

Heart Disease Prediction Using Gradient Boosting Decision Trees

Yuzhe Wu^a

Munus International Engineering School, Fuzhou University, Fuzhou, 350000, China

Keywords: Prevention, Dataset, GBDT, Machine Learning Models, Heart Disease.

Abstract: Heart disease is one of the major health challenges globally, and its prevention and treatment are crucial for ensuring people's health. This study is based on the 2020 stacked ensemble survey dataset for heart disease classification. By analyzing the relationship between various factors and heart disease, we explore the application of machine learning models in heart disease prediction. The study found that factors such as fasting blood sugar, cholesterol, and exercise-induced angina are closely related to heart disease, while the influence of resting electrocardiogram and resting blood pressure is relatively small. Among various machine learning models compared, Gradient Boosting Decision Trees (GBDT) performed the best, with high prediction accuracy and precision. However, the study also points out the limitations of the dataset and the issue of models not fully unleashing their potential. It is worth noting that this study also explores the possibility of using other machine learning models in heart disease prediction and conducts comparative analysis, providing more references for heart disease prevention and treatment.


1 INTRODUCTION

Heart disease, as a sudden and severe ailment, has always been a concern for society and humanity. According to the "China Cardiovascular Health and Disease Report 2020," there are approximately 330 million people with cardiovascular diseases, including 13 million cases of stroke, 11.39 million cases of coronary heart disease, 5 million cases of cardiomyopathy, 45.3 million cases of peripheral arterial disease, and 245 million cases of hypertension, imposing an increasing economic burden on society and becoming a significant public health issue (Cheng, 2023). Research indicates that due to its diverse and complex disease types and extremely high mortality rates, heart disease has become a formidable challenge in the field of medicine (Lin, 2019). Not only does heart disease severely impact individual health, but it also imposes a substantial burden on the socio-economic landscape. The expensive medical costs required for treating heart disease, coupled with the need for long-term medical care and rehabilitation, contribute to the profound economic impact.

To address the challenge of achieving a qualitative leap in the short term in the medical

treatment of heart disease, and considering the rapid development of information technology and artificial intelligence, people have begun to prioritize the prevention and prediction of heart disease to control the occurrence of the disease at its source. To achieve this goal, various machine learning models (such as logistic regression, decision trees, deep neural networks, etc.) are widely used to analyze and predict various factors related to heart disease. These models are utilized to infer various indicators closely related to heart disease and control these indicators to reach the critical threshold for preventing heart disease attacks.

This study is based on the "Stacked Ensemble for Heart Disease Classification" dataset from 2020 and aims to analyze the factors leading to heart disease. Firstly, feature selection and visualization techniques were applied to the dataset to better understand the data. Subsequently, a series of conclusions were drawn through the analysis of relevant factors. Finally, this study compared the performance of multiple machine learning models, demonstrating the superiority of Gradient Boosting Decision Trees in this task.

^a <https://orcid.org/0009-0007-5343-2359>

2 RELATED RESEARCH

Traditional methods for predicting heart disease often rely on the experience of doctors and manual analysis, limited by human cognitive ability and information processing speed, which can easily lead to subjective biases and misjudgments. With the rapid advancement of technology and the era, people are increasingly realizing the advantages of integrating artificial intelligence and machine learning models into heart disease prediction. Weng et al. pointed out in a prospective cohort study on predicting cardiovascular disease that machine learning algorithms significantly improved the prediction of cardiovascular disease, confirming the effectiveness and feasibility of machine learning techniques in cardiovascular disease prediction (Weng, 2017).

For example, Li Linghai compared the effectiveness of traditional feature extraction algorithms and deep learning methods in the classification of echocardiograms, finding that deep learning significantly improved the classification accuracy, especially in the classification of heart disease (Li, 2017). Cheng Z. and other scholars integrated Random Forest with SHAP for heart disease prediction, and found that it can predict heart disease more accurately (Cheng, 2023). Liu Yunlong and other scholars conducted feature selection research on heart disease prediction based on GBM, and found that this method can effectively improve the efficiency of medical diagnosis (Liu, 2023). Shafiey M G and other scholars introduced an efficient hybrid genetic algorithm and particle swarm optimization method based on Random Forest to optimize the feature extraction process, in order to select key features that enhance the accuracy of heart disease diagnosis (Shafiey, 2022).

