Article:
Huang, Z, Lei, H, Chen, G et al. (5 more authors) (2021) Parkinson's Disease Classification and Clinical Score Regression via United Embedding and Sparse Learning From Longitudinal Data. IEEE Transactions on Neural Networks and Learning Systems. ISSN 2162-237X

https://doi.org/10.1109/tnnls.2021.3052652

Parkinson's Disease Classification and Clinical Score Regression via United Embedding and Sparse Learning from Longitudinal Data

Zhongwei Huang, Haijun Lei, Guoliang Chen, Alejandro F Frangi, *Fellow, IEEE*, Yanwu Xu, *Senior Member, IEEE*, Ahmed Elazab, Jing Qin, Baiying Lei*, *Senior Member, IEEE*

***Abstract*—Parkinson's disease (PD) is known as an irreversible neurodegenerative disease that mainly affects the patient's motor system. Early classification and regression of PD are essential to slow down this degenerative process from its onset. In this paper, a novel adaptive unsupervised feature selection approach is proposed by exploiting manifold learning from longitudinal multimodal data. Classification and clinical score prediction are performed jointly to facilitate early PD diagnosis. Specifically, the proposed approach performs united embedding and sparse regression, which can determine the similarity matrices and discriminative features adaptively. Meanwhile, we constrain the similarity matrix among subjects and exploit $\ell_{2,p}$ norm to conduct sparse adaptive control for obtaining the intrinsic information of the multimodal data structure. An effective iterative optimization algorithm is proposed to solve this problem. We perform abundant experiments on the Parkinson's Progression Markers Initiative (PPMI) dataset to verify the validity of the proposed approach. The results show our approach boosts performance on the classification and clinical score regression of longitudinal data and surpasses the state-of-the-art approaches.**

***Index Terms*—Classification, clinical score prediction, embedding learning, longitudinal multimodal data, Parkinson's disease, sparse regression.**

This work was supported partly by National Natural Science Foundation of China (Nos. 61871274, 61801305, 81571758, and 61950410615), National Natural Science Foundation of Guangdong Province (Nos. 2020A1515010649 and 2019A1515111205), (Key) Project of Department of Education of Guangdong Province (No. 2019KZDZX1015), Guangdong Laboratory of Artificial-Intelligence and Cyber-Economics (SZ), Shenzhen Peacock Plan (Nos. KQTD2016053112051497 and KQTD2015033016104926), Shenzhen Key Basic Research Project (Nos. JCYJ20190808165209410, 20190808145011259, JCYJ20180507184647636, GJHZ20190822095414576, JCYJ20170302153337765, JCYJ20170302150411789, JCYJ201703 02142515949, GCZX2017040715180580, GJHZ20180418190529516, and JSGG20180507183215520), NTUT-SZU Joint Research Program (No.2020003), Hong Kong Research Grants Council (No. PolyU 152035/17E), AFF is supported by the RAEng Chair in Emerging Technologies (CiET1819/19).

Z. Huang, H. Lei, and G. Chen are with the Key Laboratory of Service Computing and Applications, Guangdong Province Key Laboratory of Popular High Performance Computers, Guangdong Province Engineering Center of China-made High Performance Data Computing System, Guangdong Laboratory of Artificial-Intelligence and Cyber-Economics, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

A. F. Frangi is with the CISTIB Centre for Computational Imaging & Simulation Technologies in Biomedicine, School of Computing and School of Medicine, University of Leeds, Leeds, UK, and with Departments of Cardiovascular Sciences and Electrical Engineering, KU Leuven, Leuven, Belgium. He is also Pengcheng Visiting Scholar at the School of Biomedical Engineering, Shenzhen University, Shenzhen, China.

Y. Xu is with Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China.

J. Qin is with Centre for Smart Health, School of Nursing, Hong Kong Polytechnic University, HungHom, Hong Kong.

A. Elazab is with Misr Higher Institute for Commerce and Computers, Computer Science Department, Mansoura City, Egypt.

A. Elazab and B. Lei are with the 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, China, 518060 (e-mail: leiby@szu.edu.cn).

