Brain Imaging and Behavior (2021) 15:1986–1996
https://doi.org/10.1007/s11682-020-00392-6

ORIGINAL RESEARCH

The exploration of Parkinson's disease: a multi-modal data analysis of resting functional magnetic resonance imaging and gene data

Xia-an Bi[1,2] · Hao Wu[1,2] · Yiming Xie[1,2] · Lixia Zhang[1,2] · Xun Luo[1,2] · Yu Fu[1,2] · for the Alzheimer's Disease Neuroimaging Initiative

Accepted: 31 August 2020 / Published online: 29 September 2020

Abstract

Parkinson's disease (PD) is the most universal chronic degenerative neurological dyskinesia and an important threat to elderly health. At present, the researches of PD are mainly based on single-modal data analysis, while the fusion research of multi-modal data may provide more meaningful information in the aspect of comprehending the pathogenesis of PD. In this paper, 104 samples having resting functional magnetic resonance imaging (rfMRI) and gene data are from Parkinson's Progression Markers Initiative (PPMI) and Alzheimer's Disease Neuroimaging Initiative (ADNI) database to predict pathological brain areas and risk genes related to PD. In the experiment, Pearson correlation analysis is adopted to conduct fusion analysis from the data of genes and brain areas as multi-modal sample characteristics, and the clustering evolution random forest (CERF) method is applied to detect the discriminative genes and brain areas. The experimental results indicate that compared with several existing advanced methods, the CERF method can further improve the diagnosis of PD and healthy control, and can achieve a significant effect. More importantly, we find that there are some interesting associations between brain areas and genes in PD patients. Based on these associations, we notice that PD-related brain areas include angular gyrus, thalamus, posterior cingulate gyrus and paracentral lobule, and risk genes mainly include C6orf10, HLA-DPB1 and HLA-DOA. These discoveries have a significant contribution to the early prevention and clinical treatments of PD.

Keywords Parkinson's disease · RfMRI · SNP · Multi-modal data fusion · Clustering evolution random forest

Introduction

Parkinson's disease (PD), also called as tremor paralysis, is a neurodegenerative disease commonly appeared in the elderly (Tysnes and Storstein 2017). According to statistics, PD is the most universal degenerative neurological disease next to Alzheimer's disease (De Virgilio et al. 2016). There are 5.7 million PD patients in the world, more than half of them are accompanied by cognitive disorders and other symptoms in the process of onset (Goldman et al. 2019). In general, PD patients have symptoms of bradykinesia, limb tremor, myotonia and discriminative posture (Bologna et al. 2016; Thenganatt and Jankovic 2016). Moreover, PD starts occultly and progresses slowly, which will cause irreversible damage to the brain. At present, the popular medical methods for PD are employing brain imaging technology and gene detection technology to detect brain disease areas and predict the risk of disease in advance (Rittman et al. 2016).

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI)- database (adni.loni.usc.edu) and the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information of PPMI on the study, visit www.ppmi-info.org.As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

✉ Xia-an Bi
bixiaan@hnu.edu.cn

[1] Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha, People's Republic of China

[2] College of Information Science and Engineering, Hunan Normal University, Changsha, People's Republic of China

In recent years, the researches on PD from neuroimaging, genetics and other fields have always been the focus of researchers (Nalls et al. 2015; Santos-García et al. 2016). Although some meaningful research results have been achieved, the nosogenesis of PD does not been completely comprehended. On the one hand, in neuroimaging studies, Wen et al. (2016) have found that fear in PD patients is closely related to dopaminergic density in their caudate and putamen nuclei, and they also find that the functional connectivity between limbic and prefrontal networks is reduced in PD patients, while the neural activity in prefrontal area is increased. For the purpose of studying the potential pathological mechanisms of postural instability in PD patients who fall frequently, Kaut et al. (2020) compare the changes in functional connections between falls, non-fallers, and healthy controls (HC), and find that functional connections between cerebellar structures in PD patients are enhanced. Furthermore, Martin et al. (2019) have detected that during motor planning, the dorsolateral prefrontal cortex of PD patients shows significant relative hyperactivity.

On the other hand, in genetic researches, Agliardi et al. (2019) have studied the regulation of snap25 single nucleotides on gene expression at different levels, and ultimately discovered that snap25 may have synergistic effect in the pathogenesis of PD. Similarly, it has been found that the adenosine A2A receptor gene is an important molecular target for PD therapeutic compounds, further indicating that the selective epigenetic mechanism targeting gene promoters is a tool for developing new therapies (Falconi et al. 2019). By studying the entire exome sequencing data from 1156 PD subjects and 1679 control subjects, Robak et al. (2017) have identified several promising new susceptible loci that have enhanced the importance of lysosomal mechanisms in the pathogenesis of PD, and they have also discovered that multiple gene mutations may work together to reduce lysosomal function, thus enhancing the susceptibility of PD. In addition, in the process of exploring the risk loci of PD, Reynolds et al. (2019) have found that the risk sites of PD do not exist in specific cell types or single brain area, but in the whole cell process that can be detected in multiple cell types. In summary, we have noticed that most of the researches on PD are focusing on the disease-causing brain areas or genes, rarely involving the combination of imaging and genetic data, but multi-modal data can make full use of multiple complementary information. If the correlation between gene and brain area can be fully utilized, we can study the pathological mechanism of PD more comprehensively. Therefore, adopting multi-modal data to PD study is the general trend.

However, in the face of limited available data and high-dimensional characteristics, classic statistical analysis methods such as logistic regression, factor analysis, and discriminant analysis (Akgun 2012), it is often difficult to make full use of existing data, especially in multi-modal data fusion analysis. At this point, machine learning methods, especially improved machine learning methods, always show a wider application prospect in such problems (Du et al. 2019; Huang et al. 2018; Su et al. 2019). Therefore, applying the improved machine learning method to the exploration of PD multi-modal data may be more effective in exploring the pathogenesis of PD.

In our study, we construct the correlations from genes and brain areas, and integrate it as the sample characteristics of multi-modal data firstly. Then, the clustering evolution random forest (CERF) method is adopted to analyze the correlations extracted from the data of genes and brain areas. By picking samples and sample characteristics stochastically, the random forest is established, and the clustering evolution idea and threshold filtering is employed for detecting the pathogenic brain areas and genes in PD disorder. Based on the resting functional magnetic resonance imaging (rfMRI) and gene data of 104 subjects, the CERF method is applied. The experimental results show that the CERF method can further improve the diagnosis of PD and healthy control compared with the several existing advanced methods. More meaningfully, in the experiment, we identify some pathogenic brain areas and genes for the prevention and diagnosis of PD, which are meaningful for further research of PD.

Materials and methods

Overview

In Fig. 1, we show the flow of the model we applied. The model has three important parts: (1) the fusion analysis of rfMRI and gene data, (2) the application of CERF method, (3) the detection of pathological mechanism of PD. We fuse rfMRI data and gene data at first. Then, based on CERF method, the fusion characteristics are training to obtain the optimal fusion characteristics. Finally, based on the analysis of optimal fusion characteristics, we detect discriminative genes and brain areas associated with PD disorders.