Mohan et al. proposed a heart disease prediction model called HRFLM, constructed using a linear mixed RF algorithm. The model improved the level of disease prediction performance, achieving an

accuracy of 88.7% (Mohan, 2019). Ali et al. in 2002 utilized an adapted form of the Health Belief Model, selecting a dataset of 178 female patients with coronary heart disease, and conducted an analysis of coronary heart disease risk factors prediction (Ali, 2002). In 2009, Avci utilized genetic algorithms to optimize parameters of the Support Vector Machine model and experimentally validated it on heart disease data. The research results indicated that this method could achieve better predictive performance (Avci, 2009). In 2013, Amin et al. proposed a hybrid model combining artificial neural networks with genetic algorithms, aimed at optimizing the connection weights of neural networks to improve the predictive performance of artificial neural networks. The model utilized 50 identified important risk factors to predict heart disease, and the research results demonstrated an accuracy of 89% for the predictive model (Amin, 2013).

Previous studies have utilized models such as Logistic Regression, Decision Trees, and Deep Neural Networks to analyze the related data, and these models indeed exhibit a certain level of accuracy and play an important role to some extent. However, when faced with large and dense datasets, these traditional models still have some limitations, particularly in discovering higher-order relationships and achieving further precision. However, GBDT, which this experiment is based on, can effectively address the shortcomings of previous models

3 METHOD

By comparing different machine learning models and using performance comparison metrics, we determine the optimal model. These models include, but are not limited to, GBDT and LR. We compare their performance in predicting heart disease and select the best-performing model as the final conclusion (Fig. 1).

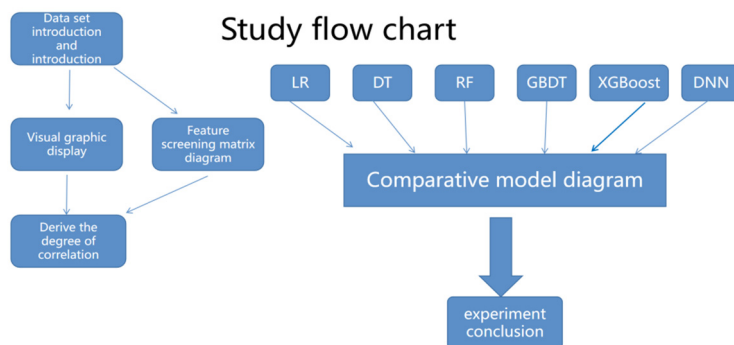


Figure 1: Research Workflow Diagram (Photo/Picture credit: Original).

3.1 GBDT

The predicted values (results of weak classifiers) are cumulatively stacked to equal the original value. Each time, the current prediction serves as the baseline, and the next classifier fits the residual of the error function on the predicted values (GBDT's weak classifiers use tree models) (Figure 2).

In a given dataset $D_i^n : \{x_i, y_i\}$ with a sample size of n , GBDT is a commonly used training method. It aims to improve predictive performance by iteratively training a series of decision tree models. Each round of training produces a weak learner that minimizes the residual loss of the current model. As the iterations progress, each new model is trained on the residuals of the previous model, eventually forming a powerful ensemble model. In the context of Federated GBDT, the CART regression algorithm is typically used to train each decision tree model.

$$L(y_i, f(x_i)) = \frac{1}{2} \sum_{i=1}^n (y'_i - y_i)^2 \tag{1}$$

In the GBDT framework, y'_i represents the predicted value, y_i represents the actual value, and $y'_i - y_i$ denotes the difference between the predicted and actual values, i.e., the gradient, also referred to as the residual.

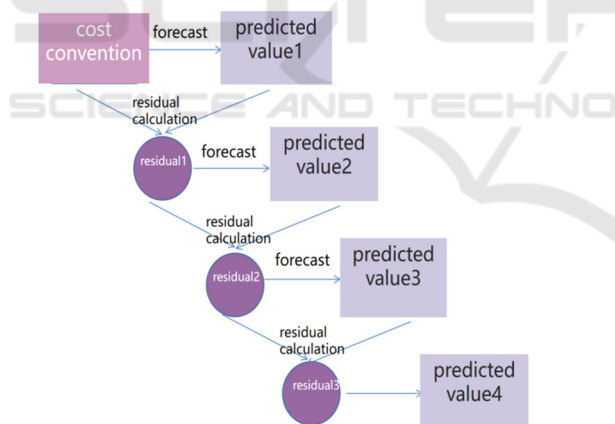


Figure 2: Flowchart of GBDT (Picture credit: Original).

