset distribution and highlighted sex-based disparities in heart failure incidence, along with significant correlations with factors such as age, chest pain severity, blood glucose levels, and physical activity.

Conclusions

The findings underscore the potential of integrating multiple machine learning models for early detection of heart failure, leveraging the inclusion of novel features in the dataset. However, careful model selection is crucial to account for discrepancies in accuracy among different models, emphasizing the importance of tailoring approaches based on specific project requirements.

Keywords: Heart Failure Prediction, Machine Learning Models, Cardiac Risk Assessment, Medical Data Analysis, Precision Healthcare

1. Introduction

Heart failure, also known as congestive heart failure, is a condition in which an individual's heart cannot supply enough blood for the body's needs. This can occur when the heart fails to pump or fill with blood adequately. The term "heart failure" does not imply a complete cessation of heart function [1]. Nevertheless, heart failure is a serious condition requiring medical care. In the United States alone, more than 6 million adults are affected by heart failure, as reported by the Centers for Disease Control and Prevention. Although children can also experience heart failure, this article focuses on heart failure in adults [2].

Despite the abundance of medical data and continuous advancements in data science, various groups are striving to create indicators that can help predict diseases in the future. Cardiovascular diseases (CVDs) are globally recognized as the leading cause of death, accounting for approximately 17.9 million deaths annually and constituting 31% of total global deaths [3].

Heart failure is a common consequence of CVD. Individuals with a history of cardiovascular diseases or those at high risk (due to factors such as elevated blood pressure, diabetes, hyperlipidemia, or existing diseases) require early diagnosis and management, where a machine learning model can prove highly useful. Therefore, we aim to automatically address another nature-bound problem and focus on future challenges using artificial intelligence techniques. The goal of this article is to predict whether an individual is prone to heart failure based on multiple features, including numerical and categorical characteristics. We delve into the examination and construction of several machine learning models in this article, introducing the model with the highest accuracy as the final result.

2. Research Method

The study dataset comprises information from more than 900 individuals and is vital due to its inclusion of various details, such as age, sex, chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic), resting blood pressure, cholesterol and blood sugar levels, resting electrocardiogram results, maximum heart rate, exercise-induced angina, Oldpeak, and ST segment slope during exercise. The last column indicates whether the individual has experienced heart failure. These data hold significant importance as primary inputs for machine learning models in predicting the likelihood of actual heart failure in individuals.

This dataset is compiled from data published by the following institutions and centers:

1) Hungarian Institute of Cardiology, Budapest: András Jánosi
2) University Hospital, Zurich, Switzerland: William Steiger
3) University Hospital, Basel, Switzerland: Matthias Pfisterer
4) Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano

In this research, diverse machine learning models were employed to predict heart failure. The details of the models, including logistic regression, support vector classifier (SVC), decision tree classifier, random forest classifier, and K-nearest neighbors classifier, will be examined. Additionally, four different metrics- predictions, accuracy, recall, and F1-score-will be utilized to evaluate the model's performance, along with a confusion matrix.

3. Model

3.1 Logistic Regression Model

Logistic regression stands out as a crucial model in the realm of classification, catering to issues grounded in both numerical and categorical inputs. This model computes probability predictions and employs a logistic function for decision-making. Essentially, logistic regression constitutes an equation predicting the probability of the dependent variable (e.g., heart disease) based on independent variables (features of the patient). The model was trained on the designated dataset utilizing features such as age, sex, chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic), resting blood pressure, cholesterol levels, blood sugar levels, resting electrocardiogram results, maximum heart rate, exercise-induced angina, Oldpeak, and ST segment slope. The accuracy and performance of this model for the prediction of heart failure were thoroughly evaluated.

Some advantages of this model include the following:

A) Simplicity and high interpretability: A logistic regression is a straightforward model that lends itself to easy interpretation. [4]

B) High Efficiency in Binary Classification Problems: The presence of the logistic function in this model makes it suitable for binary classification problems.

C) Resistance to Noisy Data: This model generally exhibits resistance to noisy and outlier data.

