Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Prediction and Feature Importance Analysis for Heart Failure using Machine Learning Techniques

Yu Zeng *

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

* Corresponding author: u3595087@connect.hku.hk

Abstract. Heart failure is a clinically complex syndrome that affects millions of people globally and posts a heavy strain on healthcare systems because of its high rates of morbidity and mortality. To manage and treat it effectively, early detection and accurate prediction are essential. Machine learning algorithms present a viable technique. This research attempts to forecast heart failure under specific circumstances, using a heart disease dataset from kaggle.com that is accessible to the public. Five machine learning models: Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), were used as they show strong performance in applications pertaining to machine. The performance of five models was evaluated by accuracy, precision, recall and f1-score with 11 symptoms. From the calculation, Random Forest shows better performance than the remaining models with an accuracy of 88% and ST_Slope_Up is tested to have the greatest impact on the prediction.

Keywords: Machine learning; natural language process; supervised learning.

1. Introduction

Heart failure happens when the heart muscle cannot pump enough blood to fulfill the oxygen and blood requirements of the body [1]. About 64 million individuals worldwide suffer from heart failure, which is a major cause of hospitalization and death [2]. It can lead to exhaustion and dyspnea, and in certain cases, persistent coughing. It might become exceedingly difficult to do daily tasks like walking, carrying groceries, or climbing stairs [1]. Several medical disorders like diabetes and high blood pressure, as well as unhealthy habits such as smoking tobacco, not exercising enough can increase the risk of heart failure [3]. As the increasing aging population and pressure from social competition among young people, this syndrome is becoming more and more problematic. Conventional methods for predicting heart failure depend on imaging, biomarkers, and clinical evaluations; however, those approaches seem to be time-consuming and less accurate when dealing with intricate multi-dimensional data. Therefore, to improve the speed and accuracy of diagnosis, better assistive technology like machine learning models can be considered in predicting heart failure.

There were some existing prediction systems. In the last decades, researchers conducted thorough studies before proposing their own prediction using different classification methods and eventually compared the performances in heart failure prediction. Various machine learning models have been used. For instance, in a comparative analysis of six machine learning methods for predicting heart disease, Dwivedhi examined Artificial Neural Network (ANN) [4], Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), and Naive Bayes (NB). The results showed that LR performed better than the other five methods. Pouriyeh et al. performed a similar analysis between DT, NB, ANN, KNN, and SVM for the prediction of heart disease, however, they discovered that SVM beats the other machine learning techniques with an accuracy rate of 84.15% [5]. Additionally, Mohan proposed that SVM performs better than DT and ANN in the prediction of coronary heart disease [6]. Gudadhe also developed a method that combines SVM and multilayer perceptron ANN. With an accuracy of 80.41%, he divided the database into two groups using the SVM to determine whether heart failure was present or absent. On the other hand, the ANN achieved a significantly greater accuracy rate than SVM, classifying the data on heart disease into five categories with a 97.5% rate [7]. Some researchers also think that RF performs better. For

instance, Thota et al. and Shaikhina et al. both reported accuracy rates of 93.0% and 88.7%, respectively, when predicting heart failure with RF [8, 9].

In this regard, this article tends to make a study of heart failure prediction, with a particular focus on the comparative analysis of several models and the discussion of the feature importance. A Kaggle heart failure dataset consisting of 918 observations with 11 attributes was used. Five machine learning models: Decision Tree, Random Forest, Logistic Regression (LR), K-Nearest Neighbor and Support Vector Machine were used as they show strong performance in applications pertaining to machine. The performance of these five models was measured by accuracy, recall, precision and f1-score. From the calculation, Random Forest shows better performance than the remaining models with an accuracy of 88%.

The remaining parts of this paper are illustrated as follows: Section 2 gives a description of the dataset as well as methods of detailed operation of the five models. Section 3 compares the results in terms of the performance scores and the feature importance, some limitations are also pointed out in this section. This paper is concluded in Section 4.

2. Method

2.1. Dataset Preparation

This paper uses a Kaggle heart failure dataset consisting of 918 observations with 11 attributes to predict potential heart failure by various machine learning techniques [10]. The 11 attributes contain Sex, Age, Cholesterol, RestingBP, ChestPainType, FastingBS, MaxHR, RestingECG, Oldpeak, Exercise Angina and ST_Slope. Heart Disease is the target variable with 1 refers to heart failure and 0 stands for normal. Therefore, this study can be seen as a binary classification with multiple features.

