Application of Machine Learning Model for Predicting Mortality Rate in Heart Failure

Tianyuan Jia*

Academy of Information Science and Engineering, Lanzhou University, Lanzhou, China

*Corresponding author: jiaty21@lzu.edu.cn

Abstract. Heart failure is among the prevalent illnesses that can give rise to perilous circumstances. Approximately 26 million individuals experience this condition annually. According to viewpoints of cardiac experts and surgeons, the timely prediction of heart failure remains a highly intricate endeavor. Fortunately, the emergence of categorization and anticipatory models provides a promising avenue for aiding the medical sector by showcasing efficient utilization of medical information. Within this study, the aim is to refine the accuracy of prognosticating heart failure risk by employing datasets derived from clinical records. To achieve this, this paper employs various machine learning techniques to interpret the information and forecast the likelihood of elevated occurrences in healthcare repositories. This study utilizes three distinct machine learning models: decision tree, logistic regression, and random forest to derive predictive accuracies for each model. The outcomes disclose that the decision tree attains the highest predictive accuracy, reaching 88.8%. Moreover, the results and comparative analyses indicate that the present study enhances the precision of earlier heart disease predictions. Incorporating the models presented in this study into healthcare systems has the potential to enable real-time forecasting of heart failure or other medical conditions using data gathered directly from patients.

Keywords: Machine Learning; Mortality Rate; Heart Failure.

1. Introduction

Arterial obstruction stands as the primary trigger for heart attacks, recognized interchangeably under various terms including cardiovascular disorder and arterial high blood pressure [1]. On a global scale, an estimated 17.9 million individuals succumb to cardiovascular illness each year, constituting about 32% of the total worldwide mortality rate [2]. Regrettably, this proportion is projected to witness a swift escalation in the impending years unless efficacious preventive actions are undertaken [3]. Alongside adopting a healthful way of life and dietary management, punctual diagnosis and all-encompassing evaluation emerge as pivotal elements that can potentially safeguard lives. Consequently, this article takes an incremental stride toward preserving the well-being of individuals with hyperlipidemia, a condition mandating patient to undergo multiple assessments. These tests could potentially impose supplementary physical exertion, time constraints, and financial obligations on the patients. Past research has indicated that prevalent origins of cardiac conditions could encompass detrimental dietary habits, tobacco usage, excessive sugar intake, being overweight, or surplus body fat [3]. Additionally, frequently encountered indications might encompass discomfort in the arms and chest. Importantly, it's noteworthy that these factors are distinct from one another; a thorough examination of such data sets has the potential to enhance the diagnostic protocol and provide valuable insights for cardiac surgical practitioners. In preceding research undertakings, an assortment of methodologies has been utilized to improve the diagnostic procedure for heart failure, encompassing methodologies such as Extreme Learning Machines [4], Machine Learning Classifiers [2], and Heart Disease Classification [5]. In light of this, the present study endeavors to enhance classifier efficacy through the execution of experiments that involve the application of multiple machine learning models. The main goal is to maximize the utilization of datasets sourced from various healthcare repositories.

The paper is organized into the subsequent sections: the following part provides an extensive outline of employing machine learning models to predict mortality linked to heart failure. Section III delivers an overview of the synthesized data, enumerates the attributes, and offers comprehensive

insights into each characteristic. Following that, Section IV delineates the procedures employed for data preparation in this study. Moreover, Sections V and VI delve into the step-by-step exploration of the experimental framework, execution, and the performance evaluation of the classifiers. Finally, Section VII encapsulates the overall findings and conclusions of this research endeavor.