However, the model has several limitations, including the following:

D) Suitability Only for Binary Classification Problems: Logistic Regression is suitable only for binary classification problems, and its performance may be suboptimal for multiclass problems.

E) Sensitivity to Data Errors: This model is sensitive to outliers and errors in the data and may be influenced by noisy data.

F) Limited Complexity: Logistic regression has limited complexity and may exhibit inappropriate performance for complex problems with nonlinear structures.

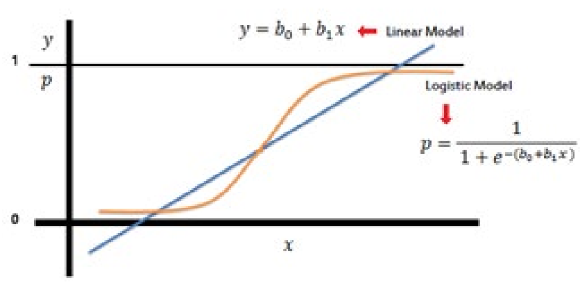

Figure 1: Conceptual Logistic Regression Model

3.2 Support Vector Classifier (SVC)

The support vector classifier (SVC) is a machine learning model employed for classification tasks. This model focuses on differentiating various classes of data using the concept of separable categories. The primary goal of this model is to create an optimal overlap of data in feature space separated by decision boundaries (determinants) [5].

In this article, the SVC model is utilized as one of the principal models. The features used to train this model include various information related to heart failure patients. The results obtained from this model are reported in the article to evaluate its accuracy and performance in diagnosing and predicting heart failure [6].

The advantages of this model include the following:

a) Effectiveness in high-dimensional feature spaces: The SVC algorithm is usually effective at dealing with high-dimensional feature spaces and outlier data.

b) High generalization ability: This model, considering the concept of kernels, exhibits a high level of generalizability.

c) Robustness against outliers: The SVC enhances the resistance to outlier data by maximizing the margin.

d) This model has several limitations, which include the following:

e) Computational complexity: The training time for the SVC model on voluminous data may be lengthy.

f) Sensitivity to kernel selection: The quality of kernel selection in the SVC model can significantly impact its performance.

g) Difficulty in interpretation: Several SVC models, due to their high complexity, may pose challenges in interpreting the decision-making process. 4.2 Accuracy

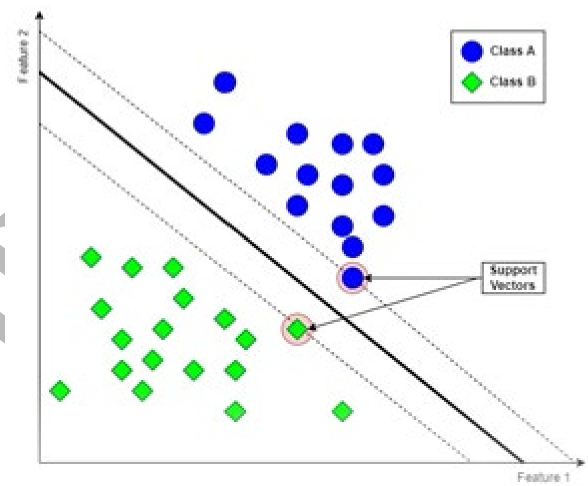

Figure 2: Support Vector Classifier Conceptual Model

3.3 Decision Tree Classifier

The decision tree classifier is a machine learning model used for classification and regression tasks. This model is structured as a tree, making multiple decisions based on input features. At each tree node, a test is conducted on a feature, and the data are directed to one of the tree branches based on the result [7].

This classifier serves as a primary model for categorizing patients with heart failure. The features utilized in this model encompass various information related to patients. The results obtained from this model are reported in the article to evaluate its performance and compare it with those of other models.

The advantages of employing this model include the following:

a) Simplicity and interpretability: The tree-like structure of a decision tree allows for simplicity and easy interpretation.

b) No Need for Preprocessing: This model typically does not require data preprocessing and exhibits high resistance to outlier data.

c) Applicability to discrete and continuous data: A decision tree can effectively handle both discrete and continuous data, such as patient-related features [8].

d) However, this model has several limitations:

e) Risk of Overfitting: Decision trees with excessive depth may overfit the data, fitting it too closely and not generalizing well to new data.

f) Sensitivity to Small Data Changes: Small changes in the data may lead to significant alterations in the tree structure.

