DOI: 10.23977/medsc.2024.050224
ISSN 2616-1907 Vol. 5 Num. 2

A study on Parkinson's disease diagnosis by random forest algorithm based on improved speech features

Zhijun Li[1,a], Jingxuan He[1,b], Di Sun[2,c], Haixia Li[3,d,*]

[1]North China University of Technology, Beijing, 100144, China
[2]Graduate College, Beijing University of Traditional Chinese Medicine, Beijing, 100029, China
[3]Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, 100053, China
[a]lzj78@ncut.edu.cn, [b]981571978@qq.com, [c]747593987@qq.com, [d]lihaixia@gamyy.cn
[*]Corresponding author

Keywords: Parkinson's disease, machine learning, deep learning, neural networks, speech diagnosis, artificial intelligence

Abstract: Neurological disorders have a serious impact on human life worldwide. Parkinson's disease, also known as idiopathic or primary Parkinson's disease, is one of the most common neurological disorders. In recent years, research related to the link between Parkinson's disease and speech has received more and more attention, and many methods to process speech signals through algorithms and thus predict the prevalence of Parkinson's have been proposed, and most of the studies have diagnosed the prevalence of Parkinson's disease by employing the speech signals of the subjects, and the results have mostly been better, successfully responding to the link between speech and Parkinson's disease. In this article, the speech signals of the subjects (554 cases in total, of which 220 cases are healthy people and 314 cases are Parkinson's patients) are collected, processed, and 12 kinds of complex speech features are extracted from them.By comparing these 12 kinds of speech features of Parkinson's patients and healthy people, three main classes of features are selected from them, which are Fundamental Frequency Perturbation Jitter Class, Amplitude Perturbation Shimmer Class, and Harmonic Signal-to-Noise Ratio (HNR) class.The speech features of the subjects are trained and tested by neural network, and comparative experiments are made on XGBoost algorithm, svm algorithm, random forest algorithm, KNN algorithm in machine learning and DNN neural network and LSTM neural network in deep learning.It is found that Random Forest Algorithm and can effectively solve the neural network's over-fitting and the problem of low accuracy and recall, and very effectively distinguish between Parkinson's patients and healthy people, with an accuracy rate of 99.3% and a recall rate of 100%.

1. Introduction

Parkinson's Disease (PD) is a prevalent neurological disorder that mainly impacts the motor system. Symptoms progress gradually and consist of tremors, stiffness, slow movement, and walking difficulties. Additionally, non-motor symptoms like depression, dementia, and sleep issues may also manifest. It's important to note that the majority of individuals with Parkinson's disease do not have a family history of the condition. Studies have shown that Parkinson's disease cannot be cured by modern medicine, but it can be treated with medication, surgery, and other treatments[1].Thus, studying the early diagnosis of PD is important for controlling the condition of PD patients and prolonging their lives[2].

In recent years, deep learning has achieved significant application results in various fields, such as speech enhancement, speech recognition, speech emotion recognition and speech pathology detection.In addition, deep learning has a wide range of applications in other fields. Particularly noteworthy is that speech pathology detection provides an application basis for recognizing Parkinson's disease patients using acoustic features.Lucijano Berus et al[3] utilized the original speech directly into the ANN neural network for classification with average results, while Shi Haobin et al[4] went directly into the Alex net network for predictive classification and obtained 87% accuracy.Huang Fangliang[5][11] et al. on the other hand, used a speech feature extraction technique to extract the voice features of the subjects and entered the extracted features into a residual neural network for classification, and although the training accuracy was 97%, probably due to the lack of data samples, the model accuracy was only 73% and the model recall was 87.5%. In this article, Parselmouth audio processing and signal processing techniques are used to extract 12 features of speech acoustics of PD patients (314 cases) and healthy people (240 cases) to form a dataset, and the set is combined with 6 algorithm models to train the model, which is expected to achieve a better understanding of PD patients and healthy people's voice acoustics from the point of view of voice features recognition. Recognition perspective to achieve early diagnosis of PD patients, and compared with Lucijano Berus et al. and residual neural network and Alex net network methods of Huang Fangliang and Shi Haobin et al. found that the AI-based random forest algorithm has very high early pd recognition model accuracy (99.3), model recall (98.6%) and model precision rate (100%). It has high reference significance for PD patients in early diagnosis methods.

