HHS Public Access

Author manuscript

*Med Image Comput Comput Assist Interv*. Author manuscript; available in PMC 2017 June 05.

Published in final edited form as:

*Med Image Comput Comput Assist Interv*. 2016 October ; 9901: 1–8. doi:10.1007/978-3-319-46723-8_1.

# Feature Selection Based on Iterative Canonical Correlation Analysis for Automatic Diagnosis of Parkinson's Disease

**Luyan Liu**[1], **Qian Wang**[1], **Ehsan Adeli**[2], **Lichi Zhang**[1,2], **Han Zhang**[2], and **Dinggang Shen**[2]

[1]School of Biomedical Engineering, Med-X Research Institute, Shanghai Jiao Tong University, Shanghai, China

[2]Department of Radiology BRIC, University of North Carolina at Chapel Hill, Chapel Hill, USA

## Abstract

Parkinson's disease (PD) is a major progressive neurodegenerative disorder. Accurate diagnosis of PD is crucial to control the symptoms appropriately. However, its clinical diagnosis mostly relies on the subjective judgment of physicians and the clinical symptoms that often appear late. Recent neuroimaging techniques, along with machine learning methods, provide alternative solutions for PD screening. In this paper, we propose a novel feature selection technique, based on iterative canonical correlation analysis (ICCA), to investigate the roles of different brain regions in PD through T1-weighted MR images. First of all, *gray matter* and *white matter* tissue volumes in brain regions of interest are extracted as two feature vectors. Then, a small group of significant features were selected using the iterative structure of our proposed ICCA framework from both feature vectors. Finally, the selected features are used to build a robust classifier for automatic diagnosis of PD. Experimental results show that the proposed feature selection method results in better diagnosis accuracy, compared to the baseline and state-of-the-art methods.

## 1 Introduction

Parkinson's disease (PD) is a major neurodegenerative disorder that threatens aged people and causes huge burdens to the society. The clinical diagnosis of PD, however, is particularly prone to errors, because the diagnosis mostly relies on substantial symptoms of the patients [1]. Computer-aided techniques can utilize machine learning for the diagnosis of PD and also for identifying biomarkers from neuroimaging data for the disease. There are several studies in the literature, which aim to distinguish PD from other similar diseases or normal subjects. In [2], Single-photon emission computed tomography (SPECT) images were analyzed automatically for PD diagnosis; while in [3], a novel speech signal-processing algorithm was proposed. Different clinical features (including response to levodopa, motor fluctuation, rigidity, dementia, speech, etc.) were evaluated in [4] for distinguishing multiple system atrophy (MSA) from PD. In [5], a novel synergetic paradigm integrating Kohonen self-organizing map (KSOM) was proposed to extract features for clinical diagnosis based on least squares support vector machine (LS-SVM).

Correspondence to: Dinggang Shen.

In this study, we propose reliable feature selection and classification models for PD diagnosis using T1-weighted MR images. Therefore, our method would be a non-invasive and reasonable solution to PD screening, which is especially important to developing countries with limited healthcare resources. Specifically, we extract numerous numbers of features from T1 MR images, which describe the volumes of individual tissues in the regions-of-interest (ROIs), such as white matter (WM) and gray matter (GM). Therefore, the features can be naturally grouped into two vectors, corresponding to WM and GM. Afterwards, we introduce an *iterative* feature selection strategy based on canonical correlation analysis (CCA) to iteratively identify the optimal set of features. Then, the selected features are used for establishing the robust linear discriminant analysis (RLDA) model to classify PD patients from the normal control (NC) subjects.

Feature selection is an important dimensionality reduction technique and has been applied to solving various problems in translational medical studies. For example, sparse logistic regression was proposed to select features for predicting the conversion from mild cognitive impairment (MCI) to the Alzheimer's disease (AD) in [6]. The least absolute shrinkage and selection operator (LASSO) was used in [7] for feature selection. Similar works can also be found in [8 9], where principal component analysis (PCA) and CCA were used, respectively.

