Medical Image Analysis 61 (2020) 101632

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Adaptive sparse learning using multi-template for neurodegenerative disease diagnosis

Baiying Lei[a], Yujia Zhao[b], Zhongwei Huang[b], Xiaoke Hao[c], Feng Zhou[d], Ahmed Elazab[a], Jing Qin[e], Haijun Lei[b,*]

[a] *National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China*

[b] *School of Computer Science and Software Engineering, Guangdong Province Engineering Center of China-made High Performance Data Computing System, Shenzhen Key Laboratory of Service Computing and Applications Shenzhen University, Shenzhen 518060, China*

[c] *School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China*

[d] *Department of Industrial and Manufacturing Systems Engineering, University of Michigan, Dearborn 48128, USA*

[e] *School of Nursing, Hong Kong Polytechnic University, Hung Hom, Hong Kong*

ARTICLE INFO

Article history:
Received 27 September 2019
Revised 17 December 2019
Accepted 20 December 2019
Available online 8 January 2020

KeyWords:
Neurodegenerative disease diagnosis
Adaptive sparse learning
Feature learning
Multi-template Multi-classification

ABSTRACT

Neurodegenerative diseases are excessively affecting millions of patients, especially elderly people. Early detection and management of these diseases are crucial as the clinical symptoms take years to appear after the onset of neuro-degeneration. This paper proposes an adaptive feature learning framework using multiple templates for early diagnosis. A multi-classification scheme is developed based on multiple brain parcellation atlases with various regions of interest. Different sets of features are extracted and then fused, and a feature selection is applied with an adaptively chosen sparse degree. In addition, both linear discriminative analysis and locally preserving projections are integrated to construct a least square regression model. Finally, we propose a feature space to predict the severity of the disease by the guidance of clinical scores. Our proposed method is validated on both Alzheimer's disease neuroimaging initiative and Parkinson's progression markers initiative databases. Extensive experimental results suggest that the proposed method outperforms the state-of-the-art methods, such as the multi-modal multi-task learning or joint sparse learning. Our method demonstrates that accurate feature learning facilitates the identification of the highly relevant brain regions with significant contribution in the prediction of disease progression. This may pave the way for further medical analysis and diagnosis in practical applications.

1. Introduction

Neurodegenerative diseases, such as Parkinson's disease (PD) (Marek et al., 2011) and Alzheimer's disease (AD) (Alzheimer's, 2015) are among the most common neurological disorders in the elderly people (Adeli et al., 2016). PD is a long-term degenerative disorder of the central nervous system that mainly affects the motor system. AD is a chronic neurodegenerative disease that destroys memory and other important mental functions. Since the symptoms of these degenerative diseases of nervous systems appear progressively, patients in middle or late stage suffer from various inconveniences and endless pains, even life-threatening problems (Aerts et al., 2012). Apart from motor symptoms, non-motor symptoms such as depression, anxiety, and sleep disorders also degrade patients' quality of life (Braak et al., 2003).

There is no known cure for neurodegenerative diseases to date (Nilashi et al., 2016; Schaffer et al., 2015). The current diagnosis mainly depends on clinical symptoms, the clinicians' knowledge and experience (Nilashi et al., 2016; Postuma et al., 2015). Meanwhile, the conventional clinical symptoms adopted for diagnosis only occur when the relevant biomarkers already show the progression of the lesions (Weiner et al., 2006). Therefore, patients diagnosed using the traditional approaches are mostly at middle or late stage of such diseases. To accurately identify different stages and improve analytical effectiveness to reduce patients' suffering, early automatic medical diagnosis is highly desirable for detecting their progression of these diseases.

Common prodromal stages of neurodegenerative diseases include scan without evidence of dopaminergic deficit (SWEDD) for PD (Marek et al., 2011) and mild cognitive impairment (MCI) for AD. MCI is further divided into light MCI (lMCI, who suffers from

[*] Corresponding author.
E-mail addresses: leiby@szu.edu.cn (B. Lei), lhj@szu.edu.cn (H. Lei).

https://doi.org/10.1016/j.media.2019.101632
1361-8415/© 2020 Elsevier B.V. All rights reserved.

light MCI) and stable MCI (sMCI, whose symptoms are stable and will not progress to AD in 18 months) (Alzheimer's, 2015). SWEDD presents the clinical symptoms without obvious dopaminergic deficit, which has high potential of PD onset.

Unlike most previous binary classification tasks, we consider a multi-class classification problem (e.g., PD vs. SWEDD vs. NC) for practical applications. From practical clinical application point of view, it is more effective to build a multi-class classifier than binary classifier as only one diagnosis decision is required. Since these neurodegenerative diseases are progressive and multiple prodromal stages may occur, multiple prodromal stage patients can be recognized for targeted intervention treatment before the nervous system gets severely damaged.

To date, numerous studies show that neuroimaging techniques are promising for computer-aided diagnosis (CAD) of these diseases (Lei et al., 2017a; Lei et al., 2017b; Lei et al., 2017c; Prashanth et al., 2014; Wei et al., 2017; Zhu et al., 2016a, 2016b). For example, magnetic resonance imaging (MRI) and diffusion-weighted tensor imaging (DTI) can reveal structural abnormalities of the brain, while positron emission tomography (PET) or functional MRI (fMRI) can capture functional abnormalities of the brain (Salvatore et al., 2014). Many recent studies utilized neuro-images to predict and assess the stage of diseases by machine learning techniques. For instance, Rana *et al.* (Rana et al., 2014) proposed a machine learning approach for PD diagnosis with T1-weighted MRI images. Fung *et al.* (Fung and Stoeckel, 2007) utilized spatial information for feature selection and classification from single-photon emission computed tomography images for AD.

Generally, a typical CAD pipeline for neurodegenerative disorders consists of data acquisition, feature extraction, feature selection, and classification (Zhang and Shen, 2012). Information provided by neuroimages is often of high dimension, which leads to the overfitting issue with a limited sample size. To tackle this issue, a feature selection method is typically applied (Jothi and Hannah, 2016; Ozcift, 2012) to find a subset of features. Feature selection is capable of simplifying the prediction model and avoids the curse of dimensionality, thus enhancing the generalization ability of the prediction model. Methods like subspace learning can also achieve this goal by transforming the original data space into a low-dimensional space (Seeley et al., 2009; Wang et al., 2016). In regard to the interpretability of brain features, feature selection methods are preferable compared to subspace learning methods. However, it is reasonable to combine feature selection and subspace learning to build an interpretable as well as accurate disease diagnosis prediction model (Zhu et al., 2016b). Motivated by this, we combine both feature selection and subspace learning into a unified framework to select the most discriminative features for automatic diagnosis.

In addition, we build a feature selection model based on the sparse least square regression. Since we may encounter multiple different classification tasks of neurodegenerative diseases, different degrees of sparseness may be required according to the specific feature relationships and properties in different tasks. Adaptive sparse learning is an appealing method since it adapts the sparseness degree to achieve a better recognition rate (Grandvalet, 2002), and an adaptive strategy is employed to control the sparseness degree in our unified model. In other words, the ratio of zeros in a weight matrix can be adjusted according to the classification task.

It is known that single-template based methods obtain the simple morphometric representation of each brain image via a certain nonlinear registration method. In contrast, multi-template based methods are more promising to discover disease status and compare group difference (Liu et al., 2016b). It is suggested in the previous studies (Jin et al., 2015; Liu et al., 2016a; Min et al., 2014) that learning with multiple templates can boost diagnosis accuracy. For example, Min *et al.* (Min et al., 2014) utilized concatenated multi-template based features of each subject and achieved promising AD classification results. Multiple templates not only represent the brain information in a comprehensive way, but also capture the disease-related discriminative information (Liu et al., 2016b). Also, multi-template based methods can extract multiple feature sets of a subject derived from different templates (Jin et al., 2015; Liu et al., 2016a; Min et al., 2014), which can effectively reduce the negative impacts of registration errors and provide distinct yet complementary information to identify different disease status. It thus leads to more promising identification performance. Also, by concatenating the multi-template based features of each subject, more promising identification results can be achieved.

Inspired by this, we use multiple atlases with different sets of regions of interest (ROIs) to extract different sets of features from the brain images. These different features are fused together to enhance classification performance by constructing a more discriminative and larger space of features with a reduced dimension. Specifically, we use an automatic anatomical labeling (AAL) atlas (Tzouriomazoyer et al., 2002) for 90 and 116 regions and Craddock's spatially constrained spectral clustering atlas (Craddock et al., 2012) for 200 regions since the AAL atlas is the most widely used atlas for brain regions extraction. The available full brain regions of AAL template are 116 ROIs and 90 ROIs with cerebellar. The 200 ROIs are obtained from Craddock's spatially constrained spectral clustering atlas (Craddock et al., 2012). More ROIs increases the interpretability since more information may be provided. Craddock offers multiple ROIs larger than 200, but we select 200 ROIs as higher numbers of ROIs increases the difficulty of efficiently extracting features. Finally, we fuse these features together by linear concatenation.

On this note, we propose a multi-template based adaptive feature selection method to build a reliable classification model. Also, we integrate linear discriminative analysis (LDA) (Lin et al., 2010) and locally preserving projections (LPP) (Zhu et al., 2016b) to construct the most informative subspace with an adaptive sparse regularization (Zhang et al., 2011). LDA considers the global information by weighing the proportion of within-class-variance and between-class-variance, while LPP reflects the local information by finding the similarity relevance within each feature. With the help of global and local information in data, we select the most discriminative features and discard those irrelevant features to enhance the classification performance in the learned feature subspace (Zhang and Ye, 2011). Different from existing methods focusing only on binary classification task with single template, we simultaneously classify multiple different clinical statuses using multiple templates for practical clinical application.

The rest of this paper is organized as follows. Section II reviews various feature selection and subspace learning methods for neurodegenerative diseases recognition. Section III introduces the methodology of the proposed method. Experimental results are presented in Section IV. Discussions and conclusions are provided in Section V and VI, respectively.

2. Related work

2.1. Feature selection

Due to the challenge of high dimensionality and limited sample size, the overfitting problem could occur in data- driven analysis (Kong et al., 2014). To address this problem, most existing methods design a feature selection process to select most discriminative neuroimaging features or a sample selection process to discard the redundant samples (Fung and Stoeckel, 2007; Lei et al., 2017a). A l_1-regularizer (i.e., a sparse term) was introduced in the estimation model for feature selection when the sample size is significantly smaller than the feature dimensionality

(Jothi and Hannah, 2016). Also, a group least absolute shrinkage and selection operator (LASSO) $l_{2,1}$-regularizers was also introduced to further improve the effectiveness. By setting the weight of irrelevant features to zero, the whole weight matrix can reach a sparse situation (Chen et al., 2009). However, these methods just simply regulate the weight matrix with sparse constraint, which fails to consider to the relationship constraints. To improve it, the relationships between modalities are considered in a feature selection model via multimodal multi-task learning (Zhang and Shen, 2012). Also, more data relationships are explored in a joint sparse learning model (Lei et al., 2017b). The multiple subspace learning method can take more relationship information into consideration when dealing with informative neuroimaging data by exploring relationships from many perspectives.

