Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/67/20240628

The prediction of Parkinson's disease based on Pearson coefficient feature screening and machine learning

Yuchen Tian

Tianjin No.1 High School, Tianjin, 300000, China

17627652898@163.com

Abstract. Parkinson's disease is a common and severe type of brain disease. Its incidence rate is relatively high among brain diseases. At present, there is no very effective treatment for Parkinson's disease. So researchers have focused on diagnosing Parkinson's disease. At present, machine learning methods have been applied in the medical field and have played a very positive role in the diagnosis of diseases. It has been proven that analyzing the patient's voice and trembling condition through machine learning can accurately diagnose Parkinson's disease. In this study, before training the model, we used the Pearson coefficient feature screening method to improve the accuracy of diagnosis. Then, we conducted training on six major models (Random Forest, GBDT, Adaboost, Logistic Regression, Decision Tree, XGboost) in order to find the model with the best performance. In this study, we found that the performance of Random Forest is the best in these models (Accuracy: 91.53%, recall: 100%), then is the GBDT model (Accuracy: 91.53%, recall: 97.78%). The other four models all have a great disparity on accuracy and recall, which are the two most important metrics on the detection of diseases. The research results have demonstrated that the feature selection method based on Pearson's coefficient indeed comprehensively improves the accuracy of diagnosis for Parkinson's disease. And we also found that in the process of diagnosing Parkinson's disease, the performance of the Random Forest and GBDT models is the best.

Keywords: Machine Learning, Parkinson Detection, Pearson coefficient.

1. Introduction

Parkinson's disease is a progressive disease. Its incidence rate is second only to Alzheimer's disease in brain diseases. The incidence rate of the age group over 60 years old reaches 1% [1]. The most obvious symptom of Parkinson's disease is hand tremors, which can also lead to stiffness or delayed movement. This tremor can also cause further symptoms such as writing disorders, language disorders, and delayed reactions [2]. In the later stages, it may even lead to some mental illnesses, such as depression, highly variable emotions, urinary system problems, constipation, skin problems, and more serious symptoms. Faced with such a serious disease, there is currently no effective treatment method in the medical field, and only medication can alleviate symptoms. Therefore, it is very necessary to find an efficient detection method. Machine learning is currently a very popular technology because it can benefit almost all industries, including health monitoring and healthcare. The essence of machine learning is actually artificial intelligence. It uses algorithms trained on datasets to create models that enable machines to perform tasks that only humans can perform [3], and achieve faster speed, higher

DOI: 10.54254/2755-2721/67/20240628

quality, and higher efficiency. The ability to learn and analyze disease risks from various data and independently complete diagnostic work has made machine learning very powerful and versatile in the medical field.

Nowadays, machine learning has been applied in the medical field, mainly to examine health-related data and achieve the effect of detecting diseases. Machine learning algorithms can assist in disease diagnosis by analyzing data and predicting the root cause of diseases using pathogenic variables in electronic health records. Compared with traditional biostatistics methods that analyze and integrate large amounts of complex healthcare data, machine learning is becoming increasingly popular in classification, prediction, and clustering tasks [4]. This indicates that many unpredictable diseases, such as cancer, Parkinson's disease, etc., can be effectively prevented and detected.

In the research of machine learning for predicting Parkinson's disease, some researchers have proven it to be feasible. Dina Katabi et al. used wireless sensors to collect signals reflected by PD patients and established predictive models through deep learning to evaluate the severity of PD, disease progression, and drug treatment efficacy [5]. However, in the current literature, there are few studies that involve using machine deep learning algorithms to improve the accuracy of health data detection. In this article, we took a Parkinson's dataset as the research object and used the Pearson coefficient feature selection method to improve the accuracy of the algorithm model. We also conducted comparative experiments on multiple models, hoping to find the best performing machine deep learning model to further improve the accuracy and efficiency of machine learning algorithms for Parkinson's disease detection, and help more Parkinson's disease patients receive treatment as soon as possible and alleviate symptoms, even rehabilitation.

2. Methods

2.1. Dataset

2.1.1. Introduction to the Parkinson's dataset

In this study, the dataset is a comparison of various indicators between Parkinson's patients and the general healthy population, as well as whether they are in a diseased state (diseased: 1, healthy: 0). We mainly collected and compared indicator information on the frequency of patient vocalization, the frequency of vibration, and corresponding amplitude changes, as well as the ratio of noise to tone components. This dataset comes from Kaggle [6], and the data and sources on the website are all legitimate and reliable.

