Bi XA, Xing ZX, Xu RH *et al.* An efficient WRF framework for discovering risk genes and abnormal brain regions in Parkinson's disease based on imaging genetics data. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 36(2): 361–374 Mar. 2021. DOI 10.1007/s11390-021-0801-6

An Efficient WRF Framework for Discovering Risk Genes and Abnormal Brain Regions in Parkinson's Disease Based on Imaging Genetics Data

Xia-An Bi, *Member*, *CCF*, *IEEE*, Zhao-Xu Xing, Rui-Hui Xu, and Xi Hu

College of Information Science and Engineering, Hunan Normal University, Changsha 410006, China

Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410006, China

E-mail: bixiaan@hnu.edu.cn; {xingzhaoxuhnnu, xuruihuihnnu, huxihnnu}@163.com

Received July 14, 2020; accepted February 28, 2021.

Abstract As an emerging research field of brain science, multimodal data fusion analysis has attracted broader attention in the study of complex brain diseases such as Parkinson's disease (PD). However, current studies primarily lie with detecting the association among different modal data and reducing data attributes. The data mining method after fusion and the overall analysis framework are neglected. In this study, we propose a weighted random forest (WRF) model as the feature screening classifier. The interactions between genes and brain regions are detected as input multimodal fusion features by the correlation analysis method. We implement sample classification and optimal feature selection based on WRF, and construct a multimodal analysis framework for exploring the pathogenic factors of PD. The experimental results in Parkinson's Progression Markers Initiative (PPMI) database show that WRF performs better compared with some advanced methods, and the brain regions and genes related to PD are detected. The fusion of multi-modal data can improve the classification of PD patients and detect the pathogenic factors more comprehensively, which provides a novel perspective for the diagnosis and research of PD. We also show the great potential of WRF to perform the multimodal data fusion analysis of other brain diseases.

Keywords multimodal fusion feature, Parkinson's disease, pathogenic factor detection, sample classification, weighted random forest model

1 Introduction

Parkinson's disease (PD) is an extremely universal neurodegenerative disease with a high disability rate among the elderly [1, 2]. If it can be diagnosed and treated at the initial stage, most patients can still maintain a normal state within a few years of onset. For this reason, the early diagnosis is of great significance to PD patients. With the deepening of the research on PD based on the classical single modality, the researchers find that PD may be the result of the interaction among pathogenic factors from different modalities. Therefore, with the rapid development of the detection technologies for the structure, function and genetic factors of brain diseases, the comprehensive and systematic study of PD combined with neuroimaging and genetics has attracted more and more attention [3].

Although imaging genetics is an emerging research field, it has been developing rapidly in the past few years. For example, Kim *et al.* [4] constructed a lin-

Regular Paper

Special Section on AI and Big Data Analytics in Biology and Medicine

This work was supported by the National Natural Science Foundation of China under Grant No. 62072173, the Natural Science Foundation of Hunan Province of China under Grant No. 2020JJ4432, the Key Scientific Research Projects of Department of Education of Hunan Province under Grant No. 20A296, the Degree and Postgraduate Education Reform Project of Hunan Province under Grant No. 2019JGYB091, Hunan Provincial Science and Technology Project Foundation under Grant No. 2018TP1018, and the Innovation and Entrepreneurship Training Program of Hunan Xiangjiang Artificial Intelligence Academy.

ear regression model to better predict the clinical score of PD using genetic and neuroimaging characteristics, with a higher correlation than traditional methods. Based on the diffusion tensor imaging, Won *et al.*[5] calculated the association between imaging and genetics through connectivity analysis, and established the model to describe the clinical score of the degree of depression in PD patients. Wang *et al.*[6] proposed a novel automatic learning model of time structure based on imaging genetic data, and used this model to automatically discover the relationship between longitudinal genotype and phenotype. Therefore, it has become an epidemic trend to combine genetic data and imaging data to study the pathological mechanism of PD for improving early detection and clinical decision-making.

However, the high dimensionality, group structure and mixed type of genetic and imaging data inevitably bring challenges to their fusion methods[7–9]. In previous studies, researchers usually used traditional dimension reduction methods, like independent component analysis[10] and principal component analysis[11] to reduce the dimension of multimodal data and extract the potential correlation of different modal data. Although these traditional methods are uncomplicated, it is difficult to analyze and interpret the important fusion features. In the latest studies, some improved methods have been put forward. Peng *et al.*[12] proposed a method of attribute reduction by combining autoencoder with deep neural network. A novel maximum ratio method developed by Mohammed *et al.*[13] also showed efficient performance.

Additionally, another challenge for multimodal fusion analysis of genetic and imaging data is the design of overall framework including the building of fusion features, the choosing of fusion features and the categorization of samples. At present, the majority of multimodal fusion studies concentrate on one certain aspect, lacking comprehensive researches on the overall framework. On the other hand, the lack of multimodal public databases and the small sample size further increase the difficulty of framework design. For example, Rana *et al.*[14] emphasized the application of 3D local binary pattern to construct fusion features with more recognizable ability. Gupta *et al.*[15] suggested using the optimized cuttlefish algorithm for feature selection to improve the diagnosis of PD. Zeng *et al.*[16] designed a new deterministic learning technology to recognize PD patients. If the building of fusion features, the choosing of fusion features and the categorization of samples can be integrated into the overall framework, it will be more beneficial for the early comprehensive diagnosis of PD.

Based on these issues and challenges, we perform multimodal fusion analysis of PD based on genes and functional magnetic resonance imaging (fMRI) data in this paper. The classical correlation analysis is used for the detection of the interaction between brain regions and genes, which is the multimodal fusion feature of the sample. In order to meet the challenge of high-dimensionality, a new optimized random forest is proposed in this paper. The idea of weighting each base classifier and each selected feature is introduced to reduce the negative impact of inefficient decision trees and delete redundant features. This method enhances the feature learning ability under the condition of small samples. Then, our research integrates the building of multimodal fusion features, choosing of features and categorization of samples into a framework to realize the comprehensive and all-round analysis of PD. Fig.1 shows the multimodal data analysis framework of PD. The overview of the framework is as follows:

- construction of multimodal fusion features;
- establishment of weighted random forest (WRF) model and categorization of samples;
- screening of the most discriminating features;
- recognition of risk genes and abnormal brain regions.

Eventually, the proposed method is evaluated by real multimodal data. Even under the condition of small samples, the proposed method still has robust classification performance. Furthermore, the proposed method is extended to multimodal data researches of other brain diseases, and also achieves excellent classification ability, which verifies the scalability and stability of the proposed method. In general, the framework is helpful to explore the lesions of nervous system diseases and provide reference for the diagnosis of PD.

2 Materials and Methods

2.1 Dataset Construction

Everyone is born with the "internal cause" of certain diseases, that is, disease susceptibility genes. One disease is a typical abnormal phenotype, and many diseases have the genotype corresponding to their abnormal phenotypes. Moreover, the fMRI technology can analyze the abnormal brain function and morphology of patients with diseases from the aspects of the brain structure, connection and circuit. Therefore, for fMRI and gene data, which are so closely related to diseases,

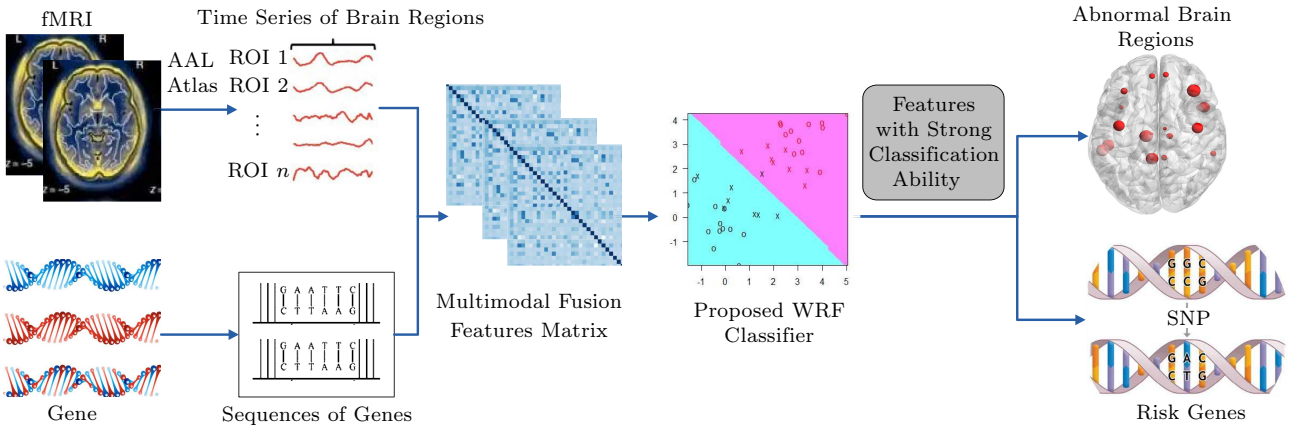

Fig.1. Overview of the proposed WRF framework.

