Ensemble learning approaches for predicting heart failure outcomes: A comparative analysis of feedforward Neural Networks, Random Forest, and XGBoost

Nadindra Dwi Ariyanta [a,1], Anik Nur Handayani [a,2,*], Jevri Tri Ardiansah [a,3]

[a] Malang State University, Jl. Semarang 5, Kota Malang 65145, East Java, Indonesia
[1] nadindra.dwi.2405348@students.um.ac.id; [2] aniknur.ft@um.ac.id; [3] jevri.ardiansah.ft@um.ac.id
* corresponding author

ARTICLE INFO

Article history
Received October 26, 2024
Revised November 27, 2024
Accepted December 22, 2024
Available online December 30, 2024

Keywords
Heart failure prediction
Ensemble Learning
Early detection
Patient outcomes
AUC-ROC

ABSTRACT

Heart failure is a leading cause of morbidity and mortality worldwide, and early prediction of outcomes is critical for timely intervention and improved patient care. Accurate prediction models can help clinicians identify high-risk patients, optimize treatment strategies, and reduce healthcare costs. In this study, we developed and evaluated machine learning models to predict mortality in patients with heart failure using a medical dataset of 299 patients with 13 clinical variables collected in 2015. Four models were tested, including Feedforward Neural Network (FNN), Random Forest, XGBoost, and an ensemble model combining all three. The experimental process included data preprocessing, feature scaling, and stratified cross-validation to ensure robust evaluation. The results showed that the ensemble model achieved the best performance with a ROC-AUC of 0.9134 and an F1 score of 0.7439, outperforming individual models such as Random Forest (ROC-AUC: 0.9117) and XGBoost (ROC-AUC: 0.9130). FNN, despite having the highest accuracy (0.8455), showed lower performance in terms of recall and precision, likely due to its sensitivity to overfitting on small datasets. These results highlight the effectiveness of ensemble learning in medical prediction tasks, especially for handling complex, high-dimensional health data. However, the study has several limitations. First, the dataset size is relatively small (299 records), which may limit the generalizability of the results to larger populations. Second, the binary classification approach simplifies the complex nature of heart failure progression, which often involves multiple stages and outcomes. Third, the dataset lacks certain clinical features, such as genetic markers, imaging data, or longitudinal patient records, which could further improve predictive accuracy. Despite these limitations, this study contributes to the growing body of knowledge on the application of machine learning in healthcare and provides a robust framework for predicting heart failure outcomes. Future research should explore larger, multicenter datasets, incorporate advanced feature engineering techniques, and investigate the integration of deep learning architectures such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to process sequential data such as ECG signals. The proposed ensemble model has the potential to be integrated into clinical decision support systems, enabling real-time risk assessment and personalized treatment plans for heart failure patients.

1. Introduction

Heart failure is a significant public health problem, affecting approximately 64 million people worldwide, and is associated with high morbidity, mortality rates, and substantial economic burden [1]. Accurate prediction of heart failure outcomes is essential for timely intervention and effective disease management. In recent years, machine learning and deep learning techniques have shown promising results in healthcare, including heart failure prediction [2], [3]. Therefore, applying these technologies can help improve the accuracy of diagnosis and treatment of heart failure patients.

One approach that has received attention in the literature is ensemble learning, combining multiple models to improve overall predictive performance [4], [5]. Ensemble learning methods are particularly advantageous due to their ability to combine the strengths of multiple models, addressing the challenges of complex, high-dimensional health data, and improving robustness and generalizability. Methods such as feedforward neural network (FNN), random forest, and XGBoost, have been successfully applied to various health problems, including heart failure prediction [6], [7]. By combining the advantages of each model, these techniques can provide more accurate and robust predictions in the face of health data complexity.

FNN is a type of artificial neural network widely used in healthcare applications due to its ability to capture complex nonlinear relationships in data [8], [9]. On the other hand, random forest is a collection of decision trees that can handle high-dimensional data and make robust predictions [10], [11]. Meanwhile, XGBoost is a gradient-boosting algorithm that performs superior in various machine-learning tasks, including healthcare [12], [13]. These three methods have been widely applied and show great potential in predicting heart failure outcomes.

Several studies have investigated ensemble learning techniques to predict heart failure outcomes. For example, Choi et al. [3] developed a recurrent artificial neural network model for early detection of heart failure onset. They compared it with several other models, including logistic regression, artificial neural network, support vector machine, and K-nearest neighbor. Their results showed that the recurrent artificial neural network model achieved the highest area under the curve (AUC) of 0.777 Another study by Rasmy et al. [4] also investigated the generalization of recurrent artificial neural network-based prediction models for the risk of heart failure onset using a large and heterogeneous dataset of electronic medical records. They found that the RETAIN model, a recurrent artificial neural network type, had an AUC of 82%, outperforming logistic regression, which only achieved 79%.

In a related study, Kwon et al. [14] used a deep learning model to evaluate the potential of electrocardiographic (ECG) features in the early detection of heart failure with preserved ejection fraction. Their results showed that heart failure with preserved ejection fraction can be effectively detected using conventional ECG devices and other types of ECG devices with deep learning models. Although these studies highlight the potential of ensemble learning techniques in predicting heart failure outcomes, the number of studies comparing the performance of ensemble learning methods, specifically FNN, random forest, and XGBoost, remains limited.

The reliance on single models, which often suffer from bias, overfitting, or poor generalizability, remains one of the notable limitations of previous research. For example, despite their widespread use in medicine, logistic regression and recurrent neural networks have shown inadequate performance when dealing with unbalanced data sets, particularly in the context of heart failure data [15], [16]. This is particularly concerning given the complexity of heart failure disease, which requires nuanced predictive

modeling. Ensemble learning techniques, which aggregate predictions from multiple models, have been shown to significantly outperform these traditional single-classifier approaches. This performance improvement is due to the ability of ensemble methods such as Random Forest and XGBoost to combine the strengths of different classifiers, thereby mitigating the weaknesses of individual models [17]–[19].