This set of data consists of 195 different data, 24 columns, and column names can be divided into the following categories: name ASCII topic name and record number. The health status of the patient (1 represents Parkinson's disease, 0 represents healthy population). Two measurement methods for measuring the ratio of noise to pitch components (NHR, HNR). Three nonlinear measurement methods (spread 1, spread 2, PPE). The average fundamental frequency, maximum fundamental frequency, and minimum fundamental frequency subsets of the human voice starting with the MDVP label, as well as their measurement methods (also starting with the MDVP label). And two nonlinear dynamic complexity measures (RPDE, D2). This dataset also provides 23 types of feature data for us to perform feature filtering.

2.1.2. Dataset processing

After importing the CSV format dataset, the data is relatively clean, with no missing values (The non-null count of the 23 columns are all “non-null”), and there is no need to clean the data.

The following figure is the data distribution maps of four sets of data with strong correlation selected through feature filtering. The sketches of these four graphs are all close to the center, with a small trend of deviation from left to right, indicating that the distribution of these four sets of data is close to or even reaches a normal distribution, and there are no obvious outliers, which will not have a significant impact on the final experimental results.

DOI: 10.54254/2755-2721/67/20240628

After analyzing the data, we imported various models by establishing a model function and made predictions, observing and comparing the results (Figure 1).

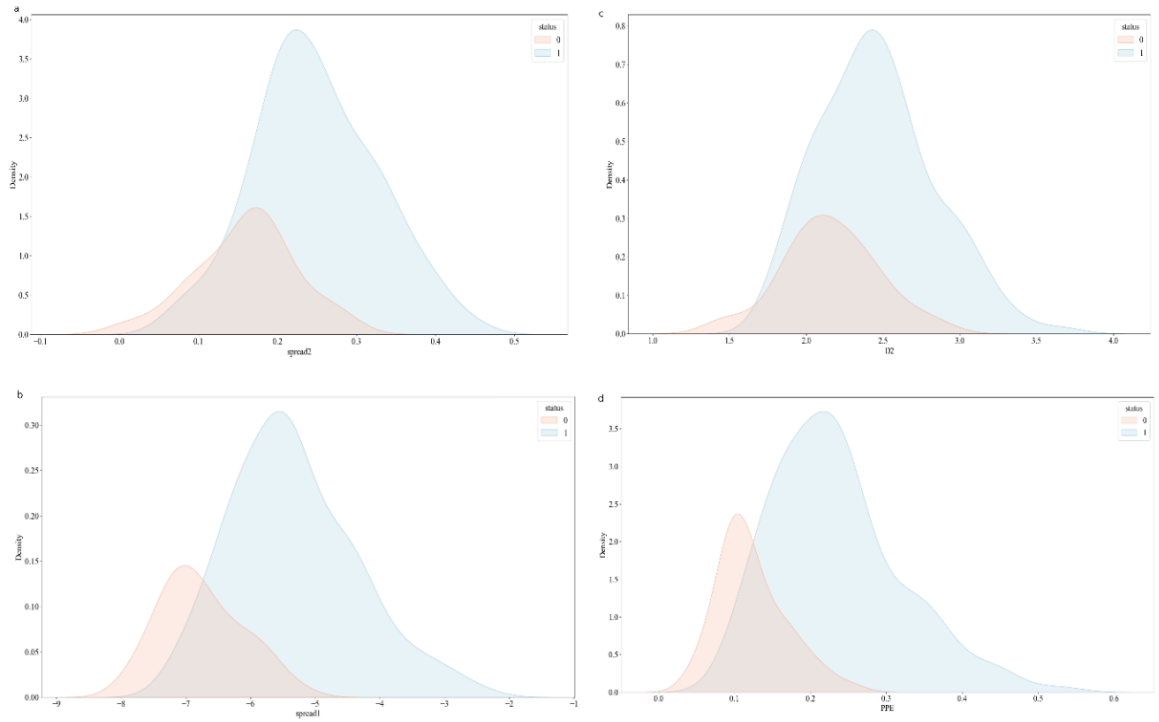

Figure 1. Data Distribution Maps. a, b, c, and d are the data distribution maps for Spread 2, Spread 1, D2, and PPE, respectively.

