Analysis and Prediction of Risk Factors for Heart Failure

Yu Wan[1], Shuya Zhang[2], Yuxuan Zhao[3,4,*]

[1]Hubei University of Economics, Wuhan, 430200, China
[2]Computer Science and Technology, Geely University of China, Chengdu, 644000, China
[3]Department of Math, University of California, Davis, Davis, 95616, United States

[4]yxuzhao@ucdavis.edu
*corresponding author

Abstract. Heart failure is a grave and progressive illness. It is associated with multiple risk factors such as age and other possible factors. Accurately identifying and quantifying these risk factors is critical to developing personalized prevention and treatment strategies. However, it is difficult for common patients to predict heart failure without the professional diagnosis of doctors, and relying on human resources to predict heart failure is more subjective. This study proposes a method to evaluate and predict the factors related to heart failure based on multiple body indicators by using random forest algorithm. The random forest prediction model was applied to assess the correlation between multiple medical indicators such as creatinine phosphokinase (CPK), serum creatinine (SCR), ejection fraction (EF), age and heart failure. CPK was found to be the most associated with heart failure. In addition, increasing follow-up period can also effectively monitor heart failure progression in patients. In this high dimension prediction problem, the prediction effect of random forest is better, and the overall accuracy is higher. The approach used in this research is important for forecasting the risk of heart failure, enhancing the survival rates of patients, and alleviating the burden on healthcare systems.

Keywords: Heart failure, machine learning algorithms, random forests, correlation analysis.

1. Introduction

The current era is characterized by the pervasive influence of data, in which all aspects of the surroundings are interconnected through data, and every facet of life is digitally documented. Machine learning has emerged as the fundamental technology in contemporary data science and is extensively utilized across various domains. Heart failure is a critical condition that seriously affects global health.

Nowdays, how to use more efficient scientific methods to predict heart failure and other diseases has become a new way of thinking. The diagnosis of heart failure primarily depends on clinical symptoms and signs [1]. There are several medical methods for detecting heart failure. For instance, heart rate variability (HRV) is a valuable indicator in the diagnosis and prevention of cardiovascular diseases [2]. However, the detection of such indicators typically requires professional equipment, which is expensive and difficult for individual patients to access.

Machine learning algorithms encompass a variety of techniques, including classification analysis, regression analysis, data clustering, association rule learning, dimensionality reduction feature engineering, and deep learning methods. These methods are commonly used in academic research and have significant implications for the field of machine learning [3]. Research examining the efficacy of disease prediction methodologies utilizing various forms of supervised machine learning algorithms has demonstrated that algorithms, including Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM), display varying degrees of performance across diverse datasets. Among them, SVM have shown high accuracy in identifying heart failure signals [4], reaching 98.81% [5]. Machine learning algorithms demonstrate great potential in heart failure risk prediction [6]. In the development of a clinical event risk prediction model for heart failure patients, Li and Zhang conducted a comprehensive analysis using univariate analysis, the Cox mortality risk model, and the PWP readmission risk model [7]. According to the literature, Xu et al mainly used Logistic regression model, Random Forest, Gradient Boosting Tree, SVM model, XGBoost, Artificial Neural Networks, which are risk prediction models, in their study of machine learning risk prediction models for heart failure occurrence [8]. In addition, several limitations of machine learning in heart failure applications have been identified, including the substantial amount of data required for computation and testing, the persistent challenge of overfitting, and the significant time and effort needed for external validation [9]. Furthermore, Lei highlighted challenges associated with deep learning models, particularly issues related to poor model interpretability and generalization, within the context of deep learning-based clinical research on heart failure [10].

The objective of this research is to establish a machine learning-based approach for the prediction of heart failure. This study will involve the collection of medical data from a large cohort of patients, with the goal of exploring the relationship between various medical indicators and heart failure and attempting to forecast the likelihood of future heart failure occurrences based on specific medical indicator data of patients. Additionally, efforts will be made to address existing challenges in machine learning-based disease prediction, with the intention of enhancing the prevention and treatment of heart failure.

2. Methodology

2.1. Data source and explanations

The authors have found some available dataset on the Internet plus, mostly self-built dataset containing patient information. The dataset contains medical indicators and personal information of patients.

Table 1. Partial content display of the dataset.

