Early Diagnosis of Parkinson's Disease based on XGBoost Algorithm and Logistic Regression

Jianbang He*, Jiyue Zhang

School of Mathematics and Statistics, Hunan Normal University, Changsha, 410000, China

* Corresponding Author Email: 2587121850hejianbang@hunnu.edu.cn

Abstract. Parkinson's disease is a common neurodegenerative disease that affects nearly millions of people in the world. Parkinson's disease is often misdiagnosed or missed because its early symptoms are similar to aging. Starting from the early dysarthria of Parkinson's disease, this paper selects 22 of 195 samples for XGBoost model. Meanwhile, in order to avoid multicollinearity, Pearson correlation coefficient is used to reduce dimension to obtain 12 sound indicators for logistics model prediction. In this paper, the two models were comprehensively compared by means of confusion matrix thermal map, AIC value and sensitivity, and it was found that the former was 95% accurate but the sensitivity was high and the model stability was poor, while the latter was 90% accurate and the model goodness of fit was good. Therefore, the method discussed in this paper provides a new idea for the early identification of Parkinson's disease and points out the direction for further research.

Keywords: Parkinson, disarticulation, pearson correlation, logistics regression, XGBoost algorithm.

1. Introduction

Parkinson's disease is a chronic neurodegenerative disease whose main clinical manifestations are limb tremor, muscle stiffness, bradykinesia, and postural instability [1]. Middle-aged and elderly people are susceptible to the disease, and younger people may also develop it. Currently, there are about 3.65 million Parkinson's disease patients in China, and it is expected to reach 5 million in 2030. As the average life expectancy continues to extend and the degree of aging deepens, neurological degenerative diseases that are closely related to aging are increasing rapidly, especially the number of patients with Parkinson's disease and Alzheimer's disease is increasing rapidly, and their related management costs and care Costs are increasing rapidly, and the related public health pressure is of particularly importance [2]. Therefore, in order to delay the development of the disease and reduce economic and social pressure, early diagnosis of Parkinson's disease is particularly important.

The pathological feature of Parkinson's disease is the degeneration and loss of dopaminergic neurons in the substantia nigra, resulting in a decrease in dopamine levels in the brain. This process is related to factors such as local neuroinflammation, mitochondrial dysfunction, oxidative stress, and abnormal protein aggregation. At present the diagnosis of Parkinson disease mainly relies on observation of clinical symptoms. However, the clinical symptoms of Parkinson's disease are complex and, mainly including motor symptoms and non-motor symptoms. In the early stages of disease development, these symptoms are often similar to symptoms of physical aging, which are easily missed and misdiagnosed [3]. When the patient enters the terminal of the disease, the clinical symptoms have become obvious, and more than half of the dopamine neurons in the substantia nigra have died, missing the optimal treatment time.

In fact, Parkinson's disease symptoms can be observed in the early stages through some physical signals, such as voice [4], Bladder changes [5], brain imaging [6], etc. Among them, speech disorders have received widespread attention because they are easy to collect, easy to observe, not easy to lose and have high user acceptance. In the 1970s, Darley et al. noted the impact of Parkinson's disease on patients' pronunciation and conducted a systematic study of the relationship between Parkinson's disease and speech [7]. Subsequent studies have further demonstrated that patients with Parkinson's disease will experience dysphonia, such as changes in vocal cord vibration patterns and increased internal respiratory noise [8]. Some studies have also found that changes in the fundamental frequency of speech in patients with Parkinson's disease can reflect the development of Parkinson's disease [9].

Therefore, early diagnosis of Parkinson's disease based on voice can help doctors detect the disease earlier and more accurately and predict the development of the disease [10].

At present, it is very common to use speech features to diagnose Parkinson's disease. At the same time, more and more researchers are combining speech features and machine learning methods to further improve the accuracy of recognition [4, 11]. Machine learning methods such as logistic regression, support vector machines, neural networks, decision trees, and XGBoost combined with feature selection methods have been widely used in early prediction, diagnosis and classification tasks of Parkinson's disease.