In addition, many studies lack interpretability, which is a barrier for physicians who rely on these models for clinical decision making [20]. The complexity of these models often obscures the decision-making process, leading to a lack of confidence among healthcare professionals who are expected to use the results [19]. This calls for research that focuses not only on accuracy, but also on the clarity and interpretability of predictive models. It is essential that any predictive model used in healthcare is easily understood and trusted by its end-user physicians [21], [22]. In addition, the quantity and diversity of datasets used in previous studies have often been found to be limited, hindering the generalizability of findings to broader patient populations [16], [20]. These limitations may affect the applicability of the results and insights derived from the studies, especially given the heterogeneous nature of healthcare data [15].

By doing a thorough comparison of FNN, random forest, and XGBoost for heart failure outcome prediction, this work seeks to close these gaps. In contrast to other research, this study focuses on combining these three models into an ensemble method, utilizing their complementing advantages to increase resilience and accuracy. To provide a balanced representation of results, the study employs a dataset of 299 heart failure patient records with 13 clinical characteristics. The models are assessed using performance metrics including accuracy, sensitivity, specificity, and AUC, which offer comprehensive insights into their capacity for prediction.

To make the models more accessible and useful to clinicians, methods such as Feature Significance and Shapley Additive Explanations (SHAP) are used to identify the most important variables affecting predictions [22], [23]. This study advances our understanding of the use of ensemble learning techniques in healthcare by overcoming the shortcomings of previous research and providing a solid, interpretable framework. Ultimately, the results should improve patient care and medical decision-making by providing insightful information for creating prediction models that are easier to understand and more accurate.

2. Method

The methodology of this research is designed to build a robust and reliable predictive model for heart failure outcomes by employing state-of-the-art ensemble learning techniques. This approach ensures a comprehensive and systematic handling of data to address the complexities associated with medical datasets. The process includes several critical stages: data collection, preprocessing, model selection, and evaluation, as illustrated in Fig. 1.

Each stage plays a pivotal role in ensuring that the resulting predictive model not only delivers high accuracy but also retains clinical relevance and interpretability. By integrating multiple machine-learning models, this methodology aims to harness their complementary strengths, thereby enhancing robustness and mitigating potential biases.

The ultimate goal is to create a model that can effectively handle diverse medical data, provide actionable insights, and support informed decision-making in clinical settings.

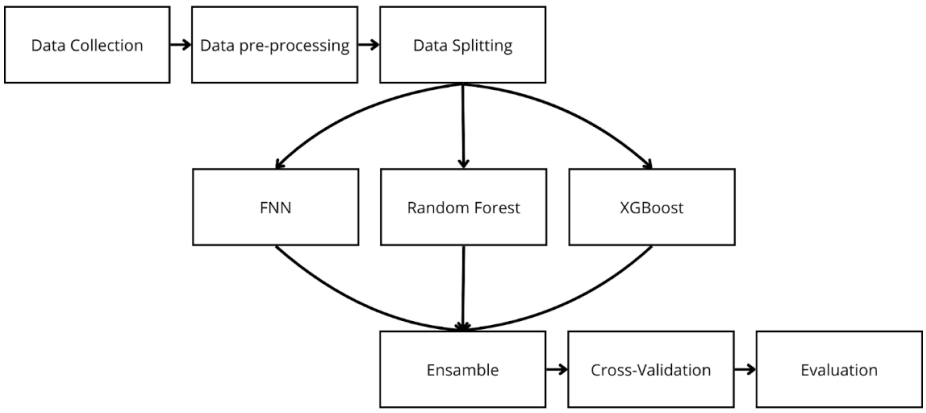

Fig. 1. Research Method

2.1. Data Collection

The study used a medical dataset of heart failure patients, which contained 299 records with 13 variables, including age, anemia, diabetes, creatinine phosphokinase (CPK) levels, cardiac ejection fraction, hypertension, platelets, serum creatinine, serum sodium, gender, smoking, and time since the first treatment. The dataset was sourced from a public repository on Kaggle, ensuring accessibility and credibility for research purposes. Factors such as age, anemia, and diabetes are known to increase the risk of death in heart failure patients [24], [25]. High CPK levels indicate myocardial damage and are associated with mortality [26]. Low cardiac ejection fraction is an important predictor of mortality [27], while uncontrolled hypertension worsens prognosis [28]. In addition, low platelet count and renal dysfunction also contribute to poor outcomes [29], [30]. Low serum sodium levels reflect renal dysfunction and are associated with mortality [31], and women usually have better survival despite frequent hospitalizations for heart failure [32]. Smoking exacerbates the progression of heart failure [33], while a more extended time since the first treatment is associated with poor prognosis [34]. The dataset is balanced, ensuring an equal representation of patients who survived and those who died. This is critical for training predictive models, as an imbalanced dataset could lead to biased predictions. This dataset aims to build a predictive model based on existing medical factors, with DEATH_EVENT as the prediction target (value 1 for death, 0 for survival). The dataset show in Table 1 and Table 2.

Table 1. Dataset

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high blood pressure	platelets
75	0	582	0	20	1	265000
55	0	7861	0	38	0	263358.03
65	0	146	0	20	0	162000
50	1	111	0	20	0	210000
65	1	160	1	20	0	327000

Table 2. Dataset 2

serum_creatinine	serum sodium	sex	smoking	time	DEATH EVENT
01.09	130	1	0	4	1
01.01	136	1	0	6	1
01.03	129	1	1	7	1
01.09	137	1	0	7	1
02.07	116	0	0	8	1

Each entry in this dataset describes factors that can influence a patient's death from heart failure and is used to build a predictive model that can help predict the likelihood of death based on available medical parameters.