2.2. Feature extraction

Encapsulation, embedding, and filtering are the three main types of feature selection methods. In order to better compare and find the data that truly has the greatest correlation, we adopted various feature filtering methods and covered the types of feature filtering in these three categories. Finally, we drew a heatmap of this dataset. In the process of feature selection, the feature selection method based on Pearson coefficient has the best effect [7].

Pearson coefficient is currently not a commonly used feature selection method. It is often used to draw heat maps. In the heat map plotted based on Pearson's coefficient, the subset names of the dataset are arranged on the X-axis and Y-axis, respectively. According to the correspondence between the X-axis and Y-axis, write the correlation between data subsets in the corresponding area in the form of -1 to +1 intervals, and represent it by color depth. The "+" and "-" symbols represent the direction of change of the dependent variable and the independent variable, respectively ("+" represents the increase of the dependent variable with the increase of the independent variable, "-" represents the decrease of the dependent variable with the increase of the independent variable). The Pearson coefficient is basically consistent with the prediction method of linear models, especially good at predicting linear relationships between subsets of data [8]. Because all three data showing strong correlation are nonlinear measurement methods, we speculated that this is a set of nonlinear correlated data (Figure 2).

DOI: 10.54254/2755-2721/67/20240628

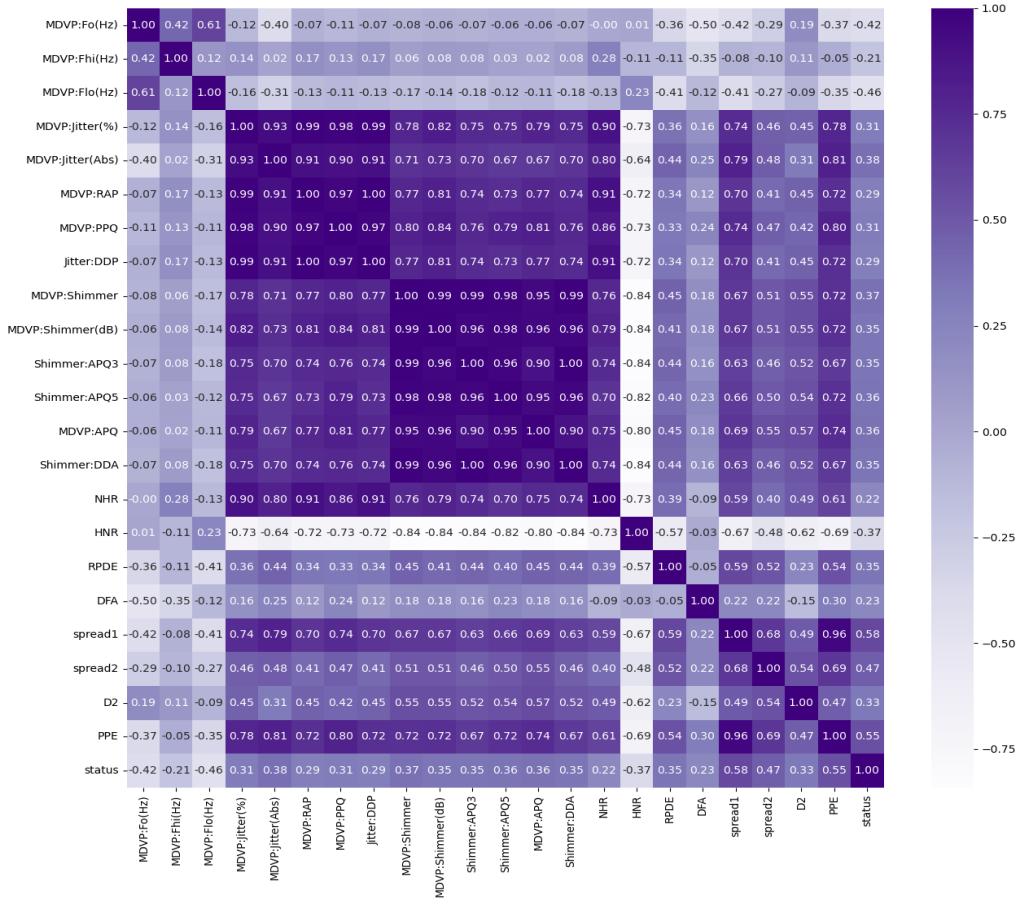

Figure 2. Heatmap of features.