2.2. Subspace learning

For feature representation, the subspace learning method has shown great potential in various prediction tasks, especially for unsupervised dimensionality reduction (Zhu et al., 2016b). For example, Wang *et al.* proposed sparse multi-view task-centralized ensemble classification method (Wang et al., 2019). Zhou et al. proposed the latent representation learning framework using three different modalities (Zhou et al., 2019a). Zhou *et al.* maximally utilized multimodal neuroimaging and genetic data via a stage-wise deep neural network to discriminate AD and its early stages (Zhou et al., 2019b). However, these methods do not fully consider the underlying structural information in multiple views or modalities. Recently, mixed kernel canonical correlation analysis via mapping the high dimensional data space into the kernel Hilbert space rather than the Hilbert space was proposed (Jia and Fu, 2016). In (Lei et al., 2017b), a self-taught dimensionality reduction with a novel joint graph sparse coding model was proposed to significantly improve the prediction effectiveness. Also, subspace learning has been applied to reveal the intrinsic relationship within data such as principal component analysis, LDA (Lin et al., 2010), locally linear embedding (Roweis and Saul, 2000), and Laplacian eigenmaps (Belkin and Niyogi, 2003). Regarding the interpretability of brain features, feature selection methods are preferable compared to subspace learning methods, particularly in neuroimaging studies, as the selected features are directly linked to the anatomy and provide an intuitive understanding. From a clinical point of view, a model for disease diagnosis should be interpretable and accurate. Hence, it is reasonable to combine feature selection and subspace learning in a systematic manner. In this study, we use LDA and LPP to construct feature subspaces using multi-template learning. By considering the feature relationships in the feature selection framework, the performance for neurodegenerative disease diagnosis can be boosted.

3. Methodology

The overview of our multi-class classification method is presented in Fig. 1. The preprocessing pipeline is the same as (Lei et al., 2017b). First, we preprocess the original brain T1-weighted MRI image by the statistical parametric mapping (SPM) tool for segmentation (Friston, 2003). Then, we extract the tissue volume in the segmented regions with AAL atlas. Then we calculate the corresponding tissue volume values as feature vectors and concatenate them linearly for feature representation.

We perform adaptive feature selection combined with subspace learning to obtain the most discriminative features on the concatenated features. We further add an adaptive p-norm regularization to the subspace learning methods (LDA and LPP) for feature selection. The clinical scores are added as the additional features to build our final feature matrix for training the classifiers. Finally, we use support vector machine (SVM) with the sigmoid kernel to classify the samples into different groups in a supervised way.

3.1. Problem formation

Given m subjects and each has n features, we start with the fundamental linear regression model $\mathbf{Y} = \mathbf{WX}$, where $\mathbf{Y} \in \mathbb{R}^{m \times c}$ is the ground truth label vector, c is the number of classes, $\mathbf{X} \in \mathbb{R}^{m \times n}$ is the input data matrix, and $\mathbf{W} \in \mathbb{R}^{n \times c}$ is the weight coefficient matrix. In this paper, we denote row vector as A_i and each column is denoted as A^j. We can obtain $\mathbf{W}$ by solving the following objective function

$$\min_{\mathbf{W}} ||\mathbf{Y} - \mathbf{XW}||_F^2, \tag{1}$$

where $\mathbf{A}_F$ is the Frobenius norm (F-norm) of $\mathbf{A}$. The F-norm is also known as the $\mathfrak{l}_2$-norm or the $\mathfrak{l}_2$-regularizer, which is defined as $\mathbf{A}_F = \sqrt{\sum_i A_{i2}^2}$. Eq. (1) is a simple and straightforward linear regression model without constraint on any variables. In addition, this equation does not take into account the properties of weight matrix, resulting in an inferior performance.

3.2. Adaptive feature selection

Feature selection and subspace learning are effective methods (Zhu et al., 2016a) to improve the classification performance. It is known that, Fisher's LDA combined with LPP makes use of both global and local information (Zhu et al., 2016b). The reason is the regularized linear regression model identifies the most discriminative and relevant features for classification performance boosting. The $\mathfrak{l}_{2,1}$-norm of X, $\mathbf{X}_{2,1} = \sum_i \mathbf{x}_{i2}$. In general, the linear prediction model is defined as

$$\min_{\mathbf{W}} \frac{1}{2} ||\mathbf{Y} - \mathbf{XW}||_F^2 + \lambda ||\mathbf{W}||_{2,1}, \tag{2}$$

where $\mathbf{W}$ represents the regression weight matrix. The first term in Eq. (2) controls the overall data fitting while the second term ensures the sparsity level of $\mathbf{W}$, and λ is the hyperparameter. However, Eq. (2) only selects the smallest features without considering the complex intrinsic relationships within the data space. Therefore, the current form of Eq. (2) cannot guarantee the class-discriminative power of the selected features and the preservation of the intrinsic structure of data points, which are vital characteristics for good classification performance. To locate the most distinct features, we construct a uniform subspace. Also, we utilize the neighborhood structure of the original data to solve this problem. Following the Fisher's LDA criterion, we add a regularization term to penalize the objective function of Eq. (2) (Zhu et al., 2016b), which is denoted as

$$Ratio = \frac{\mathbf{W}^T \sum_w \mathbf{W}}{\mathbf{W}^T \sum_b \mathbf{W}}, \tag{3}$$

where $\Sigma \boldsymbol{w}$represents the within-class variance and $\Sigma b$is the between-class variance. Minimizing Ratio ensures that we can get a $\mathbf{W}$ with relatively small within-class variance and large between-class variance. Since it is time-consuming and complex to find an optimal solution of Eq. (2) due to its non-convexity, an equivalent way can be applied to minimize Ratio (Ye, 2007) by defining the label class matrix as

$$y_{i,k} = \begin{cases} \sqrt{\frac{m}{m_k}} - \sqrt{\frac{m_k}{m}}, & if\ label(\mathbf{x}_i) = k, \\ -\sqrt{\frac{m_k}{m}}, & otherwise, \end{cases} \tag{4}$$

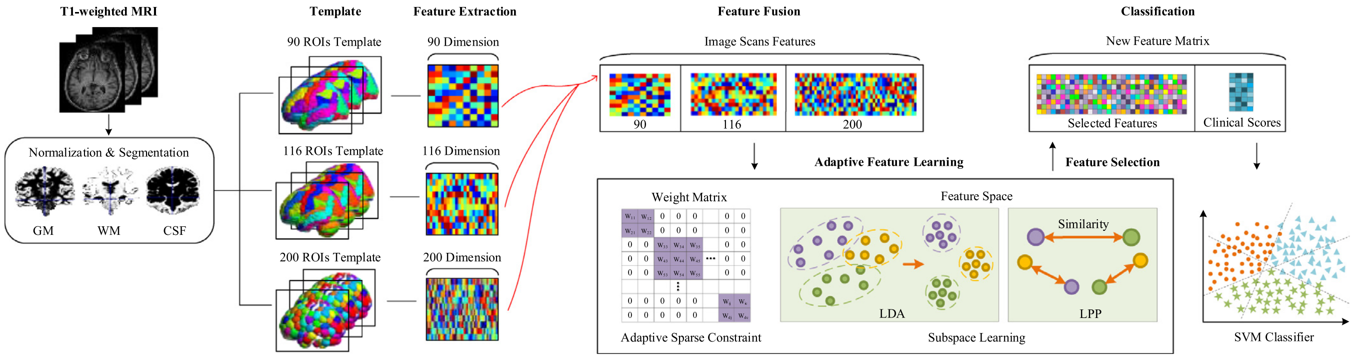


Fig. 1. Framework of our proposed method. First, we preprocess the original brain T1-weighted MRI image by the SPM. Then, we extract the tissue volume in the segmented regions with AAL. Then we calculate the corresponding tissue volume values as feature vectors. We fuse multi-tempalate features together by linear concatenation. We perform an adaptive feature selection combined with subspace learning to obtain the most discriminative features on the concatenated features. Finally, we use SVM with the sigmoid kernel to classify the samples into different groups.

where $label(\mathbf{x}_i)$ is the label of the subject $\mathbf{x}_i$ and m_k is the subject amount of class k. Using the redefined class matrix, we can utilize the global information of data distribution of original data space to construct our subspace.

For the relationship between data, LPP is used to maintain the local relation within data (Zhu et al., 2016b). We use the graph Laplacian method to define the similarity $s_{i,\,j}$ between subject $\mathbf{x}_i$ and $\mathbf{x}_j$ and the regularization term is given as

$$R_L = tr\left(\sum_{\boldsymbol{i},\boldsymbol{j}}\left(\mathbf{W}^{\boldsymbol{T}}\mathbf{x}_i - \mathbf{W}^{\boldsymbol{T}}\mathbf{x}_j\right)^2 s_{i,j}\right), \tag{5}$$

where $\mathbf{S} = [s_{i,j}] \in \mathbb{R}^{m\times m}$ denotes the affinity matrix of subjects. In LPP, it first constructs a neighborhood graph for data $\mathbf{X}$ using k-nearest neighbor and then computes the value of $\mathbf{S}$ by summing each row equals to 1. Then, we can reformulate our objective function as

$$\min_{\mathbf{W}} \frac{1}{2}||\mathbf{Y}-\mathbf{XW}||_F^2 + \lambda_1 tr\left(\sum_{\boldsymbol{i},\boldsymbol{j}}\left(\mathbf{W}^{\boldsymbol{T}}\mathbf{x}_i - \mathbf{W}^{\boldsymbol{T}}\mathbf{x}_j\right)^2 s_{i,j}\right) + \lambda_2||\mathbf{W}||_{2,1}, \tag{6}$$

where λ_1 and λ_2 are the parameters controlling the regularization terms. The features are then selected by a threshold value. Specifically, the average weight value of all features is calculated as the threshold to adjust the features, accordingly. The weights of features less than the threshold value will be set to zero. The threshold value can be initially adjusted according to average weight.