I. Introduction

Parkinson's disease (PD) is a common neurodegenerative disorder that occurs in the elderly. With the worsening of the condition, it can trigger incidents leading to death. Patients suffering from PD usually not die from the disease but accidents or complications related to the disease. For example, in advanced cases of the disease, difficulty in swallowing can cause PD patients to inhale food into their lungs, resulting in pneumonia or other pulmonary conditions. Patients with PD have four main motor symptoms: muscle rigidity, static tremor, unstable posture, and bradykinesia [1]. Except for these visual symptoms, there are also some concomitant symptoms (e.g., depression, lethargy, olfaction disorder, and cognition impairment [2]). These symptoms are mainly caused by the degeneration of dopaminergic neurons in a region of the brain called the substantia nigra [3]. However, in early PD therapeutic trials, dopaminergic imaging has found that approximately 15 percent of scans are at the normal level, namely scans without evidence of dopaminergic degeneration (SWEDD) [4]. This condition has undoubtedly augmented the difficulties of PD classification. In the early stage of this disease, it may be challenging to know if the symptoms indicate or imitate PD. Because of this, while early PD classification is essential to slow down this degenerative process from its onset, it is a quite challenging task.

Since multimodal data can offer complementary information for the classification of neurodegenerative diseases, these data have played an increasingly significant role and captured widespread attention [5, 6]. For instance, in [7], multimodal data is utilized to raise the classification performance of neurodegenerative disease based on a semi-supervised feature-subject selection approach. Gray matter (GM) in magnetic resonance imaging (MRI) is widely used to obtain information about changes in nerve cells. First eigenvalue (L1) and first eigenvector (V1) of diffusive tensor imaging (DTI) indicate the largest diffusion coefficient and its direction vector, respectively. Therefore, L1 and V1 may be more sensitive to neurodegeneration in the brain. Inspired by the above, we propose to

explore multiple modalities, namely, GM, L1, and V1, to perform PD classification. Nevertheless, for the multimodal data, the small dataset size and large feature dimension typically cause overfitting problem and renders difficulty in model generalization [8]. Though deep learning has been extensively used in the medical image fields, it is difficult to obtain excellent generalization performance on small number of subjects [9].

Feature selection is an excellent measure to discover disease-related characteristics exploiting either supervised or unsupervised approaches [10-12]. For most supervised approaches, they can be individual single-task [13, 14] or united multi-task [15, 16]. Generally, the latter has better performance since this approach combines multiple related tasks to select common features jointly. However, there exist two main limitations in existing multi-task approaches. First, the selected features obtained from these approaches are usually linearly related to multi-task goals but they ignore the learning of the structural information intrinsic within the data. Second, these multi-task approaches need additional scores and label information to learn the model. On the other hand, unsupervised approaches focus more on learning the intrinsic data structure. Most unsupervised approaches are developed based on either filtering [17, 18] or embedding [19, 20]. The latter is superior in many aspects. However, there are three main shortcomings in existing embedding approaches. First, they calculate the similarity matrix among subjects and select features, respectively. But real data in the original high-dimensional space has noise and/or redundancy, which reduces the accuracy of the similarity matrix. Second, when calculating the neighbor graph, the similarity matrix among subjects generated by conventional approaches rarely represents a proper neighbor distribution. The optimal similarity matrix among subjects ought to have *r*-connected components, where *r* equals to the number of classes. Third, many embedding approaches do not take the similarity existing among features into consideration.

