Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/112/2025.18135

Predictive Analytics in Heart Disease: Leveraging LightGBM for Improved Diagnostic Accuracy

Runxing Han[1,a,*], Ruixuan Meng[2], Qianwei Zhu[3]

[1]FedUni Information Engineering Institute, Hebei University of Science and Technology, Shijiazhuang, Hebei, 050018, China
[2]School of Information Science and Engineering, Shandong Agricultural University, Taian, Shandong, 271000, China
[3]School of Computer Science and Engineering, Guangzhou Institute of Science and Technology, Guangzhou, Guangdong, 510000, China
a. 2216030109@stu.hebust.edu.cn
**corresponding author*

Abstract: The prevalence of heart disease is significant, making it a prominent health concern, and early prediction and diagnosis are critical. The application of artificial intelligence algorithms in heart disease prediction is highly promising. This study utilizes the LightGBM algorithm for the predictive modelling of a heart disease dataset containing 1025 records and 14 health indicators. Predictive models were effectively built through data preprocessing, feature selection, and model training. The results show that the LightGBM model got 98.54% accuracy in the test set is better than the model of support vector machine (SVM) and Logistic Regression. Age, heart rate, and cholesterol levels are the main factors identified by feature analysis. The Area Under the Curve (AUC) value of the model reaches 0.985, proving its excellent performance. Despite this, the model's generalization ability may be impacted by the dataset's size and representativeness. Future research may take into account the use of larger and more heterogeneous datasets to further verify the effectiveness of the LightGBM algorithm in heart disease prediction. This study shows the effectiveness of the LightGBM algorithm in predicting heart disease, which can provide support for making medical decisions and public health management.

Keywords: LightGBM, heart disease, prediction accuracy.

1. Introduction

Heart disease is still a major cause of death worldwide and represents a significant threat to human health. According to the statistical data from the World Health Organization (WHO), cardiovascular disease (CVD) may cause 17.9 million people to die [1]. This means that the death data accounts for one-third of the total number of people deaths around the world [2]. The high morbidity and mortality rates of heart disease not only pose a huge health burden but also lead to a huge pressure on each country's economy and health system.

The factor of heart disease is always connected to various behaviors closely, such as having bad eating habits, lacking physical exercise, smoking, drinking wine, and so on [3]. Poulter's research emphasizes that the major risk factor is high blood pressure, high cholesterol, smoking, and diabetes.

DOI: 10.54254/2755-2721/112/2025.18135

These elements may affect the health of the heart by the various biological and chemical pathways [4, 5]. Heart disease is a typical multi-factor and multi-mechanism disease, that needs more and more integrated multidimensional data to conduct effective risk assessments and management.

With the development of information technology, the data-driven disease prediction model has become an important component in disease management and prevention strategy, especially in the field of artificial intelligence [6]. Machine learning technology, especially advanced algorithms like LightGBM, receives attention due to its ability to solve complex models in big data environments. Miah et al. compare and study with many machine learning models, finding that LightGBM's performance in heart disease prediction is better than other traditional models, which display a high accuracy and process speed [7]. In addition, Yang et al. verified LightGBM's high performance in heart disease datasets, especially in the efficiency and effect in dealing with large-scale and high-dimensional data [8]. Besides, Vicente explored the application of LightGBM in medical data analysis, which emphasized the advantages of interpretability and operability [9].

Although the current study in risk prediction of heart disease has improved a lot, there also exist some limitations and challenges. For example, many models don't adequately consider individual differences in the patients, such as the impact of genetic background, environmental factors, and living habits. Besides, many studies concentrated on a single population or region, lacking validation of universality across populations and regions.

In addition, the prevention of heart disease and early intervention has been proven to be the key to reducing the death rate of heart disease. Effective model prediction not only helps doctors identify high-risk individuals to implement the targeted measures for prevention but also optimizes resource allocation and improves the efficiency of medical services. Therefore, improving an efficient tool for predicting heart disease not only has important significance for medical research but also has the actual application value of public health management and policy making.

The research mainly adopts the LightGBM algorithm to make predictions and analyses for heart disease data set from the Kaggle website. The data set includes 1025 records, which cover 14 main health indicators [10]. Therefore, this study aims for in-depth analysis and applies the LightGBM algorithm to explore a more precise and personalized model for predicting heart disease, adapting to different demands from different people. In this way, the main objective is a deep comprehension of heart disease risk elements. And improve reliable predictive models to assist in making medical decisions.