CCA is able to explore the relationship between two high-dimensional vectors of features, and transform them from their intrinsic spaces to a common feature space [9]. In our study, the two feature vectors describe each subject under consideration from two views of different anatomical feature spaces, associated with WM and GM, respectively. The two feature vectors, thus, need to be transformed to a common space, where features can be compared and jointly selected for subsequent classification. Specifically, after linearly transforming the two views of features to the common space by CCA, we learn a regression model to fit the PD/NC labels based on the transformed feature representations. With the CCA-based transformation and the regression model, we are able to identify the most useful and relevant features for PD classification. In addition, PD is not likely to affect all brain regions, but rather only a small number of ROIs are relevant for classification. Therefore, in the obtained features, there could be many redundant and probably noisy features, which may negatively affect the CCA mappings to a common space. In this sense, a single round of CCA-based feature selection with a large bunch of features being discarded at the same time would probably yield suboptimal outcome. Intuitively, we develop an iterative structure of CCA-based feature selection, or ICCA, in which we propose to gradually discard features step-by-step. In this way, the two feature vectors gradually get a better common space and thus more relevant features can be selected.

Specifically, our ICCA method consists of multiple iterations for feature selection. In each iteration, we transform the features of the WM/GM views into a common space, build a regression model, inverse-transform the regression coefficients into the original space, and eliminate the most irrelevant features for PD classification. This iterative scheme allows us to gradually refine the estimation of the common feature space, by eliminating only the least important features. In the final, we utilize the representations in the common space, transformed from the selected features, to conduct PD classification. Note that, although the CCA-based transform is linear, our ICCA consists of iterative procedure and thus provides

fairly local linear operation capabilities to select features of different anatomical views. Experimental results show that the proposed method significantly improves the diagnosis accuracy and outperforms state-of-the-art feature selection methods, including sparse learning [7], PCA [8], CCA [10] and minimum redundancy-maximum (mRMR) [11].

## 2 Method

Figure 1 illustrates the pipeline for PD classification in this paper. After extracting the WM and GM features from T1 images, we feed them into the ICCA-based feature selection framework. The WM/GM feature vectors are mapped to a common space using CCA, where the canonical representations of the features are computed. The regression model, based on the canonical representations, fits the PD/NC labels of individual subjects. The regression then leads to the weights assigned to the canonical representations, from which the importance of the WM/GM features can be computed. We then select the WM/GM features conservatively, by only eliminating the least important features. The rest of WM/GM features are transformed to the refined common space by CCA and selected repeatedly, until only a set of optimal features are remained. The finally selected features are incorporated to build a robust classifier for separating PD patients from the NC subjects.

### Feature Extraction

All T1-weighted MR images are pre-processed by skull stripping, cerebellum removal, and tissue segmentation into WM and GM. Then, the anatomical automatic labeling (AAL) atlas with 90 pre-defined ROIs is registered to the native space of each subject, using FLIRT [12], followed by HAMMER [13]. For each ROI, we compute the WM/GM volumes as features. In this way, we extract 90 WM and 90 GM features for each subject. The features are naturally grouped into two vectors, which will be handled by the ICCA-based feature selection and the RLDA-based classification.

### CCA-based Feature Selection

For $n$ subjects, we record their $d$-dimensional feature vectors as individual columns in $\mathbf{X}_1 \in \mathbf{R}^{n\times d}$ and $\mathbf{X}_2 \in \mathbf{R}^{n\times d}$, corresponding to WM and GM features, respectively. The class labels for the subjects are stored in $\mathbf{y} \in \mathbf{R}^{n\times 1}$, where each entry is either 1 or 0, indicating which class (PD or NC) each corresponding subject belongs to. Let $\mathbf{X} = [\mathbf{X_1},\mathbf{X_2}] \in \mathbf{R}^{n\times 2d}$, and $\sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$ be its covariance matrix. CCA can find the basis vectors $\hat{\mathbf{B}}_1 \in \mathbf{R}^{d\times d}$ and $\hat{\mathbf{B}}_2 \in \mathbf{R}^{d\times d}$ to maximize the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$. The two basis vectors $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$ can be optimized by solving the following objective function:

$$\left(\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2\right) = \underset{(\hat{\mathbf{B}}_1,\hat{\mathbf{B}}_2)}{\arg\max} \frac{\mathbf{B}_1^T \sum_{12} \mathbf{B}_2}{\sqrt{\mathbf{B}_1^T \sum_{11} \mathbf{B}_{11}}\sqrt{\mathbf{B}_2^T \sum_{22} \mathbf{B}_2}}. \quad (1)$$

The optimal solution of $(\hat{\mathbf{B}}_1,\hat{\mathbf{B}}_2)$ is obtained by a generalized eigen-decomposition [9]. The canonical representations of all features in the common space can be computed by $Z_m = \hat{\mathbf{B}}_m^{\mathrm{T}} \mathbf{X}_m$, $m = \{1,2\}$.