Variable	Date Type	Value Range
Age	Integer	40-95
Gender	Enumeration	0(female)-1(male)
Smoking	Boolean	0(no)-1(yes)
Spanemia	Boolean	0(no)-1(yes)
CPK	Floating Point	23-7861
EF	Floating Point	14-80
SCR	Floating Point	0.5-9.4
Diabetes	Boolean	0(no)-1(yes)
Hypertension	Boolean	0(no)-1(yes)
Blood platelets	Floating Point	25100-850000
Serum Sodium	Floating Point	113-148
Follow-up period	Integer	4-285
Heart Failure	Boolean	0(no)-1(yes)

The dataset in Table 1 is sourced from the Kaggle website and contains a sample size of 299. It was collected over the past three years and contains approximately 13 indicators. There are 7 data-driven indicators and 5 categorical indicators.

Table 1 summarizes the values of each variable in the dataset, as well as the data types of each variable. The authors selected variables such as age, hypertension, blood platelets, SCR and serum sodium for data analysis. In this research, the authors will conduct an analysis of the data and employ machine learning techniques to forecast the likelihood of heart failure.

2.2. Method introduction

The Random Forest algorithm is an ensemble method that comprises numerous decision trees. In comparison to an individual decision tree, the Random Forest algorithm generally exhibits superior performance and is adept at mitigating the risk of overfitting. The pseudo-code for the Random Forest algorithm is presented in Table 2.

Table 2. The Process of random forest algorithm.

```
Algorithm 1 Random Forest
Input: Training dataset D, the number of trees $B$, minimum node size $n$, number of variables $m$ per split
Output: Ensemble of trees $\{T_b\}_1^B$
1: For $b$ = 1 to $B$:
    (a)Draw a bootstrap sample $Z^*$ of size $1 - \frac{n}{N}$ from the training dataset D.
    (b)Construct a random-forest tree $T_b$ to the bootstrapped dataset, systematically applying until the minimum node size $n$ is is attained.
        i. Randomly select $m$ variables from the $p$ variables.
        ii.Select the most appropriate variable or split-point from the set of $m$.
        iii. Divide the node into two subordinate nodes..
2: Generate the collection of decision trees..
To generate a forecast at a new data point x:
Regression:$f(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\, x$.
Classification: Let $C_b(x)$ represent the class prediction generated by the random forest tree $T_b$.
Then $C(x)$ = majority vote $\{C_b(x)\}_1^B$
```

Due to the inherent complexity of the prediction mechanism employed by the random forest model, elucidating its decision-making process directly poses significant challenges. Consequently, the prediction process of the random forest can be effectively illustrated through Figure 1.

An important feature of Random Forest is the ability to calculate the importance of each feature, thus helping to identify the risk factors that are most likely to influence heart failure. By assessing the importance of these features, it is possible to not only predict whether a patient is at high risk for heart failure, but also provide a medical interpretation of these features to inform clinical interventions. Using feature importance and decision pathways, random forests can explain the predictions of the model, thus providing medically interpretable conclusions.

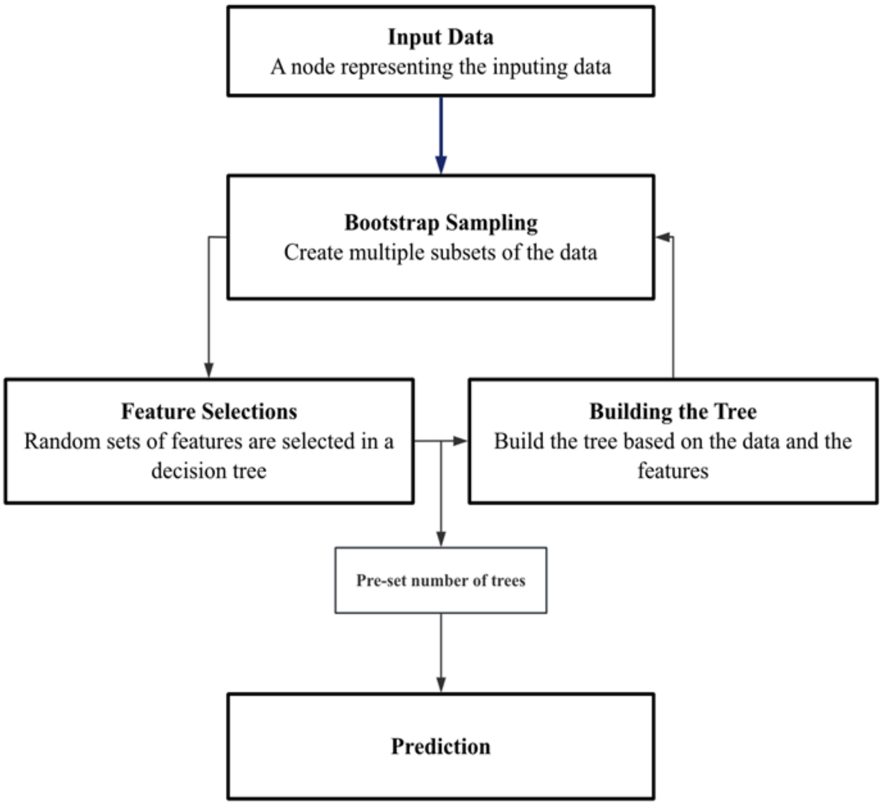
Input Data
A node representing the inputing data
Bootstrap Sampling
Create multiple subsets of the data
Feature Selections
Random sets of features are selected in a decision tree
Building the Tree
Build the tree based on the data and the features
Pre-set number of trees
Prediction

Figure 1. Brief introduction of stochastic forest algorithm prediction process.