it is of great significance to bridge the gap between them and capture the potential connections between brain regions and genetic variations. This will be helpful for researchers to extract meaningful biomarkers and improve clinical diagnosis process [17]. In biology, genes can affect the function and structure of brain through gene expressions. In the traditional research paradigm, this association can be verified by clinical experiments. In previous studies, the interaction between genes and brain regions has been detected by correlation analysis method for early diagnosis of brain diseases, and some satisfactory results have been obtained [18–20]. We are committed to using simpler and more practical methods to extract the interactions between brain regions and genes as multimodal fusion features, which is a goal of this paper.

We collect the experimental data from the Parkinson's Progression Markers Initiative (PPMI) database①, which contains fMRI data and corresponding single-nucleotide polymorphism (SNP) data. PPMI is designed to establish a comprehensive set of clinical, imaging and biosample data that will be used to define biomarkers of PD progression. Our study collects 55 PD patients and 49 healthy controls (HC), and each subject contains the fMRI data and the SNP data. The homogeneities of PD patients and HC at age and gender are confirmed by two-sample t-test and chi-square test and all subjects have signed the informed consent.

Subsequently, the fMRI data and the SNP data are preprocessed. The data processing assistant for resting-states fMRI (DPARSF) within MATLAB is applied to preprocess the fMRI data. Detailed steps include the realignment of head movement and time slice, image registration based on EPI template, image smoothing (full width at half maximum = 6 mm) and signal filtering (0.01 HZ–0.08 HZ). PLINK software② is applied to preprocess the sample gene data. The following are the settings of the specific parameters. The threshold value of sample recall rate is set at 95% for assessing the gene data's total quality. To remove the inferior SNP, the threshold values of genotyping and minimum allele frequency are set to 99.9% and 4% respectively, and we set the p-value threshold of Hardy-Weinberg equilibrium testing to 1e−4. As a result, the standardized brain images and 23 595 SNPs for each subject are obtained.

On the basis of anatomical automatic labeling (AAL) template, the preprocessed fMRI data is split into 90 brain regions, and the first 80 time points of each brain region are extracted to obtain the time series of brain regions. The preprocessed SNPs are divided into 45 groups according to the genes they belong to. The first 40 SNPs from each group are selected to represent each gene. For the retained SNP groups, four types of base A, T, C, and G in SNP are replaced by digits 1, 2, 3, and 4 respectively, and then the digital sequences of genes are obtained. In addition, the time series of brain regions and the digital sequences of genes are equal in length. Pearson correlation analysis is utilized to calculate the correlation coefficients between digital sequences and time series as the input fusion features of WRF, and 4 050 fusion features are obtained to characterize the interactions between brain regions and genes. The Pearson correlation coefficients

①PPMI database. http://www.ppmi-info.org/, Jan. 2021.

②PLINK. http://www.cog-genomics.org/plink2, Jan. 2021.

are calculated with the following equation.

$$Pea_{t,s} = \frac{l\sum B_t G_s - \sum B_t \sum G_s}{\sqrt{l\sum B_t^2 - (\sum B_t)^2}\sqrt{l\sum G_s^2 - (\sum G_s)^2}},$$

where B_t represents the functional time series of a brain region t, G_s is the digital sequence of a gene s, and l represents the length of each brain region or gene.

2.2 Weighted Random Forest

It is well known that identifying disease-related biological features in high-dimensional data is a challenging problem. Ensemble learning has unique advantages of dealing with high-dimensional feature spaces, small samples and complex data structures [21]. In this essay, we propose an original integrated learning algorithm named weighted random forest. The training process is shown in Fig.2.

To construct a base classifier, the experimental dataset $D = \{S, F\}_{n=1}^{N}$ is separated into training set $D_{\text{train}} = \{S_{\text{train}}, F_{\text{train}}\}$, validation set $D_{\text{validate}} = \{S_{\text{validate}}, F_{\text{validate}}\}$ and test set $D_{\text{test}} = \{S_{\text{test}}, F_{\text{test}}\}$ by the non-return sampling method at a ratio of 7 : 6 : 8 for each time. In specific, $S_{\text{train}} = \{s_1, s_2, \cdots, s_n\}$, where s_n indicates the n-th sample, and $F_{\text{train}} = \{+1, -1\}$ is the classification label of the sample, where "-1" denotes the PD patient and "$+1$" denotes the normal person. The training set D_{train} is used to train the decision tree to get different base learners. The validation set D_{validate} is used to get the weight of the decision tree. The test set D_{test} is used to assess the generalization ability of WRF. The procedure above is repeated certain times to form enough base classifiers. Noting that different decision trees are formed separately, and the samples are divided randomly, D_{train}, D_{validate} and D_{test} of each base classifier are different.

According to the sampling strategy above, we obtain multiple training sets and corresponding validation sets. The input features of single decision tree are also randomly selected. We assume that the total feature dimension of each sample is d. The s features are stochastically extracted as the input features according

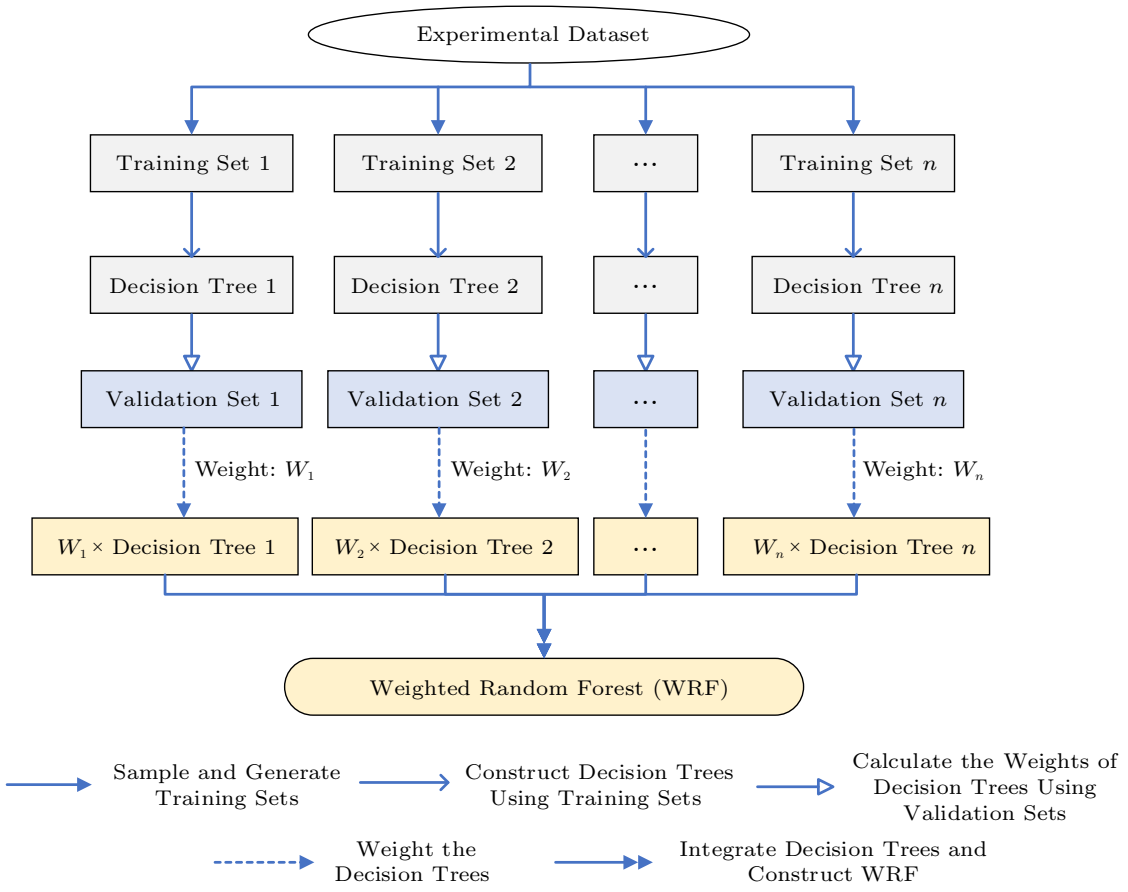

Fig.2. Training process of WRF.