Participants

The PPMI database (http://www.ppmi-info.org/) and ADNI database (http://adni.loni.usc.edu/) are large-scale public databases, which collect a large number of positron emission computed tomography data, MRI data and single nucleotide polymorphism (SNP) data of patients with PD and its related diseases (Jones-Davis and Buckholtz 2015; Marek et al. 2018; Torigian et al. 2016). The PD progress indicator program is a landmark observational clinical study sponsored by the Michael Jefferson foundation, which aim to comprehensively

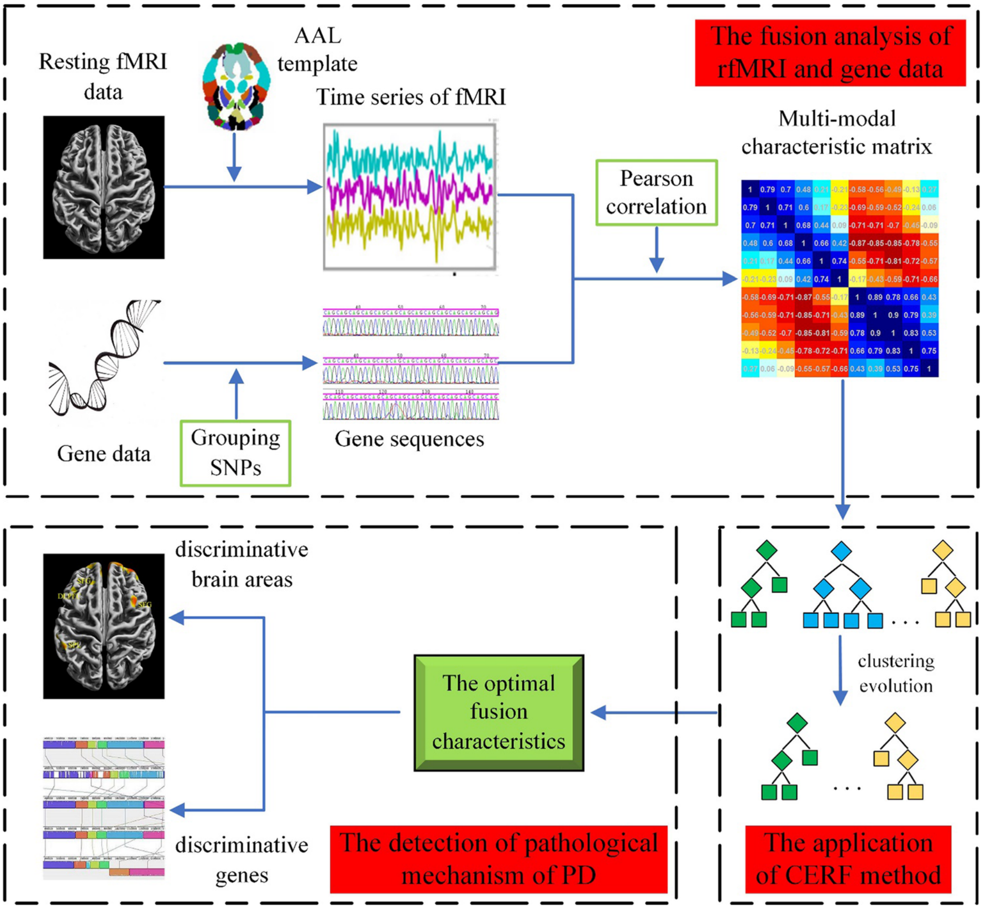
Resting fMRI data
AAL template
Time series of fMRI
The fusion analysis of rfMRI and gene data
Multi-modal characteristic matrix
Pearson correlation
Gene data
Grouping SNPs
Gene sequences
discriminative brain areas
The optimal fusion characteristics
clustering evolution
discriminative genes
The detection of pathological mechanism of PD
The application of CERF method

Fig. 1 The illustration of our model

evaluate important research objects through advanced imaging, biological samples, and clinical and behavioral assessments to identify the biological indicators of PD progress. This study collects 55 patients (18 females and 37 males, mean age: 66.9 ± 4.5 years) with PD related diseases from the PPMI database and 49 HC (25 females and 24 males, mean age: 69.3 ± 5.3 years, 35 HC from the ADNI database and 14 HC from the PPMI database) with age and gender-matched. In addition, each sample is guaranteed to have rfMRI data and gene data (Fleming 2017; Power et al. 2017). From a medical point of view, the physiology and psychology of samples meet the common standards of normal healthy people. These two databases have strict standards for data collection and processing, ensuring the homology of the data in structure. All HC are not interfered by other nervous system diseases, and all subjects have signed a written consent. What is more, the multi-modal data used in this article has been approved and authorized by PPMI and ADNI, and the data usage conforms to the standard.

In this study, chi-square test and two-sample t-test are adopted to evaluate the gender and age differences between the two groups, respectively. The test results show that there is no notable difference between the gender and age of the participants.

Multi-modal data acquisition

The rfMRI and gene data of all test samples are obtained from PPMI database and ADNI database. The 3T SIEMENS MRI scanner is applied to acquire the rfMRI data of samples. All samples are kept eyes closed and awake during fMRI scanning. The main parameters of the instrument are as follows: pulse sequence is EP, TE = 25.0 ms, TR = 2400.0 ms, field strength = 3 Tesla, time slice = 210, slice thickness = 3.2 mm, and turning angle = 80 °. The acquisition equipment of SNP data is Illumina Infinium iSelect HD chip, and blood is the raw material for all SNPs information collection of samples.

Multi-modal data preprocessing

In order to ensure the quality of image data, rfMRI data need to be preprocessed. In this experiment, DPARSF software is adopted to preprocess the data. The processes include: (1) the data format of the original image file is transformed from DICOM to NIFTI for the next preprocessing, (2) the first 10 time points are deleted to reduce the negative effect of magnetic field on the image data, (3) the time difference between each layer is corrected, (4) the head motion correction is used to specify the range of head swing, (5) the EPI template is adopted to standardize the image to make up for the difference of anatomical structure in data acquisition, (6) the standardized image is Gaussian smoothed to guarantee image quality, (7) the linear drift is removed and the pathological signals are retained, (8) the filtering noise (0.1HZ-0.8HZ) is conducted to remove the noise in specific high-frequency band of data, (9) the covariates are removed and signals that affecting subsequent experimental results are regressed.

Similarly, we need to preprocess the gene data for ensuring the quality. PLINK software is adopted to preprocess SNP data, the processes are as follows. Firstly, we set the thresholds of "sample call rate", "genotyping" and "minimum allele frequency" to 95%, 99.9% and 4%, respectively. Next, the Hardy Weinberg test threshold is set to 1e-4. Finally, we reserve 23,000 SNPs for subsequent experiments.

Multi-modal characteristic construction

At present, in the field of neuroimaging, the brain dominates the activity processes in the body and regulates the balance between the body and the surrounding environment. The genes control many important physiological processes of life activities, such as the division of cells, the synthesis of proteins and so on. Therefore, based on the feasible fusion method in the correlation of brain area and gene, multi-modal data can be better applied for its complementary information (Du et al. 2018; Du et al. 2020; Hao et al. 2016), and can provide more favorable information for PD exploration.