We aim to reduce the feature dimension to get the most discriminative features, and multiple regularizers are adopted to enhance performance. Instead of $\ell_{2,1}$norm, we introduce a generalized $\ell_{2,p}$ norm in our method for adaptive sparseness control, which is defined as

$$||\mathbf{W}||_{2,p}^p = Tr\left(\mathbf{W}^T\mathbf{W}\right)^{\frac{p}{2}}, \tag{7}$$

Note that, the $\ell_{2,p}$ norm is a general form of the trace norm to enable us to flexibly search a suitable solution by adjusting the value of p. Therefore, our final objective function is

$$\min_{\mathbf{W}} \frac{1}{2}||\mathbf{Y}-\mathbf{XW}||_F^2 + \lambda_1 tr\left(\sum_{\boldsymbol{i},\boldsymbol{j}}\left(\mathbf{W}^{\boldsymbol{T}}\mathbf{x}_i - \mathbf{W}^{\boldsymbol{T}}\mathbf{x}_j\right)^2 s_{i,j}\right) + \lambda_2||\mathbf{W}||_{2,p}^p, \tag{8}$$

Therefore, we can jointly combine subspace learning and feature selection, where each feature has a representative weight value.

Unlike traditional regularizers, $\ell_{2,p}$ norm is designed to regulate the objective function and construct a subset of most discriminative and relevant features. By minimizing $\ell_{2,p}$ norm, we can mitigate the influence of noisy and less relevant rows in $\mathbf{W}$. The parameter p is tuned to control the sparseness degree. Smaller p value indicates lower correlation of different modalities. Hence, smaller p value is suitable for feature selection since $\mathbf{W}$ approximates the low rank of the matrix. The p value is limited to 2 since there is no sparsity when p is equal to 2. The adjusted p value can help to uncover underlying information among different modalities. Hence, our model is adaptive to learn discriminative features for the classification task. Eventually, we train the regression model to select the most distinct features with the limited size.

Our method can be discriminated from the previous methods in the following aspects: First, unlike the previous sparse linear regression-based feature selection methods, the proposed method finds the class-discriminative and noise-resistant regression coefficient matrix due to the Fisher's criterion and Laplacian graph. Second, different from the subspace learning methods such as PCA, LDA, and LPP, the proposed method selects features from the original space to facilitate the investigation of the results. Third, differing in the conventional LDA in Eq. (2), the proposed method adopts the Fisher's criterion but still operates on the original feature space, and thus allows for an intuitive interpretation of the selected features. Fourth, the conventional least square regression uses traditional regularizes ($l_{2,\,1}$ norm), while our $\ell_{2,p}$ norm is designed to regulate the objective function and construct a subset of most discriminative and relevant features (Rana et al., 2014; Salvatore et al., 2014; Zhang & Shen, 2012). Therefore, our model is capable of learning discriminative features for the classification task.

3.3. Multi-classification model

Differing in the preceding methods that merely perform a binary classification, a multi-class classification is adopted for PD diagnosis in our method. In machine learning, SVM is a supervised learning model used for pattern classification. The main idea of SVM is to find the best hyperplane that can separate different class samples with the maximum margin. Hence, we choose the SVM to construct a multi-class classification model. The free and available software LIBSVM toolbox (version: libsvm-2.91) is used to perform the classification task (Luo et al., 2014).

In binary SVM classification, we make use of the outputs of the prediction function. In case of multi-class classification, the output of SVM prediction will change. For example, the dimension of decision values is equal to the number of all possible binary clas-

Algorithm 1 Solving Eq. (8).

```
Input:   X: The data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, $m$ denotes the number of subjects and $n$ denotes the dimension of features
         $\overline{\mathbf{Y}}$ : The target matrix $\overline{\mathbf{Y}} \in \mathbb{R}^{m\times k}$, $k$ denotes the number of target classes
Output:  1: Initializing $\bar{\mathbf{W}}$ as $w_{ij} = 0$ for $(i, j) \in \Omega$;
         2: repeat
         3: Calculate $\nabla f(\mathbf{W}_{(t)}) = (\mathbf{XX^T} + \lambda_1\mathbf{XLX^T})\mathrm{W}_{(t)} - \mathbf{XY^T}$.
         4: Calculate$\mathbf{W}_{(t+1)} = \arg\min_{\mathrm{W}} \mathbf{W} - \mathbf{W}_{(\boldsymbol{t})} + \frac{1}{\lambda(t)}\nabla f(\mathbf{W}_{(t)})_F^2 + \frac{\lambda_2}{\lambda(t)}\mathbf{W}_{2,p}^p$
         5: $t$=$t$+1.
         6: until convergence
         $\bar{\mathbf{W}} \in \mathbb{R}^{n\times k}$ the weight matrix
```

sification combinations. For subjects belonging to k class, we can select all the binary combination. Then, a total of n×(n-1)/2 classification models are constructed. The final performance of multi-class is obtained from the best result in all the n×(n-1)/2 binary combinations. ultimate goal is to find the best **W** in Eq. (7). Since Eq. (7) is a convex but non-smooth function, we solve it by designing a new accelerated proximal gradient method (Prashanth et al., 2014; Seeley et al., 2009). We first conduct the proximal gradient method on Eq. (7) by setting

$$f(\mathbf{W}) = ||\mathbf{Y} - \mathbf{XW}||_F^2 + \lambda_1 tr\left(\sum_{\boldsymbol{i},\boldsymbol{j}}\left(\mathbf{W^T x}_i - \mathbf{W^T x}_j\right)^2 s_{i,j}\right), \tag{9}$$

$$L(\mathbf{W}) = f(\mathbf{W}) + \lambda_2||\mathbf{W}||_{2,p}^p, \tag{10}$$

Note that, $f(\mathbf{W})$ is convex and differentiable while $\lambda_2||\mathbf{W}||_{2,p}^p$ is convex but non-smooth (Prashanth et al., 2014). To optimize **W** with the proximal gradient method, we iteratively update it using the following optimization rule

$$\mathbf{W}_{(t+1)} = \arg\min_{\mathbf{W}} \boldsymbol{P}_{\lambda(t)}\left(\mathbf{W}, \mathbf{W}_{(t)}\right), \tag{11}$$

where $\boldsymbol{P}_{\lambda(t)}(\mathbf{W}, \mathbf{W}_{(t)})\, f(\mathbf{W}_{(t)}) + \langle\nabla f(\mathbf{W}_{(t)}), \mathbf{W} - \mathbf{W}_{(t)}\rangle + \lambda(t)||\mathbf{W} - \mathbf{W}_{(t)}||_F^2 + \lambda_2||\mathbf{W}||_{2,p}^p$, $\nabla f(\mathbf{W}_{(t)}) = (\mathbf{XX}^T + \lambda_1\mathbf{XLX}^T)\mathbf{W}_{(t)} - \mathbf{XY}^T$ $\mathbf{L} = \mathbf{D} - \mathbf{S}$, and **D** is a diagonal matrix $\mathbf{D} = [d_{i,j} = \sum_j s_{i,j}] \in \mathbb{R}^{m\times m}$. The$\lambda(t)$ and $\mathbf{W}_{(t)}$ are the corresponding tuning parameter at the t-th iteration.

By ignoring the independent terms of **W** in Eq. (10), we can rewrite this equation as

$$\mathbf{W}_{(t+1)} = \boldsymbol{E}_{\lambda(t)}\left(\mathbf{W}_{(t)}\right) = \arg\min_{\mathbf{W}} ||\mathbf{W} - \mathbf{W}_{(t)} + \frac{1}{\lambda(t)}\nabla f\left(\mathbf{W}_{(t)}\right)||_F^2 + \frac{\lambda_2}{\lambda(t)}||\mathbf{W}||_{2,p}^p. \tag{12}$$

where $\boldsymbol{E}_{\lambda(t)}(\mathbf{W}_{(t)})$ is the Euclidean projection of $\mathbf{W}_{(t)}$ onto the convex set $\lambda(t)$. Then, we can find a closed form solution of each row of $\mathbf{W}_{(t+1)}$. We can finally solve Eq. (7) by the accelerated proximal gradient method to iteratively update the value of **W** (Jothi & Hannah, 2016; Nesterov, 2004). The algorithm is summarized in Algorithm 1.

4. Experiments and results

In our study, we use the two publicly available datasets, PPMI (Marek et al., 2011) and ADNI (Alzheimer's, 2015) to compare the proposed method with other widely used methods such as ElasticNet and LASSO (Tibshirani, 1996). We also compare the proposed method with other state-of-the-art feature selection methods applied for neurodegenerative disease diagnosis: multi-modal multi-task (M3T) (Zhang and Shen, 2012), joint sparse learning for classification and regression (JSL) (Lei et al., 2017b), multi-kernel SVM (mSVM) (Zhang et al., 2011) and sparse learning which has no p-norm regularization (SL). We conduct all the experiment on the same datasets and evaluation standards for those methods to achieve a fair comparison.

4.1. Experimental setup

The model parameters in our method are the tuning parameters in Eq. (7), λ_1 and λ_2, and p. We set the initial values as:$\lambda_1\{10^{-5}, \ldots, 10^5\}$,$\lambda_2 \in \{10^{-5}, \ldots, 10^5\}$, and $p \in \{0.1, \ldots, 2\}$.

The final values of the hyperparameters are decided by adjusting the values until the overall performance reaches the peak in the range intervals. We automatically choose the best values as the fine-tuned parameters. We set the C and G parameters of the SVM classifier as $C \in \{2^{-10}, \ldots, 2^{10}\}$and $G \in \{2^{-10}, \ldots, 2^{10}\}$, respectively, to automatically choose the result with the highest accuracy. We choose sigmoid kernel for SVM classifier as the kernel function. The quantitative measurements for the classification performance include classification accuracy (ACC), sensitivity (SEN), precision (PREC), specificity (SPEC), F-scores (F1), and area under the receiver operating characteristic curve (AUC). Additionally, a 10-fold cross-validation strategy is applied in our experiments to split the origin data into training and testing groups. These quantitative measurements are assessed on the classification results of testing data and the average validation results are obtained and presented.

4.2. Data preparation

The data used in this experiment is PD and AD subjects acquired from the PPMI database and ANDI database. PPMI is the first internationally recognized observational study created to identify and validate biomarkers for prediction of PD progression. ADNI is the most comprehensive effort to identify neuroimaging measures and biomarkers associated with cognitive and functional changes in healthy elderly subjects and subjects who have MCI and AD. These publicly available datasets contain an inclusive set of clinical and behavioral assessments, brain neuroimaging scans, and several biological specimens. The first category of features is from the T1-weighted MRI scans.

All the obtained original MRI images are preprocessed by the anterior commissure-posterior commissure correction. Then, the images are processed by skull-stripping and cerebellum removal if 90 brain regions are needed for subsequent processing. For MRI data, we segment the image into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) using the SPM segmentation toolbox. These three instances are the remarkable anatomical indicators in the brain images (Eusebi et al., 2017). The GM and WM are two essential and vulnerable components of the central nervous system, which plays a critical role in assessing therapeutic impacts and determining prognoses. The CSF is a fluid surrounding the brain and spinal cord that indicates pathological condition of the central nervous system.

