Heart Disease Prediction Using Machine Learning Models

Daoyi Cheng

School of Computer Science, University of California, Davis, California, United States
daoyi.cheng@outlook.com

Abstract. This study is aimed at the binary classification task of heart diseases, using 918 subjects and 11 clinical and diagnostic features from public datasets. Before training, the data underwent missing and anomaly checks, numerical features were standardized, category-specific features were encoded, and stratified sampling was used to divide the training/test sets in an 8:2 ratio. Some models enable category weights to alleviate mild class imbalance. The experiment compared Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest (RF) and on the same preprocessing process Multilayer Perceptron (MLP). Among them, the accuracy rate of KNN is 0.918 and the AUC is 0.953. The recall rate of SVM is the highest (0.951); The F1 value of RF is the highest (0.913), and it can provide feature importance. Overall, traditional machine learning methods have achieved stable and interpretable diagnostic performance on this data, among which KNN and RF perform even better in combination. The research conclusion can provide a reference for the early clinical identification of high-risk individuals. Subsequently, it can be verified on larger, cross-center data, and further combined with interpretability methods to support clinical decision-making.

Keywords: Heart Disease, Machine Learning, Random Forest, Classification, Medical Diagnosis.

1 Introduction

Cardiovascular disease remains among the leading causes of illness and mortality worldwide [1]. The traditional diagnostic method relies on doctors' manual judgment of electrocardiograms, laboratory test results and medical records, which is highly dependent on doctors' experience and work efficiency. If the number of patients increases, doctors may not be able to complete all diagnoses in a short time, resulting in an excessive workload and possibly affecting the accuracy of the diagnosis. Through data-driven techniques, for example, machine learning and neural networks, doctors can identify patients' conditions earlier, classify and predict heart diseases, thereby improving diagnostic efficiency and precision. [2].

Recent studies have consistently shown that machine learning methods are effective in predicting cardiovascular diseases [3]. For instance, some studies utilizing the hospital heart disease dataset have found that models such as logistic regression,

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_69

support vector machines, and decision trees can reach an accuracy rate of approximately 89% in identifying patients with heart disease [4]. Subsequent studies have indicated that ensemble learning algorithms (random forest and gradient boosting) perform better in prediction because they can reduce overfitting, thus enhance the generalization ability of the model [5]. However, the current research is still facing a number of problems, including the current datasets are imbalanced, lack interpretability, or even hard to generalize to a clinic [6].

It will compare various machine learning algorithms and neural network algorithms regarding accuracy, precision, recall, F1-score, and so on, with a fixed dataset related to patients. The difference between the results produced by various algorithms will be interpreted so that can better understand what affects the performance difference between algorithms. The final aim of this research is providing an efficient and accurate means, through this technological invention, for medical personnel to identify heart problems related to various diseases in advance.

The rest of the sections in this paper are presented as follows: Section 2 discusses the sources and features of the dataset, as well as how this research preprocesses data to make accuracy better. Section 2 will also discuss the five models selected for the classifications and tests. Section 3 will present the results and discuss the differences between model results. Section 4 summarizes the full text, discusses research limitations, and suggesting future work, such as applying the research results to the early diagnosis of other similar diseases to help medical workers improve their work efficiency.

2 Dataset and Methods

2.1 Dataset

Prior to training the models, the dataset underwent preprocessing to enhance data quality and ensure smoother model performance. During this process, the dataset has been checked for incomplete data or outliers in datasets and deleted incomplete or duplicate records to keep the data clean and reliable [7].

All numerical features have been standardized to keep them within a similar numerical range, thus facilitating better training. For some non-numeric variables in the dataset, such as ChestPainType and ST_Slope, this paper converts them into numeric form and processes them using the method of One-Hot Encoding. So that the model can make better use of these features [7].

In order to maintain the ratio of positive to negative cases, the dataset was then divided into training and testing subsets at an 80/20 ratio using stratified sampling.

Because of a slight imbalance in class distribution within the dataset, the parameter class_weight = "balanced" is set in some models to reduce the bias of the model towards most classes [8]. These preprocessing steps make the data more consistent and standardized, which helps the model to converge faster and improve the accuracy of prediction [9].

2.2 Models

Regarding the quantitative measurement of the predictive potential possessed by an algorithm for a heart disease prediction classification, five major commonly prevailing supervised learning algorithms have been selected and developed by utilizing the scikit-learn Python library [7]. The algorithms are applied to the same dataset, and to make the results evenhanded, all parameters are kept at their default values.

The algorithms studied in this research include: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP). Among all algorithms, the former four are categorized into traditional algorithms, which are easy to interpret and computationally efficient, whereas the last ones are neural network algorithms, which can handle a greater number of complexities among the features, and hence, they can be useful in improving the generalization capabilities of the results.