With the canonical representations, we build a sparse regression model. The regression aims to fit the class labels with the canonical representations by assigning various weights to the representations in the common space.

$$\arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y}-\mathbf{Z}\mathbf{w}\|_{\mathrm{F}}^2+\beta\|\mathbf{w}\|_1+\gamma\|\mathbf{w}\|_{\mathrm{CCA}}^2. \quad (2)$$

where $\mathbf{Z} = [\mathbf{Z}_1,\mathbf{Z}_2] \in \mathbf{R}^{n\times 2d}$ is the canonical representation matrix and $\mathbf{w} = [\mathbf{w}_1;\mathbf{w}_2] \in \mathbf{R}^{2d\times 1}$ is the regression coefficient matrix, which assigns weights to individual canonical representations; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are trade-off parameters; $\|\cdot\|_1$ denotes the $l_1$ norm, which tends to assign non-zero coefficients to only a few canonical representations; and $\|\cdot\|_{\mathrm{CCA}}$ denotes the canonical regularizer [10]:

$$\|\mathbf{w}\|_{\mathrm{CCA}}^2=\sum_{i=1}^{d}\frac{1-\boldsymbol{\lambda}_i}{\boldsymbol{\lambda}_i}\left(w_i^2+w_{i+d}^2\right), \quad (3)$$

where $\{\boldsymbol{\lambda_i}\}_{i=1:d}$ denotes a set of canonical correlation coefficients. $w_i$ and $w_{i+d}$ are the weights corresponding to a same feature index across the two views (GM and WM). Canonical regularizer enforces to select highly correlated representations across the two feature views. In other words, larger canonical correlation coefficients tend to be selected, while less correlated canonical representations across the two views (small canonical correlation coefficients) are not selected. Note that greater $\boldsymbol{\lambda}_i$ will lead to larger values in $w_i$ and $w_{i+d}$ after the optimization process.

## The Proposed ICCA-based Feature Selection Method

The CCA-based feature selection might be limited, as all features are (globally) linearly transformed to the common space and then truncated in a one-shot fashion. Therefore, we propose a novel ICCA-based feature selection method, in which we iteratively discard the most irrelevant pair of features in each iteration, and re-evaluate the new set till the best set of features are selected. Altogether, this fairly simulates a local linear operation.

In Eq. (2), we obtain the regression coefficient matrix $\mathbf{w} = [\mathbf{w}_1;\mathbf{w}_2]$, containing the weights for the canonical representations $\mathbf{Z} = [\mathbf{Z}_1,\mathbf{Z}_2]$ in the tentatively estimated common space. Since the canonical representations $\mathbf{Z}$ are linear combinations of the WM/GM features in $\mathbf{X}$ ($\mathbf{X} = [\mathbf{X}_1,\mathbf{X}_2]$), the weights in $\mathbf{w}$ are also linearly associated with the importance of WM/GM features prior to the CCA-based mapping. Therefore, the importance of WM/GM features can be computed by $\tilde{\mathbf{w}}_m = (\hat{\mathbf{B}}_m)^{-1}\mathbf{w}_m$, where $\tilde{\mathbf{w}}_m$ records the importance of the $m$-th view of the original features. Given $\tilde{\mathbf{w}}_1$ and $\tilde{\mathbf{w}}_2$, we eliminate the least important WM and GM features, respectively. Then, CCA is applied to transform the remained WM/GM features to an updated common space. This transforming-eliminating scheme is iteratively executed till the number of iterations exceeds a pre-defined threshold. In other words, the iterations are stopped, when the classification performance in the subsequent steps does not increase anymore.