2. Method

This study uses the algorithm called LightGBM to forecast heart disease and focus on predictive analytics to improve diagnostic correctness. The method part describes the data source, feature selection, and calculation method in detail.

2.1. Data source and illustration

The dataset used was sourced from Kaggle and contains 1,025 records covering 14 key health indicators considered relevant to foreseeing heart disease. The index comprises variables such as age, gender, type of chest discomfort, maximum achieved heart rate, angina induced by exercise, resting blood pressure levels, serum cholesterol levels, fasting blood sugar levels, results of the resting electrocardiogram indicating a decline in ST-segment due to exercise and relative rest periods, and peak ST segment slope during motion, the number of major blood vessels in fluorescent contrast for colouring and the result of the thallium pressure test. Table 1 provides a detailed overview of the role and significance of each attribute in heart disease. The data sets combine classification and continuous variables to provide a solid foundation for the exploitation of predictive models.

Table 1: Data attributes description

Attribute	Description	Data Type
age	Age of the patient	Integer
sex	Sex of the patient (1 = male; 0 = female)	Binary
cp	Chest pain type (values 0-3; 4 values)	Categorical
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Integer
chol	Serum cholesterol in mg/dl	Integer
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)	Binary
restecg	Resting electrocardiographic results (values 0, 1, 2)	Categorical
thalach	Maximum heart rate achieved	Integer
exang	Exercise induced angina (1 = yes; 0 = no)	Binary
oldpeak	ST depression induced by exercise relative to rest	Float
slope	The slope of the peak exercise ST segment	Categorical
ca	Number of major vessels (0-3) colored by flourosopy	Integer
thal	Thalassemia (0 = normal; 1 = fixed defect; 2 = reversible defect)	Categorical
target	Diagnosis of heart disease (1 = disease; 0 = no disease)	Binary

2.2. Feature Selection and Description

Feature selection is based on the factors that previous research indicated associated with heart disease risk, like elevated cholesterol levels, blood pressure, lack of physical activity, and genetic predispositions. An automatic feature selection technique is used in LightGBM, which evaluates feature importance based on gain and split metrics, allowing the model to be optimized by the most influential factors.

2.3. Method introduction

The analysis method is centered around LightGBM, this sophisticated machine learning technique is renowned for its high efficiency in processing large-scale and high-dimensional data. Using stratified sampling, the dataset was split into a training set (80%) and a test set (20%) to preserve the proportional representativeness of the resulting categories. The LightGBM gradient boosting framework is configured to minimize binary log loss as its objective function, which is suitable for binary classification problems (Whether or not there is heart disease). To quantify the accuracy of the model, the study used binary log loss, a measure of probabilistic prediction uncertainty comparing the probabilistic prediction to the true label, as shown in the following equation:

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot log(\hat{y_i}) + \left(1 - y_i\right) \cdot log\left(1 - \hat{y_i}\right)\right] \tag{1}$$

Where y_i is the observed value, $\hat{y_i}$ is the predicted probability, and N is the number of samples.

The model is guaranteed to be safe with this approach optimized to enhance predictive accuracy by minimizing the error in probability estimation for the classification of heart disease presence or absence.

2.3.1. Model parameters

Model setup parameters included a learning rate of 0.05, 31 leaves per tree, and unlimited tree depth. Model regularization by feature ratio (0.9), bagging ratio (0.8), and frequency 5 helped to reduce overfitting by subsampling features and data points. To prevent overfitting, we employ an early stopping mechanism that terminates the training process if there is no improvement in the validation score for ten consecutive iterations. Model performance evaluation involves utilizing metrics such as

DOI: 10.54254/2755-2721/112/2025.18135

accuracy, precision, recall, and F1 score derived from both the classification report and confusion matrix. For a comprehensive understanding of each parameter configuration, please refer to Table 2.

2.3.2. Statistical analysis

Statistical Analysis employs the Receiver Operating Characteristic (ROC) curve to assess the model's diagnostic ability in various threshold settings and the Area Under the Curve (AUC) to summarize the overall diagnostic accuracy. The trade-offs between sensitivity and specificity in disease prediction models are essential for understanding these measures.

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{t})\mathrm{dFPR}(\mathrm{t}) \tag{2}$$

When t indicates various thresholds, TPR stands for the True Positive Rate, and FPR for the False Positive Rate.