In our study, we get at least 90, 116, or 200 ROIs in the brain based on the atlas templates. We then compute the mean tissue density value of each region as features. The more regions we segment, the more feature dimensions we get. The more detailed information, the harder feature extraction. If we simply augment the dimension, there may be insufficient extracted features from the images. If we want every feature extracted from the region to be effective, the regions shall not be too large. Therefore, there are tradeoffs for the experimental settings.

In this work, we collect a total of 238 subjects including 62 NC, 142 PD, and 34 SWEDD subjects from PPMI database and a total of 814 subjects including 220 NC, 192 AD, and 402 MCI subjects from ANDI database. Among the middle progression stage MCI, we further divide the 402 subjects into 146 lMCI patients and 256 sMCI patients. The subjects are collected under the criteria that they can be successfully segmented and visualized, and features must be extracted with a complete scale. Those subjects cannot smoothly reach the final feature extraction step will be excluded. For each disease subject, we acquire a MRI scan volume. After preprocessing the MRI images, we collect 90, 116, and 200 tissue volumes for GM, WM, and CSF, respectively. We compute the mean tissue density value of each region in GM, WM, and CSF as features.

Apart from these features, we also collect the clinical assessment scores as features. The clinical scores for PD include sleep, olfaction, depression, and Montreal cognitive assessment. These scores reflect the non-motor symptoms that are unrevealed by the imaging data. The clinical scores for AD are mini-mental state examination scores measuring cognitive impairment. These clinical assessments contain information obtained from questionnaires answered by patients or their health care professionals, prevailing to estimate the severity and progression of non-motor state impairment of PD and AD.

Currently, the diagnosis adopted by most doctors in the clinical practice is based on the clinical scores [8]. The diagnosis for PD is performed using the assessment of motor symptoms such as shaking, rigidity, slowness of movement, and postural instability. The diagnosis for AD also refers to the assessment of relative clinical symptoms. However, there is a period of 5 to 20 years between the start of neurodegeneration and the exhibition of the clinical symptoms. During this period, the patients mainly show non-motor symptoms. These subtle symptoms are insufficient for disease diagnosis because neurodegeneration already exists as imaging data shows.

4.3. Feature combinations

In our experiments, we perform multi-class classification NC vs. PD vs. SWEDD, NC vs. AD vs. MCI, and NC vs. AD vs. lMCI vs. sMCI. Using the 10-fold cross-validation method, for each subset of experiment, we train the feature selection model by different feature combination sets, i.e., G + C, W + C, and G + W + C (G for GM, W for WM, C for CSF) for each subset of experiment. After the feature selection, we combine clinical scores (S) with the selected feature matrix. We run a series of experiments to analyze the effectiveness of features over different categories and different scales. We now collect MRI images scans features in different feature dimensions and clinical assessments scores features.

Table 1 lists all the possible combinations from single type to multiple types and shows the classification results with these different feature combinations. 90 ROIs means using features with a dimension of 90 for each type (i.e., G or W). 116 ROIs and 200 ROIs share the similar definition. Multi-ROIs means using fused feature vectors by linearly concatenating 90 ROIs, 116 ROIs, and 200 ROIs. We can clearly see that classification performance differs for different classification tasks.

First, we found that WM has the least supportive impact almost on all the classification tasks. The CSF has the least supportive impact on some of the AD classification tasks. This also indicates that GM has the most affecting feature to PD and AD compared with the other two neuroimaging features. Second, we observe that fea-

Table 1
Classification accuracy (mean ± standard deviation) of different groups (G for GM, W for WM, C for CSF, S for clinical scores).

Feature	NC vs. PD vs. SWEDD				NC vs. AD vs. MCI				NC vs. AD vs. lMCI vs. sMCI			
	90 ROIs	116 ROIs	200 ROIs	Multi-ROIs	90 ROIs	116 ROIs	200 ROIs	Multi-ROIs	90 ROIs	116 ROIs	200 ROIs	Multi-ROIs
G	65.27 ± 6.76	65.68 ± 6.70	61.41 ± 4.45	65.22 ± 4.81	56.89 ± 4.46	56.89 ± 4.46	60.31 ± 3.25	62.51 ± 5.00	47.54 ± 4.31	47.54 ± 4.31	51.23 ± 2.05	51.72 ± 3.10
W	60.91 ± 3.15	61.39 ± 3.43	60.86 ± 2.51	62.79 ± 3.15	54.53 ± 3.07	55.15 ± 3.65	55.39 ± 3.08	55.89 ± 5.07	44.34 ± 2.94	45.33 ± 1.35	43.73 ± 4.48	44.10 ± 2.54
C	63.06 ± 4.07	63.11 ± 3.96	61.36 ± 3.59	64.77 ± 4.66	55.42 ± 3.14	54.68 ± 3.47	56.25 ± 4.14	56.14 ± 3.97	43.49 ± 4.47	43.49 ± 4.47	45.82 ± 1.90	45.36 ± 3.19
G+W	62.35 ± 3.80	62.74 ± 2.22	62.41 ± 4.45	63.84 ± 3.27	48.40 ± 3.66	56.15 ± 3.16	60.69 ± 5.73	61.30 ± 5.73	48.40 ± 3.66	47.66 ± 2.66	50.60 ± 4.05	51.10 ± 4.54
G+C	67.84 ± 3.89	68.66 ± 3.10	65.41 ± 4.45	68.84 ± 5.09	59.95 ± 6.08	59.09 ± 5.09	61.78 ± 5.32	63.29 ± 3.81	50.48 ± 2.15	49.88 ± 3.12	50.74 ± 2.25	51.87 ± 2.31
W+C	63.88 ± 4.07	64.46 ± 4.21	64.86 ± 4.35	64.94 ± 6.76	57.49 ± 5.03	57.86 ± 4.66	58.21 ± 5.02	58.60 ± 4.38	47.67 ± 2.76	47.16 ± 4.45	46.67 ± 3.99	47.55 ± 3.79
G+W+C	66.79 ± 3.98	68.03 ± 3.52	63.89 ± 3.82	66.74 ± 3.54	60.68 ± 4.21	59.83 ± 5.76	61.79 ± 4.67	63.63 ± 5.76	51.23 ± 3.19	51.23 ± 3.42	46.67 ± 3.99	53.20 ± 4.13
S	68.37 ± 5.93	66.24 ± 4.29	63.32 ± 3.43	67.12 ± 3.72	58.98 ± 4.36	59.18 ± 3.21	62.85 ± 4.17	65.48 ± 3.26	48.32 ± 5.10	49.07 ± 3.89	53.29 ± 3.27	55.61 ± 3.54
G+W+S	75.57 ± 5.67	75.60 ± 6.30	74.65 ± 6.17	75.07 ± 7.10	66.03 ± 3.38	70.55 ± 4.94	70.63 ± 5.17	74.81 ± 5.80	62.65 ± 6.04	62.03 ± 3.22	61.05 ± 5.13	64.36 ± 3.26
G+C+S	**76.75 ± 6.31**	**77.30 ± 6.21**	**75.58 ± 8.90**	**78.23 ± 5.85**	74.55 ± 5.62	74.81 ± 4.28	72.11 ± 5.72	75.21 ± 6.02	63.27 ± 2.94	63.38 ± 3.11	60.68 ± 3.25	64.61 ± 3.54
W+C+S	75.75 ± 4.47	76.79 ± 6.23	75.06 ± 8.52	77.09 ± 5.73	73.46 ± 5.72	72.10 ± 610	70.88 ± 5.08	74.19 ± 5.58	62.65 ± 6.04	62.03 ± 3.22	61.05 ± 4.59	63.01 ± 3.01
G+W+C+S	75.10 ± 3.97	76.57 ± 5.94	76.20 ± 7.64	77.75 ± 5.70	**75.80 ± 4.80**	**76.32 ± 5.64**	**73.08 ± 4.98**	**77.48 ± 5.12**	**64.23 ± 4.51**	**63.87 ± 8.09**	**61.66 ± 2.52**	**64.97 ± 3.29**

Table 2
P-value generated from t-test strategy comparing multiple templates in different classification tasks.

Template	NC vs. PD vs. SWEDD	NC vs. AD vs. MCI	NC vs. AD vs. lMCI vs. sMCI
90 ROIs	0.0116	0.0457	0.0121
116 ROIs	0.0487	0.0251	0.0445
200 ROIs	0.0195	0.0451	0.0380
Multi-ROIs	–	–	–

ture combinations perform better than single feature as more information is incorporated in the input data. Among these 2-feature and 3-feature combinations (4–7th row of Table 1), the impact is different on PD and AD experiments as G+C is the highest for PD subjects and G+W+C is the highest for AD subjects. This indicates that WM has less impact on PD than AD when it comes to diagnosis decisions. Third, we observe that clinical scores may achieve a bit better performance compared with imaging data features (G, W, C). However, the combination of imaging data and clinical scores performs the best. These three classification tasks are all increased about 10% in terms of classification accuracy compared to those without scores. This suggests that clinical scores-guided features can significantly improve the performance since motor and neurodegeneration information is considered. The best feature combination chosen for NC vs. PD vs. SWEDD are G + C + S and the best feature combination chosen for NC vs. AD vs. MCI and NC vs. AD vs. lMCI vs. sMCI is G + W + C + S.

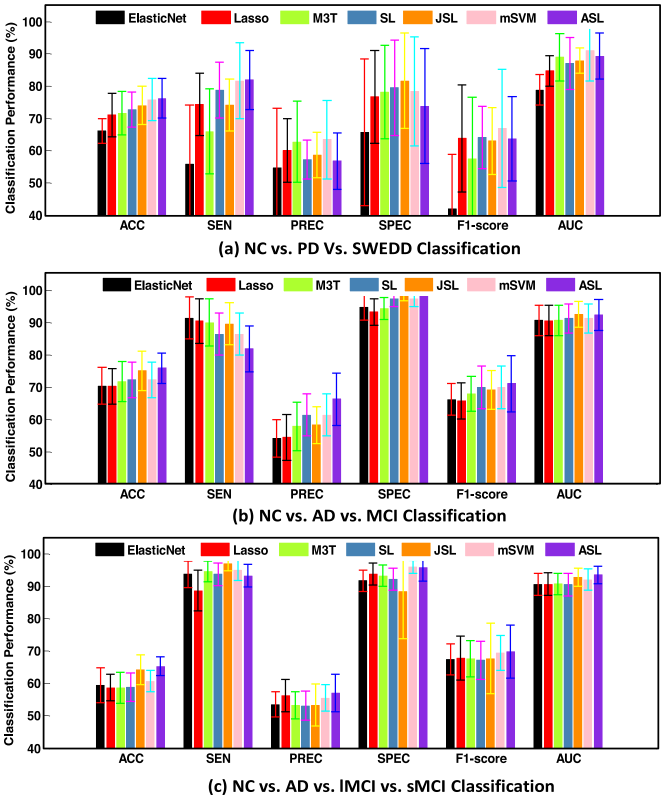
We also notice that, there are three segmentation ROI templates and a feature fusion scheme adopted in each experiment. Each modality provides feature dimensionality of 90, 116, 200, and 406. Meanwhile, it is not revealed that which ROI templates better reduce the redundancy and further boost classification performance.