This paper uses the machine learning XGBoost algorithm to analyze 195 speech sounds from 31 people (23 people with Parkinson's disease and the remaining 8 healthy controls) collected by Max Little of the University of Oxford in collaboration with the National Speech and Speech Center in Denver, Colorado. The data were statistically analyzed to establish a prediction model for early diagnosis of Parkinson's disease. The purpose of this article is to find out the speech indicators related to Parkinson's disease and predict Parkinson's disease through speech indicators.

2. Methods

2.1. Data Sources and Description

The Parkinson prediction data used in this article comes from a Parkinson patients' voice study which is being conducted by Max Little of the University of Oxford, collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals in 2007. The original data was saved in CSV format. Each of the 24 independent variables can be treated as clinical predictor.

2.2. Variable Selection

The detail meaning of the rest continuous variables are listed as Table 1 shows:

Table 1. Different types of variables

Term	Type	Range
name	Categorical	1 to 195
Status	Categorical	0- "healthy", 1- "Parkinson"
Fo (Hz)	Numeric	88 to 261
Fhi (Hz)	Numeric	102 to 593
Flo (Hz)	Numeric	65 to 240
Jitter (%)	Numeric	0 to 0.04
Jitter (Abs)	Numeric	0 to 0.001
RAP	Numeric	0 to 0.03
PPQ	Numeric	0 to 0.02
Jitter: DDP	Numeric	0 to 0.07
Shimmer	Numeric	0 to 0.12
Shimmer(dB)	Numeric	0 to 1.4
APQ3	Numeric	0 to 0.06
APQ5	Numeric	0 to 0.08
APQ	Numeric	0 to 0.14
Shimmer: DDA	Numeric	0 to 0.17
NHR	Numeric	0 to 0.4
HNR	Numeric	8 to 34
RPDE	Numeric	0.25 to 0.69
DFA	Numeric	0.57 to 0.83
Spread1	Numeric	-7.97 to -2.42
Spread2	Numeric	0 to 0.46
D2	Numeric	0 to 0.46
PPE	Numeric	1 to 3.68

The data used in this article contains 195 observations and 24 variables, both of them without any missing values. In the follow-up research and model fitting, one variable is encoded. Predictor "status" has been changed into binary factors "0" and "1", which represents "Parkinson's" and "healthy".

2.3. Method Introduction

The main methods used in this study are XGBoost (Extreme Gradient Boosting) and logistic regression through Rstudio, both of them are two effective classification algorithms to solve the problem of binary variable.

2.3.1 XGBosst

XGBoost is a popular and efficient machine learning algorithm based on gradient boosting trees. XGBoost incorporates gradient optimization, regularization techniques, and parallel processing, Multiple weak classifiers are used to train data to improve the prediction of the model ability and training speed, which can deliver high performance and will not be limited by datasets. It offers a range of hyperparameters for fine-tuning and provides tools for model interpretation and feature importance analysis. Overall, XGBoost is a powerful and versatile algorithm that has gained popularity for its accuracy and efficiency in various domains.

2.3.2 Logistics regression

Logistic regression, a generalized linear regression model, is a simple machine learning method used to solve classification problems. The independent variable can be continuous variable or subtype variable, and the output dependent variable must be subtype variable. Different from general linear regression, the fitting function of logistic regression algorithm is sigmoid function, which makes logistics regression can map the value of any domain to 0-1, and the mapping result can be used to represent the probability of prediction.

Logistic regression model uses the maximum likelihood method to estimate the parameter values, and obtains the relationship between the dependent variable and the independent variable when taking a certain value. Moreover, this model has no requirement on the distribution of variables, so it has strong practicability.

$$Logit(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + +\beta_n x_n \quad (1)$$

3. Results and Discussion

3.1. Data Preprocrss

Figure 1 shows the correlation results of Shimmer related indicators.



Fig. 1 Correlation Test of Shimmer Related Indicators

Compared with the sample size, there are more indicators in this paper, and there are only unit differences among some indicators, and the measured variables are the same. In order to prevent multicollinearity problems, this paper chooses some indicators to test Pearson correlation coefficient, and removes one of the variables whose correlation coefficient is greater than 0.7.