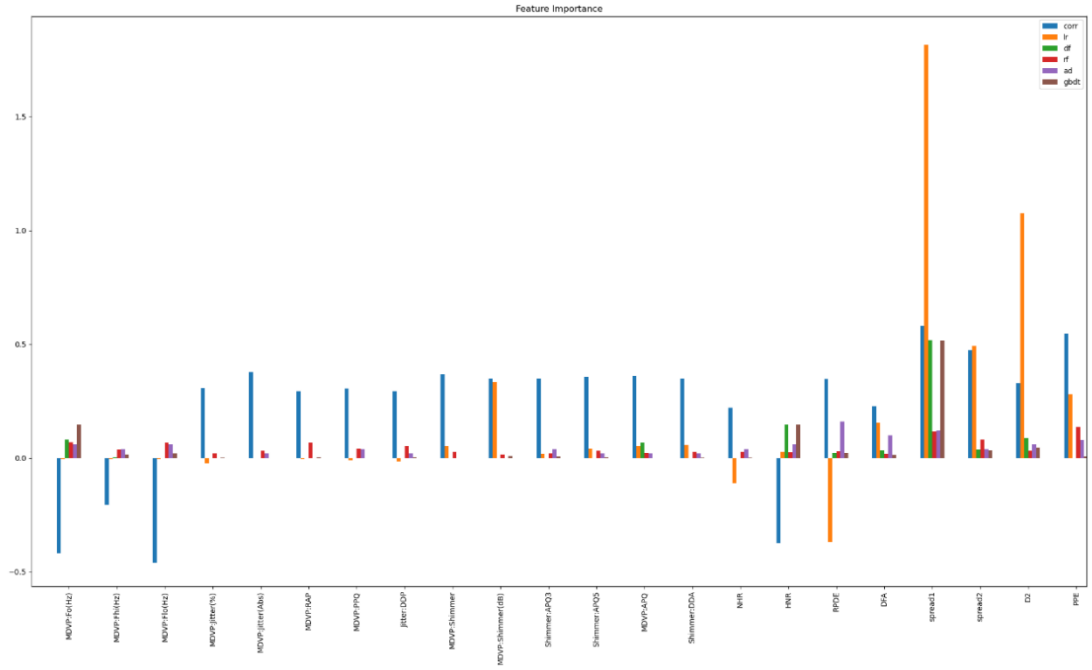

Figure 3. Results of other feature selection methods.

DOI: 10.54254/2755-2721/67/20240628

The above are the results of feature selection (Figure 3). Among them, corr represents the Pearson coefficient method for feature selection, LR represents logistic regression, DF represents decision tree, RF represents random forest, AD represents Adaboost, and GBDT represents Gradientboosting. Firstly, through the heatmap based on the Pearson coefficient selection method, we can see that the correlation between "spread 1", "spread 2", "PPE", "D2" and "status" is relatively high ("spread 1" 0.58, "spread 2" 0.47, "PPE" 0.55). And another result graph of six feature filters clearly shows that "spread 1" has the strongest correlation with the target variable "status". Then, the variables "D2", "PPE", and "spread 2" also have a certain correlation with "status", which is consistent with the results of the heatmap. Furthermore, based on the description of the dataset on the website, “spread 1," "spread 2," and "PPE" are both nonlinear measurement methods for fundamental frequency variation. "D2" is also a measure of nonlinear dynamic complexity. Therefore, based on this, we speculated that the dataset will have a non-linear relationship with the diagnosis of the health status of Parkinson's patients, thereby predicting better experimental performance compared to logistic regression models, decision trees, Adaboost, and Gradientboosting.

2.3. Models

2.3.1. Introduction to Linear Models

The logistic regression model can be said to be almost a model specifically designed for binary classification problems. Its characteristic is that there are only two values of 1 and 0 for the test set variables. The logistic regression model is also a type of classical linear model, and its image is basically linear or S-shaped curve. The training principle of logistic regression model is to approximate the S-shaped curve into a straight line and change the values of 1 and 0 in the test set to negative infinity and positive infinity. The calculation formula [9] of this model is as follows:

$$logit(p) = a + b_1x_1 + b_2x_2 + \cdots + b_ix_i \quad (1)$$

where p is the probability of death and x1, x2 ... xi are the explanatory variables. Logit (or Logic) is the function used to convert images into straight lines in model training.

