Prediction Model of Hypertension Complications Based on GBDT and LightGBM

Xinpeng Ji, Wenbing Chang, Yue Zhang, Houxiang Liu, Bang Chen, Yiyong Xiao and Shenghan Zhou*

School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

*Corresponding author email: zhoush@buaa.edu.cn

Abstract. Complications caused by hypertension include heart failure, stroke, arteriosclerosis, etc. The prediction of hypertension complications is a hot issue, and it is difficult to predict it from a medical perspective. In this study, we aim to establish a prediction model of hypertension complications based on machine learning and data mining. We first proposed a GBDT-based feature selection method, which can screen out medical indicators that affect the hypertension complications. On this basis, we established a hypertension complications prediction model based on LightGBM. The results show that after 10-fold cross-validation and comparison analysis, the accuracy, F1 and AUC of the prediction model are 0.9189, 0.8888, and 0.9233 respectively, which are significantly better than other machine learning models. Therefore, the proposed method can accurately predict hypertension complications, so as to provide effective clinical auxiliary diagnosis for doctors and help them take preventive measures to reduce the impact of hypertension complications.

Keywords: Hypertension complication; Prediction model; Medical artificial intelligence; Data mining; Machine learning; LightGBM.

1. Introduction

As the most common chronic disease, hypertension can bring serious complications to patients, such as cardiac complications (including myocardial infarction, heart failure, etc.), stroke, arteriosclerosis, hypertensive renal damage (including nephrosclerosis, renal failure, etc.), etc. Hypertension complications are potential safety hazard, and can cause death in severe cases [1]. Therefore, it is important to provide appropriate nursing interventions, implement individualized care, and reduce the impact of hypertension complications for hypertension patients. However, patients with hypertension do not have obvious clinical symptoms at the beginning of the onset, so it often leads to delays in the disease, and often serious harm to the patient once discovered. For example, spasm of arterioles throughout the body will gradually harden the arterioles as the disease progresses, and slowly damage the heart, brain, and kidney organs of the patient. Therefore, it is difficult to predict the complications of hypertension from a medical point of view.

The development of machine learning and artificial intelligence provides new ideas for the prediction of hypertension complications. On the one hand, data storage technology has made great progress, and a large amount of medical data of patients with hypertension has been accumulated. On the other hand, data mining tools are more mature, which provide technical support for the research of this article. At present, machine learning has been successfully applied in disease prediction [2], medical image recognition [3], medical diagnosis [4], and other fields [5]. In addition, many experts and scholars have tried to use data mining methods to analyze the hypertension complications. For example, G Du et al.
used Bayesian network and text processing to predict the incidence of hypertension complication and find out its influencing factors [6]. W Lee et al. applied classification algorithms, to predict the hypertension complication [7]. Junghye Lee et al. used data analysis methods to establish a process system for decision makers to prevent hypertension complications [8]. Y Liu et al. predicted hypertension outcomes based on RFECV and XGBoost [9].

This study analysed the medical data of hypertension patients from the hypertension centre of a tertiary-grade A class hospital in Beijing, and established a machine learning model to accurately predict the occurrence of hypertension complications. First, we used data pre-processing methods to clean the data. Second, the feature dimension of medical data is relatively high, but not all medical indicators have an impact on hypertension complications. Therefore, this study used feature selection based on tree model to provide high-quality feature combinations for prediction model. Third, we established a prediction model of hypertension complications based on LightGBM, obtained the best model by adjusting parameters and verified the advantages of the method.

2. Materials and Methods

2.1. Medical Data

The data comes from the hypertension centre of a tertiary-grade A class hospital in Beijing. The data involved 372 patients and 85 medical indicators. Table 1 shows some feature characteristics in dataset. There are missing values, outliers, etc. in the data. Therefore, data pre-processing is needed.

