Application of Machine Learning in Heart Failure Prediction

Hongyi Sheng*

CIC College, Dhurakij Pundit University, Bangkok, Thailand

*Corresponding author: 1811000506@mail.sit.edu.cn

Abstract. Cardiovascular diseases (CVDs) are the primary cause of death worldwide, causing an estimated 17 million deaths each year, and are characterized by myocardial infarction and heart failure. Heart failure occurs when the heart is incapable of adequately pumping blood to satisfy the body's demands. It typically stems from conditions such as diabetes, hypertension, or other cardiac disorders. Early detection and management are crucial for individuals with cardiovascular disease or those at an elevated risk of developing cardiovascular issues due to factors such as hypertension, diabetes, hyperlipidemia, or pre-existing medical conditions. In this regard, utilizing a machine learning model can offer significant benefits. This research paper will utilize machine learning models to document the symptoms, physical attributes, and clinical laboratory test results of patients. Consequently, through the analysis of this patient information, biostatistical analyses can be conducted, enabling the detection of patterns and correlations that may elude medical practitioners. The research will benefit the research on cardiovascular diseases.

Keywords: Machine learning; heart failure prediction; cardiovascular diseases.

1. Introduction

Along with the change in people's lifestyles, cardiovascular disease has become a highly prevalent type of disease and has become particularly important in prevention and early screening. Machine learning can provide an important reference for early prevention by analyzing and summarising historical data and risk factors of cardiovascular diseases, building risk-based models, and assessing and predicting risks for individual patients.

The medical field is increasingly using machine learning due to the advancements in artificial intelligence. For example, Averbuch, Sullivan, Sauer, Mamas, Voors, Gale, Metra, Ravindra and Spall discussion focuses on how ML can be utilized to speed up implementation and resolve healthcare gaps [1]. There are many more examples of AI being applied to healthcare; many heart failure patients experience persistent clinical symptoms, which require the use of risk stratification prediction models [2]. Topkara, Elias, Jain, Sayer, Burkhoff, Uriel use the training data set (70%) was used to develop ML models, while the validation data set (30%) was used to assess them using the receiver operating curve and Kaplan-Meier analysis [3]. There is a growing body of research on applying ML in HF environments due to the coexistence of big data and new advanced analytics using ML [4]. Heart disease has a high prevalence in some less developed countries, a series of data was obtained by field visits to hospitals and other places in Bangladesh and a dataset related to the prediction of heart disease was generated for the purpose of applying ML algorithms to improve the accuracy of prediction [5-6]. The problems of hypertension and heart failure have a high burden in China and worldwide, and ML can address these challenges in different ways. Studies have shown that machine learning can improve every stage of patient care, from research and development to everyday clinical practice and population health [7]. Different evaluation metrics were used to understand the data and improve prediction accuracy by using several popular machine learning models. Using RStudio and Weka software, Basha, Nassif and AlShabi simulated some of the algorithms, which helped them better understand the data and data preprocessing [8]. Machine learning plays a vital role in identifying movement abnormalities, heart disease, and other medical conditions. By predicting these conditions in advance, this information can provide valuable insights to doctors, enabling them to adjust their diagnosis and treatment approach for each individual patient [9]. A large proportion of heart disease is preventable, mainly because of inadequate preventive

measures. To enhance the accuracy of early disease onset prediction, it is essential to develop more sophisticated models that integrate diverse geographical data sources [10].

In this paper, algorithms such as Lasso, SGD Classifier, SVM, K Neighbors, Gaussian NB, Decision Tree, and Cat Boost are used to construct prediction models. For each model the performance of the model is evaluated by cross-validation and the best performing model is selected.

2. Data and data processing

2.1. Data Preprocessing

The dataset was divided into feature and target variables, and the categorical variables were coded uniquely hot and the numerical variables were standardised.

The code uses solo thermal coding to transform categorical variables. Categorical variables are features with discrete values, such as gender, type of chest pain, etc. Solo thermal coding converts each discrete value into a binary code for use in the model. This conversion removes the magnitude relationship between categorical variables and prevents the model from being incorrectly interpreted as an ordered relationship.

In the code, the numerical features are standardized, i.e., the numerical features are scaled to a distribution with a mean of 0 and standard deviation of 1. Normalization removes the difference in magnitude between the features, making the model more stable, accelerating the training process and improving model convergence.

In summary, a comprehensive data processing and feature engineering step reduces the impact of data quality issues on the model and improves the robustness and accuracy of the model.