2. Early Diagnostic Challenges in Heart Failure

Heart failure constitutes a clinical syndrome marked by impairment in ventricular filling and/or ejection due to structural and/or functional irregularities within the heart. Its main clinical manifestations include dyspnea, fatigue, peripheral edema, elevated jugular venous pressure, and increased plasma levels of natriuretic peptides. The etiology of heart failure is complex, with common causes as shown in Fig. 1. Currently, diagnostic evaluations for heart failure encompass cardiac biomarker analysis, electrocardiography (ECG), chest radiography, echocardiography, lung ultrasound, and other auxiliary tests and examinations. In recent times, even though guidelines for diagnosing and treating heart failure have incorporated specific biomarkers for evaluating the severity and future prognosis of the condition, these serum indicators are influenced by various non-cardiac factors. This underscores the importance of enhancing their sensitivity and specificity to improve their accuracy for prognostic predictions. There is an urgent need to develop convenient, non-invasive, and accurate diagnostic methods that can effectively assess patient prognosis.

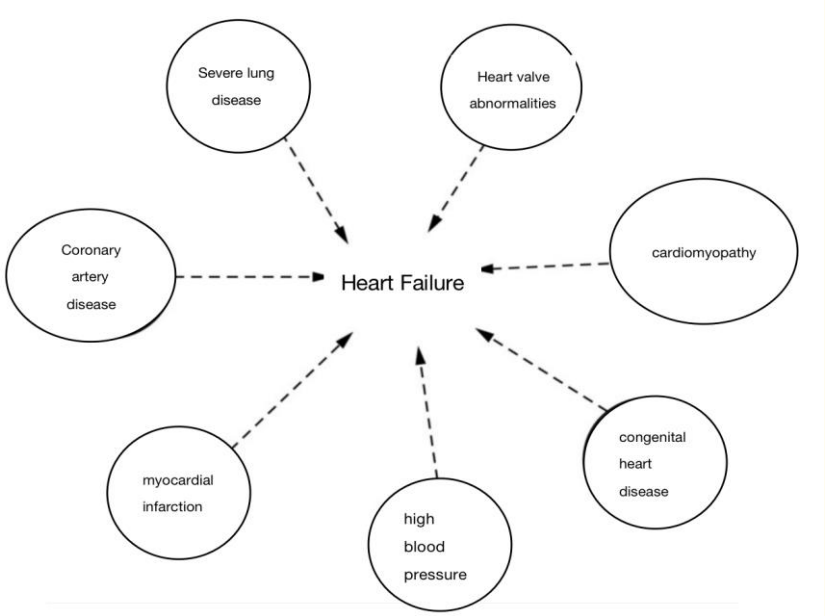

Fig. 1 Different pathological states that can lead to heart failure

Assessing whether an individual has heart disease entails a multifaceted undertaking that involves a range of particulars, laboratory examinations, and equipment [6]. The central aim of this study is not to supplant traditional methods used for diagnosing and estimating the likelihood of heart failure. Instead, this research aims to harness modern methodologies like Machine Learning to improve this process. Machine Learning is a well-established concept and has been widely adopted in diverse domains.

Bashir and colleagues made an endeavor to enhance the predictive performance for heart disease by employing a feature selection approach. Within their experimentation, diverse models including Naïve Bayes, Random Forest, and others were implemented through the Rapid Miner tool. The outcomes revealed significant accuracy improvements attributed to the utilization of the feature selection technique. Similarly, the utilization of Extreme Learning Machine techniques in combination with a feedforward neural network was applied to the Cleveland dataset comprising 300 patients, resulting in an 80% accuracy rate in the prognosis of heart disease for an individual patient [4].

In a separate study, a neural network, specifically a multi-layer perceptron operating under supervised learning, was employed. The design of this system was aimed at evaluating the potential cardiovascular risks in patients through an analysis of their past health records. Another research effort introduced the concept of the HF ratio, utilizing preserved ejection fraction, incorporating multiple factors such as strain rate, hypertensive conditions, and velocity. This approach achieved an accuracy exceeding 80% in its predictions.