Meanwhile, most existing approaches use $\ell_{2,1}$ norm to conduct sparseness control [21, 22]. This norm cannot achieve adaptive sparseness according to different cases. Most existing studies only conducted PD classification [7, 23, 24]. Relatively few studies considered another essential task of clinical score regression [25, 26]. Since clinical score regression (e.g., depression, sleep, olfaction, and cognition scores) can assist doctors in staging and treating disease, these two tasks need to be conducted simultaneously. In addition, in most existing studies, classification and regression are performed only based on the baseline data [27, 28], while the longitudinal data (i.e., multi-time points data) are ignored. Owing to the persistent exacerbation of the disease, it is imperative to learn reliable classification and prediction models that meet multi-time points [29]. We highlight our contributions:

1) We propose a novel unsupervised learning method from longitudinal multimodal data for feature selection. The united embedding learning and sparse regression are exploited to adaptively learn the low-dimensional manifold structure and select the informative features.
2) We dynamically update the similarity matrices among subjects and features. The connected number among subjects from the similarity matrix is equal to the number of classes, which can gain the intrinsic structural property of the data.
3) We conduct abundant experiments on the PPMI dataset to verify the effectiveness of the proposed approach. The results show that our algorithm effectively boosts the performance of classification and clinical score regression and surpasses other state-of-the-art approaches by taking full advantage of the longitudinal data.

The paper is organized as follows. In Section II, we discuss the most related work on feature selection. Detailed interpretation of our approach is introduced in Section III. Our results and discussions are shown in Section IV and V. Finally, several conclusions are recapitulated in Section VI.

II. Related work

In the literature, there are many supervised united multi-task and unsupervised embedding approaches. For instance, the multimodal multi-task (M3T) [30] approach based on $\ell_{2,1}$ norm learns a feature selection model to gain common relevant features of multiple tasks from every modality. The multimodal sparse learning (MMSL) [15] approach concurrently performs classification and regression prediction of PD based on a united multi-task feature selection function that considers the similarity of difference among rows and columns in response matrix. In [31], multimodal data is utilized to improve performance on the classification and regression prediction of Alzheimer's disease via relational regularization and discriminative learning. In [5], the authors perform joint learning from multiple relations and modalities to select the discriminative features for classification and prediction of PD. Multi-cluster feature selection (MCFS) [32] approach first calculates the nearest neighbor graph and then selects the discriminative features that best present the clustering information. Flexible manifold embedding (FME) [33] approach is a generalized model exploited by many unsupervised and semi-supervised embedding approaches to reduce feature dimensionality. Robust spectral feature selection (RSFS) [19] approach concurrently utilizes FME and ℓ_1 norm to robustly select the discriminative features. Joint embedding learning and sparse regression (JELSR) [34] approach conducts feature selection through embedding learning with sparse regression. These existing approaches suffer from some limitations. For instance, the supervised multi-task approaches ignore to learn the structural information intrinsic within the data. Previous unsupervised embedding approaches may also calculate an inaccurate similarity matrix due to the noise in the original feature space.

III. Methodology

A. System Overview

The overall flowchart of our approach is illustrated in Fig.1. First, we extract features from GM, L1, and V1 and concatenate them directly. We then use the proposed approach to perform feature selection. Finally, we exploit support vector classification (SVC) and support vector regression (SVR) models to conduct classification and regression prediction on longitudinal multimodal data.

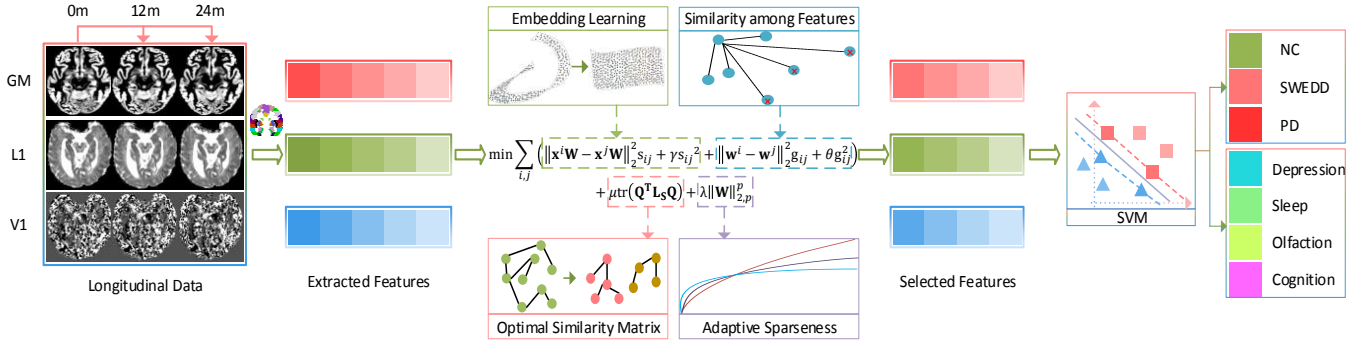


Fig. 1. Illustration of the proposed approach through united embedding and sparse regression.