The most basic linear model image is a straight line, but the logistic regression model image is approximated by an S-shaped curve as a straight line, so the distribution of the logistic regression model is actually not linear, but binomial. This is also the essential difference between logistic regression models and simple linear models, which makes logistic regression models perform better in binary classification problems.

2.3.2. Introduction to Nonlinear Models

2.3.2.1. Decision Tree

The decision tree model is actually a very common data mining method used to establish classification systems based on multiple covariates or develop prediction algorithms for target variables. This method can classify the Jiangzhong group into multiple nodes, namely branches, and construct a set of tree like structure models. This algorithm is non-linear, and its advantage lies in its ability to effectively handle complex datasets with large samples. When the sample size is large enough, the research data can be divided into training dataset and validation dataset. Use a training dataset to establish a decision tree model, and use a validation dataset to determine the appropriate tree size required to achieve the optimal final model. Figure 4 shows the working principle of a decision tree [10].

DOI: 10.54254/2755-2721/67/20240628

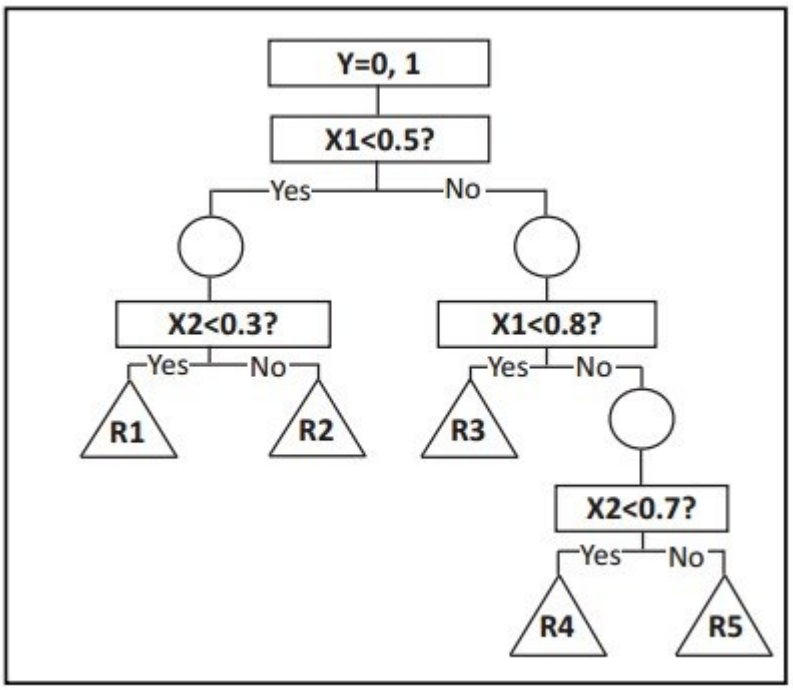

Figure 4. Working pinciple of Decision Tree

2.3.2.2. Adaboost

Adaboost is a model that can shorten the process of achieving adaptive growth, and its algorithm has many advantages compared to decision trees. Adaptive growth is a pattern in machine learning, and its advantage lies in dealing with binary classification problems, making it more suitable for this study. The main idea behind AdaBoost is to retrain the weak classifiers in the training data group, with each continuous classifier assigning more weight to the wrong data points. The final AdaBoost model is defined by combining all the weak classifiers used for training with the specified weights of the model based on their accuracy. The weaker model with the highest accuracy is determined by the higher weight, while the weaker model with the lowest accuracy is determined by the lower weight [11]. Because the problem studied in this study is a binary classification problem, we have only explained the formula for solving classification problems in the Adaboost algorithm. The calculation formula [12] of this model is as follows:

First, we need to check the data.

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{-1,1\} \tag{2}$$

$$W_0(i) = \frac{1}{m} \; for \; i = 1, \ldots, m \tag{3}$$

Then, we need to train a weak learner and process the data.

$$h_t: \mathcal{X} \to \{-1,1\} \; w.r.t. \; the \; distribution \; W_{t-1} \tag{4}$$

$$\epsilon_t = \sum_{i=1}^{m} W_{t-1}(i) 1_{h_t(x_i) \neq y_i} \tag{5}$$

$$\alpha_t = \frac{1}{2} log \frac{1-\epsilon_t}{\epsilon_t} \tag{6}$$

DOI: 10.54254/2755-2721/67/20240628

$$W_t(i) = \frac{W_{t-1}(i)\, exp\big(-\alpha_t y_i h_t(x_i)\big)}{Z_t} \tag{7}$$

$$Z_t = \sum_{i=1}^{m} W_{t-1}(i)\, exp\big(-\alpha_t y_i h_t(x_i)\big) \tag{8}$$

