Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/75/20240498

Prediction of heart failure patients based on multiple machine learning algorithms

Yuhong Mo[1,6] , Yuchen Zhang[2] , Hanzhe Li[3], Han Wang[4] , Xu Yan[5]

[1]College of engineering, Carnegie Mellon university, PA, Pittsburgh, 15213, USA
[2]Mailman School of Public Health, Columbia University, New York, NY, 10032, USA
[3]New York University, NY, Brooklyn 11201, USA
[4]Financial Mathematics(department of Mathematics), University of Southern California,
Alhambra, CA, 91803, USA
[5]College of Graduate and Professional Studies, Trine University Phoenix, 85284, USA

[6]yuhongmo@cmu.edu

Abstract. Heart failure is a common heart disease whose incidence and mortality rate are increasing year by year. In order to predict heart failure accurately, three models, LightGBM, adaboost and XGBoost, were used for training and evaluated in this paper. After data preprocessing, the data was divided into training and test sets in the ratio of 7:3 and the models were evaluated using parameters such as precision, accuracy, recall and F1 score. The results showed that the best performer in terms of prediction accuracy was the LightGBM model, which achieved 88.4% accuracy, followed by the adaboost model with 87.7% accuracy, and the XGBoost model, which also achieved 87.3% prediction accuracy. In conclusion, all three prediction models achieved more than 85% accuracy and could accurately predict a patient's heart failure. Confusion matrix results showed that each model was able to effectively identify both positive and negative samples in the test set with high sensitivity and specificity. These results indicate that these models are highly reliable and practical in practical applications, and can provide important reference information for doctors to help them better diagnose and treat heart failure patients, thus improving treatment outcomes and survival rates.

Keywords: Heart failure, XGBoost, LightGBM.

1. Introduction

Heart failure is a common cardiac disease, and its incidence and mortality are increasing year by year. Currently, research on heart failure focuses on early prediction, diagnosis and treatment [1,2]. Among them, machine learning algorithms have become a powerful tool for predicting heart failure.

Machine learning algorithms can analyse and process large amounts of data to find patterns and build models from them, and then predict what may happen in the future [3]. In heart failure prediction, machine learning algorithms can build prediction models and assess the risk of patients by analysing their clinical data, physiological indicators and imaging data [4]. For example, algorithms such as

DOI: 10.54254/2755-2721/75/20240498

support vector machine and random forest have been applied to heart failure prediction and achieved good results [5,6].

The application of machine learning algorithms in heart failure prediction is of great significance. Firstly, it can help doctors to more accurately assess the risk of patients and develop personalised treatment plans [7]. Second, it can improve the accuracy and efficiency of early diagnosis and intervention, thereby reducing patient mortality and hospitalisation rates [8]. Finally, machine learning algorithms can mine the hidden features and patterns of patients, providing new ideas and methods for the etiological research and treatment of heart failure [9].

The application of machine learning algorithms in heart failure prediction has achieved certain results, but still needs further improvement and optimisation. In this paper, based on publicly available datasets in heart failure, we use a variety of novel machine learning algorithms to predict heart failure based on various physiological indicators of patients, and compare the advantages and disadvantages of each model in terms of prediction effect. With the continuous improvement of data acquisition technology, algorithmic models and computational power, it is believed that machine learning algorithms will play an increasingly important role in heart failure prediction and treatment [10].

2. Data sources

The data used in this paper is from Kaggle's publicly available dataset on heart failure prediction. The dataset contains various physiological indicators of the patients. These are Age, Gender, Chest Pain Type, Resting BP, Cholesterol, Fasting BS, Resting ECG, MaxHR, Exercise Angina, Oldpeak, ST Slope, and Heart Disease, where Heart Disease is the final diagnostic category of the patient, with 1 indicating a diagnosis of the presence of heart disease and 0 indicating that the patient's heart is Normal. The data of some cases are shown in Table 1.

Table 1. Partial data.