B. Notation

In this study, the uppercase boldface letter $\mathbf{X}$ indicates matrix, and the lowercase boldface letter $\mathbf{x}$ indicates vector. Let the data matrix be $\mathbf{X} = [x_{ij}]$, $\mathbf{x}^i$ and $\mathbf{x}_j$ are its i-th row and j-th column, respectively. In addition, the $\ell_{2,p}$ norm of $\mathbf{X}$ is indicated as $\|\mathbf{X}\|_{2,p} = \left(\sum_i \|\mathbf{x}^i\|_2^p\right)^{\frac{1}{p}}$. The transpose and trace operators of $\mathbf{X}$ are indicated as $\mathbf{X}^T$ and $tr(\mathbf{X})$, respectively.

C. Proposed Approach

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ indicate the training data of n subjects and d features and $\mathbf{S} \in \mathbb{R}^{n\times n}$ indicate the subject similarity matrix. According to our intuitive understanding, closer subjects usually have greater similarities, and thus we can calculate the similarity $\mathbf{S}$ using the following formula:

$$\min \sum_{i,j} \left(\|\mathbf{x}^i - \mathbf{x}^j\|_2^2 s_{ij} + \mu s_{ij}^2\right), \quad s.t.\, \mathbf{s}^i\mathbf{1} = 1, 0 \le s_{ij} \le 1 \tag{1}$$

where $\mathbf{x}^i$ and $\mathbf{x}^j$ indicate the i-th and j-th subjects of $\mathbf{X}$, respectively. μ is a regularization parameter to avert the meaningless solution. The subject similarity matrix $\mathbf{S}$ built with (1) rarely has a proper neighbor distribution. The optimal subject similarity matrix ought to have r-connected components, where r equals to the number of classes. Nevertheless, it is nearly impossible to obtain $\mathbf{S}$ with (1) that satisfies the above requirements. To solve this problem, the rank of the Laplacian matrix $\mathbf{L_S}$ of $\mathbf{S}$ is ensure to be equal to $n - r$. In this way, there will be r-connected components in the subject similarity matrix [35]. We add this constraint on $\mathbf{L_S}$ to (1) as:

$$\min \sum_{i,j} \left(\|\mathbf{x}^i - \mathbf{x}^j\|_2^2 s_{ij} + \mu s_{ij}^2\right), \quad s.t.\, \mathbf{s}^i\mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(\mathbf{L_S}) = n - r \tag{2}$$

where $\mathbf{L_S}$ equals to $\mathbf{D} - \frac{\mathbf{S}^T+\mathbf{S}}{2}$ and the degree matrix $\mathbf{D}$ is a diagonal matrix whose i-th element value on the diagonal equals to $\sum_j \frac{s_{ij}+s_{ji}}{2}$. Since $rank(\mathbf{L_S}) = n - r$ relies on the subject similarity matrix $\mathbf{S}$, it is challenging to optimize (2) directly. To solve this problem, let $\psi_i(\mathbf{L_S})$ indicate the i-th minimum eigenvalue of $\mathbf{L_S}$. Because $\mathbf{L_S}$ is positive semi-definite, we obtain $\psi_i(\mathbf{L_S}) \ge 0$. Meanwhile, it can be easy to prove that $rank(\mathbf{L_S}) = n - r$ equals $\sum_{i=1}^r \psi_i(\mathbf{L_S}) = 0$. Since it is hard to solve the derivation of $\sum_{i=1}^r \psi_i(\mathbf{L_S})$, we refer to Ky Fan's Theorem [36], and we obtain:

$$\sum_{i=1}^r \psi_i(\mathbf{L}) = \min tr(\mathbf{Q}^T\mathbf{L_S}\mathbf{Q}), \quad s.t.\, \mathbf{Q} \in \mathbb{R}^{n\times r}, \mathbf{Q}^T\mathbf{Q} = \mathbf{E} \tag{3}$$