3.2 LR

Learning from a set of samples' multiple feature values to establish a model (abstracted as a formula $f(x)$), using this model to predict the outcomes of other samples, with logistic regression being the classification of predicted results.

We have established a LR model with i independent variables, denoted by $x_1, x_2, x_3 \dots \dots x_i$, to study the presence of heart disease risk. The event of absence or presence of heart disease risk is

represented by the dependent variable $y \in \{0, 1\}$, where $y=0$ indicates the absence of heart disease risk and $y=1$ indicates the presence of heart disease risk. The presence of heart disease risk is denoted by p . Therefore, the LR model is defined as follows:

$$Z = w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_ix_i \tag{2}$$

$$P = \frac{\exp(\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_ix_i)}{1 + \exp(\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_ix_i)} \tag{3}$$

Normalize the values of x , then define the weight function, activation function, and loss function for gradient descent. Finally, make predictions using the trained weight parameters w .

3.3 DT

DT is a graphical method used in decision analysis to evaluate project risks and feasibility by calculating the probability of net present value being greater than or equal to zero based on known probabilities of various scenarios. It intuitively applies probability analysis to assess the feasibility of projects.

$$H(x) = - \sum_{i=1}^n p_i \log_2 p_i \tag{4}$$

Here, $H(x)$ represents entropy, and $P(x)$ represents the probability of event x occurring. Information entropy is a critical factor in DT algorithms. Borrowing the concept of entropy, information entropy is introduced to describe the occurrence probability of discrete random events in decision tree algorithms. Typically, the more frequent an event occurs, the higher the emphasis on information entropy, indicating lower uncertainty. Conversely, higher uncertainty is associated with less frequent occurrences.

3.4 RF Model

RF is a classifier consisting of many decision trees, which can be used for both classification and regression problems, as well as for dimensionality reduction. It also has good tolerance to outliers and noise. As an ensemble learning method, the RF model, within the framework of the Bagging algorithm, utilizes decision trees as the base training models to improve the generalization ability of the model. We conducted a RF experiment, first using the Bootstrap method to randomly extract n subsets from the original sample set, and then selecting k features from multiple features as the subset of features. This process is repeated n times to obtain n subsets of features. Next, we randomly match the subsets of training sets and features to train n decision trees.

Finally, we integrate the predictions of these n decision trees using a voting method to obtain the final prediction result.

3.5 Extreme Gradient Boosting

XGBoost is an optimized distributed gradient boosting library designed to be efficient, flexible, and portable. It features boosting ensemble algorithm characteristics. The goal of boosting ensemble algorithms is to improve the overall model performance by building multiple weak classifiers and aggregating their results. The XGBoost algorithm achieves this goal by ensuring that each iteration generates the optimal new tree. To ensure that the results of multiple weak estimators are better than those of a single estimator, XGBoost uses a method of sampling with replacement. Simple description, the final predicted value is the sum of the predicted values for each tree.

3.6 DNN

DNN consists of input layers, hidden layers, and output layers. When constructing a DNN, it is necessary to choose an appropriate structure, including the number of hidden layers and the number of nodes in each hidden layer. The selection of these parameters can evaluate the complexity of the model. When the complexity of the model is insufficient to learn all the features in the samples, the network will exhibit underfitting, manifested by large errors in both the training and test sets. Conversely, when the complexity of the model is too high, the network will exhibit overfitting, manifested by small errors in the training set but large errors in the test set. The structural diagram of the DNN adopted in this experiment is as follows (Figure 3).

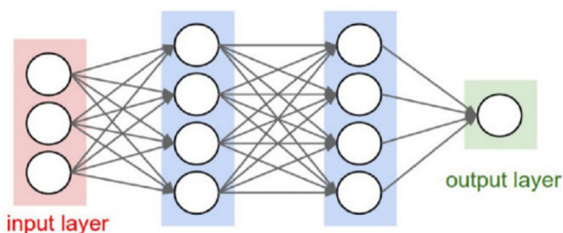


Figure 3: The structural diagram of the DNN (Picture credit: Original).

4 RESULT

4.1 Data Set

The dataset used in this experiment is sourced from the "Stacked Ensemble for Heart Disease Classification" dataset on Kaggle. This dataset comprises 14,280 data points with 12 distinct features, resulting in 1190 rows. The characteristic quantities of this experiment will be shown in the visualization section.