### Robust Linear Discriminant Analysis (RLDA)

In this study, we use the robust discriminant analysis (RLDA) [14] to classify PD from the normal subjects based on the selected features. Let $\tilde{\mathbf{X}} \in \mathbf{R}^{n\times \tilde{d}}$ be the matrix containing the $\tilde{d}$-dimensional samples possibly corrupted by noise. We know that an amount of noise can be introduced in the data, as the data acquisition and preprocessing steps are not error-free. Even a small number of noisy features or outliers can affect a model significantly. In RLDA, noisy data is modeled as $\tilde{\mathbf{X}} = \mathbf{D} + \mathbf{E}$, where $\mathbf{D}$ is the underlying noise-free component and $\mathbf{E}$ contains the outliers. RLDA learns the mapping $\mathbf{t}$ from $\tilde{\mathbf{X}}$ to fit the class labels in $\mathbf{y} \in \mathbf{R}^{n\times 1}$. RLDA decomposes $\tilde{\mathbf{X}}$ into $\mathbf{D}$ and $\mathbf{E}$, and computes the mapping $\mathbf{t}$ using the noise-free data $\mathbf{D}$, which yields the following objective function:

$$\underset{\mathbf{t},\mathbf{D},\mathbf{E}}{\arg\min}\frac{\eta}{2}\left\|\left(\mathbf{y}^{\mathrm{T}}\mathbf{y}\right)^{-\frac{1}{2}}(\mathbf{y}-\mathbf{HDt})\right\|_2^2+\|\mathbf{D}\|_*+\boldsymbol{\lambda}\|\mathbf{E}\|_1 \quad \text{s. t. } \tilde{\mathbf{X}}=\mathbf{D}+\mathbf{E}. \tag{4}$$

The normalization factor $(\mathbf{y}^T\mathbf{y})^{-1/2}$ compensates for different numbers of samples per class. $\mathbf{H} = (\mathbf{I}_n - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)$ is a centering matrix and $\mathbf{t} \in \mathbb{R}^{d\times 1}$ denotes the mapping, which is learned only from the centered noise-free data $\mathbf{HD}$. Therefore, it will avoid projecting the noise factor $\mathbf{E}$ to the output space, thus results in an unbiased estimation. After $\mathbf{t}$ is learned in the training stage, a testing data, $\tilde{\mathbf{x}}_{\text{test}}$, is projected by $\mathbf{t}$ onto the output space spanned by $\tilde{\mathbf{x}}_{\text{test}}\mathbf{t}$, then the class label of the test data can be determined by $k$-Nearest Neighbor (k-NN) algorithm. The RLDA formulation can be solved using augmented Lagrangian method, as detailed in [14].

## 3 Experimental Results

The data used in this paper were obtained from the Parkinson's Progression Makers Initiative (PPMI) database. In this paper, we use 112 subjects (56 PD, and 56 NC), each with a T1-weighted MR scan using 3T SIMENS MAGNETOM TrioTim Syngo with the following parameters: acquisition matrix = 256 × 256 $mm^2$, 176 slices, voxel size = 1 × 1 × 1 $mm^3$, echo time (TE) = 2.98 ms, repetition time (TR) = 2300 ms, inverse time = 900 ms, and flip angle = 9°.

In order to evaluate the proposed and the baseline methods, we used an 8-fold cross-validation. For each of the 8 folds, feature selection and classification models were established using the other 7 folds as the training subset, and the diagnostic ability was evaluated with the unused $8^{th}$ testing fold.

We compared our ICCA-based feature selection with three popular feature selection or feature reduction methods, including PCA, sparse feature selection (SFS), one-shot CCA and minimum redundancy-maximum relevance (mRMR). No feature selection (NoFS) was also regarded as a baseline method. Table 1 shows the performances of all the methods. As can be seen, the proposed ICCA-based feature selection method achieves the best performance. In particular, our method improves by 11.5 % on ACC and 16 % on AUC, respectively, compared to the baseline (NoFS). Moreover, compared to feature selection

through PCA, SFS CCA and mRMR, our method achieves 5.3 %, 5.4 %, 4.4 % and 3.8 % accuracy improvements, respectively.