For statistical tests, the ROC curve was used to assess the model diagnostic capabilities under different width settings. And the AUC to summarize diagnostic accuracy. These actions are important to us to understand the trade-offs between sensitivity and specificity in disease prediction models.

Table 2: LightGBM Model Parameter Description Table

arameter Name	etting Value	Description
objective	binary	Specifies a binary classification task.
metric	binary_logloss	Evaluates model performance using log loss.
boosting_type	gbdt	Employs tree-based gradient boosting.
learning_rate	0.05	Determines step size; smaller values require more trees.
num_leaves	31	Sets the maximum number of leaves per tree; higher numbers can lead to overfitting.
max_depth	-1	Unlimited tree depth.
feature_fraction	0.9	Uses 90% of features each iteration to enhance randomness.
bagging_fraction	0.8	Samples 80% of data per iteration to avoid overfitting.
bagging_freq	5	Performs bagging every five iterations.
verbose	-1	Minimizes execution output.

3. Results and Discussion

3.1. Model Evaluation and Performance

The LightGBM model used in the research performed well in the task of predicting heart disease. The result shows that the correctness of the model on the test set reached 98.54%, which displays the strong and stable predictive ability of the model. In the confusion matrix (Figure 1), it can be seen that the model correctly classified all 102 samples without heart disease (true negative cases) and correctly classified 100 out of 103 samples with heart disease (true cases), with only three misclassifications (false negative cases), indicating that the model's specificity reached 1.0 and sensitivity was 0.97. The result shows the efficiency of this model in predicting heart disease can effectively distinguish between illness and non-diseased individuals and reduce the possibility of misdiagnosis and missed diagnosis.

DOI: 10.54254/2755-2721/112/2025.18135

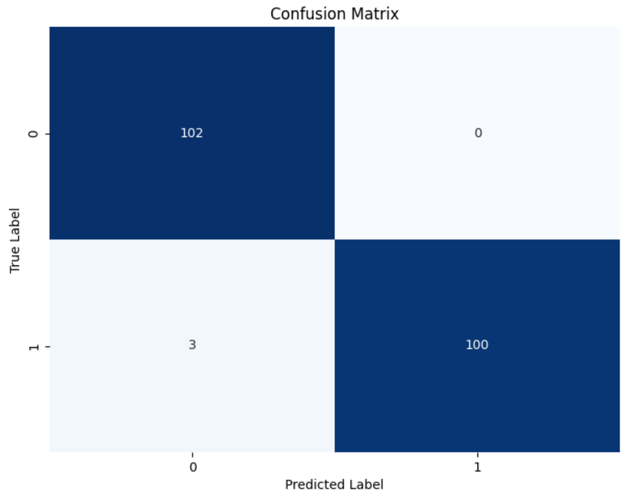

Figure 1: Confusion Matrix

In addition, the receiver of the ROC curve is shown in Figure 2, with an area AUC of 0.985, which is close to 1. The result further confirms the outstanding ability of the model to distinguish between heart disease patients and non-patients. To evaluate the performance of binary classification models, it is important to consider the value of AUC. The model's high classification accuracy and stability can be confirmed by an AUC value close to 1.

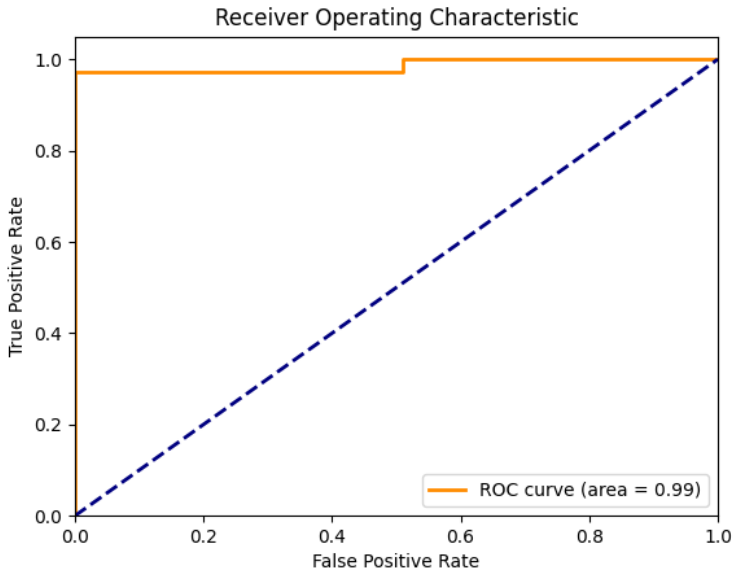

Figure 2: ROC

3.2. Feature importance analysis

The model's importance analytics demonstrate the contributions of every feature to the prediction results. As shown in Figure 3, the feature importance with the highest score is age, which is 422. It indicates age is the most critical factor in heart disease prediction. This is consistent with existing

DOI: 10.54254/2755-2721/112/2025.18135

medical research results that age is an important risk factor for heart disease, and the older the age, the higher the incidence of heart disease.