Age	Sex	Resting BP	Cholesterol	Fasting BS	Max HR	ExerciseAngina	Oldpeak	ST Slope	Heart Disease
40	M	140	289	0	172	N	0	Up	0
49	F	160	180	0	156	N	1	Flat	1
37	M	130	283	0	98	N	0	Up	0
48	F	138	214	0	108	Y	1.5	Flat	1
54	M	150	195	0	122	N	0	Up	0
39	M	120	339	0	170	N	0	Up	0
45	F	130	237	0	170	N	0	Up	0
54	M	110	208	0	142	N	0	Up	0

3. Statistical analysis of data

Each quantitative variable was statistically analysed and the maximum, minimum, median, variance and mean of each variable were calculated and the results are shown in Table 2.

Table 2. Parameter list.

Variable	Size	Max	Min	Average	Standard deviation	Median
Age	918	77	28	53.511	9.433	54
Resting BP	918	200	0	132.397	18.514	130
Cholesterol	918	603	0	198.8	109.384	223
Fasting BS	918	1	0	0.233	0.423	0
Max HR	918	202	60	136.809	25.46	138
Oldpeak	918	6.2	-2.6	0.887	1.067	0.6
Heart Disease	918	1	0	0.553	0.497	1

DOI: 10.54254/2755-2721/75/20240498

4. Model Introduction

4.1. LightGBM

LightGBM is a decision tree-based gradient boosting framework developed by Microsoft Research Asia. It is highly respected for its efficiency, accuracy and scalability.

LightGBM uses two key technologies: histogram-based decision trees and GOSS (Gradient-based One-Side Sampling). Histogram-based decision trees involve discretising continuous features into discrete values and accelerating the computation process by constructing histograms. This approach significantly reduces memory consumption and computation time, thus increasing the speed of model training and prediction.GOSS, on the other hand, is a gradient optimisation technique, which improves model accuracy by sampling samples and adjusting weights.

In addition to this, LightGBM uses a leaf-wise growth strategy, where the leaf node with the largest gradient is selected for splitting each time, resulting in a deeper and narrower tree. This method can effectively avoid overfitting and further improve model accuracy.

In practice, LightGBM has been widely used in a variety of fields, including natural language processing, image recognition, recommender systems and so on. It has not only won many awards in Kaggle competitions, but has also been used by many large companies and organisations for real-time prediction tasks in production environments.LightGBM has become one of the most popular algorithms in the field of machine learning through its unique decision tree construction method and gradient optimisation technique, as well as its efficient leaf-wise growth strategy.

4.2. Adaboost

Adaboost is an integrated learning algorithm, which is an iterative algorithm that builds a strong classifier by combining multiple weak classifiers.Adaboost is widely used in the field of machine learning because of its high accuracy and simplicity.

The principle of Adaboost is simple: it trains multiple weak classifiers by assigning a weight to each training sample, and each weak classifier minimises the weights of misclassified samples. In the next iteration, the misclassified samples will be given a larger weight, which allows the next weak classifier to better handle these difficult to classify samples. Finally, all the weak classifiers are weighted and combined to become a strong classifier.

The development of the Adaboost model has gone through three stages: the first stage is the original version of the Adaboost model, which considers only two types of data during training and uses only a single-layer decision tree as the base classifier. Although the model had some flaws, it laid the foundation for later improvements. The second stage is the multi-category Adaboost model, which allows to deal with datasets with multiple category labels and is trained using any base classifier. This version of the Adaboost model performs well in solving multi-category classification problems, but it still has some drawbacks such as sensitivity to noisy data. The third stage is the enhanced version of the Adaboost model, which addresses the shortcomings of the first two versions by introducing some improvements. The enhanced Adaboost model performs even better in handling noisy and unbalanced data and can be trained using a variety of base classifiers.