3. Results and discussion

3.1. Visual data analysis

To enhance the methodological rigor of the study and to serve as a reference for machine learning techniques aimed at predicting heart failure, the authors initially employed statistical methods to conduct a visual analysis of the original dataset. This preliminary exploration aimed to investigate the potential relationships between the characteristic variables and mortality due to heart failure, as well as to identify any underlying associations among the characteristic variables themselves.

At the beginning of the study, the authors conducted a univariate correlation analysis. For example, age was selected as one of the factors used to create a line chart of correlations with heart failure events. Figure 2 can be used as a reference for the association between age and heart failure.

Figure 2 shows that in the middle to low age group, the proportion of patients with heart failure is generally much smaller than those without heart failure. Even among people under the age of 45, almost no one has experienced heart failure. With the increase of age, the proportion of patients with heart failure also basically showed an increasing trend. Especially in people over the age of 80, most have heart failure. However, a small number of age groups do not meet the above rules, which may be due to the small number of samples in this age group.

From this figure, the authors speculate that heart failure may be related to age. However, the study of single variable is limited, and it is difficult to see the law of the whole data set. Therefore, this study conducted data analysis for multiple characteristic variables to explore the variables that may be related to heart failure.

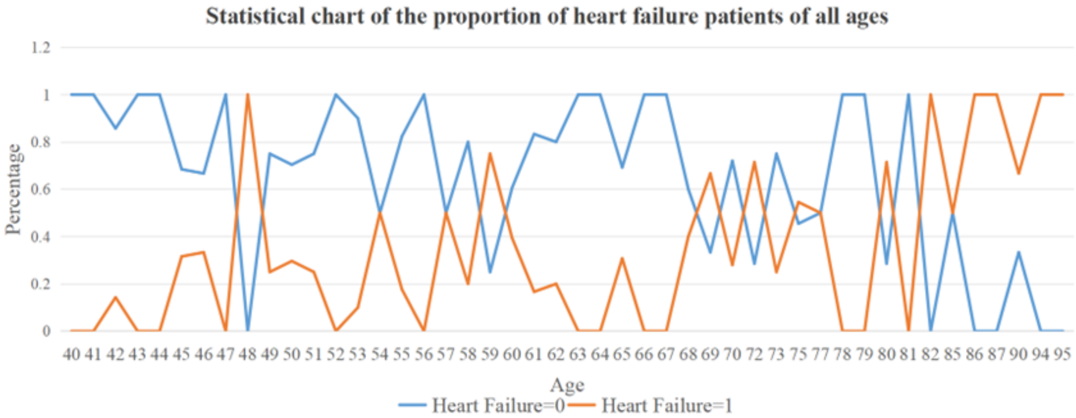

Figure 2. Statistical plot of univariate association of age with heart failure.

When multiple variables exist at the same time, samples with and without heart failure can be divided into two bar statistical charts (Figure 3). For the same variable, different colors represent "low", "normal" and "high" respectively, and the number of samples with and without heart failure can be counted respectively. The shape of the bar statistical chart can be roughly seen by comparing left and right. Only for the 5 variables in Figure 3, the figure shows that the statistical pattern of SCR and serum sodium is significantly different in whether patients suffer from heart failure, and there is also a certain difference between the left and right images of EF. However, the statistical data of CPK and blood platelets have small differences between the left and right. The authors preliminarily concluded that heart failure may have a certain correlation with SCR, serum sodium and EF. The above conclusions are only a rough statistical analysis of the data set. Further experiments are needed to draw more complete conclusions.