The process involved in training the model, the default hyperparameter settings are considered, and the random seeds are fixed to make the results reproducible. The performance of the model as a classifier was tested by checking the evaluation metrics, which included accuracy, precision, recall, F1, and AUC.

Based on this framework involving traditional algorithms combined with new approaches, research is able to study the interpretability and predictive accuracy of model results under various view perspectives, which will offer references for further medical diagnostic and medical applications.

3 Experiment

3.1 Experimental Setup

Google Colab environment was utilized to conduct all the tests. The coding atmosphere was developed utilizing Python 3.10 and the scikit-learn 1.5.1 library. Five types of supervised learning algorithms were developed to classify the patients. The aim was to identify whether a person has been a victim of heart problems or not. The critical parameters used to train each model are provided in Table 1.

Table 1. Key model parameters.

Model	Key Parameters
Logistic Regression (LR)	max_iter = 1000; class_weight = 'balanced'
K-Nearest Neighbors (KNN)	n_neighbors = 7; weights = 'distance'
Support Vector Machine (RBF)	C = 1.0; gamma = 'scale'
Random Forest (RF)	n_estimators = 200; max_depth = None; class_weight = 'balanced'; random_state = 42
Multilayer Perceptron (MLP)	hidden_layer_sizes = (64, 32); alpha = 1e-3; max_iter = 300; early_stopping = True

Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC) were the parameters measured to evaluate each model's performance. Taken collectively, the parameters provide a holistic measure of a model's robustness and effectiveness in correctly classifying a positive event.

3.2 Results and Analysis

Table 2. Model performance on the test set.

Model	Accuracy	Precision	Recall	F1	AUC
KNN	0.918	0.922	0.931	0.927	0.953
SVM (RBF)	0.891	0.866	0.951	0.907	0.949
Logistic Regression	0.897	0.888	0.931	0.909	0.930
Random Forest	0.902	0.896	0.931	0.913	0.929
MLP	0.864	0.835	0.912	0.885	0.924

Table 2 presents the overall performance of five models on the test set.

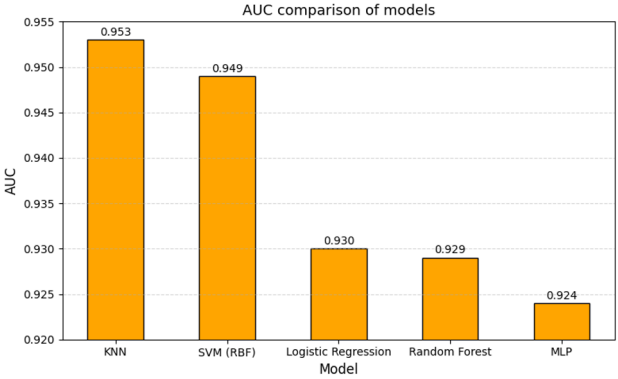

Fig. 1. AUC of models on the test set.

In terms of results, all models achieved a relatively high Accuracy rate (Accuracy exceeded 0.85) in the task of classifying heart diseases, indicating that the data preprocessing and model design were quite reasonable (see Figure 1).

Among them, the KNN model achieved the highest Accuracy and AUC value (Accuracy = 0.918, AUC = 0.953), performing the best. The Recall of SVM is the highest (0.951), and it is more sensitive in identifying diseased samples. Random Forest is slightly higher on the F1-score, indicating that it has achieved a balanced result in Precision and Recall. The performance of the MLP model is average, which may be related to the small sample size and the insufficient tuning of model parameters.

Overall, all models achieved relatively high accuracy (above 0.85), indicating that the data preprocessing and model design were effective.

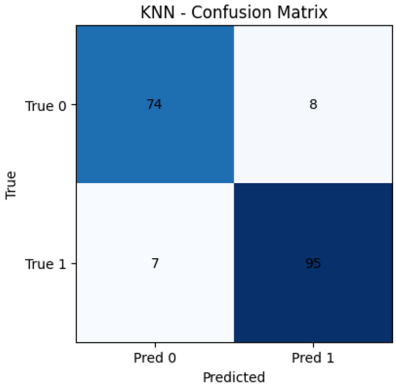

Fig. 2. Confusion matrix of the KNN model (threshold = 0.5).

This study will further demonstrate the results of KNN, as it is the best-performing among these models. Figure 2 shows its confusion matrix. It can be seen that the model can correctly distinguish between diseased and non-diseased samples in most cases.

The KNN model identified 95 True patients with heart disease (True Positive) and correctly judged 74 non-disease samples (True Negative). Meanwhile, the number of False positives and False negatives is relatively small, indicating that the model maintains a high Recall while also taking into account a certain level of Precision.

This manifestation is particularly important for medical prediction tasks, as in clinical screening, a missed diagnosis often leads to more serious consequences than a misdiagnosis.