In order to further analyze the proposed ICCA-based feature selection, we also implement another experiment, in which we select canonical representations rather than the original WM/GM features. That is, in every iteration of ICCA, we directly eliminate the least important canonical representations, instead of using the inverse-transform and elimination steps in the original WM/GM features. The left canonical representations are used for re-estimating the new common space by CCA and further selected. The last two lines in Table 1 show the experiment results, where the performance of the ICCA-based feature selection in the canonical space is not better than the performance of the selection in the original WM/GM feature space. These results indicate that canonical representations are not better than the original features (after proper selection) for distinguish PD from NC. This can be due to the fact that the CCA mapping to the common space is unsupervised, and, after the two feature vectors are mapped, they are highly correlated. As a result, there are many redundant data in the canonical representations and that could mislead the feature selection and the classification. Inverse-transforming the representations and going back to the original feature space (using our proposed ICCA framework) avoids this shortcoming.

To identify the biomarkers of PD, we further inspect the selected features, which correspond to specific WM/GM areas. Since the features selected in each cross-validation fold may be different, we define the most discriminative regions as features, which were selected at least 60 % of the times in cross-validation. The most discriminative regions, as shown in Fig. 2, include ‘precuneus’, ‘thalamus’, ‘hippocampus’, ‘temporal pole’, ‘postcentral gyrus’, ‘middle frontal gyrus’, and ‘medial frontal gyrus’. GM and WM features extracted from these regions are found to be closely associated with PD pathology, which are in line with previous clinical researches [15, 16].

## 4 Conclusion

In this paper, a novel feature selection technique was proposed to help identify individuals with PD from NC. The proposed ICCA-based feature selection framework can achieve a fairly local linear mapping capability. Moreover, it can dig deeper into the underlying structure of the feature space and explore the relationship among features. By increasing the depth of learning in ICCA framework, the two views of the selected features would be closer and closer, when mapped to their CCA common space. This also decreases the number of the selected features. The results show that the proposed ICCA feature selection framework outperforms conventional feature selection methods, and can improve the diagnosis ability based on T1 MR images.

Note that, in the proposed framework of ICCA-based feature selection, we discard a pair of features from GM and WM at each iteration. In order to avoid dropping the possibly important features, we drop out the WM/GM features conservatively in each iteration. Dropping out the features in a smarter way could optimize the whole feature selection framework. This can be pursued in the future works. Furthermore, current ICCA-based feature selection can only explore relationship between two views of the features. We will

investigate the possibility of handling more views of features simultaneously, which can effectively enhance the feasibility of the proposed method.

Besides further optimizing the efficiency and performance of the ICCA framework, the future work includes improving the classification method for PD diagnosis. RLDA is a linear classifier, which cannot model the nonlinear relationship between features and labels. Therefore, some nonlinear classifiers can probably perform at least equally or better than the linear classifier. In this study, we only used structural information in T1 MR images; we will explore the integration of other imaging modalities such as diffusion tensor imaging (DTI) and functional MRI in the future to further improve the classification performance based on the proposed framework.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (NSFC) Grants (Nos. 61473190, 61401271, 81471733).

## References

1. Calne DB, Snow BJ. Criteria for diagnosing Parkinson's disease. Ann Neurol. 1992; 32:125–127.
2. Goebel G, Seppi K, et al. A novel computer-assisted image analysis of [123I] 3-CIT SPECT images improves the diagnostic accuracy of parkinsonian disorders. Eur J Nucl Med Mol Imaging. 2011; 38:702–710. [PubMed: 21174092]
3. Tsanas A, Little MA, et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE TBME. 2012; 59:1264–1271.
4. Wenning GK, et al. What clinical features are most useful to distinguish definite multiple system atrophy from Parkinson's disease. J Neurol Neurosurg Psychiatry. 2000; 68:434–440. [PubMed: 10727478]
5. Singh G, Samavedham L. Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: a case study on early-stage diagnosis of Parkinson disease. J Neurosci Methods. 2015; 256:30–40. [PubMed: 26304693]
6. Ye J, et al. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurol. 2012; 12:1. [PubMed: 22289169]
7. Ye J, Liu J. Sparse methods for biomedical data. ACM SIGKDD Explor Newsl. 2012; 14:4–15.
8. Lu Y, et al. Feature selection using principal feature analysis. ACM-MM. 2007
9. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. Neu Comp. 2004; 16:2639–2664.
10. Zhu, X., Suk, H-I., Shen, D. Multi-modality canonical feature selection for Alzheimer's disease diagnosis. In: Golland, P.Hata, N.Barillot, C.Hornegger, J., Howe, R., editors. MICCAI 2014, Part II. LNCS. Vol. 8674. Springer; Heidelberg: 2014. p. 162-169.
11. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005; 27:1226–1238. [PubMed: 16119262]
12. Smith SM, et al. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage. 2004; 23:208–219.
13. Shen D, Davatzikos C. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE TMI. 2002; 2:1421–1439.
14. Huang, D., Cabral, RS., De la Torre, F. Robust regression. In: Fitzgibbon, A.Lazebnik, S.Perona, P.Sato, Y., Schmid, C., editors. ECCV 2012, Part IV. LNCS. Vol. 7575. Springer; Heidelberg: 2012. p. 616-630.
15. Hanakawa T, Katsumi Y, et al. Mechanisms underlying gait disturbance in Parkinson's disease. Brain. 1999; 122:1271–1282. [PubMed: 10388793]