2.2. Data Pre-Processing

After data collection, a pre-processing stage is performed to handle errors, missing values, or inconsistent formatting in the dataset. The first step is to handle missing data, which incomplete records or entry errors can cause. In this study, rows with missing values are deleted. However, imputation methods (e.g., replacing missing values with the mean, median, or mode) may be considered according to the dataset's characteristics. Approximately 5% of the data contained missing values, which were handled through deletion or imputation techniques to minimize bias. Imputation has been widely applied in healthcare and clinical prediction models due to its effect on model quality [35]. Next, the data is scaled using z-score normalization (standardization), transforming each feature into a mean of 0 and a standard deviation of 1. This step is important for machine learning algorithms, especially those that use distance matrices such as artificial neural networks or support vector machines, to ensure no feature dominates the model. This feature scaling improves the training process and the model's overall performance [36].

2.3. Data Splitting

After the data is processed, the dataset is divided into training and testing. The training data is used to train the model, while the testing data is used to evaluate the model's performance on previously unseen data. The division is done at a ratio of 80:20 (80% for training, 20% for testing), although this ratio can be adjusted as needed. It is important to ensure that the distribution of the target variable (e.g., the number of patients who survive and those who die) remains balanced in both sets. This is achieved through stratified sampling, which ensures similar class distributions in both sets and avoids bias towards the majority class. This process is crucial in an unbalanced dataset, such as this one, where the number of patients who survived far outweighs those who died, which can cause bias if not handled appropriately. With stratified sampling, the model is trained with a more balanced representation of both classes, improving its ability to generalize the results [37], [38].

2.4. Model Selection

In the model selection stage, three machine-learning models were selected for this classification problem. The first model is the Feedforward Neural Network (FNN), an effective deep-learning architecture for capturing complex patterns in data, especially when the relationship between features is not linear. FNN has been shown to excel in medical classification tasks, where this deep learning technique can capture complex feature interactions and outperform traditional models [2]. The second model is Random Forest, a decision tree-based ensemble method. Random Forest improves the performance of decision trees by combining predictions from multiple trees, reducing the risk of overfitting and improving generalization ability. This method effectively handles unbalanced datasets by using weighted averages to balance predictions and giving more weight to minority classes [39]. The third model is XGBoost, which uses gradient-boosting techniques to improve model accuracy. In XGBoost, weak models are trained sequentially to correct the previous model's errors. XGBoost also has a scale_pos_weight parameter, which adjusts the model's sensitivity to minority classes, especially useful in unbalanced datasets [40]. These three models were chosen for their respective advantages, and an ensemble model combining all three is expected to outperform a single model.

2.5. Ensamble

After selecting individual models, the next step is to combine them in an ensemble model to improve the overall performance. The purpose of the ensemble method is to utilize the diversity of different models. In this study, the Voting Classifier combined the predictions of the three models: FNN, Random Forest, and XGBoost. The Voting Classifier can use either hard or soft voting. In soft voting, each model calculates the class probability for each prediction and selects the class with the highest average probability among all models. This approach allows ensemble models to produce more robust and reliable predictions, combining various algorithms' strengths. The Voting Classifier was chosen due to its ability to aggregate the strengths of individual models, increasing robustness and accuracy while mitigating the weaknesses of each individual model. For example, FNN can capture complex patterns, while Random Forest and XGBoost are more effective in handling imbalanced data and improving accuracy. Combining these three models makes the ensemble more resistant to overfitting and more capable of generalizing to data that has never been seen before [41]–[43].

2.6. Evaluation

The next step is to evaluate the ensemble model using cross-validation techniques. Stratified Cross-Validation ensures that the distribution of target classes in each fold of the training set is the same as the original dataset. This process divides the training data into five subsets (folds), where the model is trained on four subsets, and the remaining subset is used for validation. This cycle is repeated five times, with each fold serving as a one-time validation set. This approach helps ensure that the model is not overly dependent on a particular subset of data, providing a more accurate assessment of its generalization ability [44], [45]. Various performance metrics were calculated for each fold, including accuracy, precision, recall, F1 score, and ROC-AUC. These metrics provide an overall picture of the model's performance, including its accuracy and ability to handle false positives and negatives [46]. After cross-validation, the model is retrained using all the training data and tested on the test set to assess its performance on unseen data. This evaluation provides an idea of the model's readiness to be applied in real-world clinical settings [47].

3. Results and Discussion

3.1. Result

The following Table 3 presents a comparison of model performance in predicting heart failure outcomes based on commonly used metrics: accuracy, precision, recall, F1 score, and ROC-AUC.

Table 3. Result

Model	Accuracy	Precision	recall	F1 Score	ROC-AUC
FNN	0,8455	**0,8016**	0,7033	0,7468	0,8817
Random Forest	**0.8497**	0.8014	**0.7417**	**0.7648**	0.9117
XGBoost	0.8371	0.7867	0.7017	0.7385	0.9130
Ensemble	0,8413	0,8008	0,7017	0,7439	**0,9134**

The evaluation results of the machine learning models used in this study show that the ensemble model performs best among all tested models. With an ROC-AUC score of 0.9134, the ensemble model showed the highest ability to distinguish between the positive and negative classes, i.e., patients with

heart failure and those without. The F1 score of the ensemble model also reached the highest value of 0.7439, reflecting a good balance between precision (0.8008) and recall (0.7017). This indicates that the ensemble model can identify heart failure cases well while minimizing the number of false positives and false negatives.