Then data is processed for improving the model performance. The first step of processing is to look through the entire dataset for missing and duplicate values that would cause bias. Removing the missing and duplicate values and then checking for outliers. Since this dataset has both numerical and categorical data, so the next step is to convert the categorical data into a format that machine learning algorithms can understand. Creating dummy variables which are binary (0 or 1) for each level of the categorical variable. For instance, the categorical variable "RestingECG" with three categories: Normal, ST, and LVH, one-hot encoding this variable would create two new columns: "RestingECG_Normal", and "RestingECG_ST". “1 0” in these two rows refers to Normal; “0 1” refers to ST; and “0 0” refers to LVH. The whole dataset can be regarded as a binary classification with multiple features. The data is then split into two sets: a test set and a training set. The model is trained using the training set, which teaches the algorithm how the characteristics relate to the target variable, heart disease. In contrast, the test set is reserved and utilized solely for assessment, offering a dispassionate appraisal of the predictive capacity of the model. In Python programming, data splitting is done using the train_test_split function from the sklearn module. Twenty percent of the data in this study is utilized for testing, while the remaining eighty percent is used for training.

2.2. Machine Learning Models

In this paper, five models are employed to predict heart failure and determine which model performs the best overall for forecasting patients’ heart condition. A synopsis of every model is covered in this section.

2.2.1. Logistic Regression (LR).

Logistic Regression was used to predict the probability of binary classification which is suitable for this study. In this study, LR calculates the correlation between the independent factors (the patient's 11 qualities) and the dependent variables (whether the patient will have heart failure). The equation of logistic regression is defined as below [11]:

$$f(x) = \frac{1}{1+e^{-(b0+b1x)}} \quad (1)$$

Where x is the input value, b0 is the intercept term, e is the base of natural logarithms, x is the feature matrix, and b1 is the coefficient vector for input (x). As seen in Fig 1, the result of a logistic regression is a binary number (0 or 1).

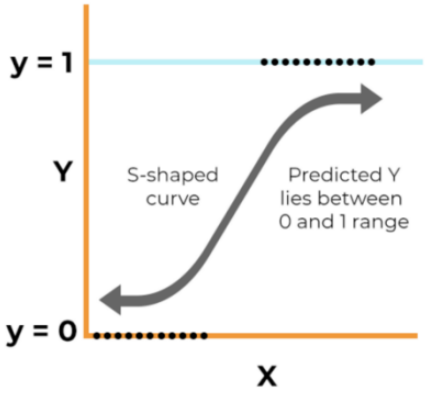

Figure 1. The schematic diagram of Logistic Regression [11].

2.2.2. Decision Tree (DT).

This classification model resembles a tree that was constructed using branches and nodes to represent the evidence gathered for each characteristic throughout the whole model-learning phase [12]. Observations described in the dataset determine how the branches and nodes are connected. The quantity of values allotted to each attribute is used in the forwarding process. Moreover, each transaction's choice was made by adhering to the guidelines specified on each branch and node. Lastly, the record will be assigned a class label based on the decision node. Repeating this process until every transaction has a class category assigned to it. An example of an entire DF is presented in Fig 2.

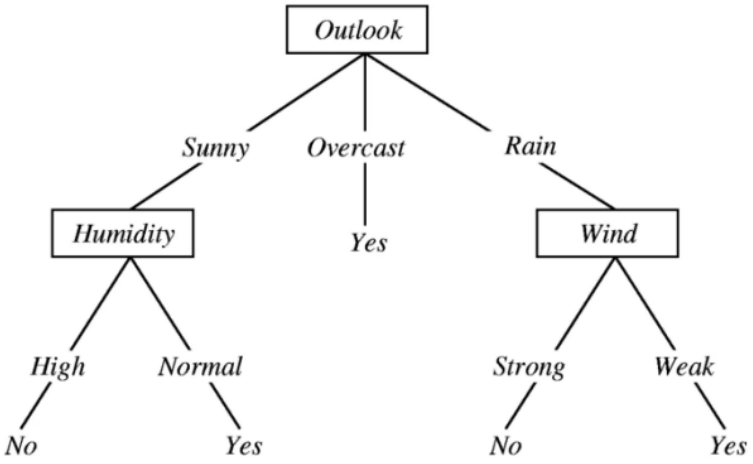

Figure 2. The structure of Decision Tree [12].