2. Datasets and Initial Processing of Data

2.1. Overall process of model diagnosis (As shown in Figures 1)

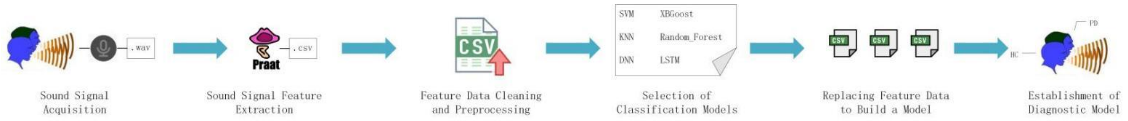

.wav
Praat
.csv
CSV
SVM
XBGoost
KNN
Random_Forest
DNN
LSTM
CSV
CSV
CSV
HC
PD
Sound Signal Acquisition
Sound Signal Feature Extraction
Feature Data Cleaning and Preprocessing
Selection of Classification Models
Replacing Feature Data to Build a Model
Establishment of Diagnostic Model

Figure 1: Overall Process of Model Diagnosis.

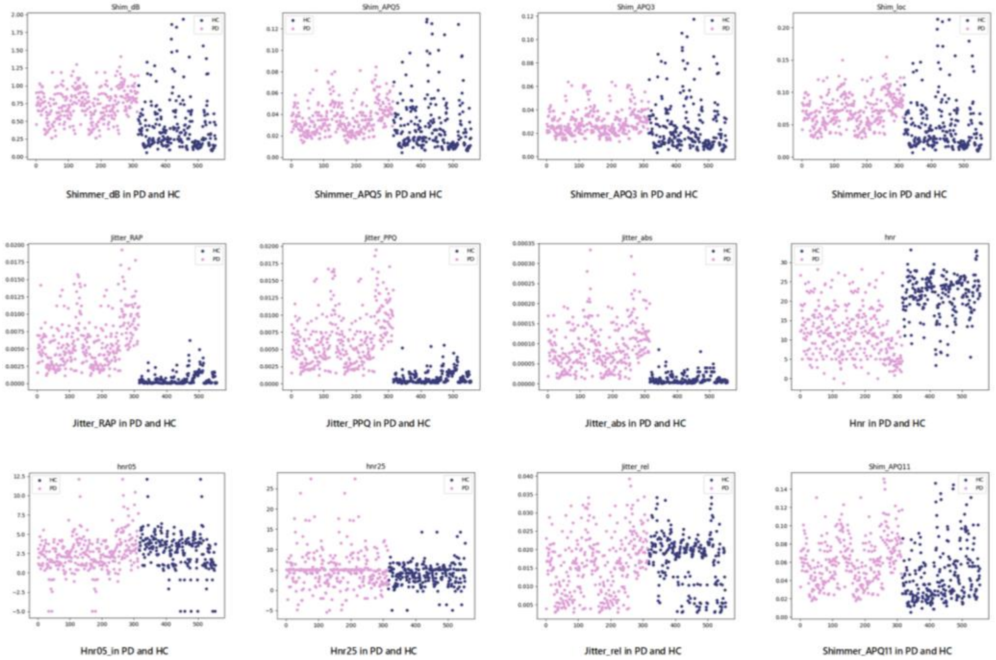
2.2. Data set collection

The data set in this article adopts 1. The voice data of 27 healthy HC people from 120 PD patients collected by the professional voice collection system provided by Guang'anmen Hospital of the China Academy of Chinese Medical Sciences. The subjects are invited to read out the “huang”,“chong”,“su”,“shi”,“gu”,“yu”,“tian”,“zhu”,“ming”,“bi”ten Chinese words. (2)King's College London, UK (MDVR-KCL)[6] In September 2017, a typical examination room with an area of about 10 square meters and a typical reverberation time of about 500 ms were used for voice recordings made via a smartphone in the form of a phone call, where subjects (16 patients with PD, 21 healthy individuals) were invited to read aloud by phone "The North Wind and the Sun "Computer applications in geography snippet" and engaged in a conversation with randomized questions, resulting in high-quality recordings.3. Istanbul University (CerrahpaŸa) Faculty of Medicine Department of Neurology[7]of 188 PD patients.The control group consisted of 64 healthy individuals.Sustained vocalizations of the vowel /a/ were collected from each subject after