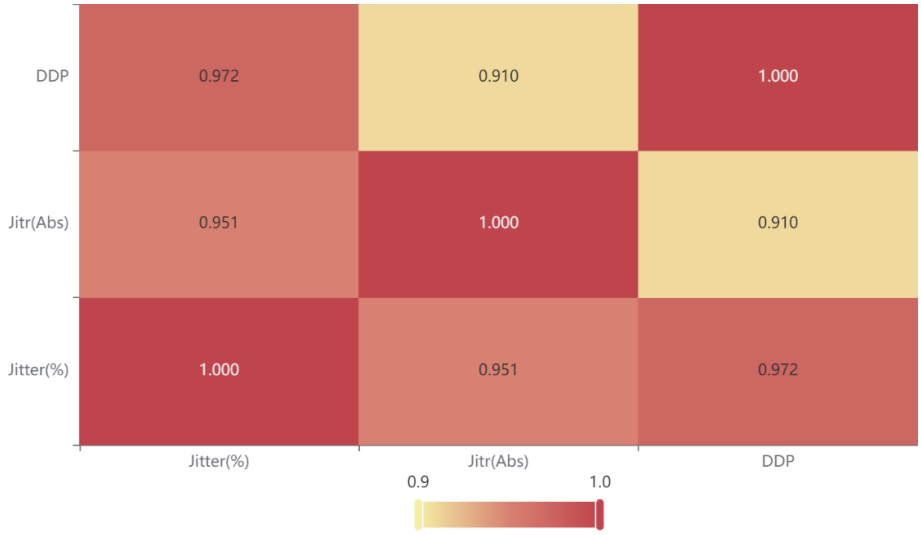


Fig. 2 Correlation Test of Jitter Related Indicators

It can be seen from the heat maps in Figure 1 and Figure 2 that there is a strong correlation between the two indicators related to Shimmer and Jitter. Therefore, in the model fitting, only shimmer and Jitter are selected as independent variables in the above indicators.

3.2. Descriptive Analysis

There is no missing data in this dataset, so there is no need to increase the missing value. The description of this article includes three parts, which are analysis of concentration trends, departure trend analysis, and related analysis of data.

Analysis of data concentration trends and departure trend analysis can be measured by maximum value, minimum value average, division, difference, bias, and peak degree. Because there are too many indicators in this article, only the relevant characteristics of the first eight indicators are displayed.

Table 2. Number of Features of Sound Indicators

Name	Minimum	Maximum	Average	Variance	SD
Fo	88.33	260.11	154.23	1713.13	41.39
Fhi	102.15	592.03	197.10	8370.70	91.49
Flo	65.48	239.17	116.32	1894.11	43.52
Jitter	0.01	0.033	0.01	0.00	0.01
HNR	8.44	33.05	21.89	19.59	4.43
RAP	0.01	0.02	0.02	0.00	0.01
PPQ	0.01	0.02	0.02	0.00	0.01
Shimmer	0.01	0.12	0.03	0.03	0.02

From the data in Table 2, it can be seen that the average values of the above indicators fall in the middle of the maximum value and the minimum value. Therefore, it can be inferred that the data of each indicator are not affected by extreme values; from the standard deviation and variance data, the distribution of Flo is relatively scattered, and the distribution of the last five indicators is relatively concentrated; the skewness and kurtosis data show the relationship between each indicator and the normal distribution. From the data in the table, it can be seen that the gap between each indicator and the normal distribution is large, which also shows that this the problem is not suitable for linear

regression fitting. Logistic regression or machine learning models should be considered that both of the models do not have requirements for data distribution for prediction.

In addition, the distribution of data can also be displayed in the figure 3.