First, this paper perform feature selection on the dataset. The workflow is illustrated in Figure 4.

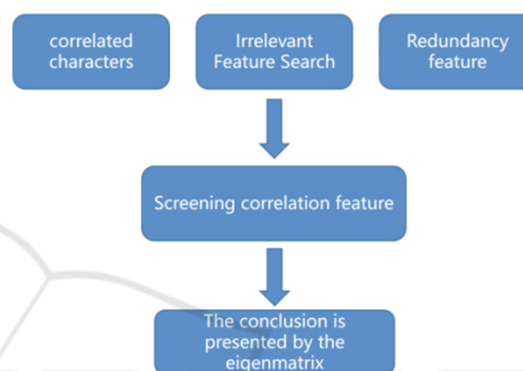


Figure 4: Workflow of Feature Selection (Picture credit: Original).

With the deepening of research, this work summarize and present the feature matrix in Figure 5, aiming to discover the relationship between each feature and the target variable.

Through this paper's analysis of the dataset, the paper have identified some notable correlations among the features (see Figure 5). Specifically, this work observe that the ST slope exhibits the highest positive correlation with the target variable, with a value of 0.51. Additionally, exercise-induced angina and chest pain type also show relatively high positive correlations, at 0.48 and 0.46, respectively. Furthermore, the maximum heart rate demonstrates the highest negative correlation with the target variable, with a value of -0.41.

4.2 Visualizations and Discussion

It is evident that the resting electrocardiogram (resting ecg) and resting blood pressure (resting bp s) exhibit relatively low correlations with the target variable, at 0.073 and 0.12, respectively. These values are close to zero and can be considered negligible. This