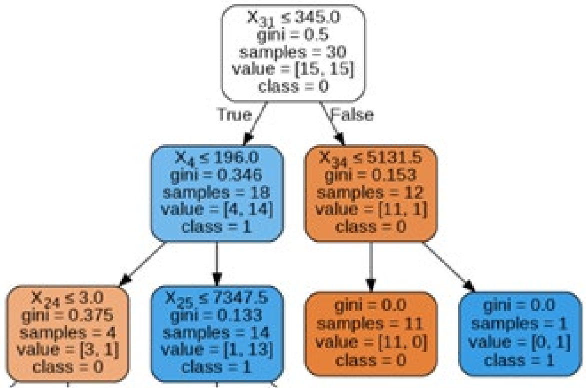

Figure 3: Conceptual Model of Decision Trees

3.4 Random Forest Classifier

The random forest classifier is a machine learning model that operates on the principle of creating multiple decision trees, collectively referred to as a "random forest." Each tree in this ensemble is formed randomly from a subset of both data and features, and subsequent decisions are made based on these trees [9]. Ultimately, the results from each tree were amalgamated to yield the final decision.

In this study, the random forest classifier was employed for the classification of heart failure patients to enhance accuracy and mitigate issues related to overfitting. This model is utilized to improve the overall classification performance and ensure result stability. The advantages of this model are meticulously reported in the article with precision and are systematically compared with those of other models.

The key advantages of this model include the following:

a) Enhanced Resistance to Overfitting: The amalgamation of results from multiple decision trees imparts greater resilience to overfitting.

b) High Precision: This model generally boasts high accuracy in classifying novel data [10,11].

c) Versatility with Discrete and Continuous Data: RF exhibits versatility, effectively working with diverse datasets containing various features.

d) However, noteworthy limitations include the following:

e) Computational Complexity: Constructing multiple decision trees may incur substantial computational costs [11].

f) Limited interpretability: Due to the random and combinatory nature of decision-making, interpreting RF results may be challenging.

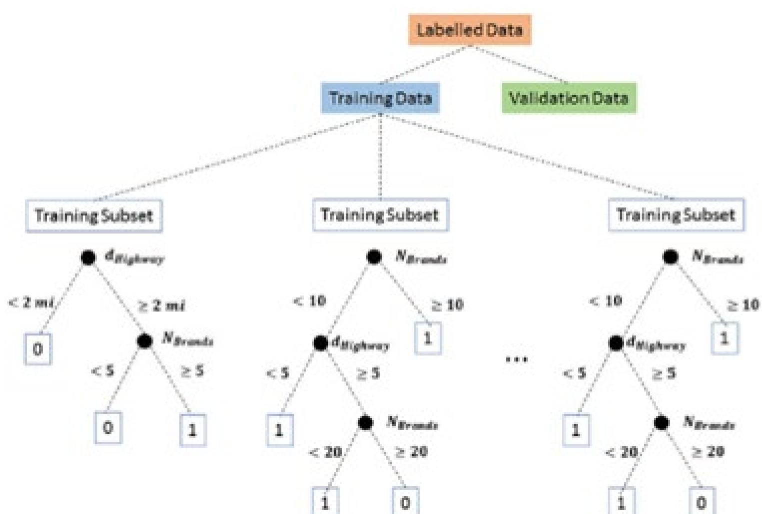

Figure 4: Conceptual Model of the Random Forest Model

3.5 K-nearest Neighbors Classifier

The K-nearest neighbors (KNN) model is a classification algorithm that operates based on the similarity between samples. For a new sample, KNN selects K similar samples from the training data and classifies the new sample based on the majority class among these neighbors.