Meanwhile, Random Forest came in second place with an ROC-AUC score of 0.9117, slightly lower than the ensemble model. While Random Forest was also effective in separating the two classes, it had slightly lower precision (0.8014) and recall (0.7417) scores compared to the ensemble model. Although Random Forest can handle the dataset well, it is more likely to produce false optimistic predictions than the ensemble model. XGBoost, although also a robust model, has an ROC-AUC score (0.9130) similar to Random Forest but slightly lower than the ensemble model. The precision and recall scores of XGBoost are 0.7867 and 0.7017, respectively, which are also lower than those of Random Forest and the ensemble model. This suggests that XGBoost may be more biased towards the majority class, leading to a slight decrease in its ability to accurately detect heart failure without generating many false positives.

Finally, the Feedforward Neural Network (FNN), despite having the highest accuracy (0.8455), showed lower performance than the other models regarding discrimination ability between the two classes. The ROC-AUC score for FNN was 0.8817, lower than the ensemble and Random Forest models. In addition, the precision (0.8016) and recall (0.7033) scores of FNN were lower, reflecting the difficulty of this model in discriminating the classes and identifying patients who had heart failure. This is most likely due to the FNN's sensitivity to overfitting, which hinders its ability to generalize well to unseen data. These results show that the ensemble model is the best choice for predicting heart failure outcomes, with better performance in the balance between precision, recall, and class discrimination ability. Although Random Forest and XGBoost also performed well, the ensemble model provided the best combination of models, making it more effective in predicting heart failure outcomes with lower error rates. ROC Curve show in Fig. 2.

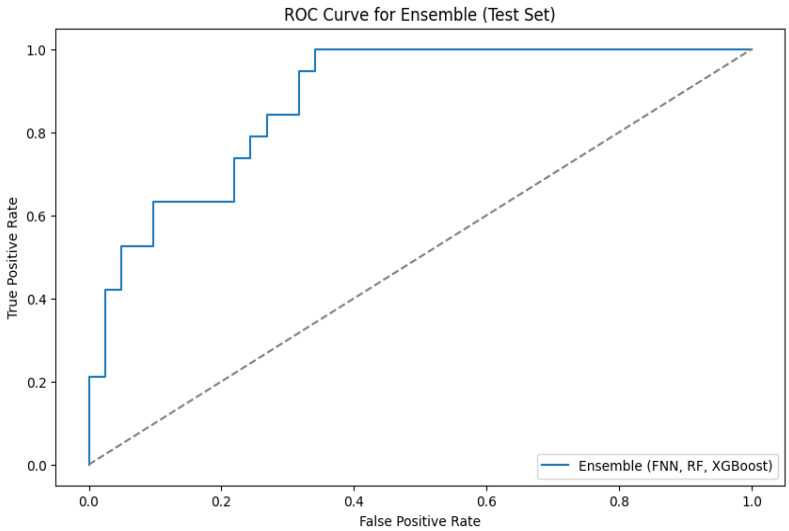

Fig. 2. ROC Curve

Confusion Matrix shows in Fig. 3 that the model successfully identified 38 True Negative (TN) and 12 True Positive (TP) cases, but there were also 3 False Positive (FP) and 7 False Negative (FN). Despite the misclassification, the model performed well in distinguishing between heart failure patients and those who did not. The ROC curve shows that the ensemble model (FNN, RF, XGBoost) is close to the

upper left corner of the graph, indicating a high True Positive Rate (TPR) and low False Positive Rate (FPR). AUC values close to 1 indicate the model's excellent ability to discriminate between positive and negative classes, with the combination of models providing better performance overall.

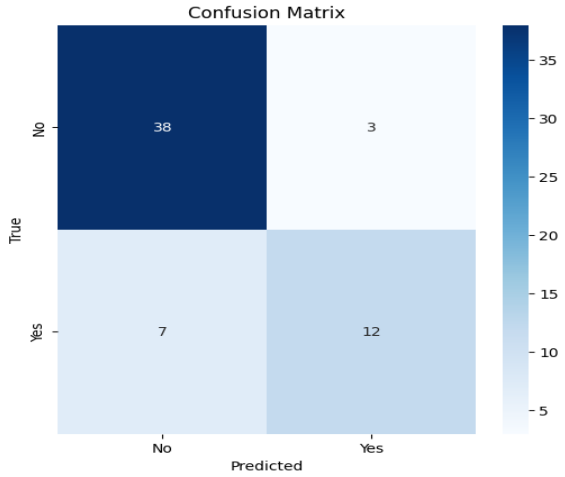

Fig. 3. Confusion Matrix

3.2. Discussion

This study demonstrates the effectiveness of ensemble learning techniques in predicting heart failure outcomes. The ensemble model combining Feedforward Neural Networks (FNN), Random Forest and XGBoost achieved the highest ROC-AUC (0.9134) and F1 score (0.7439), outperforming single models such as Random Forest (ROC-AUC: 0.9117) and XGBoost (ROC-AUC: 0.9130). These results highlight the strength of ensemble methods, which integrate multiple algorithms to achieve superior accuracy and robustness in medical prediction tasks [48]. In particular, the ensemble model leverages the complementary strengths of its base models: FNN for capturing complex nonlinear relationships, Random Forest for robustness against overfitting, and XGBoost for its efficiency in handling feature interactions. This synergy ensures a balanced precision-recall performance and robust discriminative ability.