Finally, we need to return H.

$$x \in \mathcal{X} \rightarrow sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \in y \tag{9}$$

2.3.2.3. Gradient Boosting

The GBDT model has three important elements: the loss function that needs to be optimized, a weak learner with prediction abilities, and an additive model that adds weak learners to minimize the loss function. It is precisely because of these three essential elements that the GBDT model can have its own advantages in predicting binary classification problems while handling data that presents non-linear relationships [13]. Its calculation formula is as follows:

The weak learner is as follows:

$$T\big(\vec{x}; \vec{\theta}\big) = \sum_{j=1}^{J} \gamma_j\, I\left(\vec{x} \epsilon\, R_j\right) \tag{10}$$

where, J is the number of leaves, which are defined by the disjoint regions Rj numbered by j, and γj are the values in each region. $\vec{\theta}$ denotes a set of parameters of the decision tree.

2.3.2.4. Random Forest

The main advantage of the random forest model is its unique working principle. In the process of training a model, unlike other models with a single decision tree, it can construct multiple decision trees based on the correlation of features in the dataset, and mainly focus on the decision tree of the subset with the highest correlation between the dataset and the test set. This property makes the algorithm of a decision tree no longer rigid, but changes the tree according to the needs of the test set. This working principle enables the random forest model to apply the rules of the training dataset more reasonably, with a higher probability of normal fitting and better performance on test data. Especially, in binary classification problems, the voting method is an innovation of the random forest algorithm in tree models. Aggregating the results of all trees in this way makes the final training result more accurate and stable. This principle also makes the random forest model one of the best models to improve fitting [14].

3. Model training and evaluation metrics

3.1. Evaluation metrics

In this study, the problem we are dealing with is a binary classification problem, which involves diagnosing whether a patient has Parkinson's disease. Based on this question, we have a total of four evaluation indicators. Firstly, accuracy is a very important measure, usually expressed as a percentage of the total number of correctly predicted samples. In addition, in clinical medical diagnosis, missed detections often come at a greater cost than misdiagnosis, so we need to observe the recall rate of each model (Recall, usually the percentage of samples predicted by the model to be correct out of actual correct samples, corresponding to missed detections in real life. In summary, we can see that recall is a priority to consider. In addition, another evaluation metric is Area under the Curve (AUC), which is commonly used as a model performance evaluation metric to measure the model's generalization

DOI: 10.54254/2755-2721/67/20240628

ability. It is obtained by calculating the area under the Receiver Operating Characteristic Curve (ROC) curve [15].

3.2. Training coefficient adjustment

In the process of model training, we first divided the training set and the test set (ratio: 7:3). We specifically divided the "status" into separate test sets, as the main purpose of this study is to diagnose whether patients with Parkinson's disease have Parkinson's disease based on their characteristics. We set the sample proportion of the test set to a floating-point parameter of 0.3 using the test_size parameter, and set the random_state parameter to 50. After importing the model, in order to obtain more accurate prediction results, we set the training results to display four decimal places and present them in the form of a table.

4. Results

As the training progresses, Loss gradually decreases and stabilizes after approximately 30 epochs; Accuracy, however, dramatically increases and stabilizes after approximately 35 epochs (Figure 5). Figure 5 is two line charts of Loss and Accuracy in this research. Figure a is the change of Loss; Figure b is the change of Accuracy.

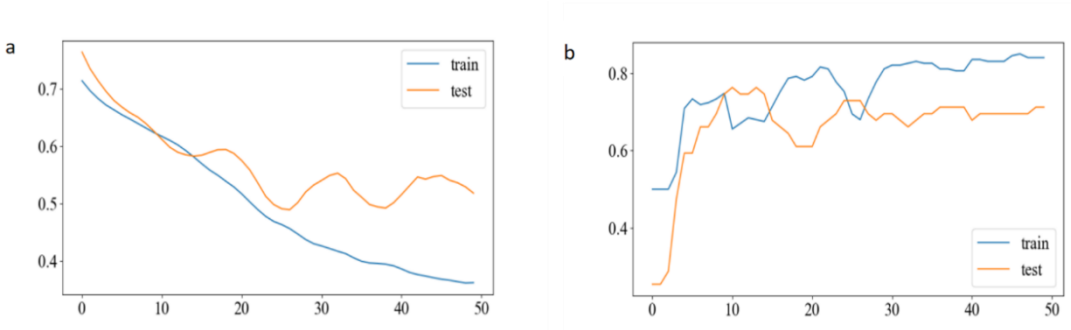

Figure 5. Line chart of Loss and Accuracy. (a) The loss curves. (b) The accuracy curves.