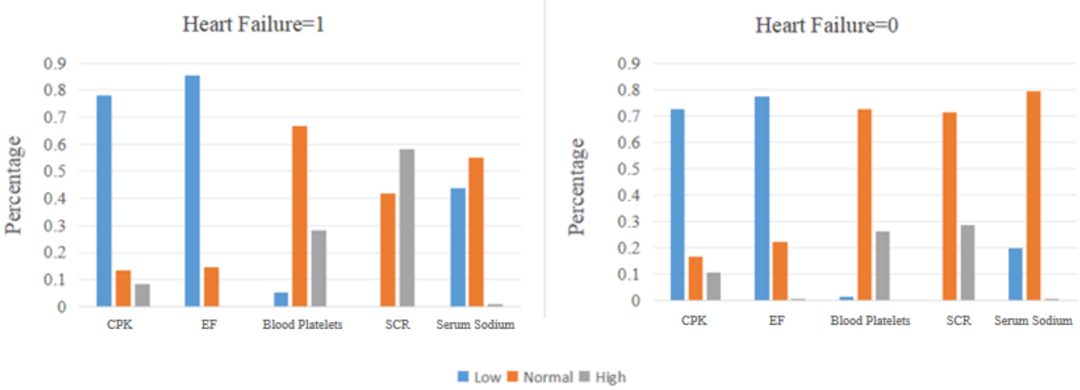

Figure 3. Statistical plot of multivariate analysis.

3.2. Random forest results

In the present study, the random forest methodology is employed to develop the model, utilizing the dataset for training purposes to derive the feature weight values. The weight values associated with various characteristic variables, including SCR, within the dataset are presented in Table 3 below.

Table 3. Weight of characteristic variables.

Characteristic Variables	Weight value
Age	0.107
Gender	0.016
Smoking	0.016

Table 3. (continued).

Spanemia	0.012
CPK	0.137
SCR	0.100
EF	0.115
Diabetes	0.016
Hypertension	0.010
Blood Platelets	0.066
Serum Sodium	0.077
Follow-up Period	0.326

In order to more intuitively see the contribution and weight of each feature variable in predicting heart failure, a characteristic variables weight graph is drawn as follows.

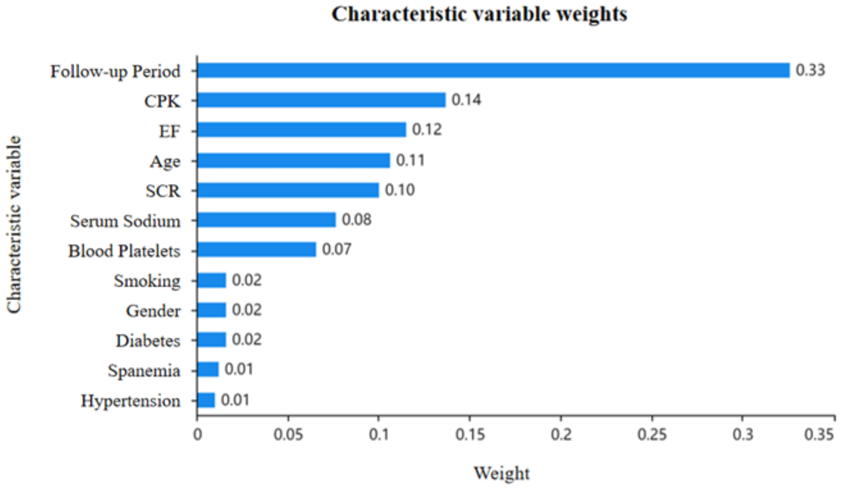

Figure 4. Characteristic variable weights of feature variables predicted by random forest algorithm partial content display of the dataset.

According to figure 4, it can be seen that the longer the follow-up period, the more likely the patient is to develop heart failure, which means that it is easier to detect and monitor the patient's heart failure. Besides the subjective factor of follow-up period, among other objective medical indicators, CPK has the greatest correlation with the occurrence of heart failure, with a characteristic weight of 0.14. Apart from this factor, SCR, EF, age, and creatine kinase are important factors influencing patient mortality.

Also, some performance indexes of the algorithm are obtained to evaluate and measure the actual effect of the algorithm on the prediction of heart failure.

In table 4, the model demonstrates an overall prediction accuracy of 83.0%., meaning that the model correctly predicted heart failure in 83.00% of cases. The 95% confidence interval represents an accuracy range of 73.5% to 90.1%. If the author always predicts the most frequent category (here category 0), the prediction accuracy is 68.2%. P-Value (Acc>NIR) indicates that this p-value represents a significantly higher model accuracy than the no information rate. A p-value less than 0.05 indicates that the predictive ability of the model is significantly better than random guessing. The Kappa coefficient is 0.611, indicating that the predictive performance of the model is Above average. Response rate indicates that the model correctly identifies category 0 (no heart failure occurred) with a proportion of 86.7%. The specificity of the model demonstrates that it accurately identifies category 1, which corresponds to the occurrence of heart failure, in 75.0% of cases. The incidence reflects that 31.8% of the dataset pertains to category 0, indicating the absence of heart failure. Furthermore, the detection rate reveals the percentage of the samples which are correctly classified by the model as belonging to class 0.