The first work of this study is to combine rfMRI data and SNP data to construct multi-modal fusion characteristics. The specific construction steps are as follows: Firstly, the rfMRI data are divided into 90 brain areas by Anatomical Automatic Labeling (AAL) template, corresponding to 90 time series. Then, we use quality control to the SNP data. According to the reference SNP number of SNPs, we group the SNPs on the basis of the belonging genes. N gene groups with a number of SNPs greater than the threshold t are retained. Then, the genes are encoded with discrete values, and their bases (A, T, C and G) are replaced by numbers (1, 2, 3 and 4) to form the digital sequence. At last, we get n gene digital sequences and 90 brain time sequences. We adopt Pearson correlation coefficient to construct the fusion characteristic between gene and brain area as the input characteristic of the model (Schober et al. 2018), so we get $N \times 90$ fusion characteristics of each sample.

The clustering evolution random forest method

Another work is that the CERF method, an improved machine learning method, is applied in this study. The following are the implementation steps.

First, the training set, verification set and test set are randomly selected according to a certain proportion. After determining the data division strategy, a randomly selected training set is used as a training sample, and 64-dimensional fusion characteristics are selected from the training sample to construct a decision tree. Through the above method, we build a single decision tree, and use the corresponding randomly selected verification set to evaluate the classification performance of the decision tree. If the accuracy is more than 50%, it is retained. We repeat the above steps of building decision tree to get multiple decision trees to build the initial random forest.

Next, clustering evolution is conducted to enhance stability and classification accuracy of the model. In this study, disagreement measure (DM) is used as the decision tree clustering similarity measure. The specific calculation method is as follows. For two arbitrary decision trees T_a and T_b, C_{ab} represents the samples number which can be classified correctly in the training set T_a and T_b. C_{a-b} represents the samples number which can be classified correctly in the training set T_a but incorrectly classified by T_b, C_{b-a} represents the samples number which are classified correctly in the training set T_b but incorrectly classified by T_a. W_{ab} represents the samples number which are incorrectly classified in the training set T_a and T_b. Thus, the DM is constructed:

$$DM_{a,b} = \frac{C_{a-b} + C_{b-a}}{C_{ab} + C_{a-b} + C_{b-a} + W_{ab}}$$

The smaller the $DM_{a,\ b}$ is, the more likely they are to belong to the same cluster. The decision tree with high similarity is further clustered into a cluster by using the linkage hierarchical clustering algorithm, and each cluster only retains the decision tree with the highest classification accuracy. The above process is a clustering evolution, which is iterated many times during the training of the model, so that the performance of the model is optimized gradually.

In this study, the majority voting method is used to get the final classification result of the CERF method. When the test sample is input into the model, each decision tree in the model will give a classification result. Next, we make statistics on the classification results, and then the category with the most votes is regarded as the final category of the sample.

Optimal fusion characteristics analysis

After multiple clustering evolutions, each decision tree in the model has a better ability of sample classification, so the fusion characteristics that these decision trees select can clearly distinguish the HC and the patient. If a fusion characteristic appears repeatedly in multiple decision trees, it means that the fusion characteristic may have a significant contribution to classification. Therefore, we count the frequency of each fusion characteristic in all reserved decision trees, and select n high-frequency fusion characteristics as important fusion characteristics. In order to find out the fusion characteristics with the strongest ability of classification, the selected *n* high-frequency characteristics are divided into several characteristic subsets, and then these subsets classification performance is tested. The fusion characteristics of the subsets corresponding to the peak classification accuracy are the optimal characteristics. Finally, the frequency of brain areas and genes in the optimal characteristic is counted. The higher the frequency of the brain area and the gene is, it shows that the brain areas and genes are more discriminative to patients and HC, so they can be used as the related factors of the disease.

Results

Construction results of fusion characteristics

The brain is divided into 90 brain areas via AAL template. 23,000 SNPs are grouped based on their corresponding genes, and 45 genes with more than 40 SNPs are preserved. Finally, Pearson correlation coefficients of 45 genes and 90 brain areas are calculated, and 45 × 90 = 4050 fusion characteristics of each sample are acquired.

Construction of clustering evolution random forest method

The optimization of the initial decision trees number and the clustering evolutions number in random forest is very necessary to improve the classification performance of the model. The results are as follows.

We set the initial decision trees number and the clustering evolutions number in random forests in the interval [300, 500] and [1, 45], respectively. We constructed 500 decision trees to form random forest. Then 45 clustering evolutions are conducted for random forest. The random forest classification accuracy is calculated after each evolution to find the optimal number of clusters.

In order to get a reliable and stable CERF method, the initial decision trees number is gradually reduced from 500 to 480, 460, 440, …, 300. According to the above method, we obtain the optimal times of clustering evolutions corresponding to the different number of initial decision trees, as shown in Fig. 2. From the figure, we can get the optimal combination of (400, 5). That is to say, the initial decision trees number is set to 400, and final CERF method classification performance is the optimal after five times clustering evolutions.

Extraction of discriminative brain areas and genes

Through the continuous clustering evolutions of random forest model, some redundant and invalid characteristics are eliminated, and the classification performance of the model is also constantly optimized, indicating that the remaining characteristics have a great contribution to the classification performance. Therefore, the frequency of the fusion characteristic in each decision tree of the final CERF method is calculated. The higher the frequency is, the greater the contribution of characteristic to the classification performance is, and the greater the difference between the normal and the patient is. In this study, the first 400 high-frequency characteristics are selected as important fusion characteristics.

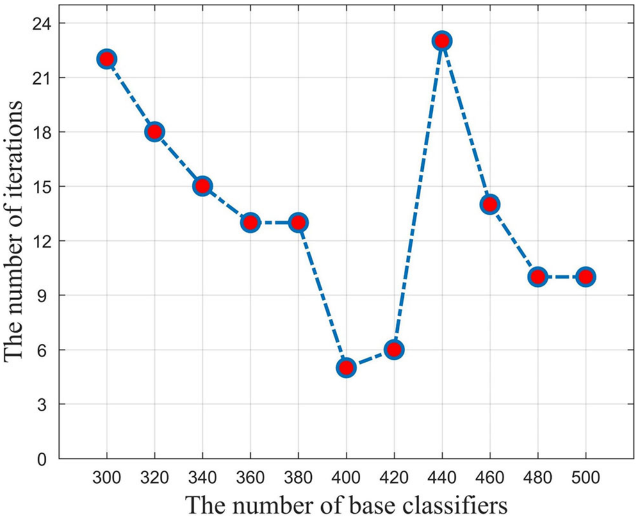

Fig. 2 The optimal number of clustering evolutions corresponding to the number of different initial decision trees

According to the step of "Optimal fusion characteristics analysis" in the section of "Materials and methods", the important fusion characteristics are divided into several subsets. The set of important fusion characteristics is superposed from 70 to 400 in 5 steps to obtain the optimal fusion characteristics. Then the traditional random forest is employed to evaluate the classification performance of these subsets, which have different fusion characteristic numbers. The result is shown in Fig. 3. When we select fusion characteristics of the top 205 frequencies from the important fusion characteristics into a subset, the classification accuracy can reach 88.6%, which is the highest accuracy. Thus, the top 205 fusion characteristics are selected as the optimal fusion characteristics. In addition, Fig. 4 shows the top 20 frequencies optimal fusion characteristics, which have significant classification ability.