Besides, the feature importance score of maximum heart rate (thalach) and serum cholesterol (chol) are 385 and 366 respectively, which also demonstrate the importance of heart disease prediction. Maximum heart rate and serum cholesterol level are both common indicators of clinical, high serum cholesterol levels are often associated with a higher risk of heart disease.

In contrast, fasting blood sugar (fbs) and sex (sex) had lower feature importance scores, which are 15 and 78 respectively. It may reflect these factors had relatively little impact on heart disease prediction in this dataset. Feature importance analysis helps identify the most critical predictor variables, providing some valuable insights in model optimization and medical research.

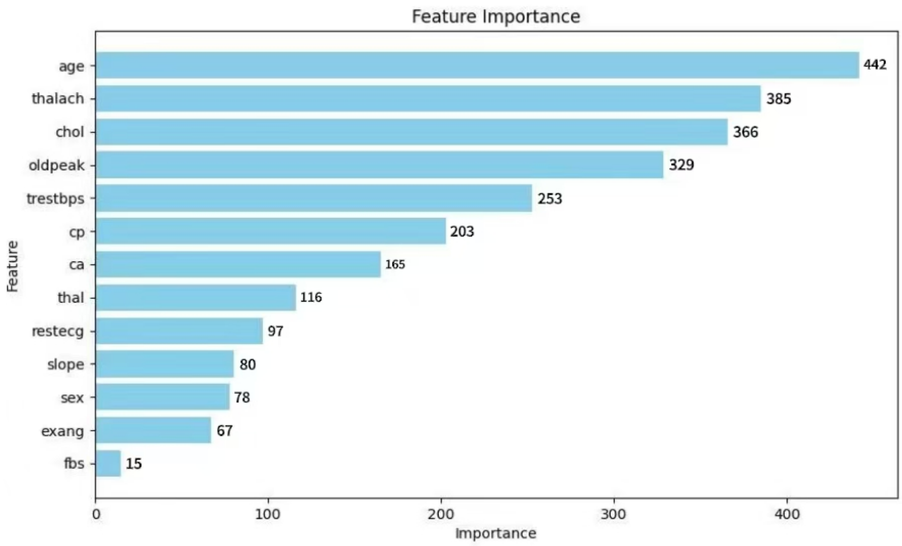

Figure 3: Feature Importance

3.3. Application of Statistical Analysis

The performance evaluation model not only calculates the accuracy, sensitivity, and specificity but also performs other statistical analyses of the performance of the comprehensive assessment model. Using the binary log loss to be one of the model evaluation criteria. The training set of the binary log loss lost about 0.0548, but the testing set of the binary log loss lost about 0.1394. These data show that the model performs well in both the testing set and training set and can keep highly predictive ability in the new data.

Besides, utilizing the ROC curve and AUC value can in-depth analyze the model's functioning. Understanding the model under different thresholds of performance can be achieved by analyzing the ROC curve. However, the AUC value can provide a conclusion about the model of the whole performance. The AUC value (0.985) approach 1 demonstrates that the model has a formidable discriminative ability and can discriminate accurately between heart disease patients and non-patients. Combining these statistical indicators, the model exhibits highly predictive accuracy and reliability.

Overall, through a comprehensive evaluation of model performance, analysis of feature importance, and detailed statistical analysis, this research uses the LightGBM model in heart disease predicted test perform excellently. It can provide effective support to the early diagnosis and prevention strategies for heart disease. This discovery not only verifies the potential applications of

DOI: 10.54254/2755-2721/112/2025.18135

machine learning models in medical prediction but can provide a strong theoretical foundation for future research.