The Adaboost model is a powerful machine learning algorithm that builds a strong classifier by iteratively combining multiple weak classifiers. Although the model has some flaws, it is still one of the most effective algorithms in many real-world application scenarios.

4.3. XGBoost

XGBoost is an integrated learning algorithm based on decision trees, which is widely used in the fields of machine learning and data mining.The principle of XGBoost consists of two main aspects: decision trees and gradient boosting. Decision tree is a classification and regression model based on recursive partitioning, which constructs a tree structure by dividing the feature space. Each leaf node represents a category or a numerical output. Gradient boosting is an iterative optimisation method that improves model performance by continuously adding new weak learners. Each weak learner is trained on the residuals of all previous weak learners and has a certain weight.

XGBoost continuously optimises the model performance by combining multiple decision trees to form a powerful integrated model with a gradient boosting algorithm. In each iteration, XGBoost calculates the error (i.e., residual) between the current model's prediction of the sample and the true result, and uses the residual as the target for the next training. Also, XGBoost regularises each decision tree to prevent overfitting.

XGBoost is a powerful machine learning algorithm that builds an efficient and accurate integrated model through decision trees and gradient boosting. With the increasing amount of data and expanding machine learning application scenarios, XGBoost is expected to play an important role in more fields in the future.

5. Experiments and Results

After data preprocessing, the data is first divided into training and testing sets in the ratio of 7:3, with 70% of the data used to train the models and 30% of the data used to test the models. LightGBM, adaboost and XGBoost models are introduced for training respectively, and the models are evaluated using parameters such as precision, collinearity, recall and F1 score, and the confusion matrix of each model is output to observe the classification prediction of the test set.

Fig. 1 shows the confusion matrix of LightGBM test set, Fig. 2 shows the confusion matrix of adaboost test set, the Figure 3 shows the confusion matrix for the XGBoost test set. The prediction accuracy of each model is shown in Table 3 and Figure 4.

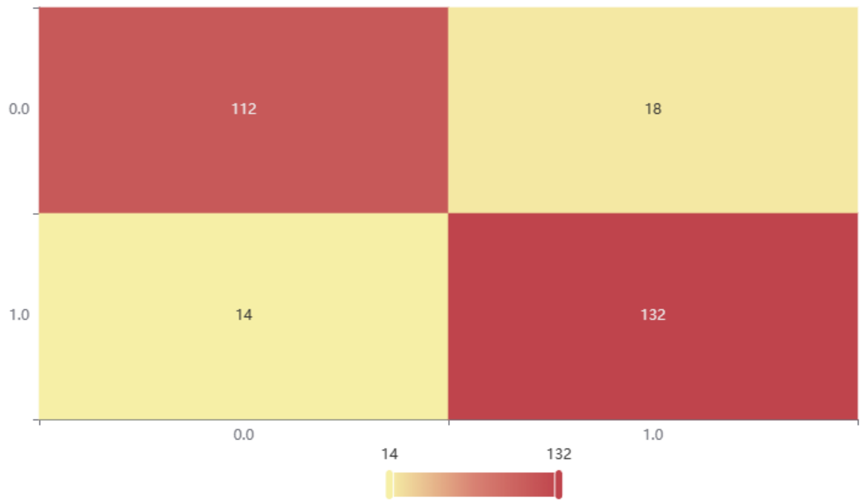

Figure 1. LightGBM confusion matrix.
(Photo credit : Original)