Physician's Check, repeated three times for 178 randomly selected cases from the 188*3 PD dataset taken from all HC healthy individuals. In total, voice data from 314 PD patients and 240 HC healthy individuals were saved, all audio data sampling rates are 44.1kHz, and the saved files are all .wav files.

2.3. Data set feature mining and selection

In order to further mine the sound characteristics of all the sorted sound data sets (554 cases in total), Praat algorithm is used in this study. Praat is a computer program used to analyze, synthesize and operate speech, which can create high-quality speech images. Praat can also do spectrum analysis, formant tracking, pitch extraction, morphological analysis, speech synthesis and other tasks. This study uses Praat for complex speech feature extraction from datasets, including fundamental frequency perturbations (Jitter type), amplitude perturbations (Shimmer type), and harmonic signal-to-noise ratio features (HNR type) [8] [9] [10] [5] [11]. The Jitter type is further divided into three sub categories, namely pitch period absolute perturbations Jitter_ ABS, Jitter with 5-point disturbance adjacent to the pitch period_ PPQ5 (later referred to as Jitter_PPQ) and three adjacent points of pitch period disturbance Jitter_ RAP has three types, Shimmer type is divided into amplitude disturbance (dB) Shimmer_ DB, amplitude disturbance: Shimmer_ Loc, amplitude adjacent to 3 points Shimmer_ APQ3, Shimmer with 5 adjacent amplitudes_ APQ5 has four categories, and the harmonic signal-to-noise ratio is characterized by an average harmonic signal-to-noise ratio (HNR) of 8 categories.Of course, the specific criteria for selecting 8 types of features from the 12 types of features extracted by Praat analysis are based on the 12 type feature indicators of 314 PD and 240 HC cases, Jitter_ abs,Jitter_ PPQ,Jitter_ RAP,Shimmer_ dB,Shimmer_ loc,Shimmer_ APQ3,Shimmer_ APQ5, HNR, and the eight feature indicators can more clearly distinguish PD and HC. The experiment analyzed the data images formed by the feature extraction dataset with sample size as the x-axis and each feature as the y-axis, as shown in Figure 2. It can be observed with the naked eye that the scatter plots of the eight feature categories (Jitter_abs, Jitter_PPQ, Jitter_RAP, Shimmer_dB, Shimmer_loc, Shimmer_APQ3, Shimmer_APQ5, HNR (HNR15)) in the first two rows can almost clearly distinguish PD and HC:



Shimmer_dB in PD and HC

Shimmer_APQ5 in PD and HC

Shimmer_APQ3 in PD and HC

Shimmer_loc in PD and HC

Jitter_RAP in PD and HC

Jitter_PPQ in PD and HC

Jitter_abs in PD and HC

Hnr in PD and HC

Hnr05_in PD and HC

Hnr25 in PD and HC

Jitter_rel in PD and HC

Shimmer_APQ11 in PD and HC

Figure 2: Scatter plots of 12 features of PD and HC.

Based on the comparison of 12 types of feature scatter plots, we ultimately selected Jitter_ abs,Jitter_ PPQ,Jitter_ RAP,Shimmer_ dB,Shimmer_ loc,Shimmer_ APQ3,Shimmer_ APQ5 and HNR eight types of feature indicators are used as feature inputs for this experiment.