The central concept underlying this discourse is to underscore how machine learning contributes to enhancing the support healthcare practitioners offer to diverse individuals afflicted by heart failure.To conclude, a noticeable deficiency pinpointed in prior research lies in the inadequate attainment of measurement precision.Hence, this section offers a synopsis of previous strategies for forecasting heart failure risks and earlier endeavors involving ML techniques in prognosticating the risk of cardiovascular ailments. The objective of this study is to advance beyond earlier efforts by utilizing the designated dataset and models, elaborated upon in the subsequent chapter. The outcomes section will evaluate the performance results of each individual model. Although the selection of models and datasets in this study is rooted in previous research, the dominant machine learning techniques embraced encompass Decision Trees, Naïve Bayes, Random Forests, as well as Support Vector Machines and Logistic Regression. A dataset obtained via Kaggle was utilized in this investigation. Finally, a comparative study will be conducted in the results section to understand the performance of each classifier

3. Data Overview

The data employed within this research was acquired from the UCI Machine Learning Repository platform., acknowledged as the repository containing clinical records associated with heart failure [7]. This dataset is openly accessible. It comprises 13 clinical attributes that are pertinent for heart disease prediction. The database file includes records for 199 patients. A detailed representation of each attribute, alongside the respective value count for each attribute, is provided in Table 1 below.

Table 1. Overview of the data and attribute exposition

S.No.	Attribute Exposition	Measurement	Unique Values
1	Age: Patient’s age	Years	[40,...,95]
2	High blood pressure: Presence of hypertension in a patient	Boolean	0,1
3	Anaemia: Reduction in red blood cells or hemoglobin levels	Boolean	0,1
4	Time: Duration of follow-up	Days	[4,...,285]
5	Creatinine phosphokinase (CPK): Concentration of the CPK enzyme in the bloodstream	mcg/L	[23,...,7861]
6	Ejection fraction: Portion of blood being expelled	Percentage	[14,...,80]
7	Diabetes: Existence of diabetes in the patient	Boolean	0,1
8	Serum creatinine: Concentration of creatinine within the bloodstream	mg/dL	[0.50,...,9.40]
9	Platelets: Blood platelet count	Kiloplatelets/mL	[25.01,...,850.00]
10	Smoking: Patient's smoking status	Boolean	0,1
11	Serum sodium: Sodium concentration in the bloodstream	mEq/L	[114,...,148]
12	Sex: Female or male	Binary	0,1
13	(Target) Death event: Occurrence of patient's demise within the follow-up duration	Boolean	0,1

4. Data Preprocessing

Data preprocessing holds significance as a crucial phase in refining data and rendering it apt for any machine learning or data mining endeavor [8]. Within this research, the chosen dataset underwent a series of preprocessing procedures. First, the data was visualized and by Fig. 2 it was found that the dataset was unbalanced and did not have enough positive targets compared to the negative targets, which was not enough to implement the machine learning approach. As mentioned in reference, the scale of the machine learning dataset introduces bias and also impacts the outcomes produced by the machine learning model. Henceforth, employing the SMOTE (Synthetic Minority Oversampling Technique) technique aids in equilibrating the dataset, thereby imparting a positive impact on classifier efficacy. The demonstration of this effect is presented in the results section. To succinctly put it, equilibrium is established between the positive and negative targets within the dataset. Following this, a scrutiny for absent values in every column is conducted. Yet, a thorough search reveals the absence of any missing values within this dataset.

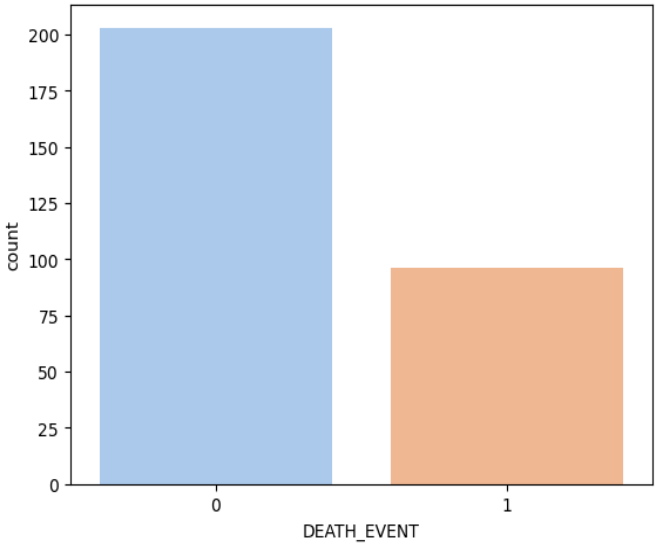

Fig. 2 The contrast between the count of fatalities and the count of recuperations