Table 4. Performance index of random forest algorithm for predicting heart failure.

Term	Numerical value
Accuracy	0.830
95% Confidence Interval	(0.735, 0.901)
Kappa Coefficient	0.611
Incidence	0.318
Specificity	0.750
No Information Rate	0.682
P-Value [Acc > NIR]	0.001
Response Rate	0.867
Detection Rate	0.591

This algorithm enhances predictive accuracy by integrating the outcomes of several decision trees. In contrast to an individual decision tree, random forests mitigate the likelihood of overfitting through the aggregation of predictions from multiple models, and they are also capable of managing high-dimensional datasets effectively. Moreover, in high-dimensional dataset, random forests can also exhibit good performance. This makes the processing and computation of heart failure dataset more convenient and efficient. The random forest algorithm can analyze and predict various physiological indicators, lifestyle habits, family medical history, and other factors of patients before the onset of heart failure symptoms. It can also provide personalized treatment, helping doctors develop more accurate treatment plans based on the individual characteristics of patients. The causes and progression of heart failure may vary among different patients, and the random forest algorithm can comprehensively consider multiple factors to provide personalized treatment recommendations for each patient.

4. Conclusion

This study explores a scientific and convenient machine method to help patients predict heart failure. In this study, a multi-medical index dataset of a large number of patients was collected. Based on the random forest algorithm, the problem of finding factors related to patients' heart failure was analyzed and predicted, and the correlation degree of age, gender and other factors with heart failure was explored. The preliminary prediction results showed that CPK was the most correlated with heart failure among objective factors. SCR, EF, age, and creatine kinas were also associated with heart failure. Furthermore, the duration of the subjective follow-up period exhibited a significant correlation with the identification of heart failure in patients. Based on the predictions generated by the random forest algorithm, it is imperative to closely monitor biomarkers such as CPK, SCR, EF, and creatine kinase in order to effectively mitigate the risk of heart failure and reduce mortality associated with this condition. Additionally, it is essential to maintain vigilant observation and follow-up of patients.

In addition, the authors combined the results of random forest algorithm in predicting heart failure with the results of statistical analysis, and found that the two have certain similarities. For example, the statistical map of the dataset and the operation results of the random forest algorithm all show that the three variables of age, creatinine and sodium are highly correlated with heart failure. However, individual variables such as creatine kinase show different correlations between the two, which is also a problem that the authors need to explore in further research. In the future research, the author will explore whether there is a more suitable algorithm or how to improve the algorithm to improve the accuracy of prediction.

The random forest model demonstrated commendable predictive performance in this study, achieving a high overall prediction accuracy. This finding offers valuable insights for forecasting heart failure across diverse patient populations. In the future, by collecting a large number of genetic data of heart failure patients and healthy people, combined with other clinical indicators, the random forest algorithm can be used to establish predictive models, which can find specific gene variants associated with heart failure, providing the direction for future gene therapy and precision medicine.

Authors contribution
All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Zhang M 2019 Study on diagnostic methods of heart failure. Chinese Journal of Practical Electrocardiology, 567-572

[2] Wang Y W, Wei D J, Cao H and Liang J 2024 Application of deep learning technology in heart failure detection. Online first paper system in Computer Science and Exploration.

[3] Sarker I H 2021 Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 160.

[4] Liu S Y, Shi Y J, Liu Y C, Liang X Y, Yang C G, Qiao W B and Dong G J 2023 Clinical changes in the pathogenesis of heart failure patients with preserved ejection fraction. Journal of Integrated Chinese and Western Medicine Cardio-Cerebrovascular Diseases, 5, 3.

[5] Guo C Y 2023 Research progress of worsening heart failure J. Journal of Integrative Cardio-Cerebrovascular Medicine, 5.

[6] Uddin S, Khan A, Hossain M E and Moni M A 2019 Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 281.

[7] Li J and Zhang Y Y 2019 Construction of clinical event risk prediction model for heart failure patients. Journal of Changzhou University (Natural Science Edition), 78-84.

[8] Xu Q, Xu C R and Cai X 2019 Research progress of prediction model of heart failure risk based on machine learning. Modern Medicine, 807-815.

[9] Wang Q 2022 Prediction of risk of malignant arrhythmia in hospitalized patients with heart failure based on machine learning. Working paper.

[10] Lei Y P, Liu S L and Wu Y X 2019 Clinical research progress of heart failure based on deep learning. Journal of Biomedical Engineering, 373-377+383.