Table 1. Description of certain medical indicators.

|       | AGE   | HEIGHT | WEIGHT | BMI    | HR    | PULSE | RYSBPL | RYDBPL |
|-------|-------|--------|--------|--------|-------|-------|--------|--------|
| count | 372   | 361    | 361    | 348    | 372   | 372   | 372    | 372    |
| mean  | 38.3118 | 170.315 | 79.6160 | 27.3473 | 76.4059 | 76.2795 | 151.903 | 98.5322 |
| std   | 11.4189 | 7.57151 | 16.4125 | 4.18626 | 12.7134 | 12.6526 | 22.6656 | 16.7969 |
| min   | 15    | 154    | 32.4   | 16.3265 | 49    | 49    | 95     | 57     |
| max   | 76    | 191    | 165    | 50.9259 | 121   | 121   | 230    | 160    |

2.2. Data Preprocessing

2.2.1. Missing value processing techniques. There is a certain degree of missing values in some patients and some indicators. On the one hand, missing values can cause errors in the prediction model and cannot contribute to the prediction results. On the other hand, directly deleting samples or indicators with missing values can cause a waste of valid data. Therefore, this study adopts a combination of deletion and filling, and the rules are as follows.

We delete samples and features with more than 60% missing values. For the remaining samples with missing values, we use the k nearest neighbour algorithm (KNN) to select the average and mode of k similar samples to fill in the missing continuous features and category features, respectively.

2.2.2. Data normalization. We use maximum and minimum standardization to unify data dimensions.

\[ x^* = \frac{x - \text{min}}{\text{max} - \text{min}} \] (1)

2.3. Embedded feature selection based on GBDT

The most commonly used embedded feature selection include tree models and linear models penalized with the L1 norm. We designed embedded indicator selection based on Gradient Boosting Decision Tree (GBDT). It can be used to calculate the feature importance based on impurity. It automatically performs feature selection according to the contribution of the feature to the hypertension complications results during the learning process. Features are ranked according to the importance and
irrelevant features can be discarded. According to the above principles, the feature selection process based on GBDT is designed as follows:

1. Select base classifier. Gradient Boosting Decision Tree (GBDT) is the base classifier of this research. The meaning of GBDT is a decision tree model trained with Gradient Boosting strategy. The result of the model is a set of CART Tree Ensemble: $T_1, \ldots, T_n$. Among them, $T_i$ learns the residual of the prediction results of the previous $T_{i-1}$ trees. The final output of the model is the sum of the results of a sample in each tree. The formula is as follows:

$$\bar{y} = \sum_{n=1}^{N} f_n(x), f_n \subset \Gamma$$

(2)

where $f_n$ represents the mapping of samples to tree output.

2. Calculate the impurity-based feature importance. The impurity-based feature importance is also known as the Gini importance. The rapid feature importance obtained by adding up the Gini impurity of each individual feature on all trees in GBDT is a good indicator to measure the features contribution to the predicted results. Higher Gini importance means the feature is more important. The calculation formula of Gini importance is as follows:

$$\text{Gini}(D) = 1 - \sum_{k=1}^{N} p_k^2$$

(3)

where $p_k$ represents the proportion of the sample with class $k$ in the sample set.

3. Feature importance rank. According to the calculation result of Gini importance, all medical indicators are ranked by feature importance.

4. Set the screening threshold. The threshold is the median of feature importance. Medical features with importance higher than the threshold are retained, while the others are discarded.

2.4. Prediction model based on LightGBM

We take medical indicators as the model input, and whether hypertension complication occurs as model output. We use LightGBM to train both input and output data, to establish the mapping relationship between medical indicators and hypertension complications with tree structure. In this way, when the patient's medical characteristics are input, the model can predict whether the hypertension complications occur. Specifically, LightGBM uses boosting strategy to integrate multiple decision trees. Each decision tree uses the negative gradient of the loss function as the residual approximate value to fit the new decision tree. In this process, LightGBM uses Gradient-based One-Side Sampling (GOSS) to exclude most of the samples with small gradients, and only uses the remaining samples to calculate the information gain. It also uses Exclusive Feature Bundling (EFB) to reduce feature dimensions to improve calculation efficiency. Figure 1 shows the process of modelling. There are some parameters of LightGBM, such as max depth and number of leaves. We used grid search method to adjust the parameters to get the best prediction performance.