Finally, the frequencies of brain areas and genes are counted based on optimal fusion characteristic, and the larger frequencies represent the more discriminative brain areas and genes. The locations and frequencies of PD-related discriminative brain areas are shown in Fig. 5, and the frequency of PD-related discriminative genes is displayed in Fig. 6. The discriminative brain areas include angular gyrus (ANG.L), thalamus (THA.L), posterior cingulate gyrus (PCG.L), paracentral lobule (PCL.L), etc. The discriminative genes include C6orf10, HLA-DPB1, HLA-DOA, etc.

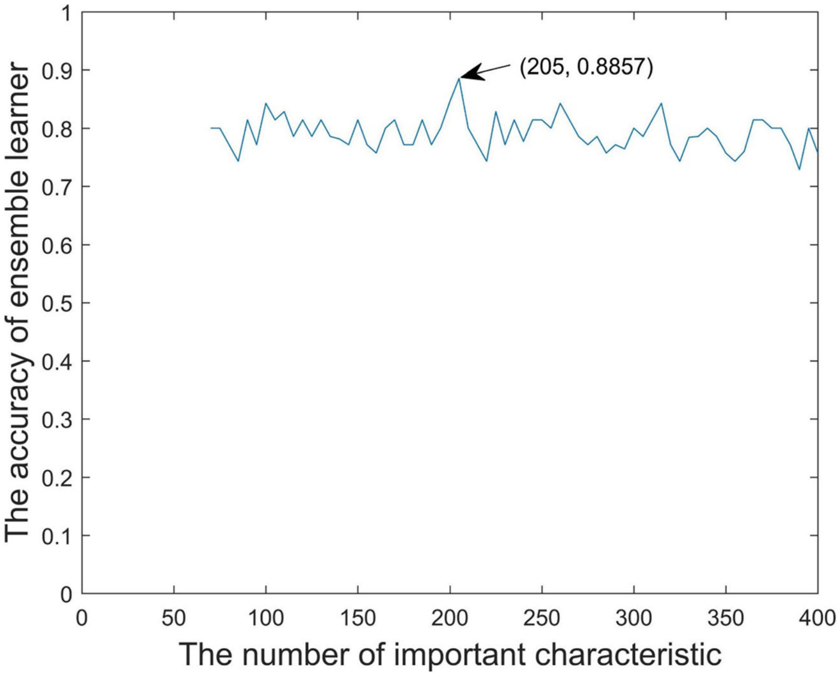

Fig. 3 The classification accuracy of different numbers of important characteristics

Comparison with existing advanced methods

In order to verify the rationality of the characteristic extracted by our method, we combine other fusion characteristic construction and selection methods to extract the optimal characteristics from different perspectives. The fusion characteristic construction methods include correlation distance and canonical correlation analysis, and the fusion characteristic selection methods are random forest, two-sample t-test and random support vector machine cluster. We calculate the optimal characteristics number corresponding to these modelsas the "Discoveries", and adopt the support vector machine (SVM) as the classifier to obtain the respective accuracy. Then the optimal characteristics extracted by different models are compared with the optimal characteristics extracted by Pearson + CERF model, and the results are shown in Table 1.

As can be seen from Table 1, the number of optimal fusion characteristics extracted by the method we apply is the least among all other existing methods, but the classification accuracy of the model based on the optimal characteristic collection is the highest. At the same time, we also find that our method intersects other methods in extracting the optimal characteristics. Based on the hypergeometric test, it is proved that overlapping is not randomized. In addition, the more numbers these optimal characteristics intersected with the optimal characteristics we extract, the higher the classification accuracy is. The rationality and reliability of Pearson + CERF model is proved.

The number of initial decision trees is set to 400, and the cluster evolution time is set to 3, 4, 5, 6, 7 and 8, respectively. We conduct 50 independent experiments to test the method performance. We also compare the classification performances of this method with two-sample t-test for single-modal and multi-modal datasets. The results are described in Fig. 7. It can be seen that as the evolution time increases, the classification effect of the CERF method is improved significantly. When peak performance is reached, if the clustering evolution continued, the method classification performance may decrease. Therefore, the clustering evolutions number is 5 and the initial decision trees number is 400, which is the optimal equilibrium between resource and performance. Finally, compared with the two-sample t-test, the CERF method has obvious advantages, and the classification accuracy of multi-modal data is higher than that of single-modal data. Therefore, it can be concluded that the Pearson + CERF model has significant classification performance in PD.

Discussion

In our study, we applied the CERF method, which can recognize some highly discriminative brain areas and genes by classifying PD patients and HC. Among them, the ANG.L and THA.L had higher frequency in the classification of PD, indicating that these two brain areas played key roles in the classification of PD. The ANG.L had no visual impairment after being injured, but people who were literate became unable to read, which was clinically called dyslexia (Manes et al. 2018). It had been reported that in PD patients, the structural and functional changes of brain area included ANG.L. This result suggested that the cognitive decline of PD was closely related