As shown in Table 1, in this article, the Jitter class is referred to as the fundamental frequency perturbation feature, the Shimmer class is referred to as the amplitude perturbation feature, and the HNR class is referred to as the harmonic signal-to-noise ratio:

Table 1: Selected 8 characteristics and basis for categorization

Feature	Classification basis
Jitter_abs	Fundamental frequency disturbance characteristics
Jitter_RAP	Fundamental frequency disturbance characteristics
Jitter_PPQ	Fundamental frequency disturbance characteristics
Shimmer_dB	Amplitude disturbance characteristics
Shimmer_loc	Amplitude disturbance characteristics
Shimmer_AQ3	Amplitude disturbance characteristics
Shimmer_AQ5	Amplitude disturbance characteristics
HNR	Harmonic signal-to-noise ratio

Through analysis, it was found that the Jitter and Shimmer features of Parkinson's patients showed higher numerical values compared to the general population, but the Jitter features were more able to present differences in speech characteristics between the two groups of people compared to the Shimmer features. The general population's Jitter features were lower and more clustered within a range. The HNR characteristics of Parkinson's patients exhibit lower values compared to normal individuals.

These parameters can be used to evaluate the characteristics and differences between normal and pathological speech.

2.4. Feature dimensionality reduction

Table 2: Feature data sets and corresponding feature schemes

Characteristic data group	Feature scheme
Characteristic data group 1	Jitter_abs,Shimmer_APQ3,HNR
Characteristic data group 2	Jitter_abs,Shimmer_APQ5,HNR
Continued from Table 2	
Characteristic data group 3	Jitter_abs,Shimmer_dB,HNR
Characteristic data group 4	Jitter_abs,Shimmer_loc,HNR
Characteristic data group 5	Jitter_PPQ,Shimmer_APQ3,HNR
Characteristic data group 6	Jitter_PPQ,Shimmer_APQ5,HNR
Characteristic data group 7	Jitter_PPQ,Shimmer_dB,HNR
Characteristic data group 8	Jitter_PPQ,Shimmer_loc,HNR
Characteristic data group 9	Jitter_RAP,Shimmer_APQ3,HNR
Characteristic data group 10	Jitter_RAP,Shimmer_APQ5,HNR
Characteristic data group 11	Jitter_RAP,Shimmer_dB,HNR
Characteristic data group 12	Jitter_RAP,Shimmer_loc,HNR

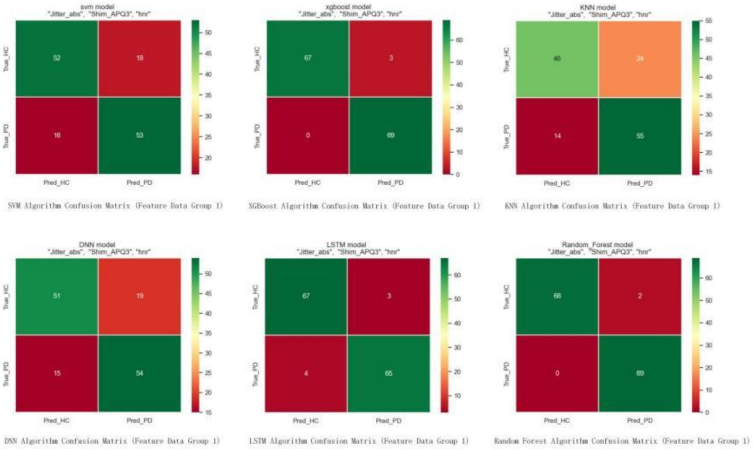
Through feature mining and selection of sounds, as well as cleaning and pre-processing of feature data, we obtained eight types of sound feature data for our study, namely Jitter_ abs,Jitter_ PPQ,Jitter_ RAP,Shimmer_ dB,Shimmer_ loc,Shimmer_ APQ3,Shimmer_ After the feature selection is completed for APQ5 and HNR, although the model can be directly trained and 8 types of features can be directly placed into the classifier of the model, the problem of large computational complexity and long training time may be caused by the large feature matrix. Therefore, reducing the dimensionality of the feature matrix is also essential. In order to reduce time and spatial complexity, and for simpler models to have stronger robustness on small feature datasets, this experiment directly divides 8 types of features into 3 types, among which Jitter_ abs,Jitter_ PPQ,Jitter_ RAP is a Jitter class, Shimmer_ dB,Shimmer_ loc,Shimmer_ APQ3,Shimmer_ APQ5 is Shimmer class, while HNR is HNR class.