to the prior knowledge, and the formula is as follows.

$$s = fix(\sqrt{d}),$$

where $fix(*)$ denotes an integral function to zero. After that, we build a decision tree by using the Gini index to look for different best categorization points of all features. The Gini index is characterized as follows.

$$GINI(D_{\text{train}}) = 1 - \sum P_a^2,$$

where P_a stands for the chance of which the categorization consequence is a. In particular, TF^j is applied for representing the j-th feature in each sample. When the feature TF^j has the value of m, the calculation formula of Gini index shows as follows.

$$\begin{aligned} &Gain_{GINI_{TF^j,m}(D_{\text{train}})} \\ &= \frac{m_1}{M} GINI(D^1_{\text{train}}) + \frac{m_2}{M} GINI(D^2_{\text{train}}), \end{aligned}$$

where M stands for the sample amount of the training set D_{train}, D^1_{train} and D^2_{train} are obtained by dividing D_{train} into two parts according to value m of feature TF^j, and m_1 and m_2 are the sample amount in the sample subsets D^1_{train} and D^2_{train} respectively. Then, the Gini index of each value of feature TF^j is calculated, and the corresponding value of the minimal Gini index is selected as the best categorization point. Moreover, the best binary categorization point of each feature is calculated by the two equations above. Therefore, one decision tree is built. By repeating the above construction step of one decision tree K times, K diverse decision trees are acquired.

Subsequently, the classification accuracies of each decision tree are acquired by means of the validation set corresponding to the training set, and the decision trees are weighted according to the corresponding classification accuracy. The weight calculation formula is characterized as

$$W_l = \frac{y_{\text{correct},l}}{Y}, \quad l = 1, 2, ..., L,$$

where $y_{\text{correct},l}$ represents the quantity of samples properly categorized by the l-th decision tree in the validation set, and Y represents the sample amount in the validation set. The classification accuracy is taken as the weight of the corresponding decision tree. It is noteworthy that the base learner does not need to be retrained in the process of weighting. Finally, these weighted decision trees are combined to construct WRF. The detailed procedure of obtaining WRF is summarized in Algorithm 1.

Algorithm 1. Framework of WRF

Input: experimental dataset $\{S, F\}$
Output: the weighted random forest
1: Initialize $\{S, F\}$, n
2: $\{S, F\}$ is experimental dataset
3: n is the number of initial decision trees
4: Partitioned $\{S, F\}$ into $\{S, F\}_{\text{tra\_1}}$, $\{S, F\}_{\text{val\_1}}$, $\{S, F\}_{\text{test\_1}}$, ..., $\{S, F\}_{\text{tra\_}n}$, $\{S, F\}_{\text{val\_}n}$, $\{S, F\}_{\text{test\_}n}$
5: **for** $k = 1$ to n
6: Select $\{S, F\}_{\text{tra\_}k}$
7: Randomly select a subset of features as $\{Features\}_{\text{tra\_}k}$
8: $\{S, F\}_{\text{tra\_}k}$ and $\{Features\}_{\text{tra\_}k} \rightarrow$ decision tree $\{Tb_k\}$
9: $\{S, F\}_{\text{val\_}k} \rightarrow$ test the classification accuracy of decision tree $\{Tb_k\}$ as weight w_k
10: weighted decision tree $\{Tb_k\} = w_k \times$ decision tree $\{Tb_k\}$
11: **end for**
12: $WRF =$ ensemble of weighted decision trees $\{Tb_1, \cdots, Tb_n\}$

2.3 Sample Classification

WRF can be applied to predict the class labels of new samples. Firstly, the unclassified samples are inputted into WRF, and the classification results of each base classifier are assembled with different weights of corresponding base classifiers, which is equal to let the base classifiers vote and decide the result. The weighted vote of category a is recorded as S_a and defined as follows.

$$S_a = \sum_{i=1}^{n} I(f_i(x) = a) \times W_i,$$

where x represents an unclassified sample, $f_i(x)$ is the prediction result of the decision tree i, W_i is the weight of decision tree i and $I(*)$ is the indicator function. If the test sample x is predicted to belong to category a by the decision tree i, the value of $I(f_i(x) = a)$ is 1; otherwise the value is 0.

For unclassified samples, category A with the most votes is selected as the final category. The calculation formula is as follows.

$$A = \arg\max(S_a).$$

To assess the overall classification performance of WRF, the classification accuracy is taken as the evaluation criterion. The calculation formula is:

$$Pre = \frac{t_{\text{true}}}{T},$$

where t_{true} represents the quantity of samples accurately classified in the test set and T represents the size of the test set.

2.4 Extraction of Pathogenic Brain Regions and Genes

In addition to the classification of samples mentioned above, WRF can also identify abnormal fusion features associated with PD, and extract risk genes and abnormal brain regions, which is another purpose of this study. The following is the specific analysis process.

The statistical weights of fusion features selected by decision trees with weights greater than 0.5 are calculated, and the features with larger weights are taken as important fusion features. The formula for calculating the weight of fusion features is characterized as:

$$BGP_v = \sum_{i=1}^{m} BGP_{i,v} \times W_i,$$

where BGP_v represents the weight of the v-th fusion feature, $BGP_{i,v}$ represents the frequency of the v-th fusion feature in decision tree i, and W_i is the weight of decision tree i. In another word, the weights of selected features are based on the weight of the base learner, and the optimal features are found by calculating the total weights of different features in multiple base learners. The weight can be employed to measure the effect of different fusion features on WRF classification performance. The greater the weight of the feature is, the more significant the effect on WRF is. This also means that these features are more distinct between normal people and PD patients, and more related to PD.

Then, we select top D "high-weight fusion features" and extract different subsets from them. Within the interval $[C, D]$, starting from C, the number of fusion features in the subset is gradually increased according to the frequency descending with b as the step size until the subset contains all D high-frequency features. Then, these feature subsets are inputted into WRF respectively to test their categorization abilities. We retain the feature subset that has optimal categorization ability. At the same time, the number of fusion features is the best. At last, we calculate the frequencies of genes and brain regions in the optimal fusion features, and the high-frequency genes and brain regions are risk genes and abnormal brain regions.

2.5 Parameter Optimization

There is one important free parameter, which needs to be optimized to achieve the best performance of WRF, that is, the initial quantity of decision trees. The optimization method is as follows: within the interval $[A, B]$, the grid search strategy is adopted. Starting from A, the initial quantity of decision trees increases gradually with a step of a, and the upper limit is B. Then, according to the accuracies of WRF with different quantities of decision trees, the quantity corresponding to the highest accuracy is the optimal initial quantity of decision trees.

3 Results

3.1 Construction of Optimal Weighted RandomForest

According to Section 2, 105 experimental data are randomly divided into a training set with 35 samples, a validation set with 30 samples and a test set with 40 samples in accordance with the proportion of $7:6:8$. The 35 samples of the training set and 57 features randomly selected from 4 050 fusion features are employed to construct a decision tree Tr_1. The number of features is obtained through many practical experiments, and it can play the advantages of WRF more effectively. After the construction, Tr_1 is verified by the corresponding validation set to get the classification accuracy, which is the weight of Tr_1. By repeating the above steps 600 times, the WRF with 600 decision trees which have their corresponding weights is formed.