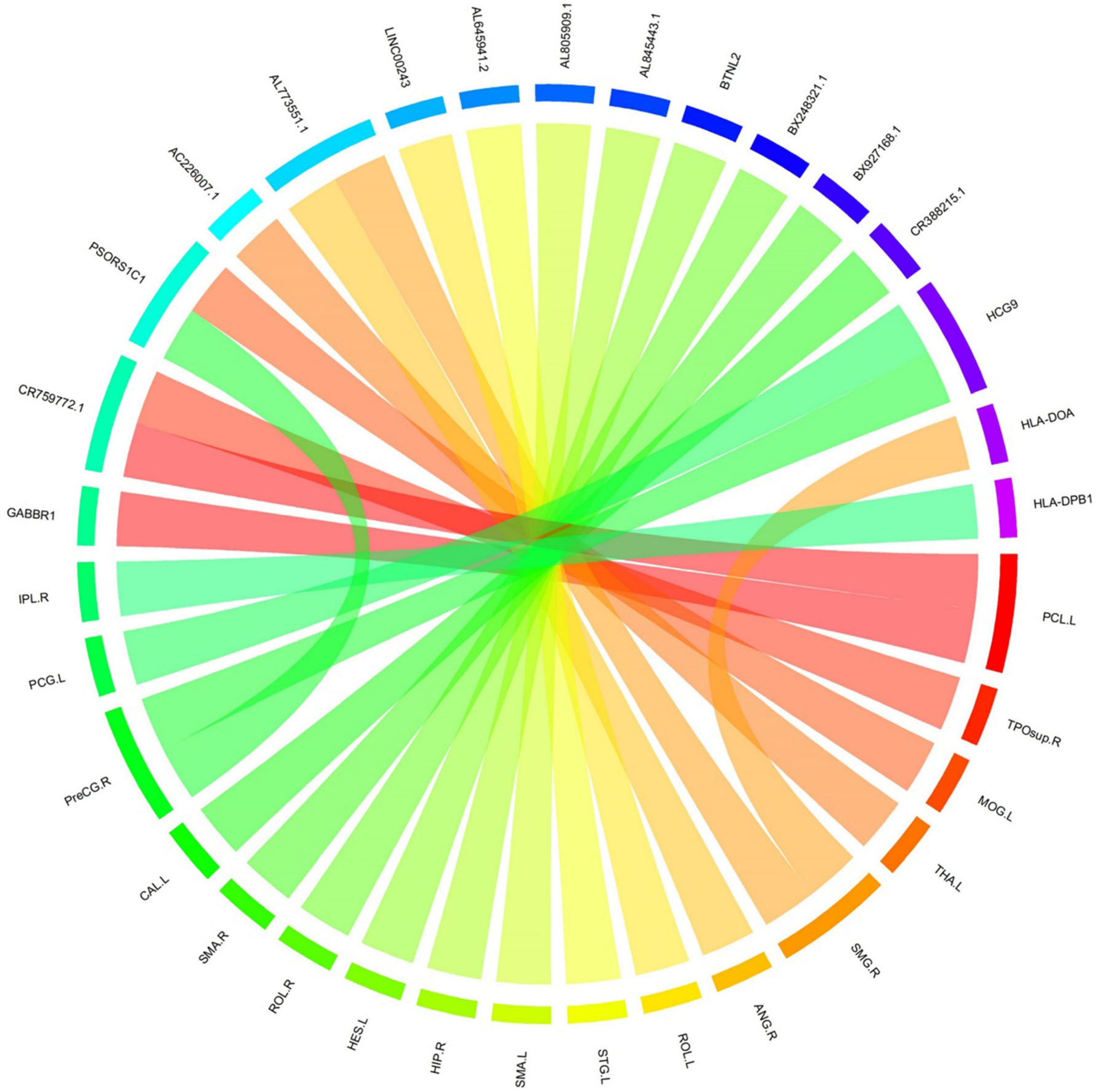

Fig. 4 The top 20 optimal fusion characteristics with the most significant classification effect

to the decline of angular structure and function, indicating the specific areas of brain atrophied. At the same time, the study found that cognitive dysfunction in patients with PD was strongly correlated with an increase in pain processing dysfunction. Pain processing dysfunction was an important aspect of the somatosensory network, and the ANG.L was one of the central areas. In addition, the left corner gyrus was particularly related to speech processing (Mihaescu et al. 2019). The THA.L was the successor of sensory conduction. It was reported that THA.L could also be used to predict the motor response of PD for deep brain stimulation (DBS) (Younce et al. 2019). Owens-Walton et al. (2019) studied the potential differences in thalamic size and shape of Parkinson's disease, as well as the relationship between morphological and functional connections and clinical variables. The results showed that the functional connectivity between some brain areas and THA.L increased. To sum up, the discovery of ANG.L and THA.L was meaningful for the diagnosis and treatment of PD.

Moreover, we also found some brain areas related to the development of PD, such as PCG.L and PCL.L. It was reported that the cingulate gyrus was related to emotional processing and cognitive functions, and the cingulate gyrus monitors sensations and stereotactic positioning and memory (Tatura et al. 2016). Wilson et al. (2019) noticed that PCL.L played a central role in integrating information and function separation in brain areas, and it was particularly prone to atrophy in PD. The studies had shown that PD patients would suffer from inconvenience in lower limb movement (Drucker et al. 2019; Khawaldeh et al. 2020). Schwartz et al. (2019) noticed that the functional orientation of the PCL.L was mainly related to the movement and sensation of the lower body. If injured, it would cause movement and sensory dysfunction of both legs,