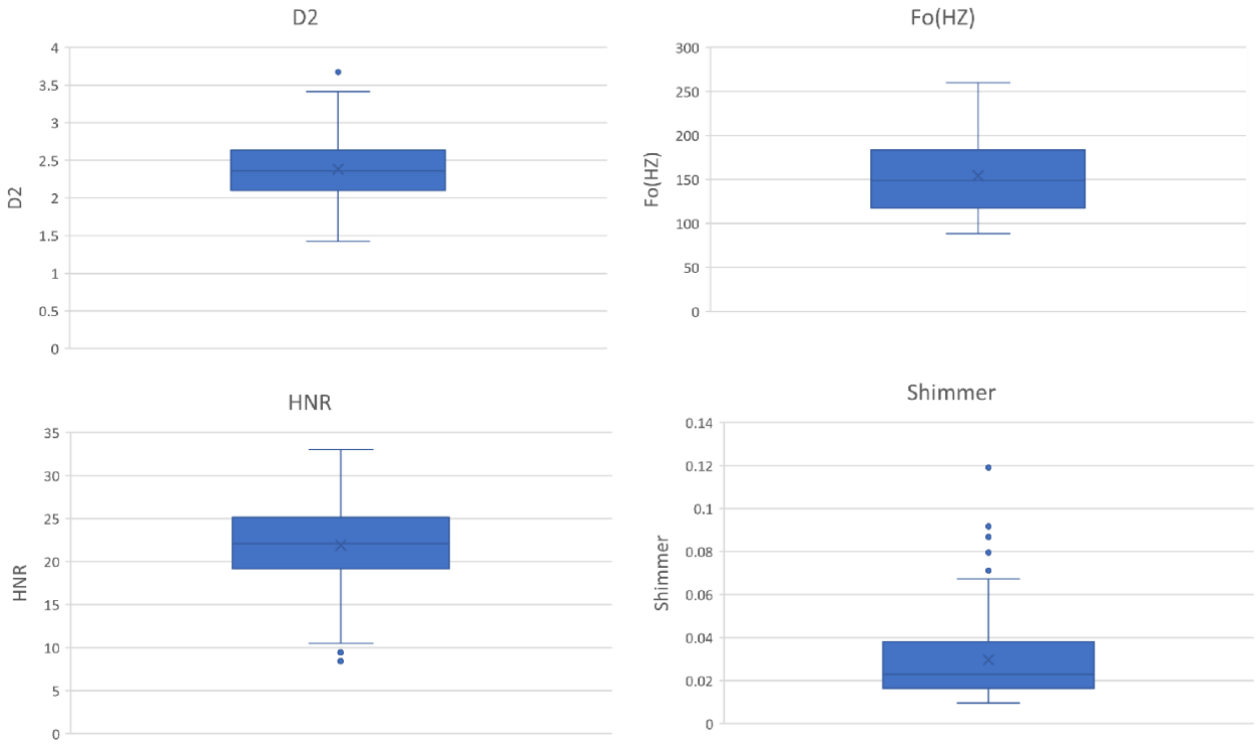


Fig. 3 Box Plot of Some Indicators

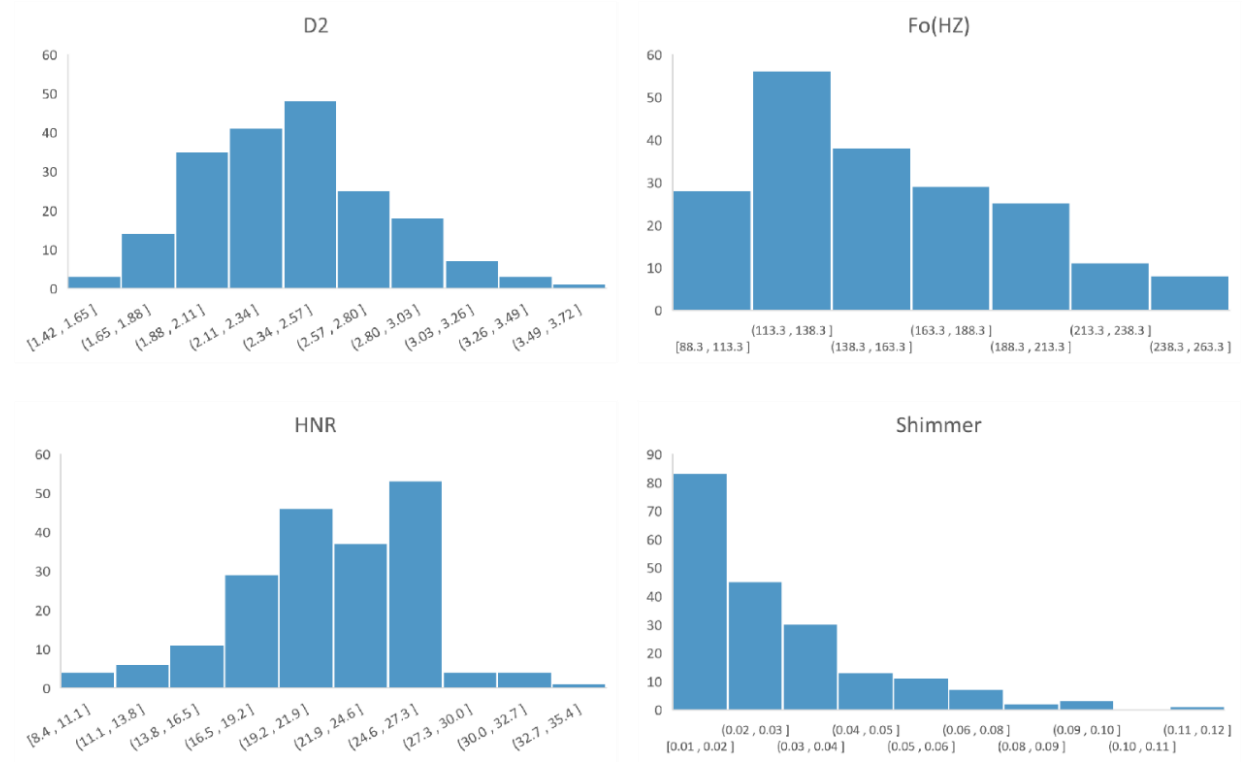


Fig. 4 Frequency distribution histogram

By figure 3, it can be seen that the data distribution of the Fo (HZ) variable is very concentrated. Most of the data is distributed between the 1/4 quantile and the 3/4 quantile; the Flo and Shimmer data show left skewness, and HNR shows right skewness. Generally speaking, the data distribution

is relatively concentrated, so the probability of outliers affecting model fitting during the modeling process is small.

Take the above variables as an example to observe the frequency histogram to further determine the distribution of the data. Frequency distribution histogram is shown in figure 4.

3.3. XGBoost Results