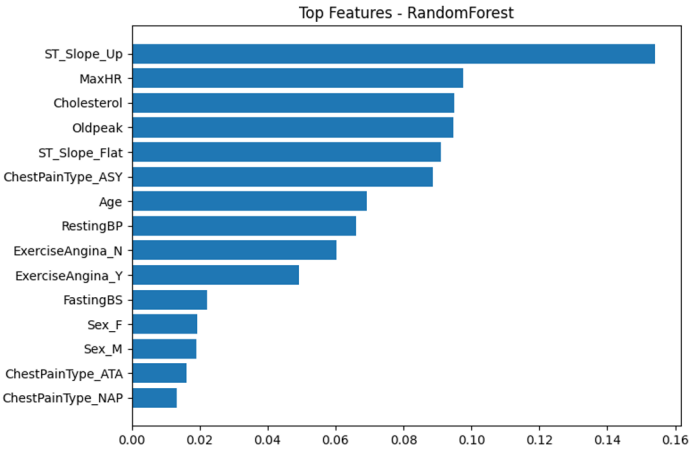

Fig. 3. The most important features identified by the Random Forest model.

Figure 3 shows the feature importance results of the Random Forest model. The results show that features such as ST_Slope, MaxHR, Cholesterol and Oldpeak have a relatively high contribution in the prediction of heart diseases.

These variables are consistent with clinical experience. Among them, ST segment changes and maximum heart rate reflect the risk of myocardial ischemia, while cholesterol levels are important indicators of cardiovascular diseases [10].

Therefore, machine learning models can not only help medical workers identify diseases, but also discover the hidden relationships between data and diseases as long as there is sufficient data, thereby providing new insights for medical research.

In summary, traditional models (such as LR, KNN, and SVM) perform better in terms of interpretability. Under conditions of high sample size and quality, they can accurately identify patients' diseases. In contrast, modern models (such as RF and MLP) have an advantage in feature extraction, being able to discover hidden connections between data and thus provide unexpected findings.

When talking about applications, varying models may be selected as required. For example, if it comes to screening, recall values may be given a higher priority to avoid overlooking a case. If a scenario involves resources or time constraints, Logistic or KNN could be a better choice due to their simplicity or lack of complexity, respectively.

4 Conclusions

This study employed several machine learning algorithms to predict heart disease using multiple variables, analyzing both the underlying factors and the corresponding results.

The accuracy rates of all models are above 0.85, and the overall performance is good. Among them, the KNN model has the best result, with an accuracy rate of 0.918 and an AUC value of 0.953. The recall rate of the SVM model is the highest, indicating that it is more sensitive in identifying patients with heart disease. Precision and recall are balanced in the Random Forest model. Its result has a certain degree of feature interpretability. The Logistic Regression model performs stably, with clear and easy-to-understand results. MLP algorithms perform below average, which may be related to the small sample size and limited feature dimensions. Overall, these models can all perform the task of predicting heart diseases quite well, providing a reliable foundation for subsequent analysis and comparison.

Although the model can predict and classify well, it still has limitations in the current situation. Firstly, the sample size of the training dataset is not that large, and it cannot be ensured that such high accuracy can be achieved outside the dataset. Furthermore, the features within the data are limited, and some unsupervised learning cannot discover deeper-level structures. If there is more data, conducting more research would be of greater help. The models used in the research were all classic ones with default parameters, and no more cutting-edge models related to research and artificial intelligence were employed. In the future, better performance can be achieved with more advanced models and more data.

Overall, this study indicates that machine learning methods can effectively support the early screening and risk assessment of heart diseases, providing doctors with faster and more accurate clinical decision-making references.

References

1. Martin, S.S., et al.: 2024 heart disease and stroke statistics: A report of US and global data from the American Heart Association. Circulation 149 (ePub ahead of print), 2024
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, 2016
3. Dwivedi, A.K.: Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications 29, 685–693, 2018
4. Kumar, A., Singh, M.: Comparative analysis of classifiers for heart disease prediction. Procedia Computer Science 132, 1393–1402, 2018
5. Rajendra, S., Kiran, P., Yadav, R.: Heart disease prediction using machine learning and deep learning techniques. IEEE Access 9, 1637–1647, 2021
6. Dua, D., Graff, C.: UCI Machine Learning Repository: Heart Disease [Data set]. University of California, Irvine, 2019. Available: https://archive.ics.uci.edu/dataset/45/heart%2Bdisease
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830, 2011
8. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284, 2009
9. Zhang, M.-L., Zhou, Z.-H.: A brief introduction to weakly supervised learning. National Science Review 5(1), 44–53, 2020
10. Goff, D.C., Lloyd-Jones, D.M., Bennett, G., Coady, S., D'Agostino, R.B., Gibbons, R., Greenland, P., Lackland, D.T., Levy, D., O'Donnell, C.J., Robinson, J.G., Schwartz, J.S., Shero, S.T., Smith, S.C., Sorlie, P., Stone, N.J., Wilson, P.W.: 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology 63(25), 2935–2959, 2014.