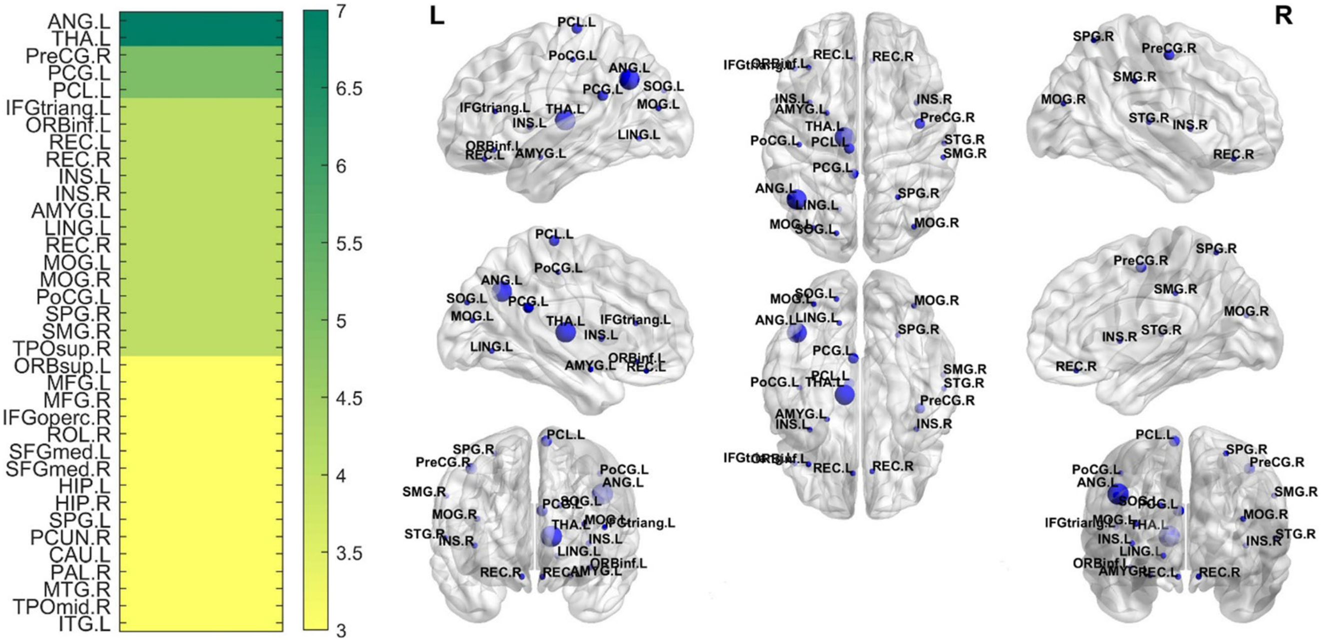

Fig. 5 The locations and frequencies of PD-related discriminative brain areas

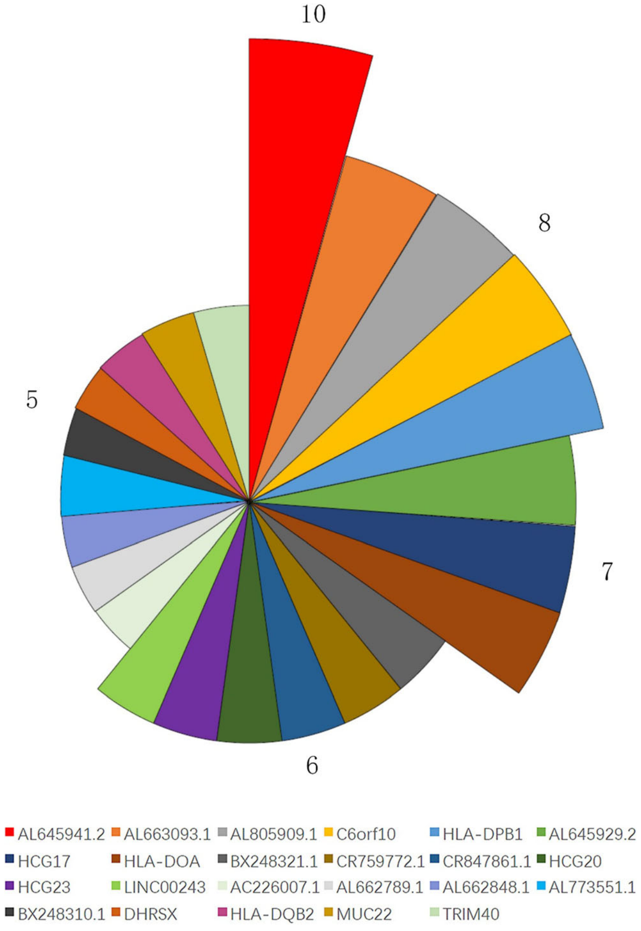

Fig. 6 The frequency information of PD-related discriminative genes

urination and discriminative defecation function. Furthermore, they also found that the degree of manic depression was relate to the PCL.L in PD patients.

We also found some discriminative genes related to PD, such as C6orf10, HLA-DPB1 and HLA-DOA. Ciani et al. (2019) showed in the article that in the recent genome-wide association studies, C6orf10 was determined to greatly affect the pathological mechanism of PD disease. Furthermore, Ghatak et al. (2018) revealed new aspects of the disease mechanism, including non-cellular autonomous events and the spread of pathogenic proteins in the brain, and found that the genetic risk variant gene for PD contains C6orf10. Based on the deoxyribonucleic acid molecular epigenetic clock,

Table 1 Comparison of our method with other existing advanced methods

Methods	Discoveries	Accuracy	Intersection
Pearson + CERF	205	88.1%	–
Pearson + RF	670	78.5%	133 ($p = 2.54358e\text{-}67$)
Pearson + RSVMC	260	61.9%	81 ($p = 2.528332e\text{-}10$)
Pearson + t-test	499	71.4%	118 ($p = 8.089311e\text{-}37$)
CCA + t-test	412	78.5%	102 ($p = 1.059997e\text{-}12$)
CD + t-test	447	73.8%	137 ($p = 6.639003e\text{-}35$)

RF, random forest; *RSVMC*, random SVM cluster; *t-test*, two-sample t-test; *CD*, correlation distance. The *p* value was gained by the hypergeometric test

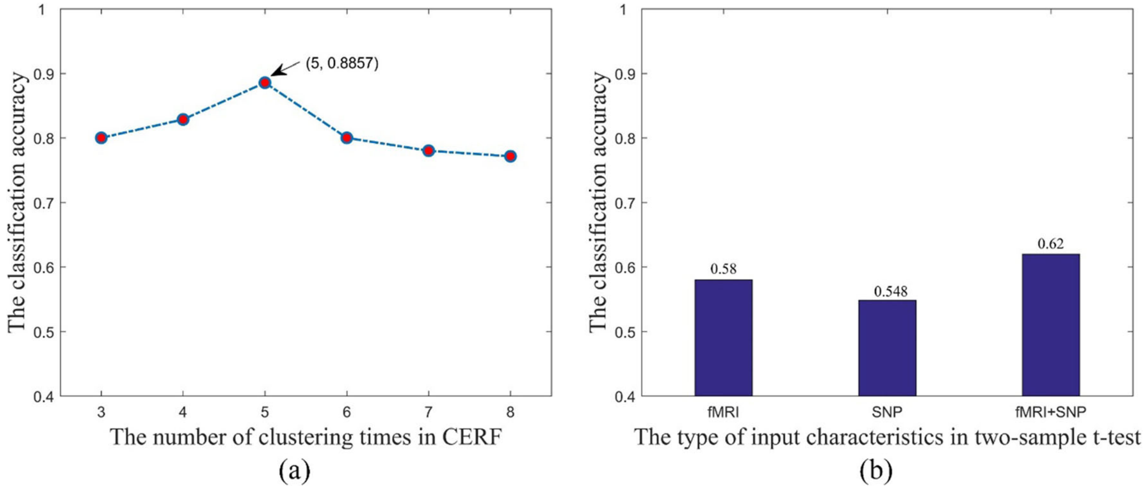

Fig. 7 The comparison between CERF and two-sample t-test. The figure (**a**) depicts the accuracy of CERF with different cluster evolution times. The figure (**b**) shows the accuracy of two-sample t-test with different input characteristics

Chouliaras et al. (2018) applied the apparent genetic age in blood to show the correlation with PD, including HLA-DPB1 in the differential methylation loci of peripheral blood monocytes related to montreal cognitive assessment. In the study of apoptosis and immune activation in response to infection, it was found that the expression of HLA-DOA and HLA-DPB1 genes were decreased in PD patients (Wu et al. 2017).

Some limitations should be declared in this study. First, some atypical pathogenic factors which are lack of relevant research are found in this paper. We will collect more data in the follow-up research work and design new algorithm to conduct in-depth analysis of PD. Second, we adopted AAL template to divide brain area, and other templates such as Broadman template could be used for matching (Joshi et al. 2004). Finally, the multi-modal data in this study were brain areas and genes, and other modal data, such as lncRNA, protein and miRNA (Chen et al. 2019, 2018; You et al. 2016), could be adopted for fusion to improve the classification accuracy, so as to deepen the discussion of PD pathological mechanism.

Conclusion

In this study, PD is explored by fusion of imaging and genetic data. The fusion characteristics are constructed based on the correlation between genes and brain areas, and further analyzed by the CERF method. The main contributions are as follows: first, a fusion scheme of rfMRI and SNP data is designed, which makes full use of the advantages of multiple discriminant characteristic fusion; second, the efficient multi-modal fusion data analysis method—CERF, is applied in the detection of PD, which can effectively identify PD patients and extract the most discriminant characteristics; third, this study identifies some landmark brain areas and genes that are vital for the prevention and diagnosis of PD. The work of this paper can provide valuable guidance for the research of PD and other similar brain diseases.

Acknowledgements This work was supported by the Natural Science Foundation of Hunan Province, China (Grant No. 2020JJ4432, the Degree & Postgraduate Education Reform Project of Hunan Province (Grant No. 2019JGYB091), the Hunan Provincial Science and Technology Project Foundation (Grant No. 2018TP1018), the National Natural Science Foundation of China (Grant No. 61502167).

PPMI – a public-private partnership–funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Avid Radiopharmaceuticals, Biogen, BioLegend, Bristol-Myers Squibb, General Electric Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal Imaging, Roche, Sanofi-Genzyme, Servier, Takeda, Teva and UCB.

Data collection and sharing for this project were also funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-20012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F.Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. Private sector contributions were facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Compliance with ethical standards

Conflict of interest The authors have declared that no competing interest exists.

Informed consent All the participants and their guardians have signed the informed consent.