2.5. Model evaluation metrics and verification method

According to the expected effect of the prediction model, we build a confusion matrix for the classification results. It can clearly indicate confusion among multiple categories, as shown in Table 2.

| Actual Class | Predicted Class |
|--------------|----------------|
| Positive     | Positive       |
| Negative     | Negative       |

Table 2. Prediction results confusion matrix.
According to the confusion matrix, we establish the following model evaluation metrics:

1. **Accuracy.** The accuracy rate is defined as the percentage of the correct results in the total sample. The formula is as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]  

2. **Precision.** The precision rate is defined as the proportion of samples that actually have hypertension complications among all samples predicted to have hypertension complications. The formula is as follows:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

3. **Recall.** The recall rate is defined as the proportion of samples with hypertensive complications predicted to occur among the samples that actually have hypertension complications. The formula is as follows:

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

4. **F1-Score.** Precision and Recall sometimes contradict each other, that is, the higher the accuracy rate, the lower the recall rate. In some scenarios, both accuracy and recall rate must be considered comprehensively. The most common method is F-Measure. The formula is as follows:

\[
F = \frac{2PR}{P + R}
\]

where \(P\) refers to Precision, \(R\) refers to Recall.

5. **Area Under Curve (AUC).** AUC is the size of the area under the Receiver Operating Characteristic curve. If AUC reaches 0.80, it means that the classifier is very accurate.

Some scholars verify the model by dividing the training set and the test set, without considering the impact of the data itself on the prediction results. Aiming to avoid this interference, and make full use of the data, the verification approach of this research adopts 10-fold cross-validation, and the main process is shown in Figure 2. The data set is divided into ten, nine of which are selected as the training set, and the remaining one is the validation set. This process is repeated ten times in sequence, and the mean of the ten results is used as the evaluation result of the prediction model.

![Figure 1. LightGBM-based Hypertension Complication Prediction Modelling Process.](image-url)
3. Results and Discussion

3.1. Data after Preprocessing

Table 3 shows part of the data after pre-processing. After pre-processing, each medical index value is mapped to the [0,1] interval, and each missing value is filled.

| NO | PSSBP1   | LA      | IVSD    | HR      | PULSE   | RYSBPL   | RYDBPL   |
|----|----------|---------|---------|---------|---------|----------|----------|
| 1  | 0.307692 | 0.346153| 0.333333| 0.402778| 0.402778| 0.259259 | 0.223301 |
| 2  | 0.316550 | 0.153846| 0.416666| 0.347222| 0.347222| 0.666667 | 0.553398 |
| 3  | 0.615384 | 0.692307| 0.416666| 0.430556| 0.430556| 0.407407 | 0.320388 |

3.2. GBDT feature selection results

The feature importance ranking of medical indicators based on Gini impurity is shown in Figure 3. Figure 3 shows that the top three features that have a greater impact on hypertension complications include right brachium-ankle pulse wave conduction velocity (BAPWVR), urinary microalbumin/creatinine (MAUCR), right upper limb systolic blood pressure (RARMSBP). After GBDT embedded feature importance, we can obtain features related to hypertension complications. Table 4 lists the selected features, meanings and their corresponding importance rankings. We screened a total of 27 features related to hypertension complications.