2.2.3. Random Forest (RF).

An ensemble of decision trees, each created with a distinct random subset of the given data, forms the basis of random forests. In the event of categorization, the final forecast of the model is decided by a majority vote, and in the case of regression, by averaging the predictions given by the forest trees. Although the variance of the model can be reduced by averaging the predictions of multiple decision trees, the bias would have a small increase [13, 14]. However, in most cases, this can improve the final performance significantly. The process of the RF is shown in Fig 3 [13].

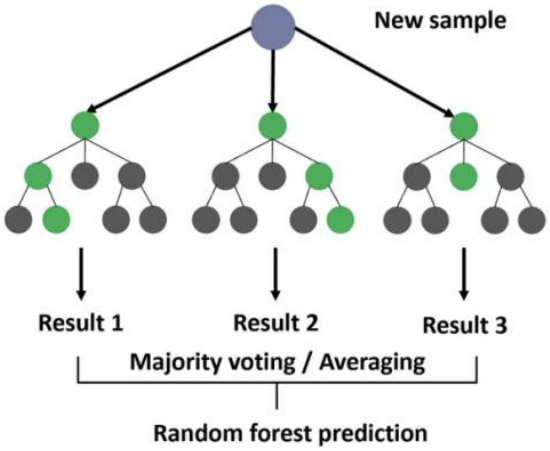

Figure 3. The structure of Random Forest [13].

2.2.4. Support Vector Machine (SVM).

Support Vector Machine is trained on labeled datasets. Both linear and nonlinear problems may be solved using SVM, and it can classify between different classes using the concept of margin [15]. To partition n-dimensional space into classes, this approach seeks to find the greatest fit line shown in Fig. 4. In this sense, fresh data points can be added in the future and placed in the appropriate class.

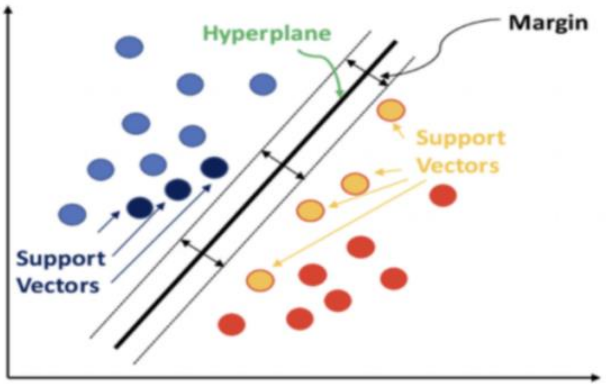

Figure 4. The schematic diagram of Support Vector Machine [15].

2.2.5. K-Nearest Neighbor (KNN).

Based on the principle that comparable objects are closer to one another, the K-Nearest Neighbor algorithm operates [16]. In this case, the distance between the new potential patient and all the observations in the dataset is calculated. Then determine which K patients are closest to the new patient in terms of distance. For instance, if K = 3, three patients are selected as they are closest to the new patients. Among these K nearest neighbors, class the new patient with a heart failure if most of the neighbors have a heart failure.

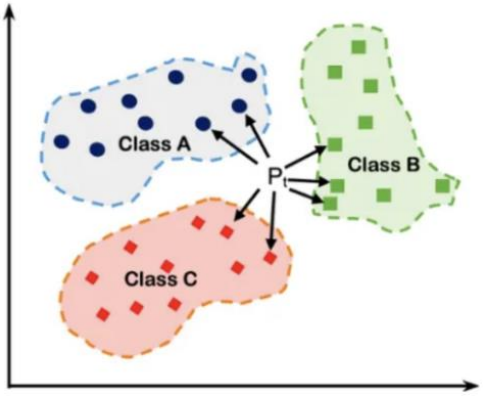

Figure 5. The schematic diagram of K-Nearest Neighbor [16].