Further, we can rewrite (2):

$$\min \sum_{i,j} \left(\|\mathbf{x}^i - \mathbf{x}^j\|_2^2 s_{ij} + \mu s_{ij}^2\right) + \sigma tr(\mathbf{Q^T L_S Q}), \quad s.t.\, \mathbf{s}^i\mathbf{1} = 1, 0 \le s_{ij} \le 1, \mathbf{Q} \in \mathbb{R}^{n\times r}, \mathbf{Q}^T\mathbf{Q} = \mathbf{E} \tag{4}$$

where σ is a model parameter that can be decreased or increased in each iteration to obtain the optimal $\mathbf{S}$ when the connected components of $\mathbf{S}$ are greater or smaller than r, respectively. $\mathbf{E}$ is the identity matrix. In (4), the subject similarity matrix $\mathbf{S}$ is calculated in the original multimodal feature space. However, the original high-dimensional data has noise and/or redundancy. To tackle this problem, we perform adaptive sparseness and embedding learning simultaneously, which is expressed as:

$$\min \sum_{i,j} \left(\|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}\|_2^2 s_{ij} + \mu s_{ij}^2\right) + \sigma tr\left(\mathbf{Q}^T\mathbf{L_S}\mathbf{Q}\right) + \lambda\|\mathbf{W}\|_{2,p}^p, \quad s.t.\, \mathbf{s}^i\mathbf{1} = 1, 0 \le s_{ij} \le 1, \mathbf{Q} \in \mathbb{R}^{n\times r}, \mathbf{Q}^T\mathbf{Q} = \mathbf{E}, \mathbf{W}^T\mathbf{W} = \mathbf{E} \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{d\times m}$ indicates the feature weight coefficient matrix of m projection dimension. λ is a weighting parameter that decreases the feature weight $\mathbf{W}$ to obtain more sparse features as the value of λ increases. Meanwhile, we use $\ell_{2,p}$ norm to carry out sparse adaptive control for selecting the most discriminative features via different scenarios. Furthermore, the data of high-dimensional features might render the covariance matrix of $\mathbf{X}$ singular, so we add the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{E}$ to obtain the discriminative features after dimension reduction.

There exists similarity among features extracted from regions of interest (ROI). If two features, e.g., $\mathbf{x}_i$ and $\mathbf{x}_j$, are similar, their weight coefficients, e.g., $\mathbf{w}^i$ and $\mathbf{w}^j$, will be similar because the i-th feature $\mathbf{x}_i$ and j-th feature $\mathbf{x}_j$ in $\mathbf{X}$ correspond to the i-th row $\mathbf{w}^i$ and j-th row $\mathbf{w}^j$ in $\mathbf{W}$, respectively. To exploit the relationship among features, we propose a regularization term considering the similarity among features, which is indicated as:

$$\min \sum_{i,j} \left(\|\mathbf{w}^i - \mathbf{w}^j\|_2^2 g_{ij} + \theta g_{ij}^2\right), s.t.\, \mathbf{g}^i\mathbf{1} = 1, 0 \le g_{ij} \le 1, \tag{6}$$

where g_{ij} indicates an element in the feature similarity matrix $\mathbf{G} \in \mathbb{R}^{d\times d}$. θ is a regularization parameter to avert the meaningless solution. Finally, we add this regularization term to (5) and then we get:

$$\min \sum_{i,j} \left(\|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}\|_2^2 s_{ij} + \mu s_{ij}^2 + \|\mathbf{w}^i - \mathbf{w}^j\|_2^2 g_{ij} + \theta g_{ij}^2\right) + \sigma tr(\mathbf{Q}^T\mathbf{L_S}\mathbf{Q}) + \lambda\|\mathbf{W}\|_{2,p}^p, \quad s.t.\, \mathbf{s}^i\mathbf{1} = 1, 0 \le s_{ij} \le 1, \mathbf{g}^i\mathbf{1} = 1, 0 \le g_{ij} \le 1, \mathbf{Q} \in \mathbb{R}^{n\times r}, \mathbf{Q}^T\mathbf{Q} = \mathbf{E}, \mathbf{W}^T\mathbf{W} = \mathbf{E} \tag{7}$$