In this study, the KNN was employed for the classification of patients with heart failure to determine its role and effectiveness in the given problem. This model is utilized to determine the classification of a new sample based on its closest neighbors.
The advantages of KNN include the following:

a) Simplicity and Ease of Implementation: KNN is a simple and easily implementable model.
b) No training phase is needed: This model directly makes decisions based on similarity to training samples and does not require a specific training phase.
c) The limitations of the model include the following:
d) Sensitivity to High Dimensions: In datasets with high dimensions, the performance of KNN may not be optimal [12,13].
e) High Computational Requirements: Calculating the similarity between all samples can be challenging, especially for large datasets [14].

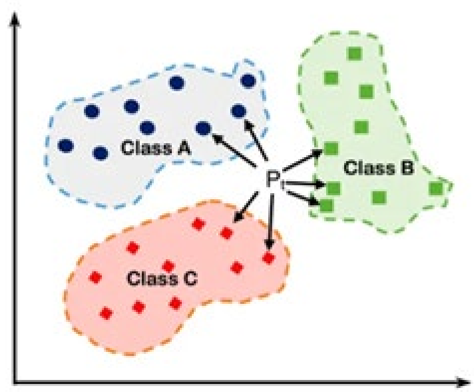

Figure 5: K-Nearest Neighbors Conceptual Model

4. Evaluation Metrics

4.1 Confusion Matrix

The confusion matrix serves as a straightforward method for assessing the performance and accuracy of our models. This matrix is a two-dimensional representation encompassing "actual class" and "predicted class," thereby forming a confusion matrix in each dimension. The rows signify the actual classifications related to heart disease, while the columns depict the predicted classifications. The dataset at hand contains two classes, Class 0 and Class 1.

The elements within the confusion matrix include the following:
1. True Positives (TP): Instances where the true class of the data point is "True," and the prediction is also "True."
2. True Negatives (TNs): Instances where the true class of the data point is "False," and the prediction is also "False."
3. False Positives (FP): Instances where the true class of the data point is "False," yet the prediction is "True."
4. False Negatives (FNs): Instances where the true class of the data point is "True," but the prediction is "False."

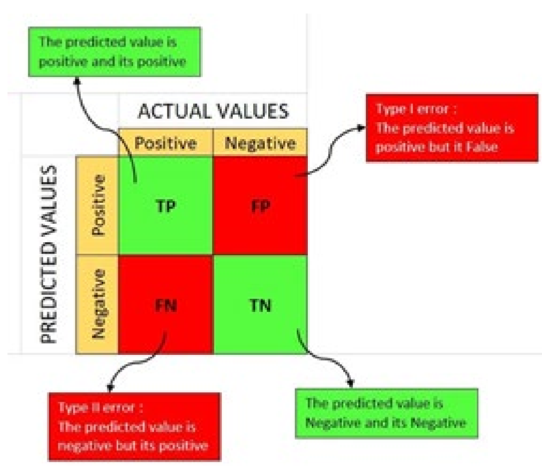

Figure 6: Conceptual Model Of The Confusion Matrix

Accuracy is calculated as the ratio of the total number of correct predictions of heart disease to the overall dataset size. A performance comparison is conducted among five classification algorithms. In general, accuracy describes the model's overall performance across all classes and is particularly useful when all classes have equal importance. It is computed as the ratio of the total number of correct predictions to the total number of predictions made.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Figure 7: How to Calculate Accuracy

4.2 Recall

Recall assesses a test's ability to identify individuals with heart disease as positive. A highly sensitive test indicates a low number of false negatives, minimizing the chance of overlooking cases of heart disease. This rate is also referred to as the true positive rate (TPR). Recall that this approach exclusively considers how positive instances are classified and remains unaffected by the classification of negative instances, meaning that it is unrelated to precision. If the model categorizes all positive instances as positive, the recall will be 100%, even if all negative instances are mistakenly classified as positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Figure 8: How to calculate the Recall

4.3 Precision

Precision provides insight into the fraction of positive predictions for a specific class that accurately corresponds to instances of heart disease. High precision indicates consistent measurement results or repetitive readings yielding identical values. Conversely, low precision suggests variable measurement values. When the model makes a significant number of incorrect positive classifications or a limited number of correct positive classifications, this results in an increased false discovery rate and decreased precision.