![Figure 2. 10-fold cross-validation process.](image)

![Figure 3. Feature importance ranking.](image)
Table 4. Features related to hypertension complications based on GBDT.

| Rank | Feature         | Explanation                                                  | Importance |
|------|-----------------|--------------------------------------------------------------|------------|
| 1    | BAPWVR          | right brachium-ankle pulse wave conduction velocity           | 0.13497985 |
| 2    | MAUCR           | urinary microalbumin/creatinine                               | 0.06486303 |
| 3    | RARMSBP         | right upper limb systolic blood pressure                     | 0.05879093 |
| 4    | ET              | endothelin                                                   | 0.03196914 |
| 5    | BAPWVL          | left brachium-ankle pulse wave conduction velocity           | 0.02850238 |
| 6    | PSDBP1          | normal diastolic blood pressure                              | 0.02771875 |
| 7    | HIGHSBP         | the highest systolic blood pressure                          | 0.02750142 |
| 8    | ZGDBP           | highest diastolic blood pressure                             | 0.02685947 |
| 9    | BMI             | body mass index                                              | 0.02486061 |
| 10   | NEUT            | percentage of neutrophils                                    | 0.02294878 |
| 11   | T4              | tetraiodothyronine                                           | 0.02242963 |
| 12   | LVPWd           | thickness of the back wall                                    | 0.0221967  |
| 13   | ABIR            | right ankle-brachium index                                    | 0.02214757 |
| 14   | WEIGHT          | weight                                                       | 0.02213694 |
| 15   | LOWSBP          | the lowest systolic blood pressure                           | 0.02124697 |
| 16   | HIGHDBP         | the highest diastolic blood pressure                         | 0.02005893 |
| 17   | ESR             | erythrocyte sedimentation rate                               | 0.0196357  |
| 18   | FT3             | serum free triiodothyronine                                  | 0.01957292 |
| 19   | TSH             | thyroid stimulating hormone                                  | 0.01919862 |
| 20   | HCY             | homocysteine                                                 | 0.01894252 |
| 21   | WBC             | white blood cell                                             | 0.01773029 |
| 22   | LA              | left atrium                                                  | 0.01510325 |
| 23   | HSCRP           | high-sensitivity C-reactive protein                           | 0.01383033 |
| 24   | K               | serum potassium                                              | 0.01377462 |
| 25   | LARMSBP         | left upper limb systolic blood pressure                      | 0.01338724 |
| 26   | LLEGDBP         | left lower extremity diastolic blood pressure                | 0.01332843 |
| 27   | RYDBPL          | left arm diastolic pressure                                  | 0.01203785 |

3.3. Hypertension Complication Prediction Model Based on LightGBM

3.3.1. Best parameters after grid search. The optimal parameters of LightGBM are shown in Table 5. Figure 4 shows that LightGBM has obtained better prediction performance after parameter adjustment, and can predict hypertension complications more accurately.

Table 5. LightGBM parameters after grid search.

| Parameter            | Value        | Parameter           | Value        | Parameter         | Value    |
|----------------------|--------------|---------------------|--------------|-------------------|----------|
| num_iteration        | 88           | min_child_samples   | 21           | lambda_l2         | 0.000001 |
| max_depth            | 3            | subsample           | 0.6          | min_split_gain    | 0        |
| num_leaves           | 4            | subsample_freq      | 0            | learning_rate     | 0.0752   |
| max_bin              | 135          | colsample_bytree    | 1            |                   |          |
| min_child_weight     | 0.001        | lambda_l1           | 0.000001     |                   |          |
3.3.2. Analysis of the importance of medical features. On the basis of the best parameters, we import all the data into the model for training. According to the information gain of features on all decision trees of LightGBM, we can obtain the degree of influence of these medical indicators on the prediction results of hypertension complications, as shown in Figure 5.

3.3.3. Prediction model performance evaluation results. We used the medical data of hypertension patients to continuously train the prediction model, and improve the prediction performance of the model through grid search parameter adjustment, thereby successfully establishing a hypertension complication prediction model. In addition, in order to verify the advantages of the method, we compared the results of the prediction model with the analysis results of the following two commonly used machine learning algorithms: cart decision tree (Cart DT), and XGBoost. Through 10-fold cross-validation, the performance evaluation results of the prediction model are shown in Table 6 and Figure 6. It can be seen from Table 6 that the AUC value of the prediction model has exceeded 0.8, indicating that the classification result of the prediction model is very accurate. The evaluation results show that the various evaluation indicators of the method proposed in this paper are very high, which are better than the other prediction methods, as shown in Figure 6.