2.2.6. Evaluation Metrics.

Accuracy, precision, recall, and f1-score were used to gauge how well the machine learning models and some formulas must be introduced.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Where TP is the true positive and TN is the true negative. FN is the false negative which means that someone is predicted not to have heart failure but in fact they have. This may cause serious consequences for the health of patients. FP is the false positive which means that someone is predicted to have a disease but in fact they are disease free. A false positive does not pose a direct threat to the health of patients, but it can lead to unnecessary emotional stress, or even unnecessary treatments.

According to the formulas, precision is the proportion of the correct heart failure predictions among all predicted heart failure cases. Recall is the percentage of correct predicted heart failures among all actual heart failure cases. The F1 score is a weighted average of recall and precision, where 0 represents the lowest number and 1 represents the highest.

3. Results and Discussion

3.1. The Performance of Various Models

The performance of five machine learning models investigated in this study is shown in Table 1 and Table 2.

Table 1. Accuracy of Five Models

Model/Result	Scheme 1	Scheme 3
Decision Tree	1.0000	0.8641
Random Forest	1.0000	0.8804
Logistic Regression	0.8706	0.8533
Support Vector Machine	0.8733	0.8424
K-Nearest Neighbor	0.8202	0.6576

Table 2. Accuracy of Five Models

Model/Result	Precision for 0	Precision for 1	Recall for 0	Recall for 1	F1 Score for 0	F1 Score for 1
Logistic Regression	0.87	0.84	0.80	0.90	0.83	0.87
Decision Tree	0.83	0.89	0.84	0.88	0.84	0.88
Random Forest	0.86	0.90	0.86	0.90	0.86	0.90
Support Vector Machine	0.86	0.83	0.79	0.89	0.82	0.86
K-Nearest Neighbor	0.62	0.68	0.59	0.72	0.60	0.70

From the above tables, K-Nearest Neighbor has the lowest accuracy of 65.76% caused by the nonstandard data. Among these five algorithms, Random Forest outperforms the others with the highest accuracy of 88.04% in predicting heart failure. The result of this study is aligned with the previous discussion [8, 9]. The K-Nearest Neighbor model has the lowest scores across all metrics, which indicates that this model is not capturing the underlying patterns in the data as effectively as the other models. This is because the KNN algorithm relies on proximity to data points, whereas the dataset used in this study had some outliers and features that were not appropriately normalized, which could have misled the KNN predictions.

3.2. The Importance of Difference Features

The feature importance is plotted in Fig. 6. From the plot, the feature ST_Slope_Up has the highest value which indicates that it has the greatest impact on the prediction model. An upward ST-slope in an ECG may be indicative of healthier heart function during exercise, therefore it is logical that it plays a significant role in the classification task. This suggests that interventions or treatments that affect the ST-slope could be particularly impactful.

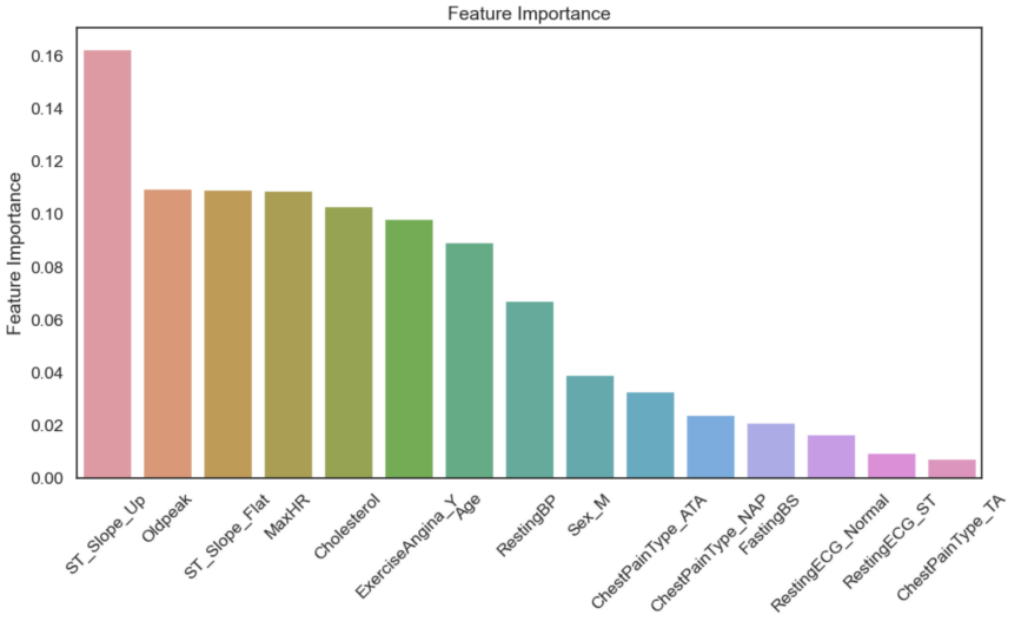

Figure 6. The Plot of Feature Importance on Random Forest Model

(Picture credit: Original).