To analyze the effectiveness of ROIs templates, all the same feature types are applied on the 90, 116, and 200 segmented brain regions experiments. It is observed that 116 ROIs achieve the highest performance among these three single templates, which indicates that 116 ROIs offer the suitable number of regions for disease recognition. Also, the performance of 200 ROIs template is concluded to be less effective for the multi-classification. The reason is that more accurate segmentation succeeds in providing larger feature dimension but fails to provide representative features. These features may not be extracted successfully in this larger template due to the smaller segmented regions. Some columns of the feature values obtained from the 200 ROIs template turn out to be zero where no feature is found. To better balance the effect of different templates, we fuse these three ROI templates together by linear concatenation. The results of feature combinations on multi-ROIs are shown in Table 1.

It is observed that our multi-template based method is marginally higher than the best single-template based method. To further validate the performance of our method, we apply the *t*-test (Student's *t*Test) on these comparison results in Table 2, where the *p*-value indicates the probability that the results from sample data occurred by chance. Lower *p*-values indicate lower chance of randomness. In most cases, a *p*-value of 0.05 is acceptable, which means the data is valid. In Table 2, all the obtained *p*-values of multi-template and single-template are lower than 0.05. Based on these results, we believe that the multi-template method is better than single-template method.

4.4. Classification performance

In this section, we apply our multi-template adaptive sparse learning multi-class classification model on the feature combinations selected from IV-C for NC vs. PD vs. SWEDD, NC vs. AD vs. MCI, and NC vs. AD vs. lMCI vs. sMCI classification tasks. We compare our method with ElasticNet, Lasso, M3T, SL, and JSL.



(a) NC vs. PD Vs. SWEDD Classification

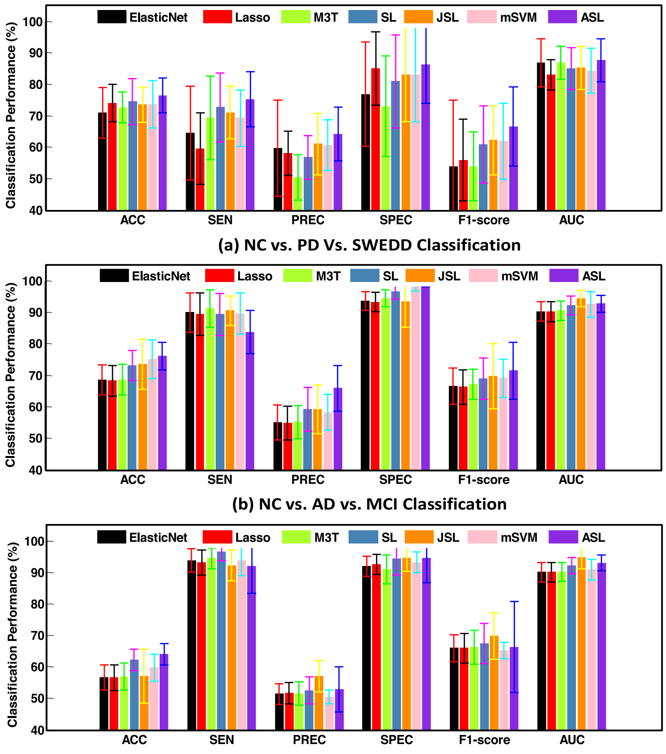
(b) NC vs. AD vs. MCI Classification

(c) NC vs. AD vs. lMCI vs. sMCI Classification

Fig. 2. Algorithm comparison performance via 90 segmented brain regions.

Figs. 2, , , and –5 show the classification results of the competing methods of different classification tasks on 90, 116, 200 regions and multi-ROIs fused features. The first row of each figure represents the result of NC vs. PD vs. SWEDD. The second and third rows of each figure are the result of NC vs. AD vs. MCI and NC vs. AD vs. lMCI vs. sMCI. In these figures, bar heights and the error lines above the bars represent mean values and standard deviation of the obtained classification performances during several cross-validations.

We observe that the proposed ASL method achieves an accuracy of 78.23% in the 3-class PD classification task with multi-ROIs features, 77.48% in the 3-class AD classification task with multi-ROIs features, and 64.97% in the 4-class classification task with multi-ROIs features. All the six evaluation measurements show similar fluctuations in AD experiments. However, the evaluation measurements in PD experiments are not always consistent as LASSO reaches high in SPEC but low in SEN in the 116 ROIs. Higher SPEC means higher true negative rate while higher SEN means higher true positive rate. Therefore, this result indicates that LASSO is more efficient on diagnosing normal controls than selecting all disease samples. We observe that the proposed method is superior to the competing methods in terms of various metrics especially ac-

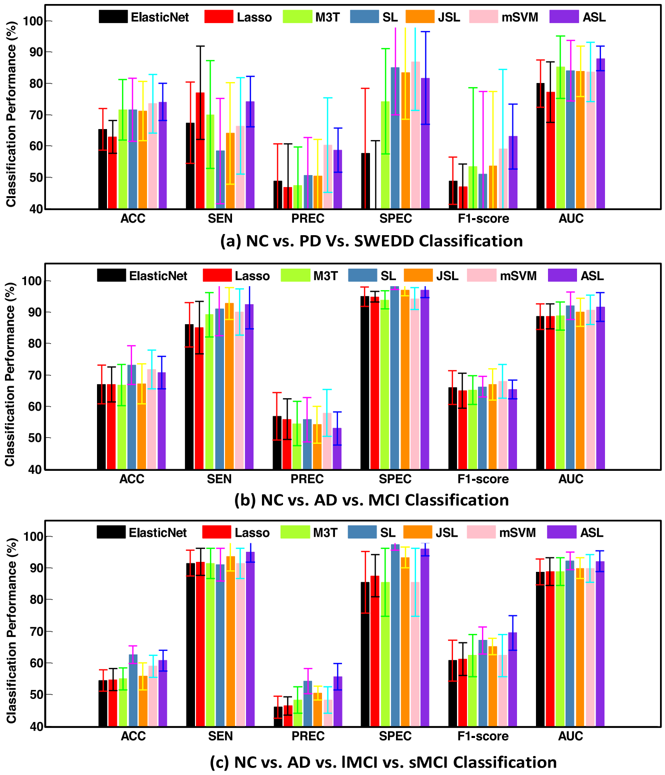


(a) NC vs. PD Vs. SWEDD Classification

(b) NC vs. AD vs. MCI Classification

(c) NC vs. AD vs. lMCI vs. sMCI Classification

Fig. 3. Algorithm comparison performance via 116 segmented brain regions.



(a) NC vs. PD Vs. SWEDD Classification

(b) NC vs. AD vs. MCI Classification

(c) NC vs. AD vs. lMCI vs. sMCI Classification

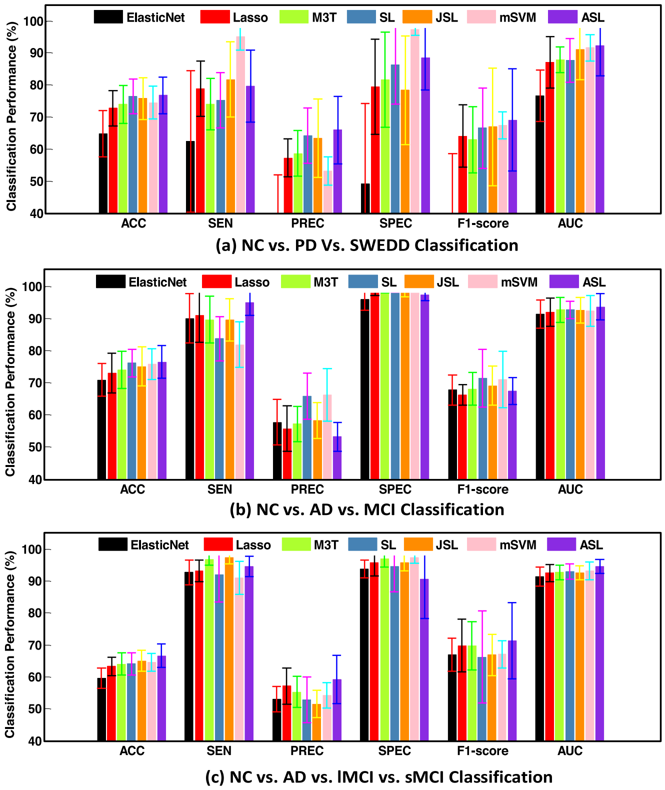
Fig. 4. Algorithm comparison performance via 200 segmented brain regions.

curacy and AUC. We also observe that more segmented regions do not have a significant impact on the classification performance. It is clear that ASL performs better than other existing methods and eventually proves that combining feature selection and subspace learning boost neurodegenerative disease diagnosis.

Experiments show that LASSO achieves the worst classification performances among all the methods. The main reason is LASSO only selects features with sparseness regularization, which are insufficient to correctly classify the neuroimaging features. On the other hand, the SL method outperforms the LASSO and ElasticNet methods, which makes it reasonable to integrate subspace learning into the feature selection framework. Moreover, the proposed method clearly outperforms both conventional feature selection and joint sparse learning methods due to the combination of two approaches.

We also plot the receiver operating characteristic curves (ROC) of each classification experiment in Fig. 6. These curves correspond to the comparison results in Fig. 5 using the multi-ROIs features. As we can see from the ROC curves, the AUC values in NC vs. PD vs. SWEDD classification indicate that our method prevails over others. Obviously, the proposed adaptive feature selection method enhances the classification performances. The corresponding ROC curves for NC vs. AD vs. MCI and NC vs. AD vs. lMCI vs sMCI are also shown in Fig. 6. Our method plots curves relatively higher than other competing methods. Our method shows a higher AUC value despite that all methods achieve a relatively promising result.