This experiment will arrange and combine each of these three types of features to form a 3 * 4 * 1 feature scheme, as shown in Table 2.

3. Experimental methods and result evaluation

3.1. Evaluation of classification methods and classification models for data features

In order to distinguish the accuracy of each model for feature data classification, this experiment first specifies the feature data as a unique group: feature data group 1 as our sample experiment, i.e., all models are classified by the feature data of feature data group 1 (Jitter_abs, Shimmer_APQ3, HNR), and the results of the experiments, as well as the experimental model evaluation indexes are as follows(As shown in Figure 3 and Figure 4)(This article uses confusion matrix evaluation metrics to evaluate model performance):



SVM Algorithm Confusion Matrix (Feature Data Group 1)

XGBoost Algorithm Confusion Matrix (Feature Data Group 1)

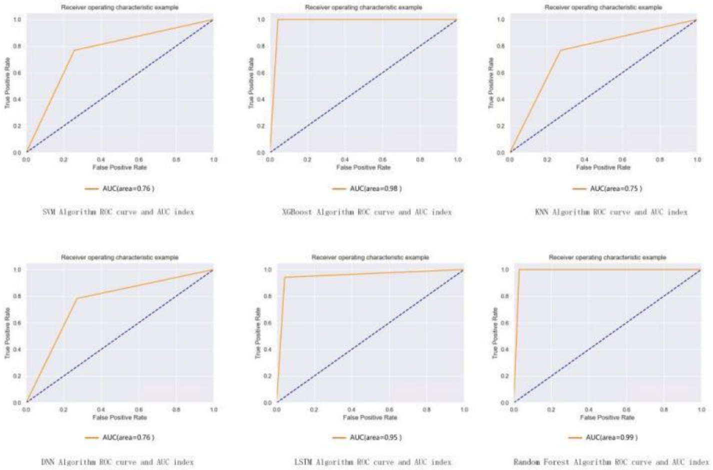
KNN Algorithm Confusion Matrix (Feature Data Group 1)

DNN Algorithm Confusion Matrix (Feature Data Group 1)

LSTM Algorithm Confusion Matrix (Feature Data Group 1)

Random Forest Algorithm Confusion Matrix (Feature Data Group 1)

Figure 3: Confusion matrix for 6 models



SVM Algorithm ROC curve and AUC index

XGBoost Algorithm ROC curve and AUC index

KNN Algorithm ROC curve and AUC index

DNN Algorithm ROC curve and AUC index

LSTM Algorithm ROC curve and AUC index

Random Forest Algorithm ROC curve and AUC index

Figure 4: ROC curves and AUC indicators for 6 different models

The ROC of the model is the subject operating characteristic curve, which reflects the relationship between the true positive rate and the false positive rate of the classifier at different thresholds, and the AUC metric of the model is defined as the area of the lower half of the ROC curve, which indicates the probability that the classifier ranks positive samples ahead of negative samples, and the larger the value of the AUC of the model, the better.

3.2. For each feature data individual algorithm model accuracy

Table 3: True Positive Rate TPR% for 12 sets of feature data predicted by 6 different models

Feature Model	Feature Data Group1	Feature Data Group2	Feature Data Group3	Feature Data Group4	Feature Data Group5	Feature Data Group6	Feature Data Group7	Feature Data Group8	Feature Data Group9	Feature Data Group10	Feature Data Group11	Feature Data Group12
SVM	76.81	76.81	76.81	76.81	76.81	76.81	76.81	76.81	76.81	76.81	76.81	76.81
KNN	76.81	76.81	78.26	76.81	76.81	78.26	76.81	76.81	76.81	76.81	77.14	76.81
XGBoost	100	100	100	100	100	100	100	100	100	100	100	100
DNN	76.81	82.60	98.55	79.71	81.15	82.60	89.86	81.15	82.60	73.91	91.30	86.96
LSTM	94.20	95.65	95.65	94.20	97.10	97.10	97.10	97.10	97.10	98.55	98.55	97.10
RF	100	100	100	100	100	100	100	100	100	100	100	100