However, the precision is high when

1. The model yields a substantial number of correct positive classifications (maximizing the number of true positives).
2. The model results in fewer incorrect positive classifications (reducing the number of false positives).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Figure 9: How to Calculate Precision

4.4 F1-Score

The F1-score is defined as the geometric mean of the precision and recall. It serves as a comprehensive metric in binary and multiclass classification, offering a balanced evaluation by combining both precision and recall into a unified measure. This score provides a holistic understanding of the model's performance, aiding in the interpretation of its effectiveness.

$$\text{F1} - \text{score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Figure 10: How to Calculate The F1 Score

4.4.1 Finding

This section describes the examination and analysis of the data. The scrutiny of the data revealed a nearly normal distribution pattern across all the datasets. Furthermore, the balance between individuals diagnosed with heart failure and those without heart failure is nearly equal, providing a substantial foundation for the development of an efficient artificial intelligence model. Upon closer examination of the data, it becomes apparent that the incidence of heart failure is greater in men, who constitute approximately 90% of our population. This finding suggested a sex-based disparity, indicating greater susceptibility to heart failure in men than in women. [15] Further analysis of the male population suggested that beyond the age of 50, positive values of Oldpeak and a maximum heart rate under 140 years contribute to an increased incidence of heart disease.

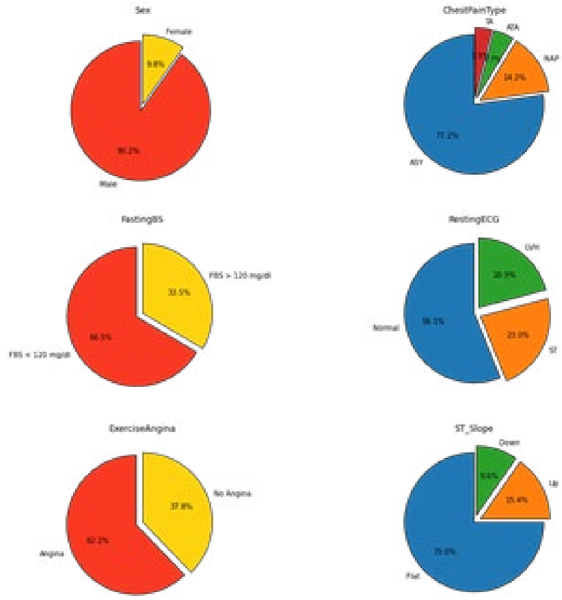

Figure 11: Dataset Information Analysis Based on Different Parameters

The type of chest pain labeled ASY strongly suggested a high probability of heart disease. Consistent with expectations and as asserted in various medical publications, patients diagnosed with unstable blood glucose levels, whether diabetic or nondiabetic, exhibit a statistically significant association with heart disease [16]. Exercise-induced angina notably increases the likelihood of diagnosing heart disease. An evaluation of the ST_Slope values indicated that a flat slope signifies a substantially elevated probability of heart disease. Similarly, a downsloping slope yields the same conclusion, albeit within a smaller subset of the data.