The 12 variables including Fo, Jitter were utilized as independent variables. Status served as the dependent variable. The training dataset was established with a proportion of 0.8, employing XGBoost machine learning algorithm and incorporating with Gradient Boosting Tree booster. Dependent variable is divided into two states: healthy and Parkinson's. The fitting model is then tested, the confusion matrix is shown in figure 5, and detail results are listed in table 3.

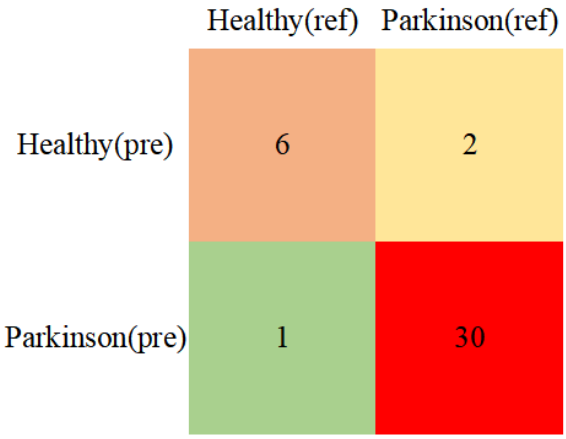


Fig. 5 Confusion matrix

As is shown in figure 5, the confusion matrix shows that 7 predicted results were "Healthy", and 8 actual results were also "Healthy". The predicted result was "Parkinson" in 32 cases and the actual result was "Parkinson" in 31 cases. This shows that the accuracy of the model is 94.87%, which means that XGBoost algorithm has a high accuracy to the test data.