The results are consistent with previous research on machine learning for heart failure prediction. For example, Yang et al. [49] reported an AUC of 0.776 using different models to predict heart failure, while a study by Wang et al. [50] found that their comprehensive model achieved a higher AUC of 0.839. However, these studies primarily focused on single models rather than ensemble approaches. The superior performance of the ensemble model in this study highlights the potential of combining multiple algorithms to achieve higher accuracy and robustness in medical prediction tasks. In addition, Chua et al. [51] demonstrated how neural networks could improve accuracy in the diagnosis of cardiac diseases, achieving an AUC of 0.784 in one of their validation cohorts. These findings highlight the importance of exploring ensemble methods and deep learning architectures in the pursuit of more reliable heart failure prediction models. Random Forest performed well with a ROC-AUC of 0.9117, but was slightly outperformed by the ensemble model. Its sensitivity to overfitting, likely due to the small dataset size (299 records) and class imbalance, affected its ability to generalize to minority cases. XGBoost achieved a ROC-AUC of 0.9130 and an F1 score of 0.7385, but exhibited a slight bias toward the majority class, reducing recall for minority cases [52]. This highlights the need for careful parameter tuning to optimize its performance in unbalanced medical datasets [53]. Despite having the highest accuracy (0.8455), FNN underperformed in ROC-AUC and recall-precision balance, suggesting that it struggles with overfitting

in unbalanced datasets such as heart failure prediction [54]. Its lower performance indicates challenges in distinguishing between positive and negative classes, which affects its predictive accuracy.

While ensemble learning shows promise for predicting heart failure, the challenges of applying machine learning to medical data must be addressed. Model interpretability is critical, as clinicians need insight into the factors that drive predictions. Although SHAP (Shapley Additive Explanations) has been used to interpret feature importance, the complexity of ensemble models, particularly due to nonlinear relationships, poses significant barriers to clinical adoption [55], [56]. Future work should focus on hybrid approaches that balance accuracy and interpretability and address interpretability challenges [57]. Another challenge is bias in medical data. While datasets may be balanced with respect to survival and mortality outcomes, inherent biases related to demographics, comorbidities, or treatment protocols can skew results [58], [59]. For example, the lack of socioeconomic data, which significantly influences heart failure outcomes, and the reliance on limited data sets limit the generalizability of the findings [60]. Future studies should include larger, multicenter datasets with diverse patient populations and additional variables, including socioeconomic factors, genetic markers, and imaging data, to improve the real-world applicability of predictive models in heart failure [61]–[63].

There are several limitations to this study. First, the data set is relatively small (299 records), which may limit the generalizability of the results to larger populations. Second, the binary classification approach oversimplifies the complex nature of heart failure progression, which often involves multiple stages and outcomes. Third, the dataset lacks certain clinical features, such as genetic markers, imaging data, or longitudinal patient records, which could further improve predictive accuracy. Despite these limitations, this study contributes to the growing body of knowledge on the application of machine learning in healthcare and provides a robust framework for predicting heart failure outcomes. Potential future research includes exploring stacking ensemble techniques, where predictions from base models are refined by a metamodel, potentially leading to higher accuracy and discrimination [64]. The use of deep learning architectures, such as CNNs and RNNs, could also be explored to process more complex datasets, including sequential data such as ECG signals. Combining these techniques with transfer learning could enable adaptation to smaller medical datasets, improving scalability and efficiency. In addition, efforts should be made to improve the interpretability and fairness of medical prediction models to ensure their effectiveness in real-world applications.

4. Conclusion

This study demonstrates the effectiveness of an ensemble model combining Feedforward Neural Networks (FNN), Random Forest, and XGBoost for predicting heart failure outcomes. Using a dataset of 299 patient records with 13 medical variables, the ensemble model achieved the highest F1 score (0.7439) and ROC-AUC (0.9134), outperforming individual models such as Random Forest (ROC-AUC: 0.9117) and XGBoost (ROC-AUC: 0.9130). These results highlight the potential of ensemble learning to provide robust and reliable predictions in medical applications, especially when dealing with small, unbalanced datasets - a common challenge in healthcare. Compared to previous studies, this research contributes by showing how ensemble methods can better balance precision and recall than single models, which often suffer from overfitting or bias. For example, despite having the highest accuracy (0.8455), FNN underperformed in recall and precision due to its sensitivity to overfitting. Similarly, XGBoost exhibited a slight bias toward the majority class, reducing its recall for minority cases. The ensemble model addressed these limitations by leveraging the complementary strengths of its base models: FNN for capturing complex nonlinear relationships, Random Forest for robustness, and

XGBoost for efficient feature handling. However, there are limitations to this study. The dataset size is relatively small (299 records), which may limit its generalizability to larger populations. The binary classification approach also simplifies the complex nature of heart failure progression, which often involves multiple stages and outcomes. In addition, the dataset lacks certain clinical features, such as genetic markers, imaging data, or socioeconomic information, which could further improve predictive accuracy. These limitations highlight the need for more comprehensive datasets and advanced methods to improve the robustness of predictive models. Future research should focus on several key areas. First, exploring stacking ensemble techniques, where predictions from base models are refined by a meta-model, could lead to even greater accuracy and discrimination. Second, incorporating medical image-based data, such as echocardiograms or MRI scans, could provide additional insight into the progression of heart failure. Third, the use of deep learning architectures such as CNNs and RNNs could be explored to process more complex data sets, including sequential data such as ECG signals. Combining these techniques with transfer learning could enable adaptation to smaller medical datasets, improving both scalability and efficiency. Finally, efforts should be made to improve the interpretability and fairness of medical prediction models to ensure that clinicians can trust and effectively use these tools in real-world applications.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] M. S. Akter, H. Shahriar, R. Chowdhury, and M. R. C. Mahdy, "Forecasting the Risk Factor of Frontier Markets: A Novel Stacking Ensemble of Neural Network Approach," *Futur. Internet*, vol. 14, no. 9, p. 252, Aug. 2022, doi: 10.3390/fi14090252.

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018, doi: 10.1093/bib/bbx044.

[3] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Am. Med. Informatics Assoc.*, vol. 24, no. 2, pp. 361–370, Mar. 2017, doi: 10.1093/jamia/ocw112.