Subsequently, the forthcoming phase involves the conversion of data values into suitable data types. Within this investigation, a range of models were employed to assess the precision of predictive outcomes. Consequently, certain attribute data types necessitate conversion in alignment with model prerequisites and preferred formatting. The experimental blueprint predominantly revolves around binary classification, a technique employed to categorize a dataset into predetermined classes. This method has been extensively utilized across a range of machine learning techniques [9]. Therefore, The identical binary classification approach was applied to the provided dataset, aptly demonstrating the precision of the chosen classifier's performance within this research. This approach aptly underscores the effectiveness of the chosen classifier within the context of the investigation.

The majority of attributes within the chosen dataset are of a categorical nature, including Anemia, High blood pressure, Diabetes, Sex, and Smoking. To illustrate, consider the Smoking attribute, which entails values based on a binary system (0, 1). Similarly, Anaemia stands as another distinct attribute in the dataset utilizing (0, 1) to indicate a drop in a patient's red blood cell count or hemoglobin, where "0" denotes a reduction and "1" indicates an augmentation. The dataset's "Target" column, also referred to as the "Category" attribute, encompasses two predefined categories, namely "0" and "1". This attribute employs other independent variables to signify the overall patient condition. Notably, a value of "0" signifies that the patient did not succumb to heart failure, while a value of "1" signifies that the patient experienced heart failure-induced demise. As an illustration, when evaluating the values of all the independent variables, a patient demonstrating characteristics like anemia, high blood pressure, and smoking tendencies is more prone to experience death due to heart failure. Conversely, the opposite scenario holds true.

5. The Experiment Preparation

This research employed the gathered dataset to prognosticate heart failure mortality via the application of three distinct models. Subsequent to the effective execution of data preprocessing protocols, the ensuing section will delve into the appraisal of each classifier's efficacy. Once the data preprocessing measures have been duly completed, this segment will elaborate on the chosen models, providing their explanations, and outlining the comprehensive approach adopted for the experimentation.

In this study, the objective was to enhance model accuracy. Consequently, distinct adjustments were implemented for varying models during the experiments, and comparisons were drawn between divergent outcomes of the same model. The following section outlines the essential aspects of the design approach adopted in this research.

In the framework aimed at forecasting heart failure conditions, this study employs a series of five stages. In the initial step (Step 1: Dataset Selection/Data Preprocessing), the investigation comprises five segments: an overview of the data, identification and removal of outliers, detection and imputation of missing data, data enhancement through random number generation, and the application of appropriate normalization methods. The subsequent step (Step 2: Model Selection) consists of two components: comprehending data values (classes) and selecting machine learning models. Moving on to Step 3: Model Implementation, this phase also comprises two segments: data importation and the implementation of all selected models. In Step 4: Performance Measurement, there exist two divisions: the computation of accuracy utilizing the "Performance" operator and the analysis of results via a Confusion Matrix. Finally, in the concluding step, Result Comparison, two sections are integrated: a comparison of accuracy across all models and the determination of the final output.

As stated earlier, a trio of machine learning models are harnessed to forecast the likelihood of a patient's risk of heart disease-related mortality, aiming to identify the most effective model among them.

5.1. Decision Tree

This model follows a tree-based classification approach, where the learning phase hinges on accumulating evidence from each attribute to construct a structure comprising nodes and branches. This study aims to provide a more accurate and interpretable method for predicting mortality rates among patients with heart failure by integrating decision tree classification and resampling techniques.