2.2. Check Values

2.2.1 Check Missing Value

The missing variable is used by this paper to calculate the number of missing values in each column using the df.isna().sum() function. Then, check if missing has a value greater than 0. If it does, it indicates the presence of missing values. If there are missing values, a fill or delete operation may be required to ensure data integrity.

2.2.2 Check Duplicate Value

The code uses df[df.duplicated()] to check for duplicate rows and outputs information if there are duplicates. If there are duplicate values, they need to be removed or disposed of.

2.3. Visualising Correlations

This paper use the heatmap function of the Seaborn library, heatmaps of correlations between features are drawn based on a matrix of correlation coefficients to make it easier to observe linear correlations between variables. Based on the correlation matrix, features with correlation greater than 0.5 with the target variable "heart disease" are selected (See Fig. 1).

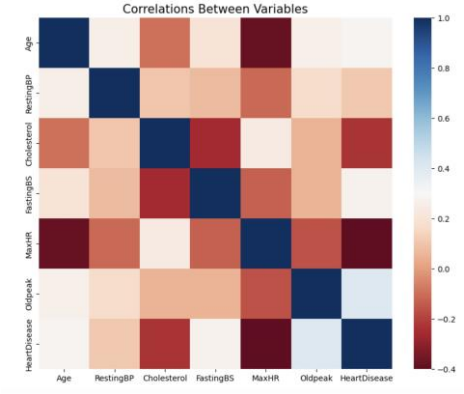


Fig. 1 Correlation Between Variables

2.4. Outlier Handling

2.4.1 Visualising outliers

This paper uses the boxplot function of the Seaborn library, boxplots of numerical features are plotted and used to identify potential outliers (See Fig. 2).