The Table 1 shows a comparison of the predicted results. We compared them with the experimental results based on the above conditions. In the prediction results of the logistic regression model, the accuracy was 81.36%, which is relatively accurate. However, in the most important recall rate, especially the probability of misdiagnosing patients with Parkinson's disease as healthy individuals is relatively high, at around 5%. But the performance of this model is better than that of decision trees. Although the accuracy of the decision tree is 0.1% higher than that of the logistic regression model, the recall rate is 0.5% lower. The random forest model has excellent performance in precision, with an accuracy of 94.74%. Consider putting it into use. Although the performance of the Adaboost model is also relatively good, each item is inferior to the random forest model. Although the Gradientboosting model may not perform well in missed detections, an error of around 0.2% is still within acceptable range. And compared to the random forest model, this model has a full 1% higher recall rate in identifying healthy individuals. The accuracy of its prediction is also very high, reaching 94.62%, so the performance of the model is also very good, and it is recommended to put it into use. The last model is XGboost, which performs too poorly in terms of recall and is insufficient for practical use.

Table 1. Results of Six Models

	LR	DT	AD	GBDT	RF	XG
Recall 1	0.96	0.91	1.00	0.96	0.98	1.00
Recall 0	0.36	0.57	0.64	0.64	0.71	0.07
Precision 1	0.83	0.87	0.90	0.90	0.92	0.76
Precision 0	0.71	0.67	1.00	0.82	0.91	1.00

DOI: 10.54254/2755-2721/67/20240628

Table 1. (continued).

F1-score 1	0.89	0.90	0.95	0.93	0.95	0.86
F1-score 0	0.48	0.62	0.78	0.72	0.8	0.13
Accuracy	0.81	0.83	0.92	0.88	0.92	0.76

5. Discussion

First, as the best performing model, the random forest model has the advantage of autonomous sampling, which can increase the model's generalization ability. In terms of outliers, the random forest model not only has strong robustness, but also can select subsets based on feature correlation to create decision trees, and autonomously select the optimal features for partitioning each decision tree node. This indicates that it is the least sensitive to outliers in weakly correlated data in the dataset and is the most suitable model for this experiment. It also means that random forest model may be the most suitable model for feature selection. Meanwhile, the high accuracy, interpretability, and feature importance of the random forest model make it the model with optimal performance.

The GBDT model, as a model that is also sensitive to outliers, performs better than the Adaboost model. We believe that the main reason for this is the three most important elements in the principles of the GBDT model, which are also its three advantages over other nonlinear models. Firstly, the loss function it uses depends on the type of problem being solved, so it is not limited by the type of problem being studied like decision tree models, which gives it a certain advantage in predicting binary classification problems. Secondly, the decision tree length of the GBDT model is relatively long, with more branches compared to Adaboost, making its predictions more accurate. Finally, the GBDT model adopts an additive model, which adds one tree at a time without changing the existing trees in the model. When adding trees, use gradient descent programs to minimize losses to the greatest extent possible. At the same time, the GBDT model has strong robustness and is less affected by noise and irrelevant features, indicating that the GBDT model focuses more on four sets of data with strong correlation, which are the data with fewer outliers.

The logistic regression model, as a model suitable for predicting linear relationships, demonstrated moderate performance in this set of experiments. We believe that the reason may be that the logistic regression model is more suitable for binary classification problems compared to other models. However, due to the non-linear relationship of the data itself, the adaptability of the logistic regression model itself is not very good, resulting in only moderate performance. The reason why the performance of the decision tree model is not very good is the limitation brought by 195 samples. The characteristic of the decision tree model is that it requires a large number of training samples, and the number of samples in this dataset is relatively small (195), which cannot support the decision tree to exert its excellent performance. But its prediction accuracy is higher than that of logistic regression models. The Adaboost model adopts an iterative algorithm, with new weak classifiers added in each round. The reason for its poor performance is the occurrence of outliers. Although there were no significant outliers observed in the four sets of strongly correlated data, there were indeed some outliers in other weakly correlated data that may affect the final output of the results. These outliers affect the final output of the strong classifier of the Adaboost model.