In the research methodology, this paper initially utilizes feature data sourced from clinical records, encompassing patients' physiological indicators and symptom information. The first section of the code involves data preprocessing, where features are extracted and divided into training, validation, and test sets. Subsequently, this paper employs the decision tree classifier as the predictive model. During the model training phase, multiple decision trees are constructed, and by adjusting the maximum depth parameter, this paper seeks the optimal model complexity. This step aids in balancing model generalization and overfitting. This study evaluates these models based on validation set accuracy, ultimately selecting the model that performs optimally on the validation set.

The chosen best model is then validated on an independent test set to assess its performance in real-world scenarios. During this stage, this study measures model accuracy and assess its efficacy in predicting patient mortality rates. Additionally, this study utilizes a decision tree visualization tool to visually depict the model's decision-making process, providing clinicians with an intuitive explanatory tool to underst

Proposes an approach with potential applications in predicting mortality rates among patients with h and the basis and rules of the model's predictions.

To summarize, this study eart failure by integrating decision tree classification and resampling techniques. Through systematic data processing, model training, and evaluation processes, this study aims to enhance predictive accuracy and interpretability, providing support for clinical decision-making and interventions. This approach holds significant practical significance in the clinical

healthcare field, promising to offer clinicians more precise patient management and treatment recommendations (See Fig. 3).

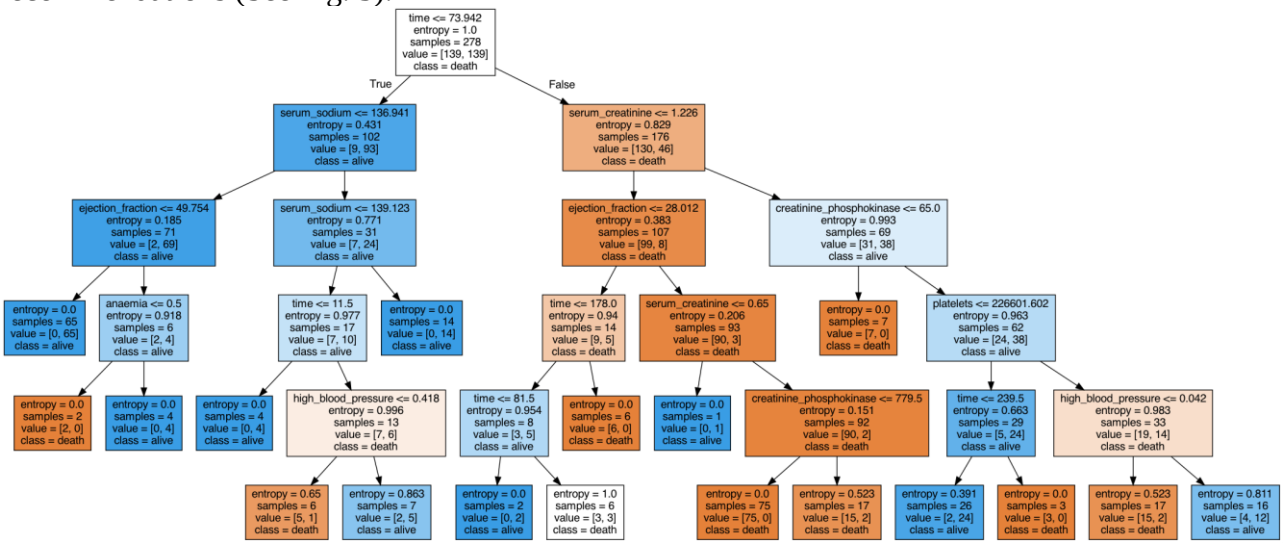

Fig. 3 Decision Tree

5.2. Logistic Regression

Logistic regression serves as an additional classification model that leverages regression analysis for learning and predicting parameters within a given dataset. The process of acquiring knowledge and making predictions is centered on assessing the probability of binary categorization. The logistic regression model necessitates class variables for facilitating binary categorization. Correspondingly, within this dataset, the "target" column encompasses two separate binary values: "0" indicating patients with heart failure who survived and "1" representing patients who passed away due to heart failure. Conversely, the independent variables can take on binary categorical, nominal, or polynomial attributes [10]. The formula for the logistic regression equation is presented below.