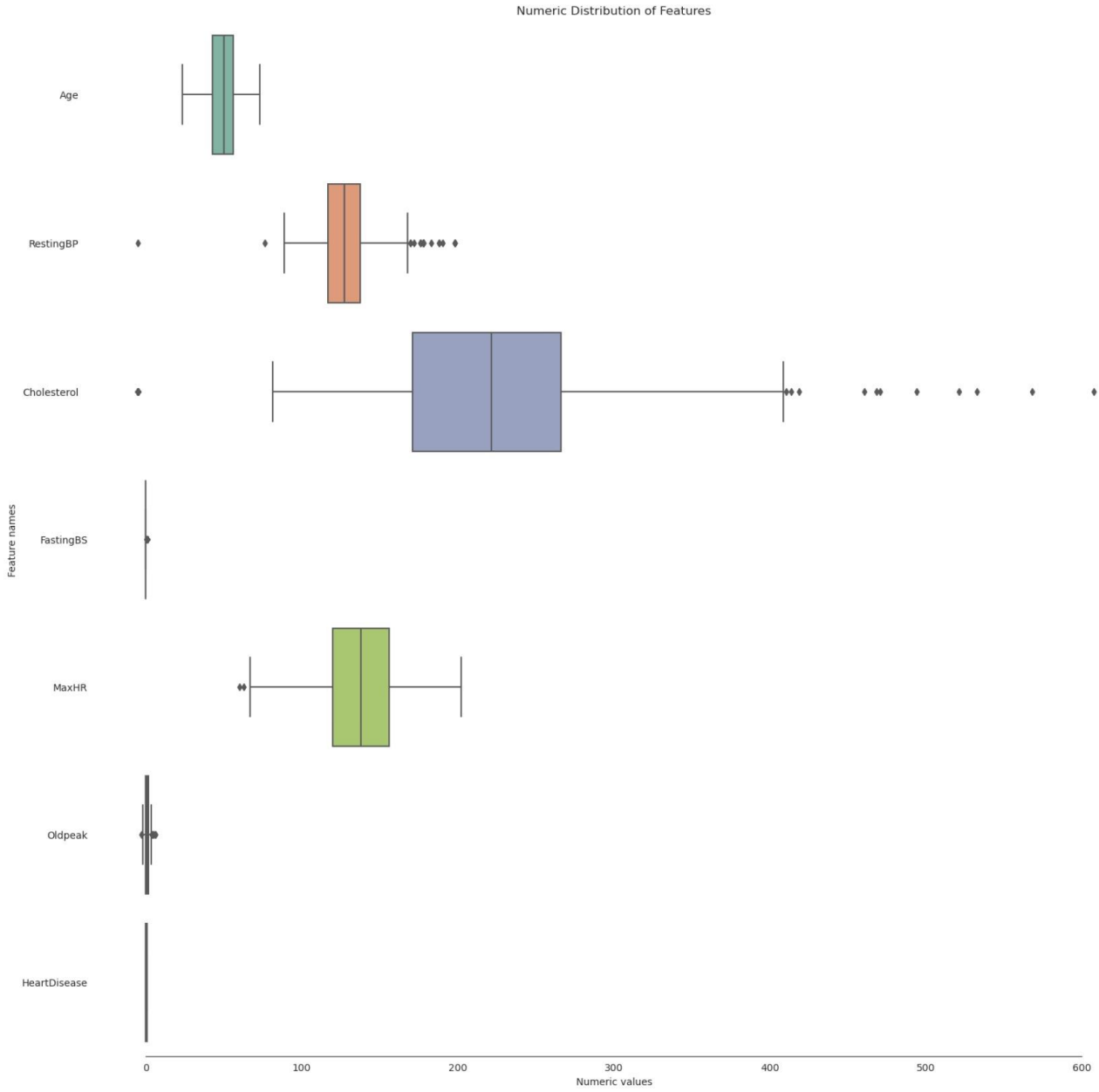


Fig. 2 Numeric Distribution of Features

2.4.2 Remove Outliers

Data points that are significantly different from the rest of the sample are known as outliers. In the code, a range of possible outliers is determined by calculating the upper and lower quartiles of each numerical feature and using the IQR (upper quartile minus lower quartile) method. The code excludes data points that fall outside this range to reduce the impact of outliers on model training. This helps to avoid the model from being affected by extreme values and improves the generalization of the model (See Fig. 3).



Fig. 3 Joint Distribution of Variables

3. Model

Multiple prediction models are constructed use different algorithms (Lasso, SGD Classifier, SVM, K Neighbors, Gaussian NB, Decision Tree, and Cat Boost). Each model uses Pipeline to contain preprocessors (solo thermal coding and normalization) and corresponding classifiers.

Lasso regression: a variant of linear regression that improves the generalization of the model by adding L1 regularisation to the coefficients. Can be used for feature selection, which can be achieved by reducing the coefficients of uncorrelated features to zero through L1 regularisation. Suitable for cases where covariance (high correlation between features) exists in the data, and can reduce model overfitting. Robust and insensitive to a small number of outliers.

SGD Classifier(Stochastic Gradient Descent Classifier): a stochastic gradient descent classifier for solving binary classification problems. Suitable for large-scale datasets, stochastic gradient descent methods can be trained efficiently. Online learning can be achieved, suitable for dynamic data.

SVM (Support Vector Machine): an algorithm commonly used for classification problems, which can be used for both binary and multiple classification. Can classify effectively in high-dimensional spaces. Suitable for non-linear problems, kernel functions are an option for mapping data into high-

dimensional space. The complexity of the model can be controlled by adjusting the kernel function and regularisation parameters.

K Neighbors (K Nearest Neighbour Classifier): Neighbour-based classification algorithm that makes predictions based on the K nearest neighbors. Simple to understand and easy to implement. Suitable for data whose distribution has no apparent structure.

Gaussian NB (Gaussian Plain Bayesian Classifier): classification algorithm based on Bayes' theorem, suitable for dealing with classification problems. Simple to compute, works better for high dimensional data. Can handle multi-class classification problems.

Decision tree (Decision Tree Classifier): an algorithm that uses a tree structure for classification. Easy to understand and interpret, generated decision trees can be visualized. Can handle both numeric and categorical features.

Cat Boost (Gradient Boosting Tree Based Classifier): a variant of the Gradient Boosting Tree algorithm, suitable for classification problems. Typically has good predictive performance and works well for complex problems. Does not require feature scaling and requires less preprocessing of the data.

All these algorithms are models used to solve classification problems. Classification problems are those that predict discrete category labels, while linear regression is typically used to solve regression problems, predicting continuous numerical outcomes.

4. Model performance comparison and validation

4.1. Model Performance Comparison

The performance of the different models is compared through training and cross-validation and the accuracy scores of each model are recorded. The scores are then visualized as bar charts for comparison (See Table 1 and Fig. 4).