References

Agliardi, C., Guerini, F. R., Zanzottera, M., Riboldazzi, G., Zangaglia, R., Sturchio, A., Casali, C., di Lorenzo, C., Minafra, B., Nemni, R., & Clerici, M. (2019). SNAP25 gene polymorphisms protect against Parkinson's disease and modulate disease severity in patients. *Molecular Neurobiology, 56*(6), 4455–4463.

Akgun, A. (2012). A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: A case study at İzmir, Turkey. *Landslides, 9*(1), 93–106.

Bologna, M., Leodori, G., Stirpe, P., Paparella, G., Colella, D., Belvisi, D., Fasano, A., Fabbrini, G., & Berardelli, A. (2016). Bradykinesia in early and advanced Parkinson's disease. *Journal of the Neurological Sciences, 369*, 286–291.

Chen, X., Wang, L., Qu, J., Guan, N.-N., & Li, J.-Q. (2018). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics, 34*(24), 4256–4265.

Chen, X., Sun, Y.-Z., Guan, N.-N., Qu, J., Huang, Z.-A., Zhu, Z.-X., & Li, J. Q. (2019). Computational models for lncRNA function prediction and functional similarity calculation. *Briefings in Functional Genomics, 18*(1), 58–82.

Chouliaras, L., Pishva, E., Haapakoski, R., Zsoldos, E., Mahmood, A., Filippini, N., Burrage, J., Mill, J., Kivimäki, M., Lunnon, K., & Ebmeier, K. P. (2018). Peripheral DNA methylation, cognitive decline and brain aging: Pilot findings from the Whitehall II imaging study. *Epigenomics, 10*(5), 585–595.

Ciani, M., Benussi, L., Bonvicini, C., & Ghidoni, R. (2019). Genome wide association study and next generation sequencing: A glimmer of light towards new possible horizons in Frontotemporal dementia research. *Frontiers in Neuroscience, 13*, 506.

De Virgilio, A., Greco, A., Fabbrini, G., Inghilleri, M., Rizzo, M. I., Gallo, A., et al. (2016). Parkinson's disease: Autoimmunity and neuroinflammation. *Autoimmunity Reviews, 15*(10), 1005–1011.

Drucker, J., Sathian, K., Crosson, B., Krishnamurthy, V., McGregor, K. M., Bozzorg, A., et al. (2019). Internally guided lower limb movement recruits compensatory cerebellar activity in people with Parkinson's disease. *Frontiers in Neurology, 10*, 537.

Du, L., Liu, K., Yao, X., Risacher, S. L., Han, J., Guo, L., et al. (2018). Fast multi-task SCCA learning with feature selection for multimodal brain imaging genetics. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 356–361).

Du, L., Liu, K., Zhu, L., Yao, X., Risacher, S. L., Guo, L., et al. (2019). Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: A longitudinal study of the ADNI cohort. *Bioinformatics, 35*(14), i474–i483.

Du, L., Liu, K., Yao, X., Risacher, S. L., Han, J., Saykin, A. J., et al. (2020). Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach. *Medical Image Analysis, 61*, 101656.

Falconi, A., Bonito-Oliva, A., Di Bartolomeo, M., Massimini, M., Fattapposta, F., Locuratolo, N., et al. (2019). On the role of adenosine A2A receptor gene transcriptional regulation in Parkinson's disease. *Frontiers in Neuroscience, 13*, 683–692.

Fleming, S. M. (2017). Mechanisms of gene-environment interactions in Parkinson's disease. *Current environmental health reports, 4*(2), 192–199.

Ghatak, S., Trudler, D., Dolatabadi, N., & Ambasudhan, R. (2018). Parkinson's disease: What the model systems have taught us so far. *Journal of Genetics, 97*(3), 729–751.

Goldman, J., Fox, S., Isaacson, S., Fredericks, D., Trotter, J., Healy, K., et al. (2019). Examining Parkinson's disease psychosis treatment outcomes in the real world: Interim year 1 findings from the INSYTE observational study. *The American Journal of Geriatric Psychiatry, 27*(3), S180–S181.

Hao, X., J. Yan, X. Yao, S. L. Risacher, A. J. Saykin, D. Zhang, et al. (2016). Diagnosis-guided method for identifying multi-modality neuroimaging biomarkers associated with genetic risk factors in Alzheimer's disease. *Biocomputing 2016: Proceedings of the Pacific Symposium*, 108-119.

Huang, J., Zhu, Q., Hao, X., Shi, X., Gao, S., Xu, X., & Zhang, D. (2018). Identifying resting-state multifrequency biomarkers via tree-guided group sparse learning for schizophrenia classification. *IEEE Journal of Biomedical and Health Informatics, 23*(1), 342–350.

Jones-Davis, D. M., & Buckholtz, N. (2015). The impact of ADNI: What role do public-private partnerships have in pushing the boundaries of clinical and basic science research on Alzheimer's disease? *Alzheimer's & dementia: the journal of the Alzheimer's Association, 11*(7), 860–864.

Joshi, S., Davis, B., Jomier, M., & Gerig, G. (2004). Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage, 23*, S151–S160.

Kaut, O., C. Mielacher, R. Hurlemann and U. Wüllner. (2020). Resting-state fMRI reveals increased functional connectivity in the cerebellum but decreased functional connectivity of the caudate nucleus in Parkinson's disease. *Neurological Research*, 1-6.

Khawaldeh, S., Tinkhauser, G., Shah, S. A., Peterman, K., Debove, I., Nguyen, T. K., et al. (2020). Subthalamic nucleus activity dynamics and limb movement prediction in Parkinson's disease. *Brain, 143*(2), 582–596.

Manes, J. L., Tjaden, K., Parrish, T., Simuni, T., Roberts, A., Greenlee, J. D., Corcos, D. M., & Kurani, A. S. (2018). Altered resting-state functional connectivity of the putamen and internal globus pallidus is related to speech impairment in Parkinson's disease. *Brain and behavior, 8*(9), e01073–e01092.