Table 3. Detailed results of XGBoost

coefficient	number
Accuracy	0.9231
95% CI	(0.7913, 0.9838)
No Information Rate	0.7949
p-value	0.00***
Kappa	0.7526
Sensitivity	0.7500
Specificity	0.9677
Pos Pred Value	0.8571
Neg Pred Value	0.9375
Prevalence	0.2051
Detection Rate	0.1538
Detection Prevalence	0.1795
Balanced Accuracy	0.8589

As is shown in table 3, the results show that the model performs well on classification model to classify the data and evaluate the performance of the model. The results show that the model performs well on classification tasks. The accuracy reached 92.31%, higher than the no information rate (0.7949), and within the 95% confidence interval (0.7913 to 0.9838). The Kappa coefficient is 0.7526,

indicating a high consistency between the classification results of the model and the actual observation.

The model shows high accuracy in different categories. The sensitivity was 0.7500, indicating that the model could correctly identify positive samples. The specificity is 0.9677, indicating that the model can correctly identify negative samples. These results show that the model has high accuracy in identifying different categories.

The prediction accuracy of the model is also very high. A positive predictive value of 0.8571 indicates the accuracy of the model in a sample with a positive prediction. A negative predictive value of 0.9375 indicates the accuracy of the model in a sample with a negative prediction. These results show that the prediction accuracy of the model is higher in different categories.

In the whole dataset, the proportion of positive samples was 0.2051. The detection rate of the model is 0.1538, indicating the ability of the model to correctly identify positive samples. The proportion of detection rate is also 0.1538, indicating the proportion of positive samples in the whole data set.

In summary, the classification model shows good performance in the classification task, with high accuracy and prediction ability.

3.4. Logistics Results

The dependent variable of logistics regression is determined as status, and the independent variable is shown in the table 4. In order to prevent the coefficient gap from being too large, this paper fits each index after standardization.

Table 4. Fit the selected variables and coefficients

Independent variable	Variable representation	coefficient
Fo	β_1	-0.686
PPE	β_2	3.133
Shimmer	β_3	1.069
HNR	β_4	0.322
PPQ	β_5	-4.928
D2	β_6	0.579
RAP	β_7	10.522
Spread2	β_8	1.057
RPDE	β_9	-0.462
DFA	β_{10}	0.222
NHR	β_{11}	0.893
Jitter	β_{12}	3.025

The logistics equation is as follows: It can be seen from the heat maps in Figure 2 and Figure 3 that there is a strong correlation between the two indicators related to Shimmer and Jitter. Therefore, in the model fitting, only shimmer and Jitter are selected as independent variables in the above indicators.

$$\ln\left(\frac{P}{1-P}\right) = 3.025 - 0.686\beta_1 + 3.133\beta_2 + \cdots + 0.893\beta_{11} + 3.025\beta_{12} \quad (2)$$