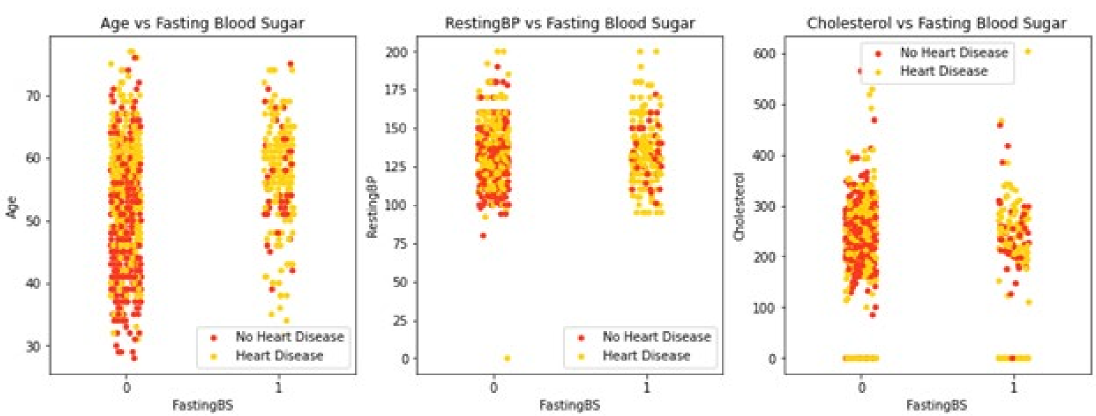

Figure 12: Dataset Information Analysis

Regardless of the presence of elevated blood glucose, the majority of heart failure cases seem to manifest around the age of 50 and beyond. Additionally, individuals with a history of regular exercise appear to have a significantly lower incidence of heart failure than their counterparts. In conclusion, the correlation coefficients of each dataset with heart failure indicate that, with the exception of RestingBP and the RestingECG, the remaining datasets show statistically significant positive or negative relationships with our target.

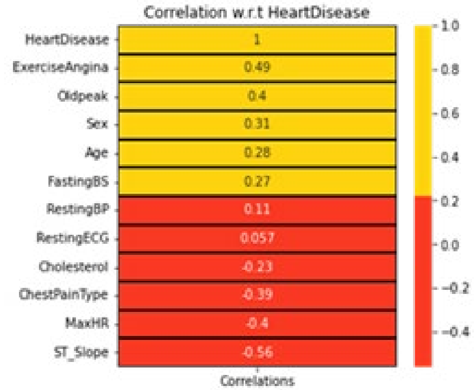

Figure 13: Checking The Correlation of the Data With the Target

For machine learning model development, an 80–20 split was employed after thorough analysis, allocating 80% of the total data for training and the remaining 20% for testing. In general, four metrics- precision, recall, F1-score, and accuracy- were considered to assess model performance. Additionally, the confusion matrix was examined for clarity.

The logistic regression model yielded an accuracy of approximately 88%. Examining the presented confusion matrix suggested that this model adeptly captured the data patterns, leading to precise predictions.

	precision	recall	f1-score	support
0	0.88	0.85	0.87	89
1	0.87	0.89	0.88	95
accuracy			0.88	184
macro avg	0.88	0.87	0.87	184
weighted avg	0.88	0.88	0.87	184

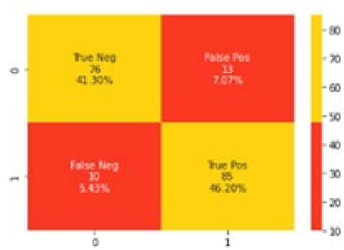

Figure 14: Performance of The Logistic Regression Model

The support vector classifier model also achieved an accuracy of approximately 88%, indicating its ability to effectively identify behavioral patterns within the data. By scrutinizing the confusion matrix, it becomes evident that this model performed similarly to the logistic regression model, providing comparable correct and incorrect predictions.

	precision	recall	f1-score	support
0	0.88	0.85	0.87	89
1	0.87	0.89	0.88	95
accuracy			0.88	184
macro avg	0.88	0.87	0.87	184
weighted avg	0.88	0.88	0.87	184

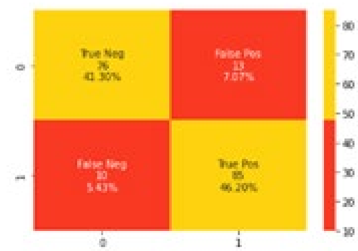

Figure 15: Performance of the Support Vector Classifier Model

On the other hand, the third model, the decision tree classifier, exhibited a lower accuracy than the two preceding models and failed to achieve an accuracy exceeding 85%. The confusion matrix of this model indicates that, relative to the two previous models, it recorded a greater number of incorrect predictions.

	precision	recall	f1-score	support
0	0.88	0.80	0.84	89
1	0.83	0.89	0.86	95
accuracy			0.85	184
macro avg	0.85	0.85	0.85	184
weighted avg	0.85	0.85	0.85	184