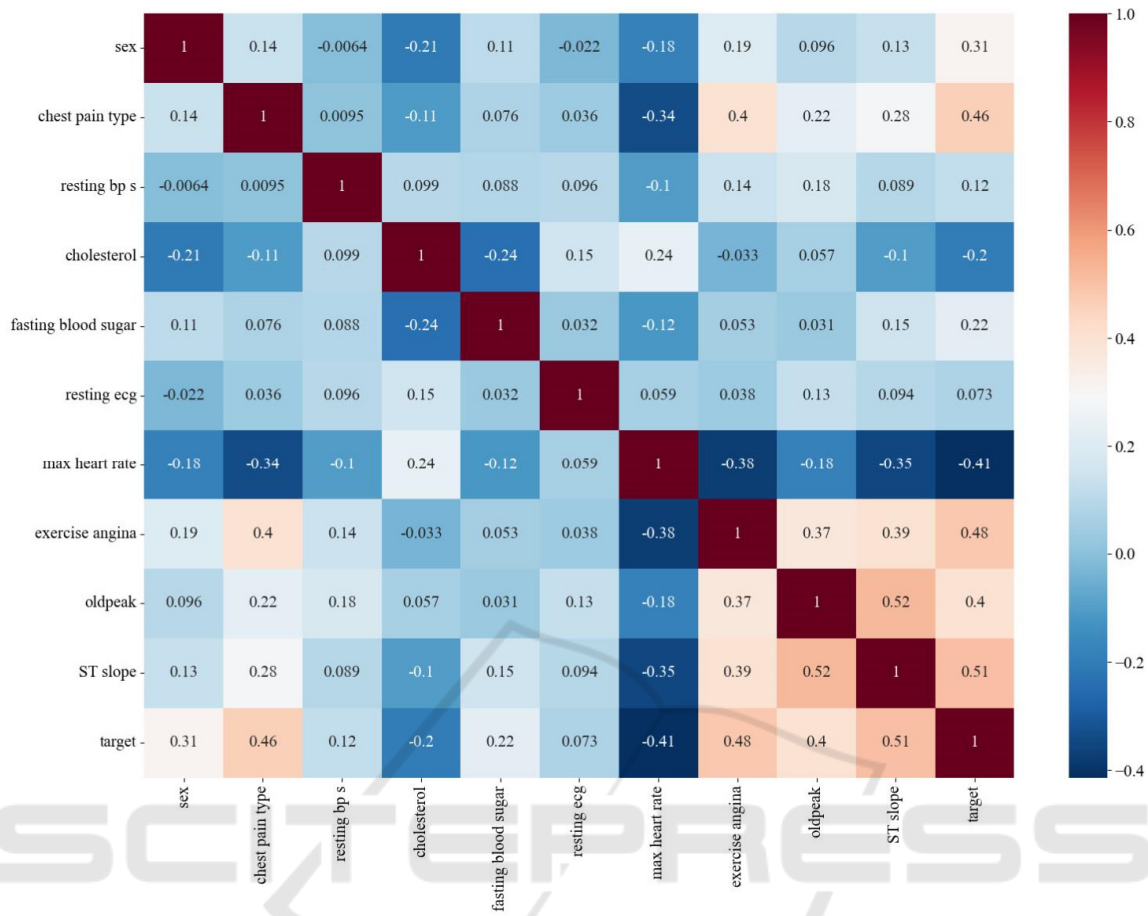


Figure 5: Feature Matrix Visualization (Photo/Picture credit: Original).

suggests that the linear relationships between these two features and the target variable are weak, potentially limiting their predictive capabilities. In the process of model training and feature selection, this paper may consider excluding these features to enhance the efficiency and accuracy of the model.

Based on the visualization in Figure 6, this paper draw the following conclusions: Firstly, the probability of maintaining heart health is higher when blood sugar levels are lower. Secondly, excessive exercise may lead to abnormalities and consequently cause heart problems. Regarding cholesterol levels, there is a higher proportion of abnormalities in the range of approximately 100 to 300, while there is also a significant proportion of abnormalities in the range of approximately -100 to 100. Additionally, males have a higher probability of abnormalities. Concerning chest pain type, asymptomatic abnormalities constitute a significant proportion in the graph. Lastly, in the relationship between maximum heart rate and heart health, the interval for

normal individuals is generally farther back compared to individuals with abnormalities.

Based on the visualization in Figure 7, the distribution of the target variable in resting ecg and resting bp s shows minimal differences overall. This suggests that the correlation between these two features and the target variable is relatively weak. This conclusion aligns with the findings from the subsequent feature matrix, which also indicated a weak correlation between resting electrocardiogram and resting blood pressure with the target variable, thereby corroborating each other.

These findings further validate the limited predictive power of resting electrocardiogram and resting blood pressure in predicting the target variable, emphasizing that they may not be the most influential predictive factors.



Figure 6: Visualization of Relationships between each Feature and the Target Variable 1 (Photo/Picture credit: Original).

To compare various models and select the most suitable one for accurately predicting this type of problem, we utilized multiple performance evaluation metrics, including AUC (Area Under the Curve), accuracy, precision, recall, and F1-score. By evaluating these metrics comprehensively, this paper were able to compare the performance of different models and ultimately determine the optimal model.

By comprehensively considering the above evaluation metrics, this paper can gain a more comprehensive understanding of the performance advantages and disadvantages of various models. This enables us to select the model that performs best in terms of accuracy, generalization capability, and robustness, thereby providing the optimal choice for predicting and studying the problem.

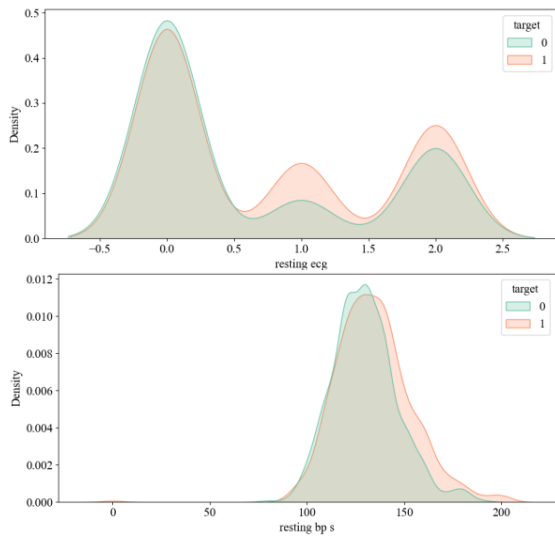


Figure 7: Visualization of the Relationship between Various Features and the Target Variable 2 (Photo/Picture credit : Original).