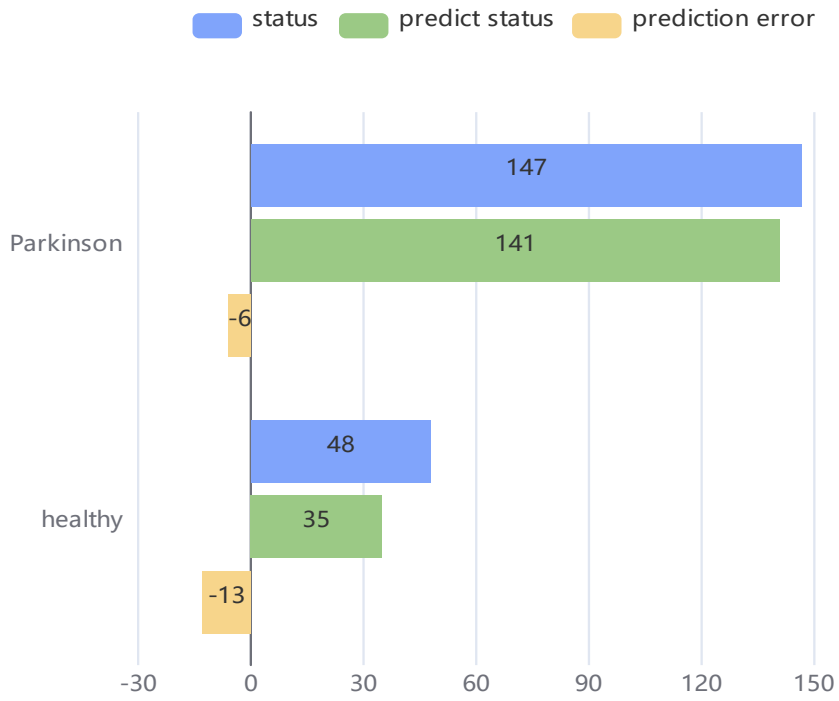


Fig. 6 Logistics forecast results show

Table 5. Evaluation of model fitting effect

P	AIC	BIC
0.000***	37.505	68.767

As can be seen from figure 6 and table 5, the prediction accuracy of this model reaches 90.3%, and a P value less than 0.05 indicates that this model is effective. The smaller the AIC and BIC values are, the better. Generally, if the values are less than 50, it can be considered that the model has a better fitting effect and the number of parameters is small, so overfitting will not occur.

4. Conclusion

This paper used data from Max Little of the University of Oxford to predict the disease status of the reference group. In general, this study conducted the Pearson correlation coefficient test on some indicators to reduce dimensionality, eliminated ten variables with correlation coefficients greater than 0.7, and standardized the remaining variables first. Logistic regression model and XGBoost model were used to fit the data, and confusion matrix was used to judge the prediction accuracy of the two methods. The results showed that the prediction accuracy of the two models for Parkinson's disease reached more than 90%, and the model fit well, but the XGBoost model had a higher prediction accuracy, reaching 92%. Therefore, XGBoost has higher early diagnosis ability of Parkinson's disease. However, due to the high sensitivity of the XGBoost model, when the model parameters change, the model prediction accuracy will be greatly affected. At this time, using the logistic model to modify the prediction results of the XGBoost model can better determine the patient's Parkinson's disease status.

For future research, more relevant data should be collected, including some clinical data to make model predictions more accurate. The model in this study has profound implications for human prediction and diagnosis of Parkinson's disease.

Authors Contribution

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

References

[1] Wang Zihao, Xia Huan, FENG Tingting et al. Screening of key genes for diagnosis of Parkinson's disease based on bioinformatics and machine learning algorithm. Neurological Diseases and Mental Health, 2019, 23(12): 837-847.

[2] Liu Changrong. Parkinson's disease with high disease burden will have the first rehabilitation "group standard". China Youth Daily, 2024.

[3] Wen Pengcheng. Research on intelligent diagnosis model of Parkinson's disease based on speech signal. Hubei University for Nationalities, China, 2023.

[4] Tuncer T, Dogan S, Acharya U R. Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. Biocybernetics and Biomedical Engineering, 2019, 40(1): 211-220.

[5] Lin Zhao, Zhao Hong, Liu Wang, et al. Progress in the treatment of bladder dysfunction in Parkinson's disease. Nerve Injury and Functional Reconstruction, 2019, 19(01): 51-54.

[6] Xu Baolei, Zhang Xiaojun, Sun Yan, et al. Influence factors of Parkinson's disease combined with depression and changes of glucose metabolism in brain. Journal of Clinical Neurology, 2019, 35(05): 324-328.

[7] Liu Aiyao. Research and system implementation of speech detection algorithm for Parkinson's disease. Shandong Agricultural University, 2023.

[8] Wang P. Research on pathologic speech detection and classification based on acoustic and kinematic characteristics. Taiyuan University of Technology, 2021.

[9] Duan Jiayi, Wei Shaohui. Effect of therapeutic communication on anxiety and depression in patients with Parkinson's disease. Nursing Research, 2019, 37(19): 3592-3596.

[10] Mou Xinguang, Tao Jiaxin, Chen Long. Parkinson's Disease diagnosis based on speech feature fusion. Digital Manufacturing Science, 2023, 21(03): 225-230.

[11] Salama A. Mostafa, Aida Mustapha, Mazin Abed Mohammed, Raed Ibraheem Hamed, N. Arunkumar, Mohd Khanapi Abd Ghani, Mustafa Musa Jaber, Shihab Hamad Khaleefah. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson’s disease. Cognitive Systems Research, 2019.