To get the optimal WRF, we adopt the optimization approach mentioned in the parameter optimization section to further optimize WRF. The initial quantity of the decision trees is first confirmed within the range of [20, 600] and the optimal value of the parameter is found. To be specific, the quantity of decision trees is gradually increased from 20 to 600 in steps of 10, and then we get the classification accuracies under different quantities of decision trees. The relationship between the quantity of decision trees and the classification accuracy is shown in Fig.3. It is obvious that when the number of base classifiers is 410, the classification accuracy of ensemble learners reaches the highest value of 0.875. [20, 600] is not the final search range for the parameter. In fact, we search the optimal parameter in a wider range, and then determine the optimal value of the parameter in interval [20, 600]. WRF corresponding to the highest classification accuracy is regarded as the

optimal model, resulting in the optimal WRF model with 410 decision trees.

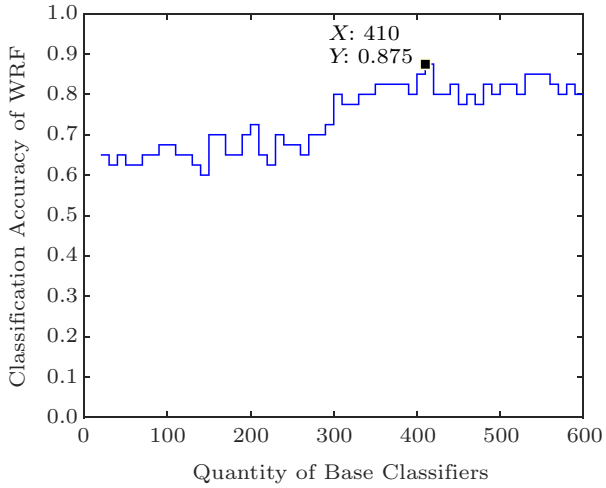

Fig.3. Relationship between the quantity of decision trees and the classification accuracy.

3.2 Optimal Fusion Features Extraction

According to the optimal WRF model, every base classifier has corresponding weight. Once the weight of a base classifier is obtained, all features in this base classifier are weighted by the same value. We sum up the weight of each feature in the base classifiers whose weights are greater than 0.5, and the weight of each feature is obtained. By sorting the features in descending order according to corresponding weights, 400 features are selected as important fusion features.

In order to ensure the performance of the ensemble learner, we regard the first 70 important fusion features as an input feature subset of WRF to classify PD patients and HC, where 57 features are randomly selected to build the base classifier, and thereby build WRF. Subsequently, in the step size of 2, we generally increase the number of important fusion features from 70 to 400. In order to ensure the performance of the ensemble learner, we choose at least the first 70 important fusion features. Fig.4 exhibits the classification accuracy of WRF with different input feature subsets. When the first 320 features are extracted, the classification accuracy of the WRF model is the highest and reaches 87.5%. Consequently, the first 320 important fusion features are the optimal fusion features.

3.3 Identification of Disease-Related Brain Regions and Genes

As mentioned in Section 2, we severally calculate the frequencies of the genes and brain regions in the optimal fusion features. Higher frequencies of genes or brain regions indicate that corresponding genes or brain regions are more abnormal. Brain regions with higher frequencies are showed in Fig.5(a), where the lowest frequency is 4. We take the frequency of the brain region as weight, and the size of the weight is graphically expressed as the size of the point in Fig.5(b). Similarly, based on the calculated frequencies, we find out the corresponding frequency for each gene, and the genes with higher frequencies are in connection with PD (Fig.6).

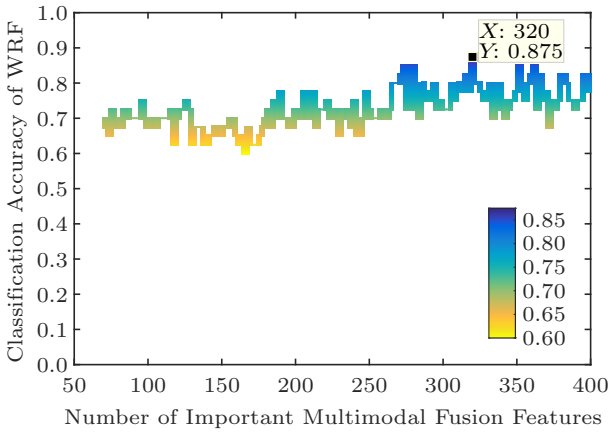

Fig.4. Relationship between the number of multimodal fusion features and the classification accuracy.

These high-frequency brain regions and genes, which include inferior frontal gyrus, opercular part (IFGoperc.R), Thalamus (THA.L), and posterior cingulate gyrus (PCG.L, C6orf10, GABBR1 and HLA-DQB2), are regarded as disease-related brain regions and genes. Our conclusions are consistent with many previous studies. For example, IFGoperc.R has the highest frequency among the found morbific brain regions. Chen *et al.*[22] studied the gray matter atrophy of PD-mild cognitive impairment (PD-MCI) and the result of one-way analysis of variance indicated the significant difference in the anatomical location of IFGoperc.R. Guimarães *et al.*[23] also detected the grey matter loss of IFGoperc.R for moderate PD and severe PD patients compared with HC through the voxel-based morphometry. Similarly, Hou *et al.*[24] combined fMRI and the graph theory approach to research the topological structure of PD and the experimental results showed that abnormal nodal centralities of IFGoperc.R were found out in PD compared with HC. IFGoperc.R was proved to be related to memory[25], language[26] and motion[27], abnormalities of which could lead to the occurrence of PD. PCG.L is also significantly abnormal brain region. Reijnders *et al.*[28] carried out a morpho-

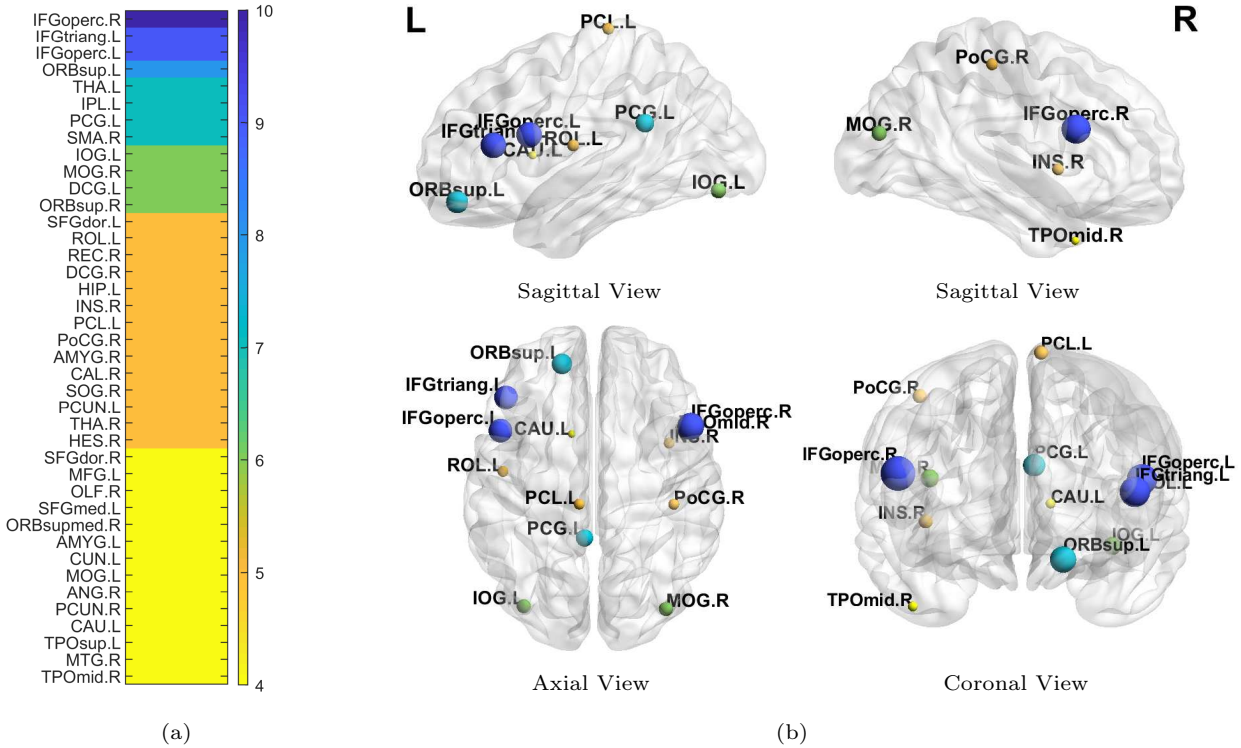

Fig.5. (a) Frequencies, and (b) locations and sizes of abnormal brain regions. The color of the point in (b) corresponds to the color bar in (a).