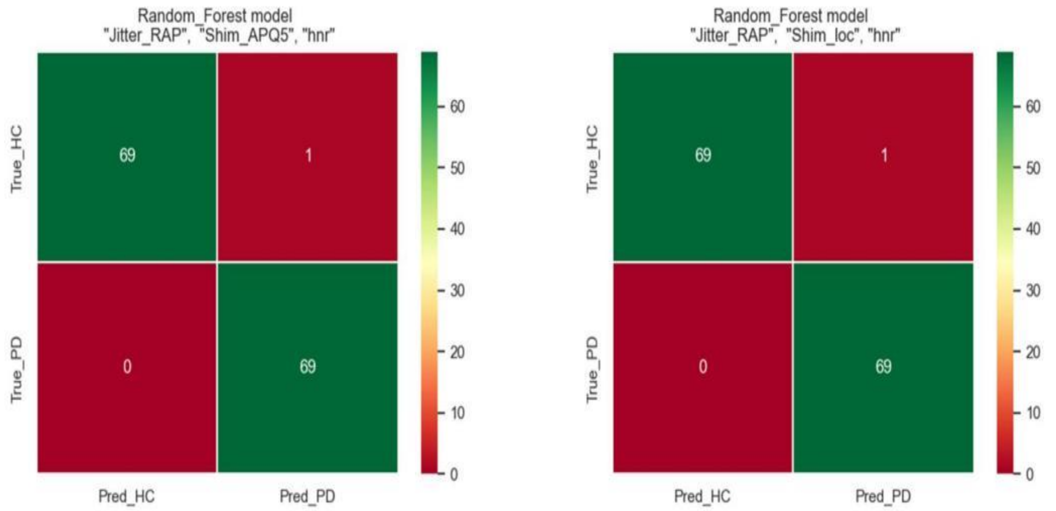
Table 4: Model accuracy ACC% of 12 sets of feature data predicted by 6 different models

Feature Model	Feature Data Group1	Feature Data Group2	Feature Data Group3	Feature Data Group4	Feature Data Group5	Feature Data Group6	Feature Data Group7	Feature Data Group8	Feature Data Group9	Feature Data Group10	Feature Data Group11	Feature Data Group12
SVM	75.53	75.53	75.53	75.53	75.53	75.53	75.53	75.53	75.53	75.53	75.53	75.53
KNN	74.28	74.28	75.53	74.28	74.28	75.53	74.28	74.28	74.28	74.28	75.0	74.28
XGBoost	97.84	98.56	97.84	98.56	97.12	96.40	96.40	97.84	98.56	97.12	97.12	98.56
DNN	77.70	72.66	77.70	74.82	82.01	84.89	79.86	77.70	85.61	84.17	79.14	74.10
LSTM	94.96	94.96	94.24	94.96	94.24	94.24	94.24	94.24	94.24	94.96	94.96	94.24
RF	98.56	96.40	95.68	96.40	96.40	98.56	96.40	97.84	97.84	99.28	98.56	99.28

Table 5: False Positive Rate FPR% for 12 sets of feature data predicted by 6 different models

Feature Model	Feature Data Group1	Feature Data Group2	Feature Data Group3	Feature Data Group4	Feature Data Group5	Feature Data Group6	Feature Data Group7	Feature Data Group8	Feature Data Group9	Feature Data Group10	Feature Data Group11	Feature Data Group12
SVM	25.71	25.71	25.71	25.71	25.71	25.71	25.71	25.71	25.71	25.71	25.71	25.71
KNN	27.14	27.14	27.14	27.14	27.14	27.14	27.14	27.14	27.14	27.14	27.14	27.14
XGBoost	4.28	2.86	4.28	2.86	5.71	7.14	7.14	4.28	2.86	5.71	5.71	2.86
DNN	21.43	37.14	42.85	30	17.14	12.86	30	25.71	11.43	5.71	32.86	38.57
LSTM	4.28	5.71	7.14	4.28	8.57	8.57	8.57	8.57	8.57	8.57	8.57	8.57
RF	2.86	7.14	8.57	7.14	7.14	4.28	8.57	4.28	4.28	1.43	2.86	1.43