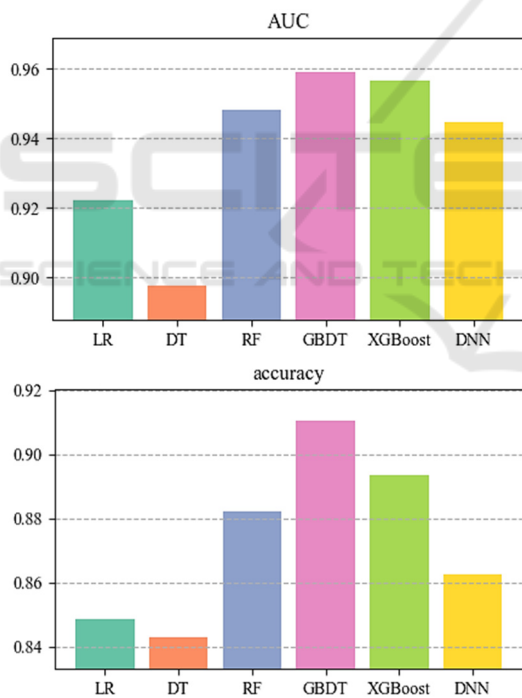


Figure 8: Evaluation of AUC and accuracy for each model (Photo/Picture credit: Original).

Table 1: Numerical Statistics 1 (Values rounded to 3 decimal places).

Methodology	AUC	accuracy
LR	0.922	0.849
DT	0.898	0.843
RF	0.949	0.882
GBDT	0.959	0.910
XGBoost	0.957	0.892
DNN	0.945	0.862

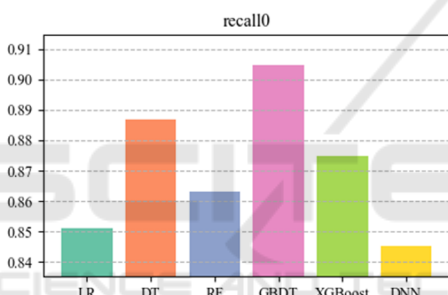
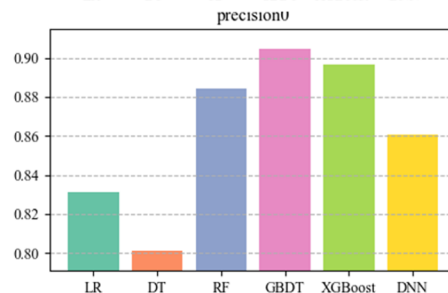
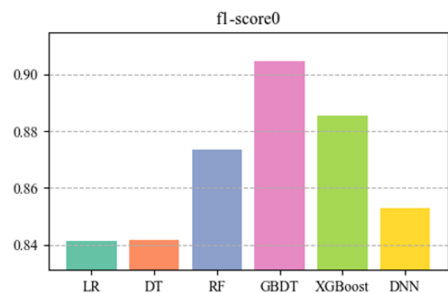
Based on the results from Figure 8 and Table 1, the differences among various models in terms of AUC and accuracy are not significant, mostly ranging from 80% to 90%. However, it is evident that GBDT exhibits a prominent advantage compared to other models. This indicates that in our study, the GBDT model demonstrates superior performance in terms of predictive accuracy and AUC values.

Based on the performance evaluation results from Figure 9 and Table 2, it is evident that GBDT consistently outperform other models, whether before or after feature selection. This indicates that the GBDT model exhibits outstanding performance across different conditions in our study.

The superiority of GBDT can be attributed to its faster prediction speed and higher accuracy. Its parallel prediction capability enables excellent performance on information-dense and large-scale datasets. Moreover, GBDT demonstrates excellent interpretability and robustness, automatically discovering high-order relationships between data without requiring additional preprocessing operations. Additionally, the model exhibits stronger adaptability to complex datasets, effectively handling complex and high-order relationships between data. Therefore, it is more suitable for datasets with diverse influencing factors, dense information, and complex relationships. Overall, despite the large scale and diversity of influencing factors in the experimental dataset, GBDT still holds significant advantages in handling such data types.

In conclusion, based on the comparative research results, we conclude that GBDT exhibit outstanding advantages compared to other models. In this paper’s experiments, both before and after feature selection, the GBDT model demonstrates excellent performance across various performance evaluation metrics, including AUC, accuracy, precision, recall, and F1-score. This indicates that GBDT provides higher accuracy and reliability in predicting heart disease, making it suitable for addressing such problems.