Another finding is the differences between PD and AD experiment. We see that PD experimental results of different methods fluctuate larger than AD experiment especially the ROC curves. The different data and feature structure affect the performance of all methods. In the PD experiment, our method outperforms others despite the fluctuation, and also proves that our method with

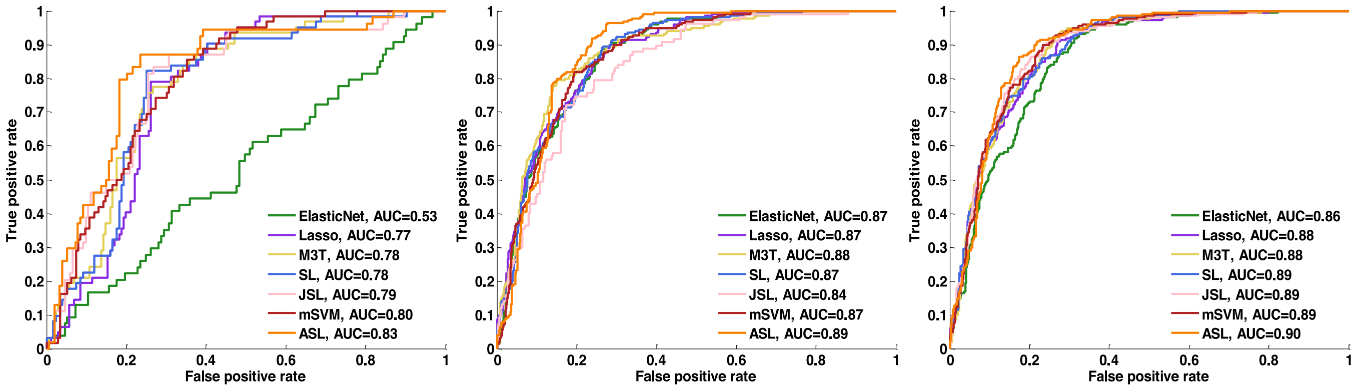


(a) NC vs. PD Vs. SWEDD Classification

(b) NC vs. AD vs. MCI Classification

(c) NC vs. AD vs. lMCI vs. sMCI Classification

Fig. 5. Algorithm comparison performance via multi-ROIs fused features.



(a) NC vs. PD Vs. SWEDD Classification

(b) NC vs. AD vs. MCI Classification

(c) NC vs. AD vs. lMCI vs. sMCI Classification

Fig. 6. ROC plots comparison of competing methods on different classification tasks using multi-ROIs fused features.

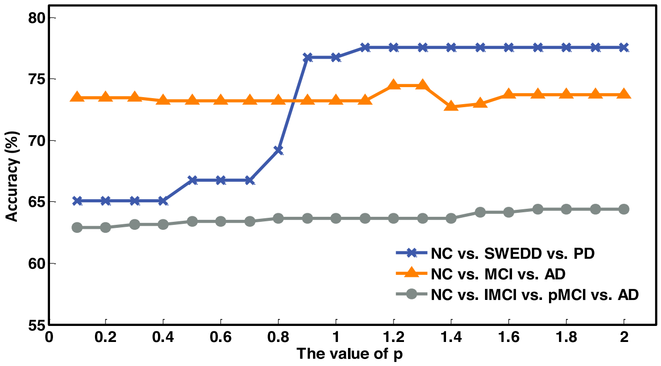


Fig 7. Classification performance of our method ASL on different classification tasks with different p values using multi-ROIs fused features.

feature selection and subspace learning benefits the classification task.

In this experiment, we also observe that the differences between multi-classification and binary classification. The best classification accuracy is only around 76%, but the multiple binary classification accuracy results are at least 90%. The main reason is that classifying samples into multiple classes increases the difficulty of prediction models. Furthermore, we conduct experiments on other methods, which are originally used, for binary classification and the results show superiority of our method over others.

4.5. The performance of adaptiveness

We perform the experiment of the multi-class classification NC vs. PD vs. SWEDD and NC vs. AD vs. MCI, and NC vs. AD vs. lMCI vs. sMCI to validate the effectiveness of adaptive p-norm regularization. The classification accuracy of each task is compared with the p-value range {0.1,..,2} with a step size of 0.1. The experimental results are obtained from the multi-ROIs features. In Fig. 7, we clearly see that the p values significantly affect the overall classification performance. Different classification tasks require different sparseness regularization even for different segmented regions to adapt itself to the environments. In the NC vs. PD vs. SWEDD classification task, the classification accuracy fluctuates larger with an overall uprising trend compared with the other two tasks. On the other hand, p-value fluctuates less and goes to a steady value. The best p-value is then used in the feature selection model for higher classification accuracy. The adaptive sparseness allows us to select the best sparse degree to control the sparse feature learning according to specific tasks. Overall, our method proves to be superior to the previous methods.

4.6. Related ROIs

The clinical symptoms traditionally used for neurodegenerative disease diagnosis start to appear when relevant biomarkers shows the progression of the lesions after more than 60% of dopaminergic neurons are lost (Hall et al., 2015; Schaffer et al., 2015).

For PD, the motor symptoms however begin to occur in very late stages of the disease when the dopamine concentration is significantly reduced by up to 80% (Emrani et al., 2017). Early diagnosis is therefore a main challenge in the field of neurodegenerative diseases therapeutics. In fact, the absence of a validated indicator of disease is the major impeding factor in understanding disease progression to develop treatments that can delay, prevent, or reverse disease progress. Identification of diagnostic biomarkers and progression monitoring are of highly significance in early diagnosis. On this note, we move forward to use the adaptive sparse learning multi-classification model to study the most predictive features of the progression and their correspondences, i.e., the most relevant and discriminative brain regions related to neurodegenerative diseases.

To find the most predictive and important neurodegenerative disease-related brain regions, we use the best features listed in Table 2 to identify the most overall important disease-related brain regions. The experiments are performed on the multi-ROIs fused features for NC vs. PD vs. SWEDD and NC vs. AD vs. MCI and NC vs. AD vs. lMCI vs. sMCI classification tasks. We obtain the weight assigned to each feature from the learned multi-class classification model. We utilize the weight coefficient matrix $\bar{\mathbf{W}}$ generated in feature selection process. The weight matrices present the corresponding significance of all the 90 brain regions.

We choose the top 10 regions with the highest 10 weight values in 90 ROIs after cerebellum removal. The selected brain regions are numbered by the brain section index and can be further investigated for clinical practice. We choose the top 10 regions with the highest weight value in a descending order and delete the repeated regions. The top 10 relevant brain regions selected from multi-class classification are visualized in Fig. 8 in the three different planes (sagittal, axial, and coronal). Detailed brain regions information is presented in Table 3.

Fig. 8 shows the top 10 chosen features distribution with the highest top 10 weights; color range indicates the feature location from the most important to the 10th lowest important. Table 3 summarizes all the chosen brain ROIs in medical terms from 3 classification tasks. We have the following observations regarding the feature contributions for prediction of neurodegenerative disease progression: 1) Hippocampus left and right are the most important brain regions for NC vs. PD vs. SWEDD and NC vs. AD vs. MCI. 2) Putamen is the most important brain regions for

Table 3
Indices and names of recognized brain ROIs from our classification tasks using ALL template with 116 ROIs.

NC vs. PD vs. SWEDD		NC vs. AD vs. MCI		NC vs. AD vs. lMCI vs. sMCI	
ROIs index	ROI names	ROIs index	ROI names	ROIs index	ROI names
38	Hippocampus_R	37	Hippocampus_L	74	Putamen_R
80	Heschl_R	68	Precuneus_R	37	Hippocampus_L
41	Amygdala_L	40	ParaHippocampal_R	43	Calcarine_L
44	Calcarine_R	14	Frontal_Inf_Tri_R	68	Precuneus_R
28	Rectus_R	39	ParaHippocampal_L	39	ParaHippocampal_L
86	Temporal_Mid_R	71	Caudate_L	40	ParaHippocampal_R
79	Heschl_L	38	Hippocampus_R	14	Frontal_Inf_Tri_R
87	Temporal_Pole_Mid_L	52	Occipital_Mid_R	38	Hippocampus_R
77	Thalamus_L	55	Fusiform_L	9	Frontal_Mid_Orb_L
42	Amygdala_R	67	Precuneus_L	48	Lingual_R

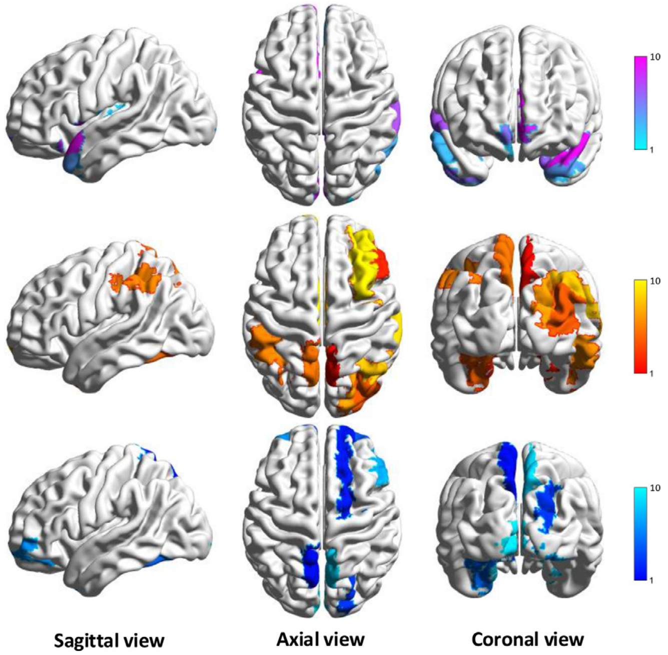


Fig. 8. Top 10 important brain regions, disease-related ROIs overall distribution from the NC. vs. PD. vs. SWEDD classification task (first row), NC vs. AD vs. MCI classification task (second row), NC vs. AD vs. lMCI vs. sMCI classification task (third row).

NC vs. AD vs. lMCI vs. sMCI. 3) Among the three groups of the selected regions, the common ROIs confirmed is Hippocampus. This indicates that Hippocampus is the most predictive brain region of the neurodegenerative disease and it has been verified in existing medical studies (Blennow et al., 2010; Hall et al., 2015). 4) Beside Hippocampus, Amygdala is also the confirmed predictive region for PD (Huang et al., 2015). We also discover the potential brain regions of PD such as Heschl, Calcarine, and Rectus.

In the experiment of AD, regions chosen for the two classification tasks are different from each other despite some similarities. Caudate, Occipital_Mid (middle occipital gyrus), and Fusiform are found in the 3-class classification but not found in the 4-class classification. Putamen, Calcarine, Frontal_Mid_Orb (superior frontal gyrus and medial orbital) and Lingual are found in the 4-class classification but not in the 3-class classification. These regions are distinctive for the different classification task even in the same AD experiments. It indicates that special attention shall be paid on these brain regions that have significant contribution on the prediction of disease progression. It requires different focus on the brain regions when differentiating sample with different scale of stages. From the distribution of Fig. 9, we can observe the chosen features for PD experiment gather closely with each other. The gathering region may be closely related with the dysfunction of movement control and non-motor problems such as depression and anxiety. For AD experiment, the feature distribution is almost consistent within these two classification tasks with minor differences.