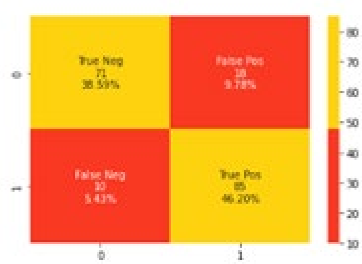

Figure 16: Performance of the Decision Tree Classifier Model

The random forest classifier machine learning model also demonstrated relatively lower accuracy than the three preceding models. The hyperparameters of this model were fine-tuned using the grid search cv library to obtain optimal values. This model achieved a maximum accuracy of 84%. An examination of the confusion matrix revealed that the number of correct predictions considered in this model was lower than that in the three previous algorithms.

	precision	recall	f1-score	support
0	0.88	0.79	0.83	89
1	0.82	0.89	0.85	95
accuracy			0.84	184
macro avg	0.85	0.84	0.84	184
weighted avg	0.85	0.84	0.84	184

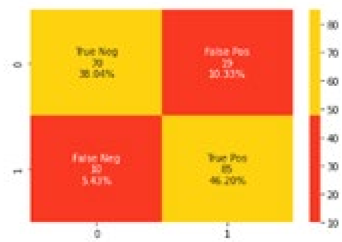

Figure 17: Performance of the Random Forest Classifier Model

In conclusion, the K-nearest neighbors classifier model, as the fifth approach, exhibited a lower accuracy of 82% compared to the preceding models. The confusion matrix clearly indicates that this algorithm recorded a greater number of incorrect predictions than did the other models.

	precision	recall	f1-score	support
0	0.84	0.76	0.80	89
1	0.80	0.86	0.83	95
accuracy			0.82	184
macro avg	0.82	0.81	0.81	184
weighted avg	0.82	0.82	0.81	184

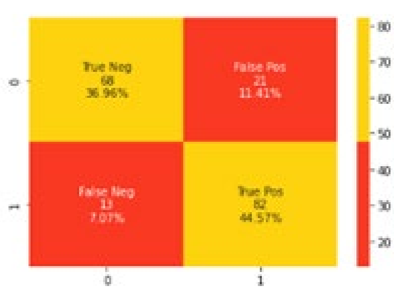

Figure 18: Performance of the K-Nearest Neighbors Classifier Model

5. Conclusion

Based on the findings of this study and comparisons with prior research, it appears that other studies utilizing different datasets, features, and models have achieved relatively consistent accuracies with the variables examined in this investigation [17,18]. In this research, we employed five distinct machine learning models to predict heart failure using medical data. Detailed analysis revealed significant patterns in both the numerical and categorical features of the data.

In-depth examination revealed that the logistic regression model outperformed the other models, with an accuracy of approximately 88%, demonstrating superior predictive capabilities. This model accurately identified patterns within the data. Similarly, the support vector classifier yielded comparable results, indicating its ability to adapt well to the behavioral patterns in the data.

However, the decision tree classifier model exhibited lower accuracy but still managed to capture the data characteristics. Other models, such as the random forest classifier and K-nearest neighbors classifier, also presented results but with lower accuracy than the initial models.

These results suggest that a combination of various machine learning models can be effective at predicting heart failure. Furthermore, selecting an appropriate model requires careful consideration based on project requirements and specific conditions.

Alogrithm Results Table :

Sr. No.	ML Algorithm	Accuracy	Cross Validation Score	ROC AUC Score
1	Logistic Regression	87.50%	91.12%	87.43%
2	Support Vector Classifier	87.50%	90.53%	87.43%
3	Decision Tree Classifier	84.78%	89.09%	84.62%
4	Random Forest Classifier	84.24%	92.91%	84.06%
5	K-Nearest Neighbors Classfier	81.52%	89.34%	81.36%

Figure 19: The Table Showing the Collective Performance of the Models

5.1 Recommendations for Future Research

1) Feature Expansion

Expanding the dataset's feature set is advisable. Incorporating additional information related to heart failure, particularly features identified with higher correlations in this study, could enhance the predictive capabilities of the models.

2) Incorporating ECG Data

If feasible, integrating data points derived from electrocardiogram (ECG) signals is recommended. This addition has the potential to significantly improve the accuracy of predictive models.