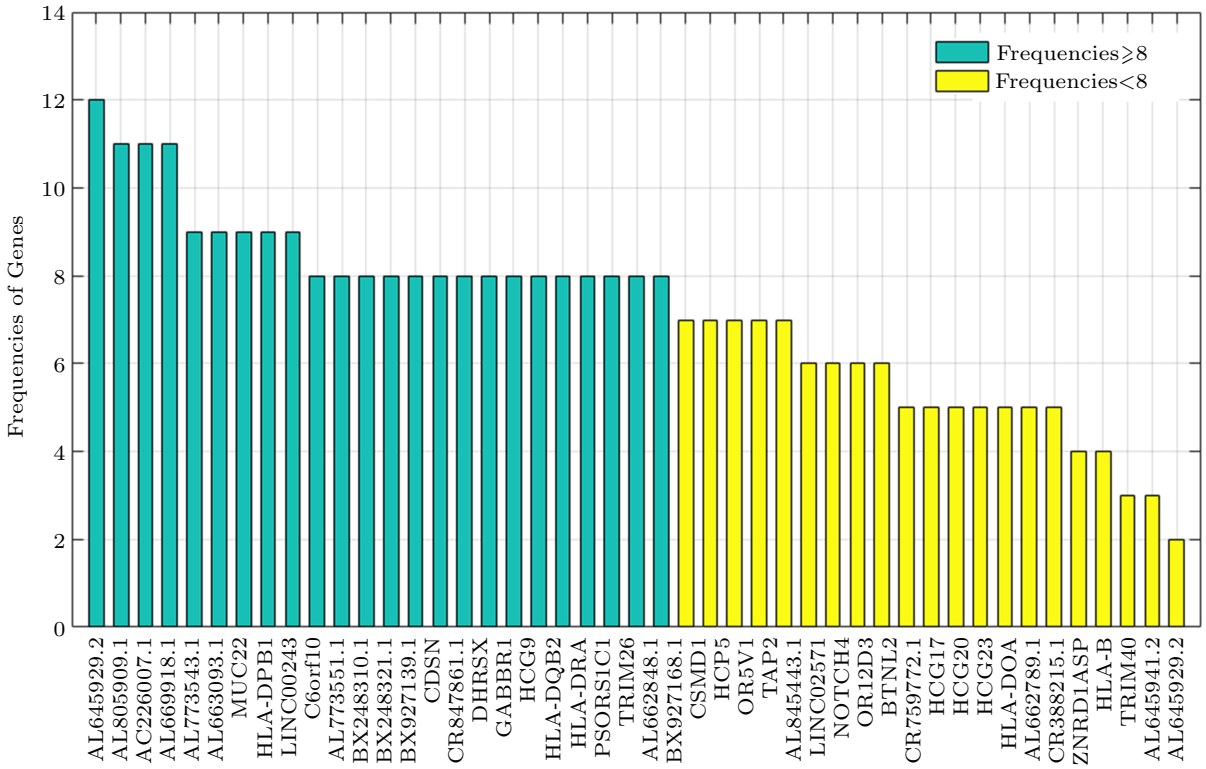

Fig.6. Frequencies of genes. The total number of candidate genes examined in this study is 45, and the high-frequency genes are more likely to be risk genes for PD.

logical MRI study, which proved that PD patients were associated with the gray matter density value of posterior cingulate gyrus. The voxel-based morphometry and multivariate linear regression were used by Melzer *et al.*[29] for studying the connection between cognitive and grey matter concentration of PD and the results exhibited that PD with dementia had reduced grey matter volume in the brain region of PCG.L. Moreover, de Schipper *et al.*[30] and Evangelisti *et al.*[31] found out the increased functional connectivity of PCG.L in PD patients. Therefore, the findings of abnormal brain regions provide evidence for diagnosis of PD.

Additionally, some risk genes are found out in our study, which is in line with existing findings. For example, the genes of C6orf10, GABBR1 and HLA-DQB2 are likely to be connected with PD, which may cause brain functional and structural changes through differential expression in the brains of PD patients[11,32–34].

3.4 Performance Comparison and Verification

In order to verify the advantages of "Pearson + WRF" framework proposed in this paper, we use Pearson correlation analysis, distance correlation (DC), canonical correlation analysis (CCA) and Kendall to build multimodal fusion features, and utilize the two-sample t-test, decision tree, random forest (RF) and WRF to abstract optimal fusion features. The parameter settings of the other methods are consistent with that of the method proposed in this paper to keep the objectivity. The classification performance of the optimal multimodal features abstracted by diverse frameworks is evaluated by support vector machine (SVM). The results are shown in Table 1.

Firstly, we are able to observe that "Pearson + WRF" abstracts the least quantity of optimal fusion features among all frameworks in Table 1, but has the highest classification accuracy. At the same time, the features extracted by other frameworks always partly overlap with those extracted by "Pearson + WRF" framework. The non-randomness of these overlaps is verified by hypergeometric test. It is noteworthy that the more the overlaps between "Pearson + WRF" and other frameworks, the higher their classification accuracy, indicating that the optimal features extracted by "Pearson + WRF" are rational. On the basis of the analysis above, we find that the "Pearson + WRF" framework extracts the fewest fusion features and is the most dependable of all frameworks since it can effectively refrain from the false positive. Consequently, the optimal fusion features abstracted by the "Pearson + WRF" have the best reasonability. Moreover, in order to observe the comparison results more intuitively, we adopt two evaluation indexes, namely true positive rate (TPR) and false positive rate (FPR) (Fig.7). The "Pearson + WRF" framework proposed in this paper has area under curve (AUC) value of 0.875, ranking in the top and indicating the best performance among all frameworks.

Secondly, in the preprocessing step, we intercept

Table 1. Comparison Results of Comprehensive Frameworks

Framework	Number of Optimal Multimodal Fusion Features	Classification Accuracy of SVM	Overlaps with Our Method
Pearson + WRF	320	0.850	–
Pearson + two-sample t-test	499	0.700	158 ($p = 2.158\,422e{-}20$)
CCA + two-sample t-test	412	0.625	125 ($p = 4.841\,794e{-}19$)
DC + two-sample t-test	447	0.625	120 ($p = 6.862\,647e{-}28$)
Kendall + two-sample t-test	518	0.625	121 ($p = 1.312\,745e{-}42$)
Pearson + decision tree	700	0.725	160 ($p = 7.600\,548e{-}62$)
CCA + decision tree	590	0.675	148 ($p = 4.006\,157e{-}43$)
DC + decision tree	370	0.650	127 ($p = 1.215\,09e{-}11$)
Kendall + decision tree	470	0.650	138 ($p = 9.618\,019e{-}24$)
Pearson + RF	670	0.775	184 ($p = 5.371\,748e{-}41$)
CCA + RF	565	0.625	112 ($p = 5.439\,128e{-}60$)
DC + RF	445	0.650	134 ($p = 5.370\,451e{-}21$)
Kendall + RF	495	0.675	145 ($p = 3.330\,824e{-}25$)
CCA + WRF	560	0.775	187 ($p = 2.901\,748e{-}19$)
DC + WRF	485	0.750	176 ($p = 4.865\,552e{-}12$)
Kendall + WRF	560	0.800	215 ($p = 1.482\,712e{-}10$)