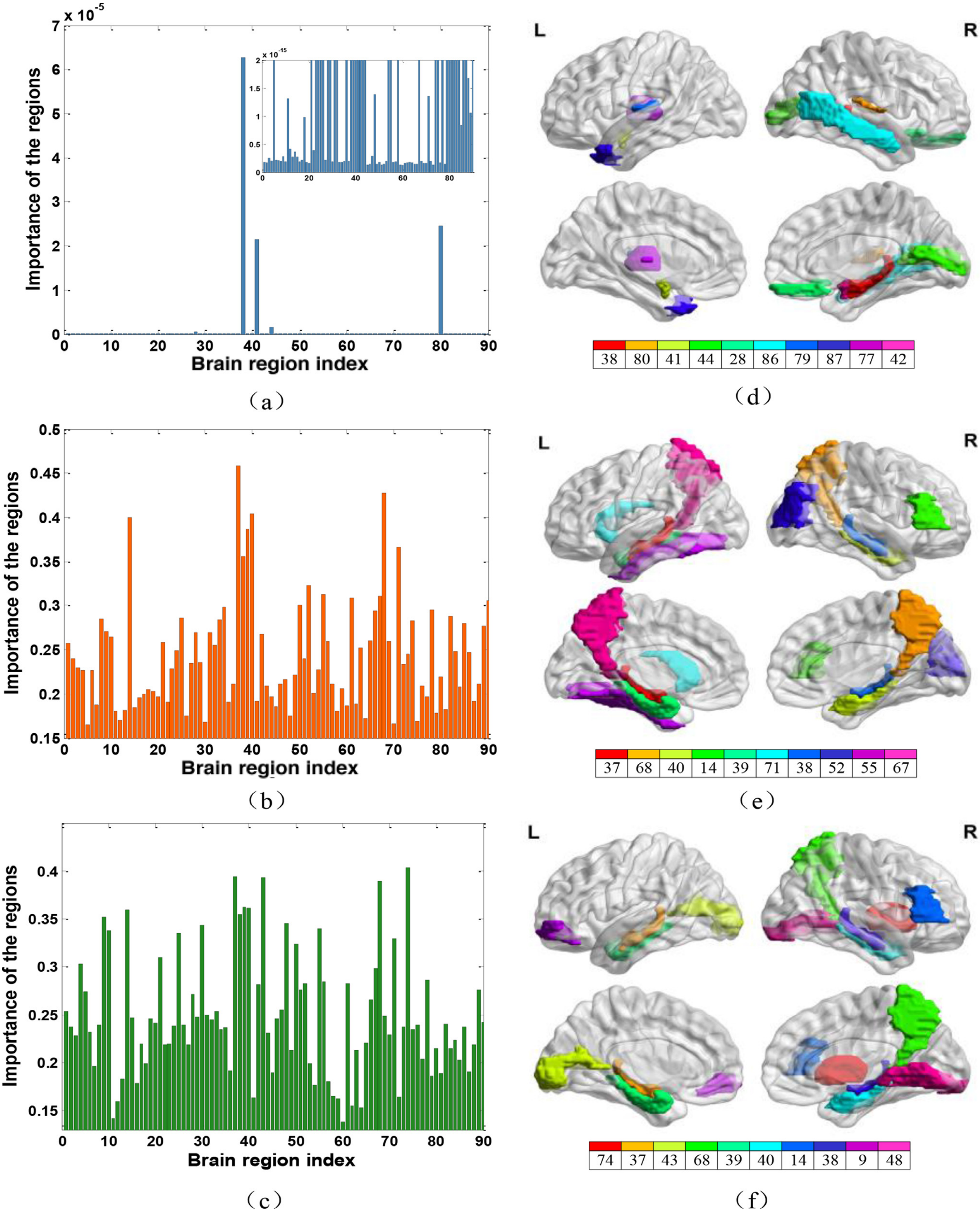
We further visualize the importance of all 90 regions (cerebellum removal) and their detailed position in Fig. 9. The left bar plot in the left column of Fig. 9 shows the actual weight value extracted from the weight matrix **W**. The actual value may be different due to the classification tasks and different datasets. The corresponding top 10 important brain regions are shown in the right column of Fig. 9. The chosen brain regions are the same results presented in Table 3 and Fig. 8. We observe that there is a huge difference between PD experiment and AD experiment. The average value of corresponding weight is much smaller than that of AD experiments. The highest value for weight value in PD experiment is around 0.6×10^{-5} while the highest weight value in AD experiment is around 0.45. The explanation is the selected brain images and the clinical assessment scores are totally unrelated evaluation. In addition, we obtain four different clinical scores for PD patients and only one clinical score for AD experiment. Corresponding obtained weight average values discrepancy will not affect the comparison of feature diagnostic importance.

We focus on the relative feature significance on the different three classification tasks. We observe that the weight values fluctuations are greatly distinctive where the highest values stand out much ahead. The detailed value distribution is plotted in the small box in the bar plot of PD experiment. The large difference of that value observed in PD experiment indicates the discriminative ability of PD-related features are more distinctive than those of AD-related features.

It is also clear that the importance is slightly different between NC vs. AD vs. MCI and NC vs. AD vs. lMCI vs. sMCI classification. The reason is that the contribution of these regions may differ for 3-class and 4-class classifications with similar overall trend. The selected brain regions are numbered by the brain section index and can be further investigated and supervised in clinical practice.

5. Discussions

We investigate the importance of the brain regions via the frequency of the selected ROIs by the proposed method using MRI images. To further study the relationship between brain regions and neurodegenerative disease, we attempt to identify the other top brain regions that are most correlated with other brain regions under the assumption that disease-related ROIs affect each other. We use the weighting matrix $\bar{\mathbf{W}}$ to calculate the Pearson correlation coefficient to represent the correlation among different regions. The results are visualized in Fig. 10. We plot the 116 brain regions and the arcs between them indicate the correlation be-



（a）

（d）

（b）

（e）

（c）

（f）

Fig. 9. The importance of 116 brain regions and selected top 10 disease-related ROIs from NC. vs. PD. Vs. SWEDD classification task (first row), NC vs. AD vs. MCI classification task (second row), NC vs. AD vs. lMCI vs. sMCI classification task (third row). The subfigure of (a) is the zoom of the original figure for better observation.

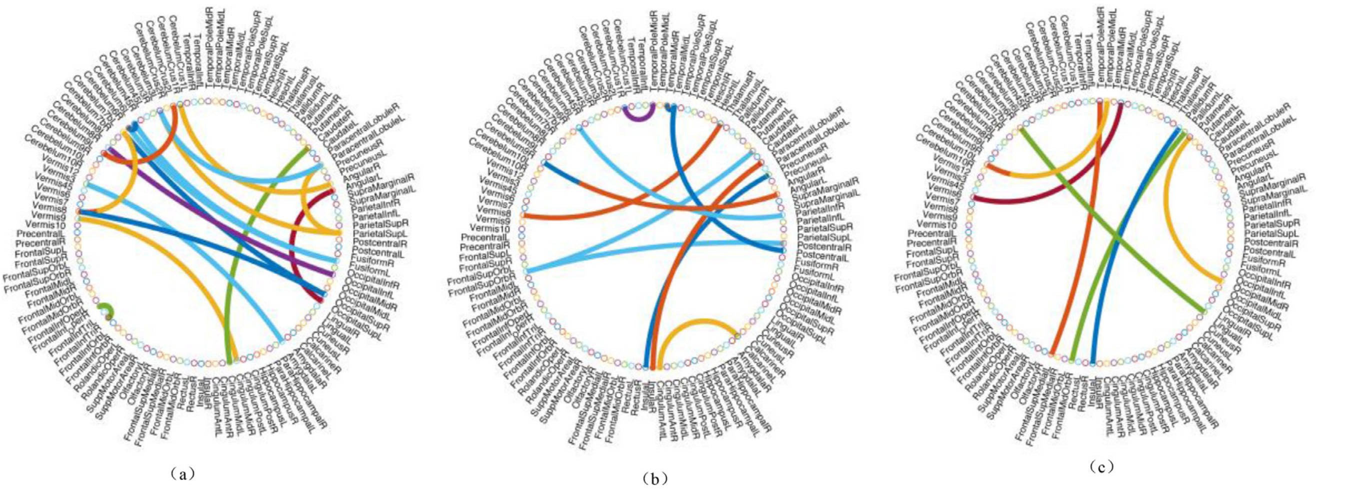


Fig. 10. (a) Selected ROIs and their connections from NC vs. SWEDD vs. PD classification task with threshold 0.99. (b) Selected ROIs and their connections from the NC vs. MCI vs. AD classification task with threshold 0.95. (c) Selected ROIs and their connections from NC vs. lMCI vs. sMCI vs. AD classification task with threshold 0.95.

tween these two regions. We screen these connections by setting a feasible threshold according to the visualization effect. Moreover, suffix ‘L’ indicates the left brain, suffix ‘R’ indicates the right brain, and different colors indicate different brain region connections.

From the correlation plot, the connections with Hippocampus cannot be found in the NC vs. PD vs. SWEDD experiment. Amygdala connection is not found in the NC vs. AD vs. MCI classification and Putamen is found in the NC vs. AD vs. lMCI vs. sMCI classification. The explanation is the former top brain regions are chosen as independent feature without considering the relationship of each other. It does not provide credit for the appearance of feature connections in this plot. However, the connections indicate that there exist some connections between certain brain regions. For example, connection between Amygdala_L and Vermis3 are found and Amygdala_L is the third important brain region found in the PD experiment. The connection between Precuneus_R and Insula_L is also found and Precuneus_R is the second important brain region found in the NC vs. AD vs. MCI experiment. Similarly, connection between Putamen_R and Occipital_Mid_R are also found and Putamen_R is the first important brain region found in the NC vs. AD vs. lMCI vs. sMCI experiment. These connections provide a broad view of analyzing the disease progression and thus helps improve the effectiveness of computer-aided diagnosis (Schaffer et al., 2015; Wei et al., 2017).

From the clinical viewpoint, our method may benefit the diagnosis of neurodegenerative disorders in two ways: 1) The automatic computer-aided diagnosis method could relieve doctor’s workload burden and help them to make clinical decisions, and further develop treatments that can delay or prevent disease progression. 2) The selected brain features indicate the potential disease-related ROIs, which may deserve special attention. They may present some new disease-related regions that have not yet been investigated.

Our method shows potential capability in differentiating samples into different disease progression stages. Beyond the impressive classification results, there are still few limitations. First, we only focus on single imaging modality from MRI and the dataset size is not so large. Second, the extracted features are concatenated linearly without considering the contribution of each feature’s significance. Hence, feature fusion may be investigated in our future work. For the future direction of this work, we also seek to obtain a larger dataset to train our classifier to improve the performance and consider multimodal images.