The conclusion is that, based on the comparison of model performance, we can conclude that Gradientboosting and Random Forest models have the best performance. Through preliminary research, it can be concluded that these two models are sufficient for practical diagnostic work. So, through this experimental result, we can also verify the hypothesis we proposed during feature selection: the relationship between the data in this dataset is non-linear. By visualizing the Loss and Accuracy in the prediction results as images, the non-linear relationship between the two can also be observed.

The limitation of this study is that, apart from commonly used models, it did not use well-known neural network techniques such as CNN for prediction, and the predicted results may be biased. Meanwhile, for these models, besides using feature selection methods to optimize them, no other

DOI: 10.54254/2755-2721/67/20240628

optimization methods were used, so the results of this study may not be the best predictive results. Finally, the dataset used in this study was relatively small, with only 195 data samples, which directly led to a decrease in the predictive performance of some models. However, the population of Parkinson's disease is relatively large, and perhaps under such a large population, the performance of some models will be improved, even surpassing the better performing models in this study.

In future research, the training model can be further optimized based on this study, and the prediction results of neural networks can be added to make the results more complete. You can also search for datasets with a larger number of features and more data samples to improve the prediction results. Meanwhile, we hope that in the future, the results of this study can lay a theoretical foundation for the prediction of Parkinson's disease and make preliminary attempts, contributing to more accurate prediction of Parkinson's disease in the medical and health field.

6. Conclusion

The research conclusion of this study was that under the Pearson coefficient based feature selection method, the GBDT model and random forest model were more suitable for predicting whether the subject population has Parkinson's disease. This study found an effective diagnostic method for Parkinson's disease and helped healthcare professionals take timely action to stabilize the condition before the disease worsens, which can help prevent and treat Parkinson's disease in the medical field.

References

[1] Kurmi A, Biswas S, Sen S, Sinitca A, Kaplun D and Sarkar R 2022 An Ensemble of CNN Models for Parkinson's Disease Detection Using DaTscan Images Diagnostics 12 1173
[2] Hayes M T 2019 Parkinson's Disease and Parkinsonism Am. J. Med. 132 802-7
[3] Raffaele P, Stefano R and Riccardo Mi 2021 Machine learning-based approach: global trends, research directions, and regulatory standpoints Data Sci. Manag. 4 19-29
[4] An Q, Rahman S, Zhou J and Kang JJ 2023 A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges Sensors 23 4178
[5] Liu Y, Zhang G, Tarolli CG, et al. 2022 Monitoring gait at home with radio waves in Parkinson's disease: A marker of severity, progression, and medication response Sci. Transl. Med. 14 eadc9669
[6] Little MA, McSharry PE, Roberts SJ, Costello DAE and Moroz IM 2007 Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection BioMedical. Eng. OnLine 6
[7] Pudjihartono N, Fadason T, Kempa-Liehr AW and O'Sullivan JM 2022 A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction Front. Bioinform. 2 927312
[8] Schober P, Boer C and Schwarte LA 2018 Correlation Coefficients: Appropriate Use and Interpretation Anesth. Analg. 126 1763-8
[9] Bewick V, Cheek L and Ball J 2005 Statistics review 14: Logistic regression Crit. Care. 9 112-8
[10] Song YY and Lu Y 2015 Decision tree methods: applications for classification and prediction Shanghai Arch. Psy. 27 130-5
[11] Hatwell J, Gaber MM and Atif Azad RM 2020 Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences BMC Med Inform Decis Mak. 20 250
[12] Perceval Beja-Battais, Centre Borelli, ENS Paris-Saclay and Université Paris-Saclay 2023 Overview of AdaBoost: Reconciling its views to better understand its dynamics, arXiv 2310.18323
[13] Seto H, Oyama A, Kitora S, et al. 2022 Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data Sci. Rep. 12 22599
[14] Mariana Belgiu and Lucian Drăguţ 2016 Random forest in remote sensing: A review of applications and future directions ISPRS J. Photo Rem. Sen. 114 24-31

DOI: 10.54254/2755-2721/67/20240628

[15] Rainio O, Teuho J and Klén R 2024 Evaluation metrics and statistical tests for machine learning Sci. Rep. 14 6086