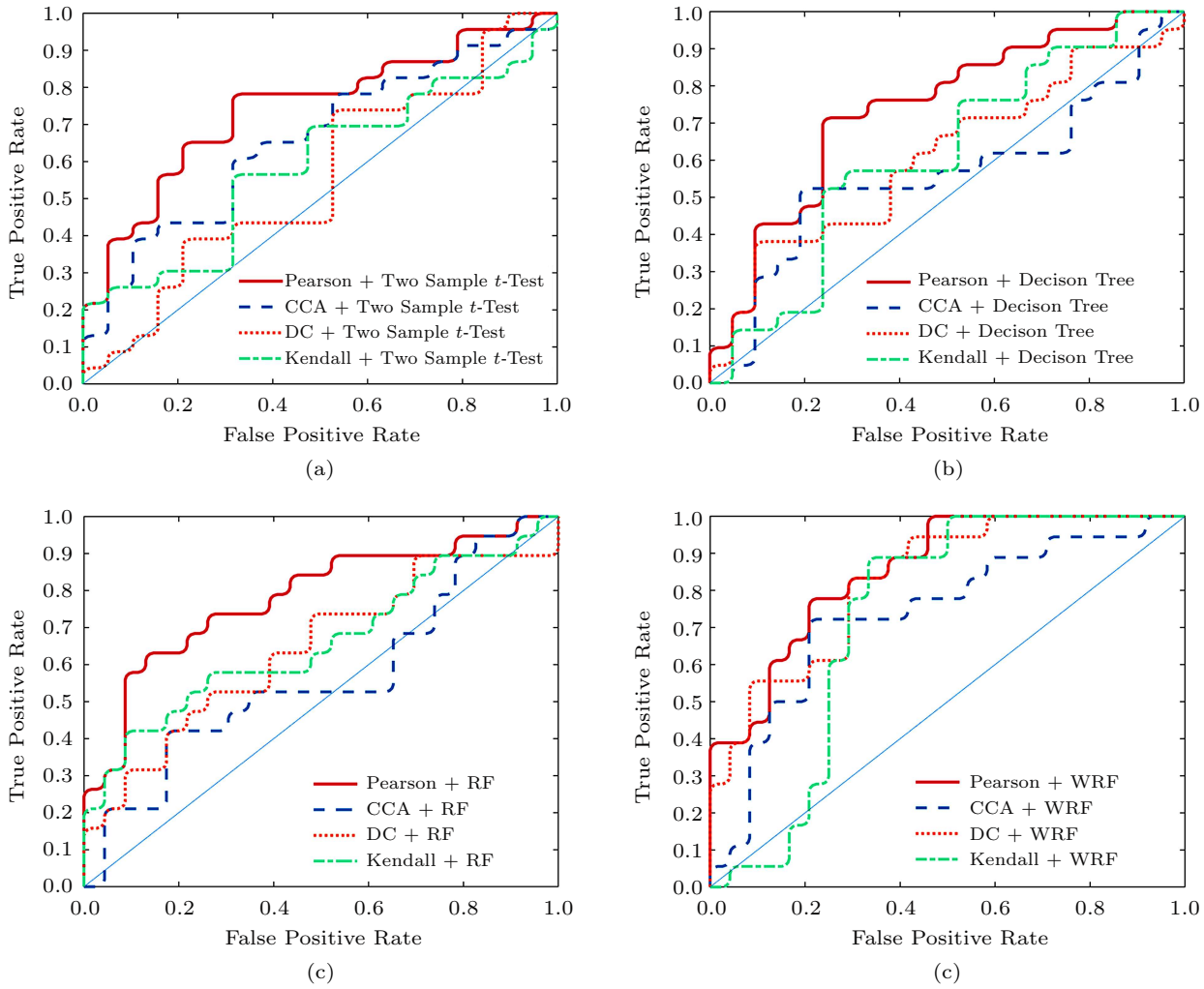

Fig.7. Comparison of TPR and FPR under different frameworks. (a) Two-sample t-test with Pearson, CCA, DC and Kendall respectively. (b) Decision tree with Pearson, CCA, DC and Kendall respectively. (c) RF with Pearson, CCA, DC and Kendall respectively. (d) WRF with Pearson, CCA, DC and Kendall respectively.

the time series of brain regions and digital sequences of genes to reach the same length of 80, and then calculate the correlation coefficients between them by Pearson correlation analysis to construct fusion features. In order to verify that the interception length of 80 is optimal, that is, l in calculation equation of Pearson correlation coefficients, we take different values of l respectively to build the fusion features. The rest of the steps to build the WRF model are the same to prevent the interference of other factors. Subsequently, we use the constructed model in the classification of the PD & HC dataset. As can be seen in Fig.8(a), when l is 80, "Pearson + WRF" shows an exclusive advantage and performs much better than the frameworks with other interception lengths.

Finally, when we extend the "Pearson + WRF" framework to multimodal data fusion researches of Alzheimer's disease (AD) & HC and early MCI (EMCI) & HC datasets, it also performs quite well in that it provides the relatively high AUC values (see Fig.8(b) and Fig.8(c)). This fully verifies the robustness and generalization ability of the proposed model. Besides, it should be noted that we get the optimal interception length through repeated experiments; therefore its values are different in the other two datasets (see Fig.8(b) and Fig.8(c)).

4 Discussion

Besides the comparison of the above methods, we summarize the current popular diagnostic methods for PD [35]. For example, Sivaranjini and Sujatha [36] tried to use deep learning neural network to classify magnetic resonance images of healthy control group and PD pa-

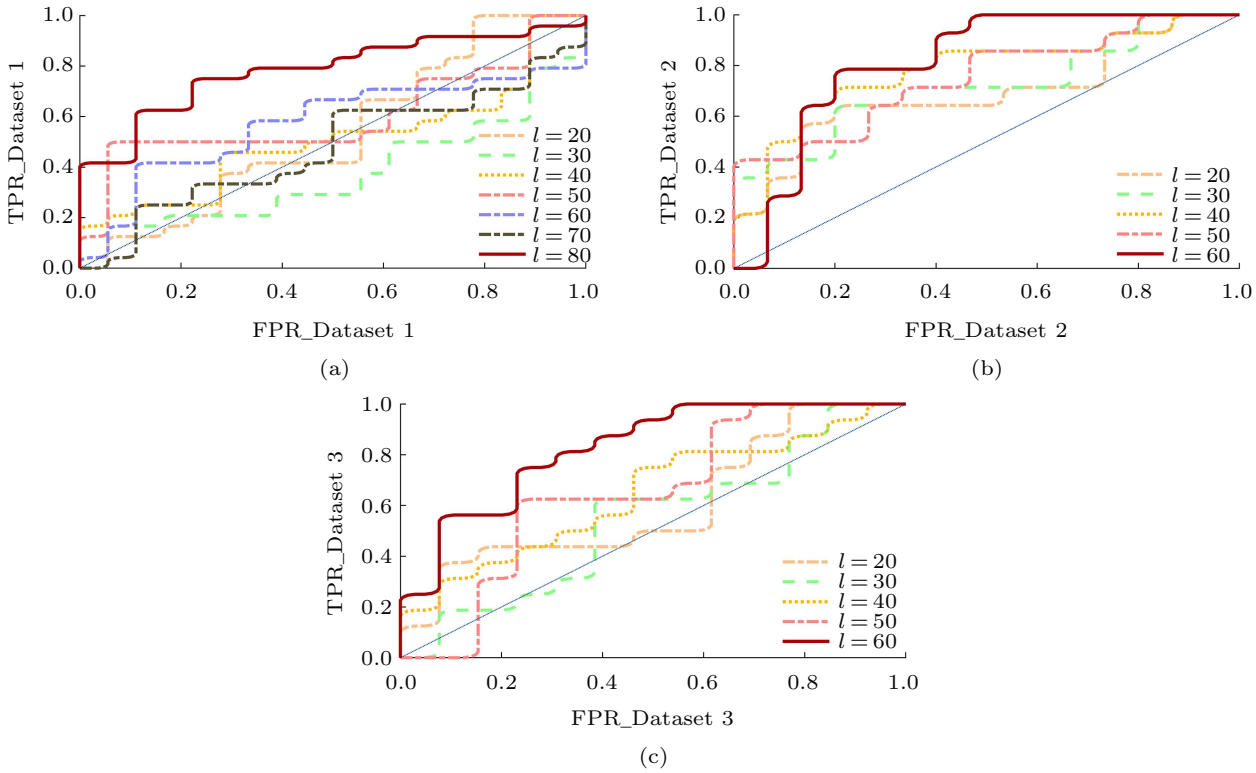

Fig.8. (a) ROC curve for dataset 1 (55 PD + 50 HC). (b) ROC curve for dataset 2 (37 AD + 36 HC). (c) ROC curve for dataset 3 (37 EMCI + 36 HC). Both dataset 2 and dataset 3 are acquired from Alzheimer's Disease Neuroim-aging Initiative.