$$Logit(p) = ln(\frac{p}{1-p}) = \frac{prob.of\ presence\ of\ characteristics}{prob.of\ absence\ of\ characteristics} \quad (1)$$

5.3. Random Forest

The subsequent model incorporated into this study is the Random Forest, which was selected for analysis. Belonging to the classification realm, this model falls under the umbrella of supervised learning algorithms. In the learning stage, the model initiates by creating numerous random trees known as a "forest". For instance, when dealing with a dataset containing "x" attributes, the model proceeds to randomly choose a subset of features, denoted as "y". By employing all of these selected features ("y"), it constructs nodes utilizing the best-split approach. Additionally, the algorithm assembles a complete forest through the repetition of the preceding steps. Subsequently, in the prediction stage, the procedure seeks to combine the trees by employing the estimation outcomes and a voting mechanism. The underlying aim of this fusion process within the forest is to identify the tree yielding the highest prediction, thus enhancing the overall predictive accuracy for the final dataset.

6. Model Performance and Comparative Evaluation

Through the experimentation involving the aforementioned three models, the outcomes are presented in Table 2. It's evident that the most elevated predictive accuracy is achieved when employing a five-layer decision tree, attaining a remarkable 88.8%. For logistic regression, this study first took the Chi- square test to get the five optimal features, and when the data contained in these five features are used as the training set the result obtained is called LR_1, and the prediction accuracy is 83.6%, and when all the data contained in the features are used as the training set the result this

study obtained is called LR_2, and the prediction accuracy is 65.3%. Finally for the random forest, the model constructed in this experiment is max_depth=8,n_estimators=250,random_state=0. The prediction accuracy of the results obtained is 85.1%. For this study, the best approach should be to take the most appropriate depth of the decision tree.

Table 2. Models and prediction accuracy

Model	Prediction Accuracy
Decision Tree	88.8%
LR_1	83.6%
LR_2	65.3%
Random Forest	85.1%

7. Conclusions

The number of people impacted by heart failure is continuously increasing. In a bid to combat this hazardous trend and mitigate the likelihood of mortality among heart failure patients, the present study necessitates a system capable of devising rules or categorizing data through machine learning techniques. Consequently, this research deliberates, proposes, and enacts a machine learning model, specifically singled out after a comparison of various models. Ultimately, the model achieving the highest predictive accuracy in this study is the decision tree, boasting an accuracy rate of 88.8%. It's important to highlight that the dataset utilized in this research is relatively modest in size. This suggests that future researchers may leverage larger datasets to facilitate a more comprehensive exploration.

References

[1] Maxwell, A. E., Warner, T. A., & Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. International journal of remote sensing, 2018, 39(9): 2784-2817.

[2] BISWAS, Niloy, et al. Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. BioMed Research International, 2023.

[3] BENJAMIN, Emelia J., et al. Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation, 2019, 139(10): e56-e528.

[4] Ismaeel, S., Miri, A., & Chourishi, D. Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis. In 2015 IEEE Canada International Humanitarian Technology Conference, 2015: 1-3.

[5] Ekız, S., & Erdoğmuş, P. Comparative study of heart disease classification. In 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, 2017: 1-4.

[6] Chen, K., Mudvari, A., Barrera, F. G., Cheng, L., & Ning, T. Heart murmurs clustering using machine learning. In 2018 14th IEEE International Conference on Signal Processing, 2018: 94-98.

[7] Chicco, D., & Jurman, G. Heart failure clinical records data set. UCI Machine Learning Repository, 2020.

[8] Al-Mudimig, A., & Saleem, F. A framework of an automated data mining systems using ERP model. International Journal of Computer and Electrical Engineering, 2009, 1(5).

[9] Kumari, R., & Srivastava, S. K. Machine learning: A review on binary classification. International Journal of Computer Applications, 2017, 160(7).

[10] HOSMER JR, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. Applied logistic regression. John Wiley & Sons, 2013.