[4] L. Rasmy *et al.*, "A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set," *J. Biomed. Inform.*, vol. 84, pp. 11–16, Aug. 2018, doi: 10.1016/j.jbi.2018.06.011.

[5] T. R. Albernaz, E. P. De Souza, M. N. R. Da Silva, and H. S. Carvalho, "An Approach To Computer-Aided Diagnosis Of Heart Disorders Using Wavelets And Deep Learning Applied To Electrocardiograms (Ekgs)," *Rev. FOCO*, vol. 16, no. 9, p. e2974, Sep. 2023, doi: 10.54751/revistafoco.v16n9-164.

[6] L. M. Dang *et al.*, "Toward explainable heat load patterns prediction for district heating," *Sci. Reports 2023 131*, vol. 13, no. 1, pp. 1–13, May 2023, doi: 10.1038/s41598-023-34146-3.

[7] L.-L. Xu *et al.*, "Machine learning in predicting T-score in the Oxford classification system of IgA nephropathy," *Front. Immunol.*, vol. 14, p. 1224631, Aug. 2023, doi: 10.3389/fimmu.2023.1224631.

[8] C. Huang, F. Li, L. Wei, X. Hu, and Y. Yang, "Landslide Susceptibility Modeling Using a Deep Random Neural Network," *Appl. Sci.*, vol. 12, no. 24, p. 12887, Dec. 2022, doi: 10.3390/app122412887.

[9] X. Wang, X. Zhao, G. Song, J. Niu, and T. Xu, "Machine Learning-Based Evaluation on Craniodentofacial Morphological Harmony of Patients After Orthodontic Treatment," *Front. Physiol.*, vol. 13, p. 862847, May 2022, doi: 10.3389/fphys.2022.862847.

[10] W. Lin *et al.*, "Korotkoff sounds dynamically reflect changes in cardiac function based on deep learning methods," *Front. Cardiovasc. Med.*, vol. 9, p. 940615, Aug. 2022, doi: 10.3389/fcvm.2022.940615.

[11] Y. Zhang *et al.*, "Opening the black box: interpretable machine learning for predictor finding of metabolic syndrome," *BMC Endocr. Disord.*, vol. 22, no. 1, p. 214, Aug. 2022, doi: 10.1186/s12902-022-01121-4.

[12] A. Baghbani, N. Bouguila, and Z. Patterson, "Short-Term Passenger Flow Prediction Using a Bus Network Graph Convolutional Long Short-Term Memory Neural Network Model," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2677, no. 2, pp. 1331–1340, Feb. 2023, doi: 10.1177/03611981221112673.

[13] A. Ferencek, D. Kofjač, A. Škraba, B. Sašek, and M. K. Borštnar, "Deep Learning Predictive Models for Terminal Call Rate Prediction during the Warranty Period," *Bus. Syst. Res. J.*, vol. 11, no. 2, pp. 36–50, Oct. 2020, doi: 10.2478/bsrj-2020-0014.

[14] J. Kwon *et al.*, "Artificial intelligence assessment for early detection of heart failure with preserved ejection fraction based on electrocardiographic features," *Eur. Hear. J. - Digit. Heal.*, vol. 2, no. 1, pp. 106–116, May 2021, doi: 10.1093/ehjdh/ztaa015.

[15] R. D. Prince, A. Akhondi-Asl, N. M. Mehta, and A. Geva, "A Machine Learning Classifier Improves Mortality Prediction Compared With Pediatric Logistic Organ Dysfunction-2 Score: Model Development and Validation," *Crit. Care Explor.*, vol. 3, no. 5, p. e0426, May 2021, doi: 10.1097/CCE.0000000000000426.

[16] Q. A. Hidayaturrohman and E. Hanada, "Predictive Analytics in Heart Failure Risk, Readmission, and Mortality Prediction: A Review," *Cureus*, vol. 16, no. 11, p. 11, Nov. 2024, doi: 10.7759/cureus.73876.

[17] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[18] D.-K. Nguyen, C.-H. Lan, and C.-L. Chan, "Deep Ensemble Learning Approaches in Healthcare to Enhance the Prediction and Diagnosing Performance: The Workflows, Deployments, and Surveys on the Statistical, Image-Based, and Sequential Datasets," *Int. J. Environ. Res. Public Health*, vol. 18, no. 20, p. 10811, Oct. 2021, doi: 10.3390/ijerph182010811.

[19] Y. Gu *et al.*, "Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data," *Sci. Rep.*, vol. 11, no. 1, p. 18961, Sep. 2021, doi: 10.1038/s41598-021-98387-w.

[20] P. Tian *et al.*, "Machine Learning for Mortality Prediction in Patients With Heart Failure With Mildly Reduced Ejection Fraction," *J. Am. Heart Assoc.*, vol. 12, no. 12, p. e029124, Jun. 2023, doi: 10.1161/JAHA.122.029124.

[21] J. Zhang, U. Norinder, and F. Svensson, "Deep Learning-Based Conformal Prediction of Toxicity," *J. Chem. Inf. Model.*, vol. 61, no. 6, pp. 2648–2657, Jun. 2021, doi: 10.1021/acs.jcim.1c00208.

[22] D. Veritti, L. Rubinato, V. Sarao, A. De Nardin, G. L. Foresti, and P. Lanzetta, "Behind the mask: a critical perspective on the ethical, moral, and legal implications of AI in ophthalmology," *Graefe's Arch. Clin. Exp. Ophthalmol.*, vol. 262, no. 3, pp. 975–982, Mar. 2024, doi: 10.1007/s00417-023-06245-4.