Pre-Selection Phase



shortlisted

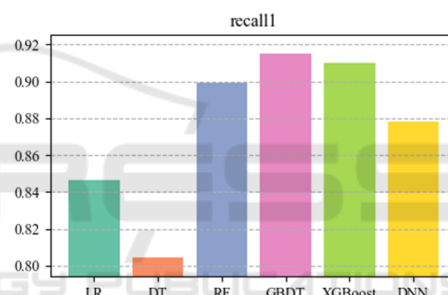
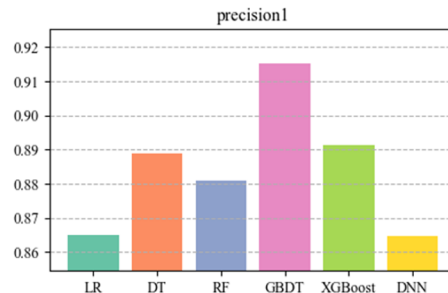


Figure 9: Evaluation Metrics of Various Models Before and After Selection (Photo/Picture credit: Original).

Table 2: Numerical Statistics 2 (Values Rounded to Three Decimal Places).

Methodology	F1-score0	Precision0	Recall0	Precision1	Recall1
LR	0.841	0.831	0.851	0.865	0.843
DT	0.841	0.800	0.887	0.889	0.802
RF	0.872	0.882	0.862	0.881	0.900
GBDT	0.903	0.902	0.904	0.915	0.914
XGBoost	0.883	0.898	0.875	0.891	0.910
DNN	0.852	0.860	0.845	0.865	0.879

5 CONCLUSION

To address the increasing prevalence and unpredictable nature of heart disease in recent years, this study developed a GBDT model for predicting heart disease, which demonstrated more precise analysis and prediction capabilities. By using the GBDT model, this work could accurately identify risk factors for heart disease onset and provide more

reliable prediction results. This achievement provides new pathways and methods for better and more systematic discovery and prevention of heart disease.

According to the research findings, heart disease onset correlates significantly with factors such as fasting blood sugar, cholesterol, exercise-induced angina, ST slope, gender, chest pain type, a certain comparative parameter, and maximum heart rate; whereas it correlates less with resting ecg and resting bps.

Through comparing the performance of various machine learning models, GBDT outperforms others in terms of AUC, accuracy, precision, recall, and F1 score. With an AUC close to 0.96 and an accuracy of approximately 0.91, GBDT dominates over other models.

Lastly, it is important to note that although the paper used a comprehensive dataset in this study, data collection may still not fully represent all scenarios, leading to potentially inaccurate analyses and inability to address specific outliers. To mitigate this issue, efforts can be made to improve the coverage of the dataset to better reflect real-world scenarios. Additionally, while this paper have made some progress in this study, I recognize that I have not fully exploited all the potential of the GBDT model. Therefore, one of the future research directions is to further explore and enhance the model's capabilities to improve its effectiveness and reliability in predicting and preventing heart disease.

Weng S, Reys J, Kai J, et al. 2017. Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? *PLoS ONE*, 12(4): 1-14.

REFERENCES

- Ali N S. 2002. Prediction of coronary heart disease preventive behaviors in women: a test of the health belief model. *Women & health*, 35(1): 83-96.
- Amin S U, Agarwal K, Beg R. 2013. Genetic neural network based data mining in prediction of heart disease using risk factors, *2013 IEEE conference on information & communication technologies*. 1227-1231.
- Avci E. 2009. A new intelligent diagnosis system for the heart valve diseases by using geneticSVM classifier. *Expert Systems with Applications*, 36(7): 10618-10626.
- Cheng Z., Zhang B., Cai Y., et al. 2023. Research on Heart Disease Prediction and Feature Analysis Based on Fusion of Random Forest and SHAP. *Intelligent Computer and Applications*, 13(11): 172-179.
- Li L..2017. Research on Heart Disease Detection Based on Deep Learning. *Modern Computer (Professional Edition)*, (9): 91-93, 110.
- Lin Z.. 2019. Research on Heart Disease Prediction Based on Decision Tree Algorithm. *Electronic Manufacturing*, 370(6): 25-27.
- Liu Y., Zhou Y., Luo C..2023. Research on Feature Selection in Heart Disease Prediction Based on GBM. *Modern Electronics Technology*, 46(19): 101-106.
- Mohan S. Thirumalai C, Srivasta G. 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7 : 81542-81554.
- Shafiey E. Hagag M G, et al. 2022. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. *Multimedia tools and applications*, 81(13): 18155-18179.