tients, with the accuracy of 88.9%. Martinez-Murcia *et al.*[37] proposed a deep convolution automatic encoder (CAE) architecture, which was tested on neuroimaging dataset of PD, and achieved an accuracy of more than 90%. Obviously, the accuracy of the PD diagnosis using deep learning was not much different from that of the proposed WRF. But WRF can not only classify PD well, but also extract disease-related factors, which is the unique advantage compared with deep learning. In addition, some machine learning methods are also extremely popular in researches for PD diagnosis. For example, Gao *et al.*[38] proved that the model-free machine learning analytical methods could provide a more reliable classification accuracy (70%–80%) for falls in PD patients, compared with the model-based analytical methods. Abós *et al.*[39] used the randomized logistic regression and SVM to carry out feature selection and distinguish PD patients with MCI from those without MCI, and the mean accuracy achieved 82.6%. Generally, our method shows relatively high comprehensive performance among the current popular PD diagnostic technologies, and its good robustness and scalability are conducive to the development of clinical medicine.

In this paper, the Weighted Random Forest method is proposed to classify PD patients and HC by fusing the complementary information from different modal data, and the classification accuracy is 87.5%. There are mainly two reasons for the good effect. The first reason is that we find out the optimal quantities of base classifiers and multimodal fusion features to achieve the best performance of WRF. When WRF based on the optimal parameters reaches the peak performance, the time complexity and the spatial complexity are more balanced, and the model's classification effect is better. The second reason is that we construct an overall framework of multimodal data fusion, sample classification and feature extraction, which can make up for the defects of most previous researches focusing on one aspect. Additionally, a framework can make better use of the information complementarity of multimodal data to improve classification performance. Furthermore, our data are obtained from the PPMI database. The collection and the processing of the data in this database have a strict standard, which ensures that the data are homologous in structure.

Although our method improves the diagnostic effect of PD, there are still some potential limitations in this study. On the one hand, the study utilizes the AAL atlas, which is a popular and accepted method in this field, to divide the brain regions, and the cor-

relations between brain regions and genes are used as features, which lead to the result that complex brain compartmentalization is not detailed enough. Therefore, we can use other brain templates to match images in the next work, such as Broadman. On the other hand, this study mainly fuses genetic data and fMRI data to explore brain diseases. There may be a combination of data fusion better than the fusion of these two datasets [40, 41], which is also a focus of our follow-up work. We have confirmed the rationality of most typical pathogenic factors through many existing studies, which shows the effectiveness of the method. Still, there are few atypical pathogenic factors lack of relevant research. In the follow-up work, we will collect more data, design new algorithms for in-depth analysis, and better explain their role in the pathogenesis of PD. Furthermore, these factors provide a reference for further exploration of PD. We plan to cooperate with clinicians to jointly explain the role and rationality of these factors.

5 Conclusions

In this study, the resting-state fMRI data and the genetic data were used to accomplish the multimodal fusion, and WRF was introduced to accurately distinguish PD patients and HC. The main contributions of this paper are as follows. Firstly, we applied practical correlation analysis to construct multimodal fusion features, which showed better identification ability than classical single modal features. Secondly, WRF was proposed and innovatively used with sample classification and feature filtering. Finally, we detected the lesion brain regions and risk genes from the select optimal multimodal fusion features. Our efforts are conducive to understanding the pathogenesis of PD, and introduce a valuable perspective for the diagnosis of brain diseases.

References

[1] Arkinson C, Walden H. Parkin function in Parkinson's disease. *Science*, 2018, 360(6386): 267-268. DOI: 10.1126/science.aar6606.

[2] Lv D J, Li L X, Chen J, Wei S Z, Wang F, Hu H, Xie A M, Liu C F. Sleep deprivation caused a memory defects and emotional changes in a rotenone-based zebrafish model of Parkinson's disease. *Behavioural Brain Research*, 2019, 372: Article No. 112031. DOI: 10.1016/j.bbr.2019.112031.

[3] Koros C, Simitsi A, Stefanis L. Genetics of Parkinson's disease: Genotype-phenotype correlations. *International Review of Neurobiology*, 2017, 132: 197-231. DOI: 10.1016/bs.irn.2017.01.009.

[4] Kim M, Kim J, Lee S H, Park H. Imaging genetics approach to Parkinson's disease and its correlation with clinical score. *Scientific Reports*, 2017, 7: Article No. 46700. DOI: 10.1038/srep46700.

[5] Won J H, Kim M, Park B Y, Youn J, Park H. Effectiveness of imaging genetics analysis to explain degree of depression in Parkinson's disease. *PLoS ONE*, 2019, 14(2): Article No. e0211699. DOI: 10.1371/journal.pone.0211699.

[6] Wang X, Yan J, Yao X *et al.* Longitudinal genotype-phenotype association study through temporal structure auto-learning predictive model. *Journal of Computational Biology*, 2018, 25(7): 809-824. DOI: 10.1089/cmb.2018.0008.

[7] Hao X, Li C, Yan J, Yao X, Risacher S L, Saykin A J, Shen L, Zhang D, Alzheimer's Disease Neuroimaging Initiative. Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. *Bioinformatics*, 2017, 33(14): i341-i349. DOI: 10.1093/bioinformatics/btx245.

[8] Min W, Liu J, Zhang S. Edge-group sparse PCA for network-guided high dimensional data analysis. *Bioinformatics*, 2018, 34(20): 3479-3487. DOI: 10.1093/bioinformatics/bty362.

[9] Hua K, Zhang X. Estimating the total genome length of a metagenomic sample using k-mers. *BMC Genomics*, 2019, 20(2): Article No. 183. DOI: 10.1186/s12864-019-5467-x.

[10] Calhoun V D, Liu J, AdalıT. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, 2009, 45(1, Supplement 1): S163-S172. DOI: 10.1016/j.neuroimage.2008.10.057.

[11] Hamza T H, Zabetian C P, Tenesa A *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nature Genetics*, 2010, 42(9): 781-785. DOI: 10.1038/ng.642.

[12] Peng J, Guan J, Shang X. Predicting Parkinson's disease genes based on Node2vec and autoencoder. *Frontiers in Genetics*, 2019, 10: Article No. 226. DOI: 10.3389/fgene.2019.00226.

[13] Mohammed A, Zamani M, Bayford R, Demosthenous A. Toward on-demand deep brain stimulation using online Parkinson's disease prediction driven by dynamic detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2017, 25(12): 2441-2452. DOI: 10.1109/TNSRE.2017.2722986.

[14] Rana B, Juneja A, Saxena M, Gudwani S, Kumaran S S, Behari M, Agrawal R K. Relevant 3D local binary pattern based features from fused feature descriptor for differential diagnosis of Parkinson's disease using structural MRI. *Biomedical Signal Processing and Control*, 2017, 34: 134-143. DOI: 10.1016/j.bspc.2017.01.007.

[15] Gupta D, Julka A, Jain S, Aggarwal T, Khanna A, Arunkumar N, De Albuquerque V H C. Optimized cuttlefish algorithm for diagnosis of Parkinson's disease. *Cognitive Systems Research*, 2018, 52: 36-48. DOI: 10.1016/j.cogsys.2018.06.006.

[16] Zeng W, Liu F, Wang Q, Wang Y, Ma L, Zhang Y. Parkinson's disease classification using gait analysis via deterministic learning. *Neuroscience Letters*, 2016, 633: 268-278. DOI: 10.1016/j.neulet.2016.09.043.

[17] Huang Y A, Huang Z A, You Z H, Hu P, Li L P, Li Z W, Wang L. Precise prediction of pathogenic microorganisms using 16S rRNA gene sequences. In *Proc. the 15th International Conference on Intelligent Computing*, August 2019, pp.138-150. DOI: 10.1007/978-3-030-26969-2_13.