According to 6 different models for 12 different feature data groups, this article did 72 groups of experimental studies, the ACC, TPR, FPR indexes of the specific models are shown in Table 3, Table 4, Table 5.We found that the model evaluation indexes of the 70th group of experiments and 72nd group of experiments are the best.This article draws them out and analyzes them separately,found that the data feature group 10 and data feature group 12 are the best, whose confusion matrix is shown in Figure 5:

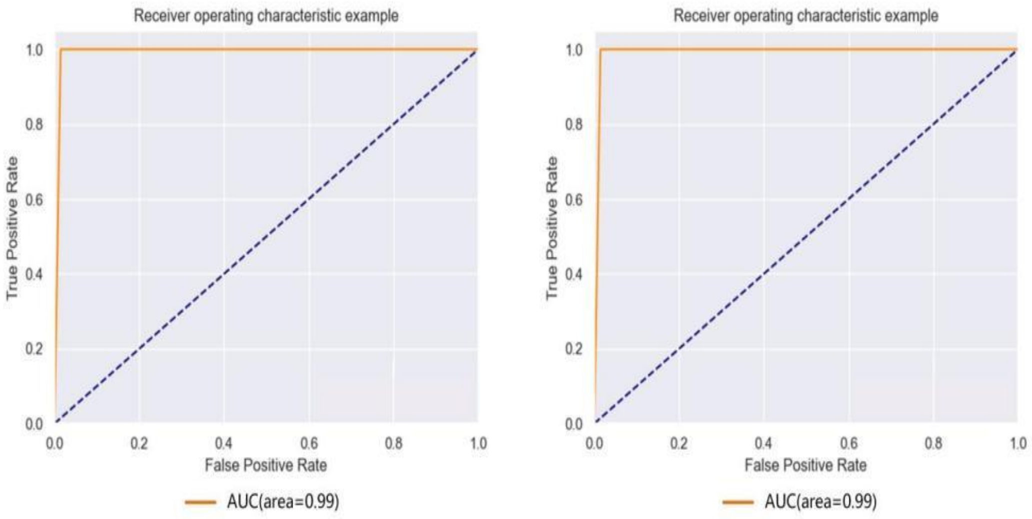


Random Forest Algorithm Confusion Matrix (Feature Data Group 10)

Random Forest Algorithm Confusion Matrix (Feature Data Group 12)

Figure 5: Model confusion matrix of characteristic data group 10 and 12 of random forest algorithm

The ROC curve and AUC indicators are shown in Figure 6:



Random Forest Algorithm ROC curve and AUC index (Feature Data Group 10)

Random Forest Algorithm ROC curve and AUC index (Feature Data Group 12)

Figure 6: ROC curve and AUC index of random forest algorithm characteristic data group 10 and 12 models

It can be found that the AUC of the classification and diagnostic models trained by the Random Forest algorithm for both feature data set 12 and feature data set 10 reached 0.99, and their ACCs also reached 99.28%, TPR=1, FPR=0.014.

4. Conclusion

This article uses machine learning algorithms and deep learning neural networks to classify the 8 types of sound features extracted from sound signals. Through training on 12 sets of feature data sets composed of 6 models and 8 types of sound features, the comparison of the above two evaluation dimensions shows that the 6 prediction models generally have good prediction accuracy for PD patients and HC healthy individuals. The reasons for this are investigated, this article extracted 12 feature dimensions of speech through data mining, and observed and analyzed them. It was found that there were significant differences in the three categories of speech features: Jitter, Shimmer, and HNR, but the differences in the other categories were negligible. Therefore, this article first screened the originally extracted 12-dimensional data and extracted the features with significant differences between PD and HC. After extracting the 8-dimensional features, we reduce the number of features in 8-dimensional feature data without losing too much information, categorize them into three categories.Finally we arrange and combine them to greatly reduce the storage space and calculation time of the data.The method proposed in this article greatly improve the generalization ability and stability of the model, eliminate the correlation between features, and improve the interpretability of the model.