3.4. Comparison of Model Accuracy

This section compares and analyzes the accuracy of three machine learning models that are SVM, logistic regression, and LightGBM to evaluate their performance in heart disease prediction tasks. The following Table 3 summarizes the accuracy results of these models on the same test set and next, the analysis process will be outlined:

As mentioned in 3.1, LightGBM performs well in the research, with an accuracy rate as high as 98.54%, which is significantly better than other models. This model constructs shallower trees and uses histogram-based to improve the method. It effectively enhances the capability to handle large-scale data and improves the model's predictive accuracy.

SVM is a powerful linear classifier that is based on interval maximization and kernel techniques. In the heart disease prediction dataset of this study, the accuracy of SVM was 88.78%. The result shows that SVM can effectively handle linearly separable problems. But in the face of nonlinear complex data structures, SVM is slightly insufficient.

The primary objective of Logistic Regression, a statistical learning method, is to estimate the probabilistic relationship between binary variables. In this research, the accuracy of the Logistic Regression Model is 79.51%, which shows the applicability and limitations in datasets with strong feature linear correlations.

Comprehensively comparing the performance of the models together, LightGBM performs the best in solving heart disease prediction tasks compared to other models. The highly computational performance and outstanding predictive accuracy that it can significantly outperform SVM and Logistic Regression in solving high-dimensional and complex data structures. Although SVM and Logistic Regression can provide good results under certain conditions, the performance of these models is limited by the characteristics of itself and the complexity of the dataset.

Table 3: Model Accuracy Comparison

Model	Accuracy
LightGBM	98.54%
SVM	88.78%
Logistic Regression	79.51%

4. Conclusion

This study explores a more accurate and personalized heart disease prediction model adapted to the needs of different populations through in-depth analysis and application of the LightGBM algorithm. Using the dataset, which includes 1025 records and 14 key health indicators from the Kaggle website, this research not only can validate the potential of machine learning models in medical prediction but also provide a strong theoretical foundation for future research. Although the LightGBM algorithm performs an accuracy of up to 98.54% in heart disease prediction, which is significantly superior to traditional models, the research does not completely exhibit the speed advantage when using the LightGBM algorithm to solve large-scale data due to the relatively small scale of the dataset.

This study especially emphasized the importance of preventing heart disease and early intervention. Effective model prediction not only helps doctors identify high-risk individuals and implement targeted preventive measures but also optimizes resource allocation and improves the efficiency of medical services. Besides, the importance of identifying factors such as age, maximum heart rate, and serum cholesterol levels for predicting heart disease by analyzing the characteristics of heavy drugs

DOI: 10.54254/2755-2721/112/2025.18135

is beneficial for optimizing the model and deepening the understanding of risk factors for heart disease.

All in all, this research indicates that applying the LightGBM algorithm in heart disease prediction and early diagnosis can significantly improve prediction accuracy, provide support for medical decision-making, and bring some practical application values to public health management and policy development. The research can further explore the generalizability of the model in different people and regions in the feature, as well as the potential applications in other disease predictions, and more comprehensively assess LightGBM's performance benefits by using larger datasets.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] *Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, Md. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. BioMed Research International, 2023(1).*

[2] *Yuan, X., Chen, J., Zhang, K., Wu, Y., & Yang, T. (2022). A stable AI-based binary and multiple class heart disease prediction model for IoMT. IEEE Transactions on Industrial Informatics, 18(3), 2032–2040.*

[3] *El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature analysis of coronary artery heart disease data sets. Procedia Computer Science, 65, 459–468.*

[4] *Poulter, N. (1999). Coronary heart disease is a multifactorial disease. American Journal of Hypertension, 12(S6), 92S-95S.*

[5] *Rethik, K., Singh, A., Singh, D., & Rakhra, M. (2023). Comparative analysis of different heart disease prediction models. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. 992–996).*

[6] *Mack, M. J. (2011). Risk scores for predicting outcomes in valvular heart disease: How useful? Current Cardiology Reports, 13(2), 107–112.*

[7] *Miah, J., Ca, D. M., Md Abu Sayed, Ehsanur Rashid Lipu, Mahmud, F., & Yasir Arafat, S. M. (2023). Improving cardiovascular disease prediction through comparative analysis of machine learning models: A case study on myocardial infarction. arXiv.*

[8] *Yang, H., Chen, Z., Yang, H., & Tian, M. (2023). Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison. IEEE Access, 11, 1–1.*

[9] *Garcez Vicente, A. L., Malaquias Junior, R. D., & Romero, R. A. F. (2024). Explainable LightGBM approach for predicting myocardial infarction mortality. arXiv.*

[10] *Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310.*