[18] Du L, Liu K, Zhang T, Yao X, Yan J, Risacher S L, Han J, Guo L, Saykin A J, Shen L, Alzheimer's Disease Neuroimaging Initiative. A novel SCCA approach via truncated ℓ1-norm and truncated group lasso for brain imaging genetics. *Bioinformatics*, 2017, 34(2): 278-285. DOI: 10.1093/bioinformatics/btx594.

[19] Du L, Liu K, Zhu L, Yao X, Risacher S L, Guo L, Saykin A J, Shen L, Alzheimer's Disease Neuroimaging Initiative. Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: A longitudinal study of the ADNI cohort. *Bioinformatics*, 2019, 35(14): i474-i483. DOI: 10.1093/bioinformatics/btz320.

[20] Du L, Liu K, Yao X, Risacher S L, Shen L. Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach. *Medical Image Analysis*, 2020, 61: Article No. 101656. DOI: 10.1016/j.media.2020.101656.

[21] Wei L, Su R, Luan S, Liao Z, Manavalan B, Zou Q, Shi X. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*, 2019, 35(23): 4930-4937. DOI: 10.1093/bioinformatics/btz408.

[22] Chen F X, Kang D Z, Chen F Y, Liu Y, Wu G, Li X, Yu L H, Lin Y X, Lin Z Y. Gray matter atrophy associated with mild cognitive impairment in Parkinson's disease. *Neuroscience Letters*, 2016, 617: 160-165. DOI: 10.1016/j.neulet.2015.12.055.

[23] Guimarães R P, Arci Santos M C, Dagher A *et al.* Pattern of reduced functional connectivity and structural abnormalities in Parkinson's disease: An exploratory study. *Frontiers in Neurology*, 2017, 7: 243. DOI: 10.3389/fneur.2016.00243.

[24] Hou Y, Wei Q, Ou R, Yang J, Song W, Gong Q, Shang H. Impaired topographic organization in cognitively unimpaired drug-naïve patients with rigidity-dominant Parkinson's disease. *Parkinsonism & Related Disorders*, 2018, 56: 52-57. DOI: 10.1016/j.parkreldis.2018.06.021.

[25] Zhao L, Wang E, Zhang X *et al.* Cortical structural connectivity alterations in primary insomnia: Insights from MRI-based morphometric correlation analysis. *BioMed Research International*, 2015, 2015: Article No. 817595. DOI: 10.1155/2015/817595.

[26] Meunier D, Stamatakis E A, Tyler L K. Age-related functional reorganization, structural changes, and preserved cognition. *Neurobiology of Aging*, 2014, 35(1): 42-54. DOI: 10.1016/j.neurobiolaging.2013.07.003.

[27] Li H F, Yang L, Yin D, Chen W J, Liu G L, Ni W, Wang N, Yu W, Wu Z Y, Wang Z. Associations between neuroanatomical abnormality and motor symptoms in paroxysmal kinesigenic dyskinesia. *Parkinsonism & Related Disorders*, 2019, 62: 134-140. DOI: 10.1016/j.parkreldis.2018.12.029.

[28] Reijnders J S A M, Scholtissen B, Weber W E J, Aalten P, Verhey F R J, Leentjens A F G. Neuroanatomical correlates of apathy in Parkinson's disease: A magnetic resonance imaging study using voxel-based morphometry. *Movement Disorders*, 2010, 25(14): 2318-2325. DOI: 10.1002/mds.23268.

[29] Melzer T R, Watts R, MacAskill M R, Pitcher T L, Livingston L, Keenan R J, Dalrymple-Alford J C, Anderson T J. Grey matter atrophy in cognitively impaired Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 2012, 83(2): 188-194. DOI: 10.1136/jnnp-2011-300828.

[30] De Schipper L J, Hafkemeijer A, van der Grond J, Marinus J, Henselmans J M L, van Hilten J J. Altered whole-brain and network-based functional connectivity in Parkinson's disease. *Frontiers in Neurology*, 2018, 9: Article No. 419. DOI: 10.3389/fneur.2018.00419.

[31] Evangelisti S, Pittau F, Testa C *et al.* L-dopa modulation of brain connectivity in Parkinson's disease patients: A pilot EEG-fMRI study. *Frontiers in Neuroscience*, 2019, 13: Article No. 611. DOI: 10.3389/fnins.2019.00611.

[32] Wang Q, Li W X, Dai S X, Guo Y C, Han F F, Zheng J J, Li G H, Huang J F. Meta-analysis of Parkinson's disease and Alzheimer's disease revealed commonly impaired pathways and dysregulation of NRF2-dependent genes. *Journal of Alzheimer's Disease*, 2017, 56(4): 1525-1539. DOI: 10.3233/JAD-161032.

[33] International Parkinson Disease Genomics Consortium. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: A meta-analysis of genome-wide association studies. *The Lancet*, 2011, 377(9766): 641-649. DOI: 10.1016/S0140-6736(10)62345-8.

[34] Ahmed I, Tamouza R, Delord M *et al.* Association between Parkinson's disease and the HLA-DRB1 locus. *Movement Disorders,* 2012, 27(9): 1104-1110. DOI: 10.1002/mds.25035.

[35] Bao W, Jiang Z, Huang D S. Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinformatics*, 2017, 18(16): Article No. 543. DOI: 10.1186/s12859-017-1968-2.

[36] Sivaranjini S, Sujatha C M. Deep learning based diagnosis of Parkinson's disease using convolutional neural network. *Multimedia Tools and Applications*, 2019, 79(3): 15467–15479. DOI: 10.1007/s11042-019-7469-8.

[37] Martinez-Murcia F J, Ortiz A, Gorriz J M, Ramirez J, Castillo-Barnes D, Salas-Gonzalez D, Segovia F. Deep convolutional autoencoders vs PCA in a highly-unbalanced Parkinson's disease dataset: A DaTSCAN study. In *Proc. the 13th International Conference on Soft Computing Models in Industrial and Environmental Applications*, June 2018, pp. 47-56. DOI: 10.1007/978-3-319-94120-2_5.

[38] Gao C, Sun H, Wang T *et al.* Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. *Scientific Reports*, 2018, 8(1): Article No. 7129. DOI: 10.1038/s41598-018-24783-4.

[39] Abós A, Baggio H C, Segura B, García-Díaz A I, Compta Y, Martí M J, Valldeoriola F, Junqué C. Discriminating cognitive status in Parkinson's disease through functional connectomics and machine learning. *Scientific Reports*, 2017, 7: Article No. 45347. DOI: 10.1038/srep45347.

[40] Niu Y W, Wang G H, Yan G Y, Chen X. Integrating random walk and binary regression to identify novel miRNA-disease association. *BMC Bioinformatics*, 2019, 20(1): Article No. 59. DOI: 10.1186/s12859-019-2640-9.

[41] Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics*, 2019, 35(22): 4730-4738. DOI: 10.1093/bioinformatics/btz297.

Xia-An Bi received his Ph.D. degree in computer applications from Hunan University, Changsha, in 2012. He is currently a professor in the Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, and College of Information Science and Engineering in Hunan Normal University, Changsha. His current research interests include machine learning, brain science, and artificial intelligence. He is the author or coauthor of several technical papers including several ESI Highly Cited Papers according to 2020 Clarivate Analytics ESI report, and also a very active reviewer for many international journals and conferences. He is a member of CCF and IEEE.

Zhao-Xu Xing received his B.E. degree in computer science and technology from Hunan Normal University, Changsha, in 2019. He is currently pursuing his M.S. degree in College of Information Science and Engineering, Hunan Normal University, Changsha. His research fields include data mining, brain science and artificial intelligence.

Rui-Hui Xu received his B.E. degree in mechatronic engineering from Nanyang Normal University, Nanyang, in 2019. He is currently pursuing his M.S. degree in College of Information Science and Engineering, Hunan Normal University, Changsha. His main research fields include data mining, brain science and artificial intelligence.

Xi Hu received her B.E. degree in information engineering from Hunan Institute of Science and Technology, Yueyang, in 2019. She is currently pursuing her M.S. degree in College of Information Science and Engineering, Hunan Normal University, Changsha. Her main research fields include data mining, brain science and artificial intelligence.