16. Burton EJ, McKeith IG, et al. Cerebral atrophy in Parkinson's disease with and without dementia: a comparison with Alzheimer's disease, dementia with Lewy bodies and controls. Brain. 2004; 127:791–800. [PubMed: 14749292]

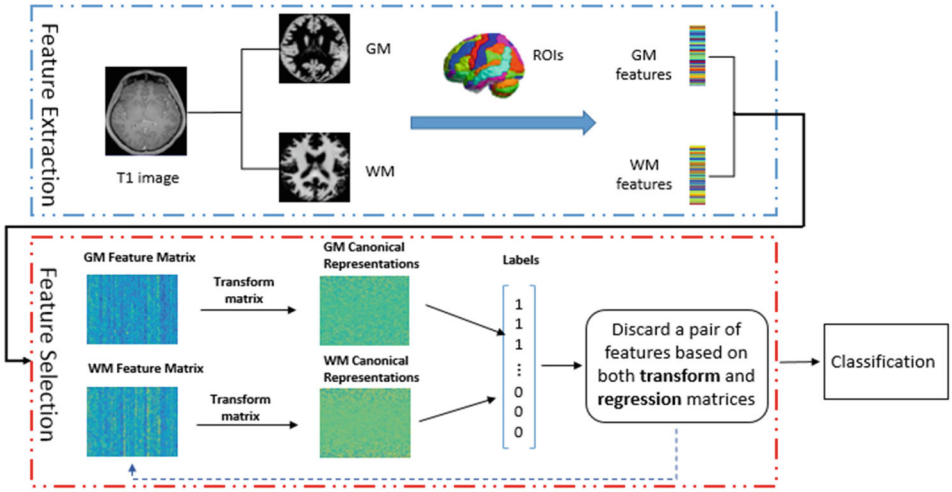


**Fig. 1.**
Pipeline of the ICCA-based feature selection and PD classification.

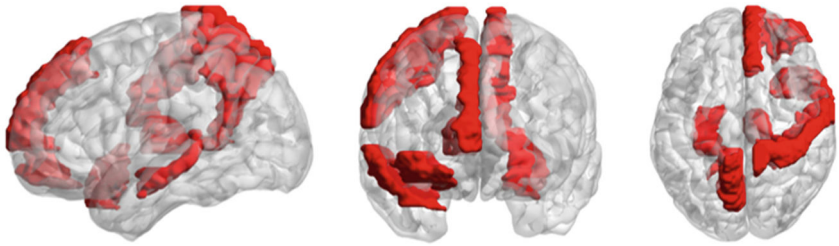

**Fig. 2.**
The most discriminative ROIs for automatic diagnosis of PD.

**Table 1**

PD/NC classification comparison (ACC: Accuracy; AUC: Area Under ROC Curve).

| | ACC (%) | AUC (%) |
|---|---|---|
| NoFS | 58.0 | 55.1 |
| SFS | 65.1 | 64.5 |
| PCA | 65.2 | 59.8 |
| CCA | 66.1 | 64.4 |
| mRMR | 66.7 | 65.6 |
| Proposed (Select in canonical feature space) | 68.8 | 69.3 |
| Proposed (Select in WM/GM feature space) | **70.5** | **71.1** |