Table 6. Comparison of the machine learning model results.

| Criteria | Model  | Accuracy | F1     | AUC   |
|----------|--------|----------|--------|-------|
|          | Cart DT | 0.7179   | 0.4205 | 0.6296|
|          | XGBoost | 0.8201   | 0.7142 | 0.8570|
|          | LightGBM| 0.9189   | 0.8888 | 0.9233|

Figure 4. Comparison before and after grid search tuning.

Figure 5. LightGBM feature importance rank.

Figure 6. Prediction performances comparisons.
4. Conclusion
The prediction of hypertension complications is a hot problem, and it is difficult to analyse it from a medical perspective. From a data-driven perspective, we established a hypertension complications prediction model based on GBDT and LightGBM in this study. We first preprocessed the data to eliminate data impurities. After that, this paper established an embedded feature selection method based on Gradient Boosting Decision Tree. This method can screen out the medical features that have an impact on the hypertension complications, and judge the importance of each medical feature, thereby reducing the data dimension. On this basis, this study established a LightGBM-based hypertension complication prediction model, and improved the prediction performance through grid search tuning. After 10-fold cross-validation, the accuracy, F1 value and AUC value of the prediction model are 0.9189, 0.8888, and 0.9233 respectively. The verification results show that the method can accurately predict the hypertension complications, thereby providing effective assistance for doctors in clinical diagnosis. In addition, we compared the prediction performance of the proposed method with other two common machine learning models, which confirmed the superiority of the proposed method. The results of this study have the following significance. First, this study established a hypertension complication prediction model, which can accurately predict the patient’s hypertension complication. Second, this study can efficiently predict patient complications from a data perspective. When the patient's medical indicators are input, the model can output whether the hypertension complications will occur. Third, this study can analyse the factors that affect the complications of hypertension, so that doctors can conduct in-depth analysis from the perspective of pathology. Fourth, since the complications of hypertension can be predicted in advance, doctors can take preventive measures in advance, which can help improve the survival rate of patients and reduce the impact of hypertension complications.

Acknowledgments
This study is supported by the National Natural Science Foundation of China (Grant No.71971013 & 71871003). The study is also sponsored by the Fundamental Research Funds for the Central Universities (YWF-20-BJ-J-943) and the Graduate Student Education & Development Foundation of Beihang University.

References
[1] "Medical care expenditures for hypertension, its complications, and its comorbidities” Hodgson, T. A., & Cai, L. Medical care, 2001: 599-615.
[2] "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique” Venkatesh R, Balasubramanian C, Kaliappan M. Journal of medical systems, 2019, 43(8): 272.
[3] “Deep learning for medical image processing: Overview, challenges and the future” Razzak M I, Naz S, Zaib A. Classification in BioApps. Springer, Cham, 2018: 323-350.
[4] “Machine learning for medical diagnosis: history, state of the art and perspective” Kononenko I. Artificial Intelligence in medicine, 2001, 23(1): 89-109.
[5] “Steel Surface Defect Classification Using Deep Residual Neural Network” Konovalenko I, Maruschak P, Brezinová J, et al. Metals, 2020, 10(6): 846.
[6] “Risk prediction of hypertension complications based on the intelligent algorithm optimized Bayesian network” Du G, Liang X, Ouyang X, et al. Journal of Combinatorial Optimization, 2019: 1-22.
[7] “Prediction of hypertension complications risk using classification techniques” Lee W, Lee J, Lee H, et al. Industrial Engineering & Management Systems, 2014, 13(4): 449-453.
[8] “A data-driven procedure of providing a health promotion program for hypertension prevention” Lee J, Kwon R H, Kim H W, et al. Service Science, 2018, 10(3): 289-301.
[9] “A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data” Chang W, Liu Y, Xiao Y, et al. Diagnostics, 2019, 9(4): 178.