In the research, this article found that Random Forest algorithm has more obvious advantages than other models in the identification and classification of PD patients and HC healthy people.Analyze the reason, for the research object is audio, video, image, text and other data sets, which has a large number of features, each feature and the final result may have a relationship but not so obvious.But after feature selection and feature dimensionality reduction, and then after the arrangement and combination of features to form a feature group, not only reduces the number of feature dimensions, but also does not lose too much information. The Random Forest algorithm based on traditional data mining methods performs classification and regression by constructing multiple decision trees, which happens to be more adept at handling tasks with fewer features and a more obvious relationship between the features and the results, and has a strong generalization ability, learns the interactions between the features and is not prone to overfitting.

Parkinson's patients and healthy people have a lot of different physiological differences and differences.This article from the perspective of voice features, feature dimensionality reduction and machine learning and deep learning neural networks to study the differences between the medical Parkinson's patients and healthy people.This article also provides new ideas and methods for early diagnosis of Parkinson's disease.The method proposed in this article, based on traditional data mining and feature dimensionality reduction of the Random Forest Algorithm, achieves a diagnostic model accuracy of up to 99.28%.The next step is to use WeChat applet, to make "based on voice recognition Parkinson's patient diagnosis system", for early detection of Parkinson's disease contribute to the value.

References

[1] Connollybs, Langae. Pharmacological treatment of Parkinson disease: a review[J].Jama, 2014,311(16): 1670-1683.

[2] Pereira Cr, Webersa, Hook C, et al.Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics [C]. Proceedings of the 29th SIBGRAPI Conference on Graphics, Sao Paulo: Patterns and Images (SIBGRAPI), IEEE, 2016: 340-346

[3] Berus L, Klancnik S,Brezocnik M.Classifying Parkinson's disease based on acoustic measures using artificial neural networks[J]. Sensors, 2019, 19(1): 1-15.

[4] Shi Hao-Bin.Research on the diagnosis of speech disorders in Parkinson's disease based on convolutional neural network [D]. Qinhuangdao: Yanshan University, 2017.

[5] Huang Fang-Liang, Xu Huan-Qing, Shen Tong-Ping, Jin Li, Yu Lei. A study on Parkinson's disease identification combining residual neural network and speech diagnosis [J]. Journal of Qilu University of Technology, 2022, 36(01): 36-43. DOI:10.16442/j.cnki.qlgydxxb.2022.01.006.

[6] Hagen Jaeger, Dhaval Trivedi and Michael Stadtschnitzer.(2019). King's College London (MDVR-KCL) Mobile device recordings in patients with early and late onset Parkinson's disease and healthy controls [Dataset]. Zenodo.https://doi.org/10.5281/zenodo.2867216

[7] Sakar, C.O., Serbes, G., Gunduz, A., Tunc, H.C., Nizam, H., Sakar, B.E., Tutuncu, M., Aydin, T., Isenkul, M.E. and

Apaydin, H., 2018. Speech signal processing for Parkinson's disease classification A comparative analysis of algorithms and the use of adjustable Q-factor wavelet transform. Applied Soft Computing.

[8] Zhang Tao, Peipei Jiang, Zhang Yajuan, et al. Research on the analysis method of speech impairment in Parkinson's disease based on local statistics in time-frequency mixed domain [J]. Journal of Biomedical Engineering, 2021, 38(1): 21-29.

[9] Xu Jing. Exploring the quantitative assessment method of hypokinetic dysarthria Parkinson's disease based on acoustic features [D]. Guangzhou: Jinan University, 2020.

[10] Liu Feng. Diagnosis and prediction of Parkinson's disease based on hand mapping and speech [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2019.

[11] F. Huang, H. Xu, T. Shen and L. Jin, "Recognition of Parkinson's Disease Based on Residual Neural Network and Voice Diagnosis," 2021 IEEE 5th Information Technology ,Networking, Electronic and Automation Control Conference (ITNEC), Xi'an, China, 2021, pp. 381-386, doi: 10.1109/ITNEC52019.2021.9586915.