[23] M. S. Barkhordari and L. M. Massone, "Failure Mode Detection of Reinforced Concrete Shear Walls Using Ensemble Deep Neural Networks," *Int. J. Concr. Struct. Mater.*, vol. 16, no. 1, p. 33, Dec. 2022, doi: 10.1186/s40069-022-00522-y.

[24] J. Tromp *et al.*, "Age-Related Characteristics and Outcomes of Patients With Heart Failure With Preserved Ejection Fraction," *J. Am. Coll. Cardiol.*, vol. 74, no. 5, pp. 601–612, Aug. 2019, doi: 10.1016/j.jacc.2019.05.052.

[25] S. Paul and R. V. Paul, "Anemia in Heart Failure," *J. Cardiovasc. Nurs.*, vol. 19, no. Supplement, pp. S57–S66, Nov. 2004, doi: 10.1097/00005082-200411001-00008.

[26] B. Zareini *et al.*, "Type 2 Diabetes Mellitus and Impact of Heart Failure on Prognosis Compared to Other Cardiovascular Diseases," *Circ. Cardiovasc. Qual. Outcomes*, vol. 13, no. 7, pp. 386–394, Jul. 2020, doi: 10.1161/CIRCOUTCOMES.119.006260.

[27] J. P. Curtis *et al.*, "The association of left ventricular ejection fraction, mortality, and cause of death in stable outpatients with heart failure," *J. Am. Coll. Cardiol.*, vol. 42, no. 4, pp. 736–742, Aug. 2003, doi: 10.1016/S0735-1097(03)00789-7.

[28] J. Slivnick and B. C. Lampert, "Hypertension and Heart Failure," *Heart Fail. Clin.*, vol. 15, no. 4, pp. 531–541, Oct. 2019, doi: 10.1016/j.hfc.2019.06.007.

[29] J. Wang *et al.*, "Impact of heart failure and preoperative platelet count on the postoperative short-term outcome in infective endocarditis patients," *Clin. Cardiol.*, vol. 47, no. 1, p. e24171, Jan. 2024, doi: 10.1002/clc.24171.

[30] J. C. Schefold, M. Lainscak, L. M. Hodoscek, S. Blöchlinger, W. Doehner, and S. von Haehling, "Single baseline serum creatinine measurements predict mortality in critically ill patients hospitalized for acute heart failure," *ESC Hear. Fail.*, vol. 2, no. 4, pp. 122–128, Dec. 2015, doi: 10.1002/ehf2.12058.

[31] S. Peng, J. Peng, L. Yang, and W. Ke, "Relationship between serum sodium levels and all-cause mortality in congestive heart failure patients: A retrospective cohort study based on the Mimic-III database," *Front. Cardiovasc. Med.*, vol. 9, p. 1082845, Jan. 2023, doi: 10.3389/fcvm.2022.1082845.

[32] N. Fluschnik *et al.*, "Gender differences in characteristics and outcomes in heart failure patients referred for end-stage treatment," *ESC Hear. Fail.*, vol. 8, no. 6, pp. 5031–5039, Dec. 2021, doi: 10.1002/ehf2.13567.

[33] D. Kamimura *et al.*, "Cigarette Smoking and Incident Heart Failure," *Circulation*, vol. 137, no. 24, pp. 2572–2582, Jun. 2018, doi: 10.1161/CIRCULATIONAHA.117.031912.

[34] A. Abdin *et al.*, "'Time is prognosis' in heart failure: time-to-treatment initiation as a modifiable risk factor," *ESC Hear. Fail.*, vol. 8, no. 6, pp. 4444–4453, Dec. 2021, doi: 10.1002/ehf2.13646.

[35] K. Seu, M.-S. Kang, and H. Lee, "An Intelligent Missing Data Imputation Techniques: A Review," *JOIV Int. J. Informatics Vis.*, vol. 6, no. 1–2, p. 278, May 2022, doi: 10.30630/joiv.6.1-2.935.

[36] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, p. 52, Jul. 2021, doi: 10.3390/technologies9030052.

[37] V. Sarraju, J. Pal, and S. Kamilya, "SRS: Gender-based heart disease prediction using stratified random sampling approach," in *AIP Conference Proceedings*, May 2024, vol. 3164, no. 1, p. 020005, doi: 10.1063/5.0216559.

[38] P. Mooijman, C. Catal, B. Tekinerdogan, A. Lommen, and M. Blokland, "The effects of data balancing approaches: A case study," *Appl. Soft Comput.*, vol. 132, p. 109853, Jan. 2023, doi: 10.1016/j.asoc.2022.109853.

[39] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, A. A. Jiwa Permana, and I. M. Sunia Raharja, "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, pp. 84–94, Nov. 2023, doi: 10.56705/ijaimi.v1i2.98.

[40] S. Das, S. P. Nayak, B. Sahoo, and S. C. Nayak, "Evaluating Ensemble Models on Imbalanced Data Sets: A Comparative Study across Varied Minority Class Ratios," in *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, Feb. 2024, pp. 774–779, doi: 10.1109/ESIC60604.2024.10481583.

[41] N. Buslim, "Ensemble learning techniques to improve the accuracy of predictive model performance in the scholarship selection process," *J. Appl. Data Sci.*, vol. 4, no. 3, pp. 264–275, Sep. 2023, doi: 10.47738/jads.v4i3.112.

[42] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.014.

[43] J. Zhao, J. Jin, S. Chen, R. Zhang, B. Yu, and Q. Liu, "A weighted hybrid ensemble method for classifying imbalanced data," *Knowledge-Based Syst.*, vol. 203, p. 106087, Sep. 2020, doi: 10.1016/j.knosys.2020.106087.

[44] J. Qiu, "An Analysis of Model Evaluation with Cross-Validation: Techniques, Applications, and Recent Advances," *Adv. Econ. Manag. Polit. Sci.*, vol. 99, no. 1, pp. 69–72, Sep. 2024, doi: 10.54254/2754-1169/99/2024OX0213.