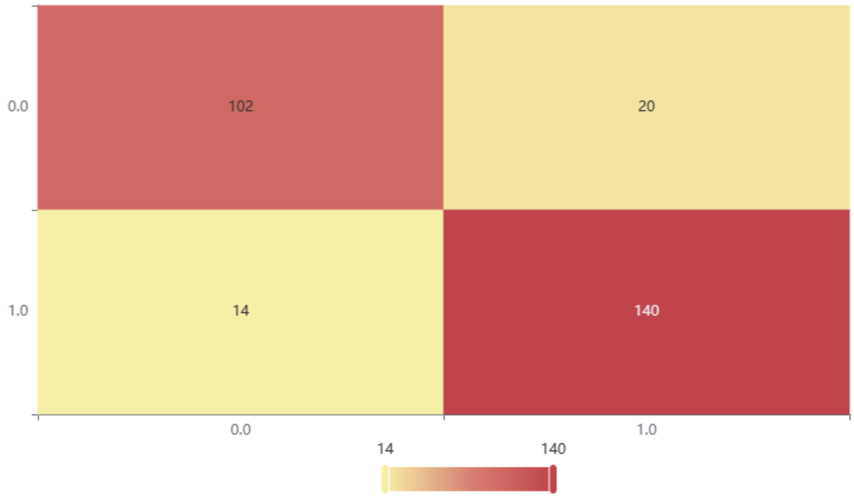

Figure 2. Adaboost confusion matrix.
(Photo credit : Original)

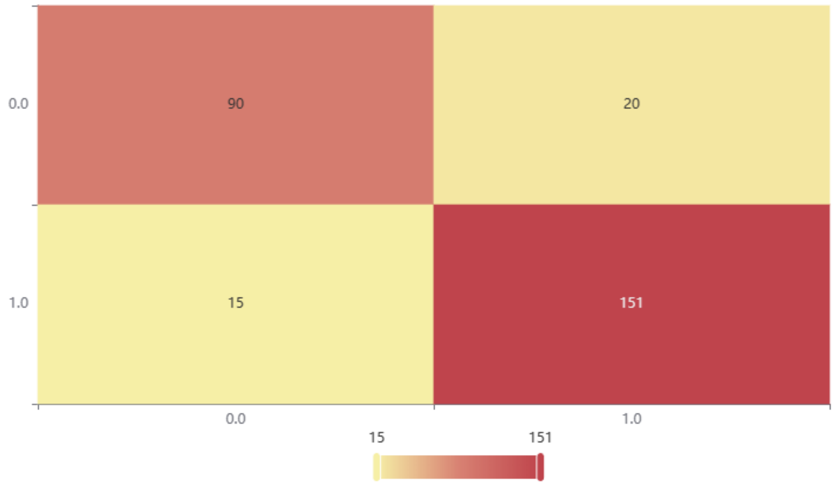

Figure 3. XGBoost confusion matrix.
(Photo credit : Original)

Table 3. Modelling evaluation.

Model	Accuracy	Precision	Recall	F1 Score
LightGBM	88.4	88.4	88.4	88.4
adaboost	87.7	87.7	87.7	87.6
XGBoost	87.3	87.3	87.3	87.3

DOI: 10.54254/2755-2721/75/20240498

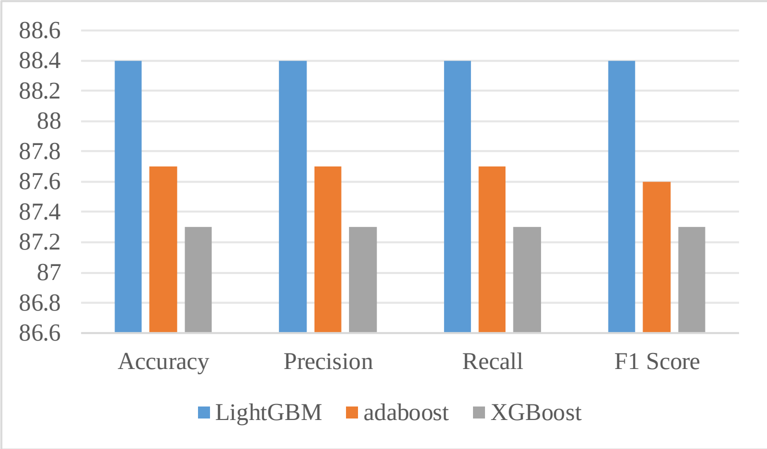

Figure 4. Modelling evaluation.
(Photo credit : Original)

As can be seen from the confusion matrix and model evaluation indexes, the best performance in terms of accuracy of heart failure prediction was achieved by the LightGBM model with an accuracy of 88.4%, followed by the adaboost model with an accuracy of 87.7%, and the XGBoost model with a prediction accuracy of 87.3%. In conclusion, all three prediction models achieved more than 85% accuracy and can accurately predict a patient's heart failure.