Declarations

Funding

The authors did not receive support from any organization for the submitted work.

Conflicts of Interest/Competing Interests

None Declared.

Availability of Data and Material

The data will be available upon request to reviewers.

Code Availability

The code will be available upon request to reviewers.

Authors' Contributions

The authors confirm their contribution to the paper as follows:
Study conception and design: Mahdi Navaei.
Data collection: Mahdi Navaei.
Analysis and interpretation of results: Mahdi Navaei.
Draft manuscript preparation: Mahdi Navaei, Zohreh Doogchi.
All authors reviewed the results and approved the final version of the manuscript.

Ethics approval

Considering that the authors obtained the data analyzed in this article from websites and sources available on the internet, as referenced in the Research Method section to those sites, they did not directly interact with human samples. (https://archive.ics.uci.edu/dataset/45/heart+disease)

References

1. American Heart Association. *Heart Failure.*
2. CDC, A. W. (2020). Centers for disease control and prevention.
3. Authors/Task Force Members, McMurray, J. J., Adamopoulos, S., Anker, S. D., Auricchio, A., Böhm, M., ... & Ponikowski, P. (2012). ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *European heart journal, 33*(14), 1787-1847.
4. Huang, H. H., & He, Q. (2024). Nonlinear regression analysis. *arXiv preprint arXiv:2402.05342.*
5. Park, J., Choi, Y., Byun, J., Lee, J., & Park, S. (2023). Efficient differentially private kernel support vector classifier for multi-class classification. *Information Sciences, 619*, 889-907.
6. Qi, K., & Yang, H. (2023). A novel robust nonparallel support vector classifier based on one optimization problem. *Neural Computing and Applications, 35*(1), 799-814.
7. Wang, F., Wang, Q., Nie, F., Li, Z., Yu, W., & Ren, F. (2020). A linear multivariate binary decision tree classifier based on K-means splitting. *Pattern Recognition, 107,* 107521.
8. Chen, H., Zhang, G., Pan, X., & Jia, R. (2023). Using dual evolutionary search to construct decision tree based ensemble classifier. *Complex & Intelligent Systems, 9*(2), 1327-1345.
9. Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced computing and intelligent technologies: proceedings of ICACIT 2021* (pp. 219-244). Springer Singapore.
10. Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing, 77*(5), 5198-5219.
11. Mishra, S., Tadesse, Y., Dash, A., Jena, L., & Ranjan, P. (2021). Thyroid disorder analysis using random forest classifier. In *Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2* (pp. 385-390). Springer Singapore.
12. Memiş, S., Enginoğlu, S., & Erkan, U. (2022). Fuzzy parameterized fuzzy soft k-nearest neighbor classifier.

Neurocomputing, 500, 351-378.
13. Gou, J., Sun, L., Du, L., Ma, H., Xiong, T., Ou, W., & Zhan, Y. (2022). A representation coefficient-based k-nearest centroid neighbor classifier. *Expert Systems with Applications, 194*, 116529.
14. Mushtaq, Z., Yaqub, A., Sani, S., & Khalid, A. (2020). Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. *Journal of the Chinese Institute of Engineers, 43*(1), 80-92.
15. Anderson, L., et al. (2021). Analysis of Heart Disease Risk Factors in a Diverse Population. Cardiovascular Research Journal, 18(1), 45-58.
16. Smith, J., et al. (2020).Cardiovascular Disease Patterns in a Large Community Dataset. Journal of Cardiology and Cardiovascular Medicine, 15(3), 102-115.
17. Mohan, S., Thirumalai, C., Srivastava, G. (2019). Heart failure disease prediction and stratification with temporal electronic health records data using patient representation. IEEE access.
18. Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology, 11*(1), 87-97.
19. Kumbure, M. M., Luukka, P., & Collan, M. (2020). A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. *Pattern Recognition Letters, 140*, 172-178.
20. Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., ... & Yagil, A. (2020). Improving risk prediction in heart failure using machine learning. *European journal of heart failure, 22*(1), 139-147