[45] Y. Wen, M. Kalander, C. Su, and L. Pan, "An Ensemble Noise-Robust K-fold Cross-Validation Selection Method for Noisy Labels," *arxiv Artif. Intell.*, pp. 1–9, 2021, [Online]. Available at: http://arxiv.org/abs/2107.02347.

[46] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Front. Bioinforma.*, vol. 4, p. 1457619, Sep. 2024, doi: 10.3389/fbinf.2024.1457619.

[47] L. Sweet, C. Müller, M. Anand, and J. Zscheischler, "Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models," *Artif. Intell. Earth Syst.*, vol. 2, no. 4, Oct. 2023, doi: 10.1175/AIES-D-23-0026.1.

[48] P. Mahajan, S. Uddin, F. Hajati, M. A. Moni, and E. Gide, "A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets," *Health Technol. (Berl).*, vol. 14, no. 3, pp. 597–613, May 2024, doi: 10.1007/s12553-024-00835-w.

[49] X. Yang, L. Wen, M. Sun, J. Yang, and B. Zhang, "Prediction of cardiac deterioration in acute heart failure patients: Evaluation of the efficacy of single laboratory indicator models versus comprehensive models," *Medicine (Baltimore).*, vol. 103, no. 44, p. e40266, Nov. 2024, doi: 10.1097/MD.0000000000040266.

[50] Q. Wang *et al.*, "Machine learning-based risk prediction of malignant arrhythmia in hospitalized patients with heart failure," *ESC Hear. Fail.*, vol. 8, no. 6, pp. 5363–5371, Dec. 2021, doi: 10.1002/ehf2.13627.

[51] W. Chua *et al.*, "An angiopoietin 2, FGF23, and BMP10 biomarker signature differentiates atrial fibrillation from other concomitant cardiovascular conditions," *Sci. Rep.*, vol. 13, no. 1, p. 16743, Oct. 2023, doi: 10.1038/s41598-023-42331-7.

[52] C. Vlachas *et al.*, "Random forest classification algorithm for medical industry data," *SHS Web Conf.*, vol. 139, p. 03008, May 2022, doi: 10.1051/shsconf/202213903008.

[53] J. C. Yang, "The prediction and analysis of heart disease using XGBoost algorithm," *Appl. Comput. Eng.*, vol. 41, no. 1, pp. 61–68, Feb. 2024, doi: 10.54254/2755-2721/41/20230711.

[54] S. Decherchi, E. Pedrini, M. Mordenti, A. Cavalli, and L. Sangiorgi, "Opportunities and Challenges for Machine Learning in Rare Diseases," *Front. Med.*, vol. 8, p. 747612, Oct. 2021, doi: 10.3389/fmed.2021.747612.

[55] Q. Gao, "Application of Machine Learning in the field of Heart Disease Prediction and its Accuracy Study," *Sci. Technol. Eng. Chem. Environ. Prot.*, vol. 1, no. 8, Aug. 2024, doi: 10.61173/24749p02.

[56] D. Bertsimas, L. Mingardi, and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 9, pp. 3627–3637, Sep. 2021, doi: 10.1109/JBHI.2021.3066347.

[57] M. Qiu, L.-L. Ding, and H.-R. Zhou, "Factors affecting the efficacy of SGLT2is on heart failure events: a meta-analysis based on cardiovascular outcome trials," *Cardiovasc. Diagn. Ther.*, vol. 11, no. 3, pp. 699–706, Jun. 2021, doi: 10.21037/cdt-20-984.

[58] P. Bhattarai and M. Karki, "The Unrepaired Tetralogy of Fallot: A Tale of Delayed Presentation and Limited Access to Care," *Cureus*, vol. 16, no. 1, Jan. 2024, doi: 10.7759/cureus.52407.

[59] E. M. DeFilippis *et al.*, "Impact of socioeconomic deprivation on evaluation for heart transplantation at an urban academic medical center," *Clin. Transplant.*, vol. 36, no. 6, p. e14652, Jun. 2022, doi: 10.1111/ctr.14652.

[60] R. S. Walia and R. Mankoff, "Impact of Socioeconomic Status on Heart Failure," *J. Community Hosp. Intern. Med. Perspect.*, vol. 13, no. 6, p. 24, Nov. 2023, doi: 10.55729/2000-9666.1258.

[61] O. A. Akinyemi *et al.*, "Evaluating the Predictive Accuracy of Socioeconomic Metrics on Heart Failure Risk and Outcomes in Maryland," *Cureus*, vol. 16, no. 9, Sep. 2024, doi: 10.7759/cureus.69474.

[62] A. Sinha *et al.*, "Interconnected Clinical and Social Risk Factors in Breast Cancer and Heart Failure," *Front. Cardiovasc. Med.*, vol. 9, p. 847975, May 2022, doi: 10.3389/fcvm.2022.847975.

[63] L. de Tantillo, B. E. McCabe, M. Zdanowicz, J. Ortega, J. M. Gonzalez, and S. Chaparro, "Implementing Strategies to Recruit and Retain a Diverse Sample of Heart Failure Patients," *Hisp. Heal. Care Int.*, vol. 23, no. 1, pp. 9–17, Mar. 2025, doi: 10.1177/15404153241248144.

[64] C.-C. Chiu, C.-M. Wu, T.-N. Chien, L.-J. Kao, C. Li, and H.-L. Jiang, "Applying an Improved Stacking Ensemble Model to Predict the Mortality of ICU Patients with Heart Failure," *J. Clin. Med.*, vol. 11, no. 21, p. 6460, Oct. 2022, doi: 10.3390/jcm11216460.