6. Conclusion

Heart failure is a common heart disease, and its incidence and mortality are increasing year by year. In order to improve the prediction accuracy of heart failure, three machine learning models such as LightGBM, adaboost and XGBoost are introduced in this paper for training and the models are evaluated using the parameters such as precision, collinearity, recall and F1 score.

After data preprocessing, we divided the training set and test set in the ratio of 7:3 and trained them using the three machine learning models. By analysing the evaluation metrics of the models, we found that the best performer was the LightGBM model, which achieved a prediction accuracy of 88.4%. This is followed by the adaboost model with an accuracy of 87.7% and the XGBoost model with a prediction accuracy of 87.3%. This indicates that all three prediction models are able to predict heart failure accurately and with high reliability in practical applications.

The confusion matrix allows us to observe the classification prediction of the test set. We found that in the LightGBM model, both true positive examples (TP) and true negative examples (TN) are high, while false positive examples (FP) and false negative examples (FN) are low. This indicates that the model is able to accurately predict whether a patient is suffering from heart failure or not.

In summary, all three machine learning models are able to accurately predict heart failure and are highly reliable in practical applications. Among them, the LightGBM model performed the best, with a prediction accuracy of 88.4%. Therefore, in future clinical practice, we can use these machine learning models to help doctors more accurately determine whether a patient is suffering from heart failure and take effective treatment measures in a timely manner, so as to improve the therapeutic effect and survival rate.

References

[1] Zang, Hengyi, et al. "Evaluating the Social Impact of AI in Manufacturing: A Methodological Framework for Ethical Production." Academic Journal of Sociology and Management, vol. 2, no. 1, 2024, pp. 21-25.

[2] Dong, Xinqi, et al. "The Prediction Trend of Enterprise Financial Risk based on Machine Learning ARIMA Model." Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, 2024, pp. 65-71.
[3] Zang, Hengyi, et al. "Evaluating the Social Impact of AI in Manufacturing: A Methodological Framework for Ethical Production." Academic Journal of Sociology and Management, vol. 2, no. 1, 2024, pp. 21-25.
[4] Liu, Shun, et al. "Financial Time-Series Forecasting: Towards Synergizing Performance And Interpretability Within a Hybrid Machine Learning Approach." 2023. arXiv preprint arXiv:2401.00534.
[5] "A Deep Learning-Based Algorithm for Crop Disease Identification Positioning Using Computer Vision." International Journal of Computer Science and Information Technology, vol. 1, no. 1, 2023, pp. 85-92.
[6] "Machine Learning Model Training and Practice: A Study on Constructing a Novel Drug Detection System." International Journal of Computer Science and Information Technology, vol. 1, no. 1, 2023, pp. 139-146.
[7] Liu, B. (2023). "Based on Intelligent Advertising Recommendation and Abnormal Advertising Monitoring System in the Field of Machine Learning." International Journal of Computer Science and Information Technology, 1(1), 17-23.
[8] Liu, B., Yu, L., Che, C., Lin, Q., Hu, H., & Zhao, X. (2023). "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872.
[9] Tianbo, Song, Hu Weijun, Cai Jiangfeng, Liu Weijia, Yuan Quan, and He Kun. "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition." In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 834-837. IEEE, 2023. DOI: 10.1109/mce.2022.3206678.
[10] Rammal F H ,Emam Z A .Heart Failure Prediction Models using Big Data Techniques[J].International Journal of Advanced Computer Science and Applications (IJACSA),2018,9