3.3. Limitations

Apart from the above findings, this study also has some limitations.

3.3.1. Sampling Bias.

The dataset used in this study is from 5 countries, the regional data are not representative of all heart failure patients in general. Therefore, the prediction model may be biased and cannot generalize to a wider group of patients.

3.3.2. Measurement Bias.

Measurement errors or inaccuracies seems to show up in the data for training the model. For example, in this study, some zeros appear in attributes “RestingBP” and “Cholesterol” which are impossible in practice, can lead to biased model predictions.

3.3.3. Data is Not Scaled.

As mentioned before, in this study, data is not standardized. Although feature scaling is not necessary, for instance, tree-based models such as Decision Tree and Random Forest are not affected by feature

scaling. However, for many other models, especially those involving distance calculation like K-Nearest Neighbor, feature scaling is essential and can improve the accuracy of the prediction.

4. Conclusion

Medical officers can make early predictions for healthcare management reasons with less time and effort when they use machine learning. As heart failure is rapidly increasing nowadays, the number of deaths is also increasing, machine learning algorithms are extremely essential in predicting heart failure effectively and accurately. In this investigation, Random Forest seems to achieve the highest performance score when compared to other models. It can assist in developing a workable plan for managing the disease that slows the progression of the illness. And ST_Slope_Up is tested to have the greatest impact on the prediction. To increase accuracy, new methods, such as hybrid models that incorporate optimization algorithms and machine learning techniques can be employed in subsequent research.

References

[1] American Heart Association. What is Heart Failure, March 22, 2023. Available at: https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure.

[2] Shahim S, et al. Global Public Health Burden of Heart Failure: An Updated Review, 2024. Available at: https://www.cfrjournal.com/articles/global-public-health-burden-heart-failure-updated-review.

[3] Centers for Disease Control and Prevention. About Heart Disease, January 5, 2023. Available at: https://www.cdc.gov/heartdisease/heart_failure.htm.

[4] Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing & Applications, 2018, 29 (10): 685 – 693.

[5] Pouriyeh S, et al. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. IEEE Symposium on Computers and Communications, 2017: 204 - 207.

[6] Mohan V. Comparative Analysis of Classification Function Techniques for Heart Disease Prediction. International Journal of Innovative Research in Computer and Communication Engineering, 2013, 1 (3): 735 - 741.

[7] Gudadhe M. Decision support system for heart disease based on support vector machine and Artificial Neural Network. International Conference on Computer and Communication Technology, 2010: 741 - 745.

[8] Shaikhina T, et al. Decision Tree and Random Forest Models for Outcome Prediction in Antibody Incompatible Kidney Transplantation. Biomedical Signal Processing and Control, 2019, 52: 456 - 462.

[9] Thota L, Nimmala S, Manasa K. Heart Disease Prediction Using Random Forest Algorithm. Global Journal of Engineering Science and Research, 2018, 5 (8): 248 - 252.

[10] Kaggle. Heart Failure Prediction dataset. Available at: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

[11] Kanade V. What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices, April 18, 2022. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression.

[12] Prince Yadav. Decision Tree in Machine Learning, November 14, 2018. Available at: https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96.

[13] Roi Yehoshua. Random Forests, March 25, 2023. Available at: https://medium.com/@roiyeho/random-forests-98892261dc49.

[14] Qiu Y, Hui Y, Zhao P, Cai CH, Dai B, Dou J, Bhattacharya S, Yu J. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy. 2024 Mar 7: 130866.

[15] Datatron Blog. What is a Support Vector Machine? 2024. Available at: https://datatron.com/what-is-a-support-vector-machine/.

[16] Sachinsoni. K Nearest Neighbours — Introduction to Machine Learning Algorithms, June 11, 2023. Available at: https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2.