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C. M., Trojanowski, J. Q., Shaw, L. M., Seibyl, J., Schuff, N., Singleton, A., Kieburtz, K., Toga, A. W., Mollenhauer, B., Galasko, D., Chahine, L. M., Weintraub, D., Foroud, T., Tosun-Turgut, D., Poston, K., Arnedo, V., Frasier, M., Sherer, T., the Parkinson's Progression Markers Initiative, Bressman, S., Merchant, M., Poewe, W., Kopil, C., Naito, A., Dorsey, R., Casaceli, C., Daegele, N., Albani, J., Uribe, L., Foster, E., Long, J., Seedorff, N., Crawford, K., Smith, D., Casalin, P., Malferrari, G., Halter, C., Heathers, L., Russell, D., Factor, S., Hogarth, P., Amara, A., Hauser, R., Jankovic, J., Stern, M., Hu, S. C., Todd, G., Saunders-Pullman, R., Richard, I., Saint-Hilaire, H., Seppi, K., Shill, H., Fernandez, H., Trenkwalder, C., Oertel, W., Berg, D., Brockman, K., Wurster, I., Rosenthal, L., Tai, Y., Pavese, N., Barone, P., Isaacson, S., Espay, A., Rowe, D., Brandabur, M., Tetrud, J., Liang, G., Iranzo, A., Tolosa, E., Marder, K., Sanchez, M., Stefanis, L., Marti, M., Martinez, J., Corvol, J. C., Assly, O., Brillman, S., Giladi, N., Smejdir, D., Pelaggi, J., Kausar, F., Rees, L., Sommerfield, B., Cresswell, M., Blair, C., Williams, K., Zimmerman, G., Guthrie, S., Rawlins, A., Donharl, L., Hunter, C., Tran, B., Darin, A., Venkov, H., Thomas, C. A., James, R., Heim, B., Deritis, P., Sprenger, F., Raymond, D., Willeke, D., Obradov, Z., Mule, J., Monahan, N., Gauss, K., Fontaine, D., Szpak, D., McCoy, A., Dunlop, B., Payne, L., Ainscough, S., Carvajal, L., Silverstein, R., Espay, K., Ranola, M., Rezola, E., Santana, H., Stamelou, M., Garrido, A., Carvalho, S., Kristiansen, G., Specketer, K., Mirlman, A., Facheris, M., Soares, H., Mintun, A., Cedarbaum, J., Taylor, P., Jennings, D., Slieker, L., McBride, B., Watson, C., Montagut, E., Sheikh, Z., Bingol, B., Forrat, R., Sardi, P., Fischer, T., Reith, D., Egebjerg, J., Larsen, L., Breysse, N., Meulien, D., Saba, B., Kiyasova, V., Min, C., McAvoy, T., Umek, R., Iredale, P., Edgerton, J., Santi, D., Czech, C., Boess, F., Sevigny, J., Kremer, T., Grachev, I., Merchant, K., Avbersek, A., Muglia, P., Stewart, A., Prashad, R., & Taucher, J. (2018). The Parkinson's progression markers initiative (PPMI)–establishing a PD biomarker cohort. *Annals of clinical and translational neurology, 5*(12), 1460–1477.

Martin, J. A., Zimmermann, N., Scheef, L., Jankowski, J., Paus, S., Schild, H. H., Klockgether, T., & Boecker, H. (2019). Disentangling motor planning and motor execution in unmedicated de novo Parkinson's disease patients: An fMRI study. *NeuroImage: Clinical, 22*, 101784.

Mihaescu, A. S., Masellis, M., Graff-Guerrero, A., Kim, J., Criaud, M., Cho, S. S., Ghadery, C., Valli, M., & Strafella, A. P. (2019). Brain degeneration in Parkinson’s disease patients with cognitive decline: A coordinate-based meta-analysis. *Brain Imaging and Behavior, 13*(4), 1021–1034.

Nalls, M. A., McLean, C. Y., Rick, J., Eberly, S., Hutten, S. J., Gwinn, K., et al. (2015). Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: A population-based modelling study. *The Lancet Neurology, 14*(10), 1002–1009.

Owens-Walton, C., Jakabek, D., Power, B. D., Walterfang, M., Velakoulis, D., Van Westen, D., et al. (2019). Increased functional connectivity of thalamic subdivisions in patients with Parkinson’s disease. *PLoS One, 14*(9), e0222002.

Power, J. D., Plitt, M., Laumann, T. O., & Martin, A. (2017). Sources and implications of whole-brain fMRI signals in humans. *Neuroimage, 146*, 609–625.

Reynolds, R. H., Botía, J., Nalls, M. A., Hardy, J., Taliun, S. A. G., & Ryten, M. (2019). Moving beyond neurons: The role of cell type-specific gene regulation in Parkinson’s disease heritability. *NPJ Parkinson's disease, 5*(1), 1–14.

Rittman, T., Rubinov, M., Vértes, P. E., Patel, A. X., Ginestet, C. E., Ghosh, B. C., et al. (2016). Regional expression of the MAPT gene is associated with loss of hubs in brain networks and cognitive impairment in Parkinson disease and progressive supranuclear palsy. *Neurobiology of Aging, 48*, 153–160.

Robak, L. A., Jansen, I. E., Van Rooij, J., Uitterlinden, A. G., Kraaij, R., Jankovic, J., et al. (2017). Excessive burden of lysosomal storage disorder gene variants in Parkinson’s disease. *Brain, 140*(12), 3191–3203.

Santos-García, D., Mir, P., Cubo, E., Vela, L., Rodríguez-Oroz, M. C., Martí, M. J., et al. (2016). COPPADIS-2015 (COhort of patients with PArkinson’s DIsease in Spain, 2015), a global–clinical evaluations, serum biomarkers, genetic studies and neuroimaging–prospective, multicenter, non-interventional, long-term study on Parkinson’s disease progression. *BMC Neurology, 16*(1), 26–39.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763–1768.

Schwartz, F., Tahmasian, M., Maier, F., Rochhausen, L., Schnorrenberg, K. L., Samea, F., Seemiller, J., Zarei, M., Sorg, C., Drzezga, A., Timmermann, L., Meyer, T. D., van Eimeren, T., & Eggers, C. (2019). Overlapping and distinct neural metabolic patterns related to impulsivity and hypomania in Parkinson’s disease. *Brain Imaging and Behavior, 13*(1), 241–254.

Su, R., Liu, X., Wei, L., & Zou, Q. (2019). Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods, 166*, 91–102.

Tatura, R., Kraus, T., Giese, A., Arzberger, T., Buchholz, M., Höglinger, G., & Müller, U. (2016). Parkinson's disease: SNCA-, PARK2-, and LRRK2-targeting microRNAs elevated in cingulate gyrus. *Parkinsonism & Related Disorders, 33*, 115–121.

Thenganatt, M. A., & Jankovic, J. (2016). The relationship between essential tremor and Parkinson's disease. *Parkinsonism & Related Disorders, 22*, S162–S165.

Torigian, D. A., Kjær, A., Zaidi, H., & Alavi, A. (2016). PET/MR imaging: Clinical applications. *PET clinics, 11*(4), xi–xii.

Tysnes, O.-B., & Storstein, A. (2017). Epidemiology of Parkinson’s disease. *Journal of Neural Transmission, 124*(8), 901–905.

Wen, M. C., Chan, L., Tan, L., & Tan, E. (2016). Depression, anxiety, and apathy in Parkinson's disease: Insights from neuroimaging studies. *European Journal of Neurology, 23*(6), 1001–1019.

Wilson, H., Niccolini, F., Pellicano, C., & Politis, M. (2019). Cortical thinning across Parkinson's disease stages and clinical correlates. *Journal of the Neurological Sciences, 398*, 31–38.

Wu, C., Xu, G., Tsai, S.-Y. A., Freed, W. J., & Lee, C.-T. (2017). Transcriptional profiles of type 2 diabetes in human skeletal muscle reveal insulin resistance, metabolic defects, apoptosis, and molecular signatures of immune activation in response to infections. *Biochemical and Biophysical Research Communications, 482*(2), 282–288.

You, Z.-H., Zhou, M., Luo, X., & Li, S. (2016). Highly efficient framework for predicting interactions between proteins. *IEEE transactions on cybernetics, 47*(3), 731–743.

Younce, J. R., Campbell, M. C., Perlmutter, J. S., & Norris, S. A. (2019). Thalamic and ventricular volumes predict motor response to deep brain stimulation for Parkinson's disease. *Parkinsonism & Related Disorders, 61*, 64–69.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.