6. Conclusions

In this paper, we introduce a multi-template adaptive sparse learning along with a multi-class classification model for neurodegenerative disease diagnosis. We use multiple brain parcellation atlases with different sets of regions of interest to fuse different features together. Specifically, we integrate the feature selection and subspace learning with a *p*-norm regularization. In the constructed subspace, we jointly consider the global and local information in the data space. To further identify the disease type for clinical application, we perform a multi-class classification task. We verify our method using neuroimaging data and clinical assessment scores from PPMI and ANDI datasets with the automatically-selected discriminative features. Experiments show that our method is able to classify different categories simultaneously with promising results, which can benefit medical diagnosis in the long run. Furthermore, we generate relevant ROIs according to their weighing values to demonstrate the important brain regions for further diagnosis.

Declaration of Competing Interest

There are no conflicts of interest.

Acknowledgements

This work was supported partly by National Natural Science Foundation of China (Nos. 61871274, 61801305, 61806071 and 81571758), the Integration Project of Production Teaching and Research by Guangdong Province and Ministry of Education (No. 2012B091100495), Key Laboratory of Medical Image Processing of Guangdong Province (No. K217300003), Guangdong Province Key Laboratory of Popular High Performance Computers (No. 2017B03031407), Guangdong Pearl River Talents Plan (2016ZT06S220), Shenzhen Peacock Plan (Nos. KQTD2016053112051497 and KQTD2015033016 104926), Shenzhen Key Basic Research Project (Nos. JCYJ20180507184647636, JCYJ20170818142347251, JCYJ20170818 094109846, and JCYJ20170302153337765), Hong Kong Research Grants Council(No. PolyU 152035/17E), and Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 201900043).

References

Adeli, E., Shi, F., An, L., Wee, C.-Y., Wu, G., Wang, T., Shen, D., 2016. Joint Feature-Sample Selection and Robust Diagnosis of Parkinson's Disease From MRI Data. Neuroimage 141, 206–219. doi:10.1016/j.neuroimage.2016.05.054.

Aerts, M.B., Esselink, R.A., Post, B., Bp, V.D.W., Bloem, B.R., 2012. Improving the Diagnostic Accuracy in Parkinsonism: A Three-Pronged Approach. Practical Neurology 12 (2), 77–87.

Alzheimer's, A., 2015. Alzheimer's disease facts and figures. Alzheimer's & dementia: the j. the Alzheimer's Association 11 (3), 332.

Belkin, M., & Niyogi, P. (2003). Laplacian Eigenmaps for dimensionality reduction and data representation: MIT Press.

Blennow, K., Hampel, H., Weiner, M., Zetterberg, H., 2010. Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. Nat. Rev. Neurol. 6 (3), 131–144.

Braak, H., Del, T.K., Rüb, U., de Vos, R.A., Jansen Steur, E.N., Braak, E., 2003. Staging of brain pathology related to sporadic parkinson's disease. Neurobiol. Aging 24 (2), 197–211.

Chen, X., Pan, W., Kwok, J.T., Carbonell, J.G., 2009. Accelerated gradient method for multi-task sparse learning problem. Paper presented at the IEEE International Conference on Data Mining.

Craddock, R.C., James, G.A., Iii, P.E.H., Hu, X.P., Mayberg, H.S., 2012. A whole brain FMRI atlas generated via spatially constrained spectral clustering. Hum. Brain Mapp. 33 (8), 1914–1928.

Emrani, S., McGuirk, A., Wei, X., 2017. Prognosis and diagnosis of Parkinson's disease using multi-task learning. International Conference on Knowledge Discovery and Data Mining Paper presented at the ACM SIGKDD.

Eusebi, P., Hansson, O., Paciotti, S., Orso, M., Chiasserini, D., Calabresi, P., Parnetti, L., 2017. Cerebrospinal fluid biomarkers for the diagnosis and prognosis of Parkinson's disease: protocol for a systematic review and individual participant data meta-analysis. BMJ Open 7 (11), e018177.

Friston, K.J., 2003. Statistical Parametric Mapping. In Neuroscience Databases. Springer, pp. 237–250.

Fung, G., Stoeckel, J., 2007. SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. Knowl. Inf. Syst. 11 (2), 243–258. doi:10.1007/s10115-006-0043-5.

Grandvalet, Y., 2002. Adaptive scaling for feature selection in SVMs.Paper presented at the. International Conference on Neural Information Processing Systems.

Hall, S., Surova, Y., A, Ö., Zetterberg, H., Lindqvist, D., Hansson, O., 2015. CSF biomarkers and clinical progression of parkinson's disease. Neurology 84 (1), 57–63.

Huang, P., Min, X., Gu, Q., Yu, X., Xu, X., Luo, W., Zhang, M., 2015. Abnormal amygdala function in Parkinson's disease patients and its relationship to depression. J. Affect. Disord. 183, 263–268.

Jia, C., Fu, Y., 2016. Low-rank tensor subspace learning for RGB-D action recognition. IEEE Trans. Image Process. 25 (10), 4641–4652.

Jin, Y., Wee, C.Y., Shi, F., Thung, K.H., Ni, D., Yap, P.T., Shen, D., 2015. Identification of infants at high-risk for autism spectrum disorder using multiparameter multiscale white matter connectivity networks. Hum. Brain Mapp. 36 (12), 4880–4896.

Jothi, G., Hannah, I.H., 2016. Hybrid Tolerance Rough Set-Firefly Based Supervised Feature Selection for MRI Brain Tumor Image Classification. Elsevier Science Publishers B. V..

Kong, Y., Deng, Y., Dai, Q., 2014. Discriminative clustering and feature selection for brain MRI segmentation. IEEE Signal Process Lett. 22 (5), 573–577.

Lei, B., Peng, Y., Wang, T., Chen, S., Dong, N., 2017a. Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis. IEEE Trans. Cybern. (99) 1–12.

Lei, H., Huang, Z., Zhang, J., Yang, Z., Tan, E.L., Zhou, F., Lei, B., 2017b. Joint detection and clinical score prediction in Parkinson's disease via multi-modal sparse learning. Expert Syst. Appl. 80 (1), 284–296.

Lei, H., Zhao, Y., Wen, Y., Lei, B., 2017c. Adaptive sparse learning for neurodegenerative disease classification. Paper presented at the 2017. IEEE International Symposium on Multimedia (ISM).

Lin, G.C., Wang, W.J., Wang, C.M., Sun, S.Y., 2010. Automated classification of multi--spectral MR images using linear discriminant analysis. Comput. Med. Imaging Gr. 34 (4), 251–268.

Liu, M., Zhang, D., Adeli, E., Shen, D., 2016a. Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer's disease diagnosis. IEEE Trans. Biomed. Eng. 63 (7), 1473–1482. doi:10.1109/TBME.2015.2496233.

Liu, M., Zhang, D., Shen, D., 2016b. Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans. Med. Imaging 35 (6), 1463–1474. doi:10.1109/TMI.2016.2515021.

Luo, J., Vong, C.M., Wong, P.K., 2014. Sparse Bayesian extreme learning machine for multi-classification. IEEE Trans. Neural Net. Learn. Syst. 25 (4), 836–843.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Taylor, P., 2011. The Parkinson progression marker initiative (PPMI). Prog. Neurobiol. 95 (4), 629–635. doi:10.1016/j.pneurobio.2011.09.005.

Min, R., Wu, G., Cheng, J., Wang, Q., Shen, D., 2014. Multi-atlas based representations for Alzheimer's disease diagnosis. Hum. Brain Mapp. 35 (10), 5052–5070.

Nesterov, Y., 2004. Introductory lectures on convex optimization. Appl. Optim. 87 (5) xviii,236.

Nilashi, M., Ibrahim, O., Ahani, A., 2016. Accuracy improvement for predicting Parkinson's disease progression. Sci. Rep. 6, 34181. doi:10.1038/srep34181.

Ozcift, A., 2012. SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. J. Med. Syst. 36 (4), 2141–2148.

Postuma, R.B., Berg, D., Stern, M., Poewe, W., Olanow, C.W., Oertel, W., … Lang, A.E., 2015. MDS clinical diagnostic criteria for Parkinson's disease. Movement Disorders 30 (12), 1591–1599.

Prashanth, R., Dutta Roy, S., Mandal, P.K., Ghosh, S., 2014. Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging. Expert Syst. Appl. 41 (7), 3333–3342.

Rana, B., Juneja, A., Saxena, M., Gudwani, S., Kumaran, S., Behari, M., Agrawal, R., 2014. A machine learning approach for classification of Parkinson's disease and controls using T1-weighted MRI. Mov. Disord. 29, S88–S89.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., … Quattrone, A., 2014. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy. Journal of Neuroscience Methods 222, 230–237.

Schaffer, C., Sarad, N., Decrumpe, A., Goswami, D., Herrmann, S., Morales, J., Osborne, J., 2015. Biomarkers in the diagnosis and prognosis of Alzheimer's disease. J. Assoc. Lab. Autom. 20 (5), 589–600.

Seeley, W.W., Crawford, R.K., Zhou, J., Miller, B.L., Greicius, M.D., 2009. Neurodegenerative diseases target large-scale human brain networks. Neuron 62 (1), 42–52.

Tibshirani, R.J., 1996. Regression shrinkage and selection via the LASSO. J R Stat Soc B. 58, 267–288.

Tzouriomazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15 (1), 273–289.

Wang, J., Wang, Q., Zhang, H., Chen, J., Wang, S., Shen, D., 2019. Sparse multiview task-centralized ensemble learning for ASDdiagnosis based on age-and Ssex-related functional connectivity patterns. IEEE Trans. Cybern. 49 (8), 3141–3154.

Wang, K., He, R., Wang, L., Wang, W., Tan, T., 2016. Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 38 (10), 2010–2023.

Wei, X., Wei, X., Wei, X., 2017. Prognosis and diagnosis of Parkinson's disease using multi-task learning. Paper presented at the ACM SIGKDD.

Weiner, W. J., Oksana Suchowersky, M., &. Rajesh Pahwa, M. (2006). Diagnosis and prognosis of new-onset Parkinson's disease. Patient care for the nurse practitioner.

Ye, J., 2007. Least squares linear discriminant analysis. In: Paper Presented at the Proceedings of the 24th International Conference on Machine Learning, Corvalis. Oregon, USA.

Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 59 (2), 895–907.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.

Zhang, Z., Ye, N., 2011. Locality preserving multimodal discriminative learning for supervised feature selection. Knowl. Inf. Syst. 27 (3), 473–490.

Zhou, T., Thung, K.-H., Liu, M., Shen, D., 2019a. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. IEEE Trans. Biomed. Eng. 66, 165–175.

Zhou, T., Thung, K.-H., Zhu, X., Shen, D., 2019b. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum. Brain Mapp. 40, 1001–1016.

Zhu, X., Suk, H.I., Lee, S.W., Shen, D., 2016a. Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. Brain Imaging Behav. 10 (3), 1–11.

Zhu, X., Suk, H.I., Lee, S.W., Shen, D., 2016b. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. IEEE Trans. Biomed. Eng. 63 (3), 607–618.