Table 1. Score of the model

	Algorithms	Score
0	Lasso	0.240474
1	SGD Classifier	0.786026
2	SVM	0.899563
3	K Neighbors	0.864629
4	Gaussian NB	0.882096
5	Decision Tree	0.820961
6	Cat Boost	0.864629

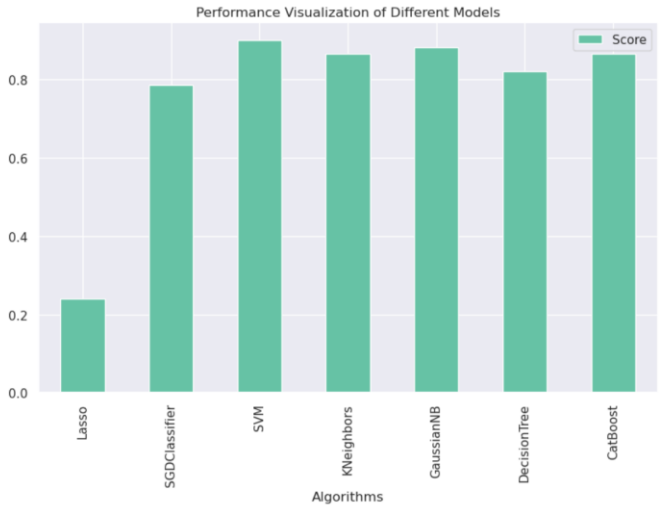


Fig.4 Performance Visualization of Different Models

From the output scores and Figure, it can be seen that SVM is the highest-scoring algorithm, so SVM is the best algorithm.

4.2. Cross-validation

Select the best performing model and perform a final evaluation of that model use cross-validation. The evaluation of the model's performance on unseen data is conducted by measuring the mean accuracy of cross-validation.

Overall, this code implements a prediction for a heart disease dataset that goes through the steps of data processing, feature engineering, and model construction to select the best-performing predictive model and evaluate the accuracy of that model on new data using cross-validation. Such a process makes full use of various data processing and machine learning algorithms, resulting in a final model with good generalisation and predictive performance.

5. Conclusion

Machine learning is being utilized extensively across various domains worldwide, including the healthcare sector. It plays a critical role in predicting the occurrence of movement abnormalities, heart disease, and other medical conditions. In this paper, data is collected to predict potential human cardiac conditions using various classification algorithms, such as Decision Tree, Naive Bayes, Logistic Regression, SVM, and Random Forest, through cross-validation, it can be learned that SVM is the best algorithm, and the complexity of the model can be controlled by adjusting the kernel function and regularisation parameters. Such as patient-specific symptoms, physical characteristics, clinical laboratory test values, and other data to help the doctors in diagnosis and treatment. Subsequently, it conducts biostatistical analyses of this patient data, unveiling patterns and correlations that might elude human physicians. The findings of this research hold great promise for advancing our understanding and management of cardiovascular diseases.

References

[1] Averbuch, T., Sullivan, K., Sauer, A., et al. Applications of artificial intelligence and machine learning in heart failure. European Heart Journal-Digital Health, 2022, 3(2): 311-322.

[2] Mpanya, D., Celik, T., Klug, E., & Ntsinjana, H. Machine learning and statistical methods for predicting mortality in heart failure. Heart failure reviews, 2021, 26(3): 545-552.

[3] Topkara, V. K., Elias, P., Jain, R., Sayer, G., Burkhoff, D., & Uriel, N. Machine learning-based prediction of myocardial recovery in patients with left ventricular assist device support. Circulation: Heart Failure, 2022, 15(1): e008711.

[4] Gautam, N., Ghanta, S. N., Clausen, A., et al. Contemporary applications of machine learning for device therapy in heart failure. Heart Failure, 2022, 10(9): 603-622.

[5] Chowdhury, M. N. R., Ahmed, E., Siddik, M. A. D., & Zaman, A. U. Heart disease prognosis using machine learning classification techniques. In 2021 6th International Conference for Convergence in Technology, 2021: 1-6.

[6] Alotaibi, F. S. Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 2019, 10(6).

[7] Cai, A., Zhu, Y., Clarkson, S. A., & Feng, Y. The use of machine learning for the care of hypertension and heart failure. JACC: Asia, 2021, 1(2): 162-172.

[8] Basha, Y. H., Nassif, A. B., & Al-Shabi, M. Predicting heart failure disease using machine learning. In Smart Biomedical and Physiological Sensor Technology XIV, 2022, (12123): 75-84

[9] Shah, D., Patel, S., & Bharti, S. K. Heart disease prediction using machine learning techniques. SN Computer Science, 2020, 1: 1-6.

[10] Sharma, V., Yadav, S., & Gupta, M. Heart disease prediction using machine learning techniques. In 2020 2nd international conference on advances in computing, communication control and networking, 2020: 177-181.