Algorithm 1: Solution to (12)

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{L_S} \in \mathbb{R}^{n\times n}$, $\mathbf{L_G} \in \mathbb{R}^{d\times d}$, λ;
Output: $\mathbf{W} \in \mathbb{R}^{d\times m}$;
Initialize $t = 0$, $\mathbf{Z}(t)$ by an identity matrix;
Repeat
  Under the current $\mathbf{Z}(t)$, the optimal solution $\mathbf{W}(t+1)$ of (12) is the m eigenvectors corresponding to the m minimum eigenvalues of $\mathbf{X^T L_S X} + \mathbf{L_G} + \lambda\mathbf{Z}$.
  Update matrix $\mathbf{Z}(t+1)$ by solving (11);
Until the convergence or stop condition is satisfied.

D. Optimization

Since (7) contains $\ell_{2,\mathrm{p}}$ norm and four variables, it is difficult to solve this problem directly. Thus, we exploit an alternative approach to tackle this problem. Meanwhile, regarding the optimization of parameter $\mathbf{W}$, it is necessary to use Laplacian matrices to transform (7) into trace forms. Next, we fix the similarity matrix $\mathbf{S}$ among subjects and the similarity matrix $\mathbf{G}$ among features and then update $\mathbf{W}$. Equation (7) is transformed into:

$$\min tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{X}\mathbf{W}) + tr(\mathbf{W}^T\mathbf{L_G}\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,p}^p, \quad s.t.\,\mathbf{W}^T\mathbf{W} = \mathbf{E} \tag{8}$$

The objective function with the Lagrange multiplier for (8) is:

$$\mathcal{F}(\mathbf{W},\mathbf{\Lambda}) = tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{X}\mathbf{W}) + tr(\mathbf{W}^T\mathbf{L_G}\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,p}^p + tr\big(\mathbf{\Lambda}(\mathbf{W}^T\mathbf{W} - \mathbf{E})\big), \tag{9}$$

where $\mathbf{\Lambda} \in \mathbf{R}^{m\times m}$ is the diagonal matrix indicating the Lagrange multiplier. We set the derivative of (9) on $\mathbf{W}$ to zero and get:

$$\frac{\partial\mathcal{F}(\mathbf{W},\mathbf{\Lambda})}{\partial\mathbf{W}} = \mathbf{X}^T\mathbf{L_S}\mathbf{X}\mathbf{W} + \mathbf{L_G}\mathbf{W} + \lambda\mathbf{Z}\mathbf{W} + \mathbf{W}\mathbf{\Lambda} = 0, \tag{10}$$

where $\mathbf{Z} \in \mathbf{R}^{d\times d}$ is a diagonal matrix and its i-th diagonal element is defined as:

$$\mathrm{z}_{ii} = \frac{p}{2\left(\mathbf{w}^i\mathbf{w}^{i^T} + eps\right)^{\frac{2-p}{2}}} \tag{11}$$

where we add a floating number *eps* equal to 2^{-52} to the denominator of (9) since $\mathbf{w}^i\mathbf{w}^{i^T}$ can be zero in theory. The solution of (10) equals to:

$$\min tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{X}\mathbf{W}) + tr(\mathbf{W}^T\mathbf{L_G}\mathbf{W}) + \lambda tr(\mathbf{W}^T\mathbf{Z}\mathbf{W}), \quad s.t.\,\mathbf{W}^T\mathbf{W} = \mathbf{E} \tag{12}$$

where we can exploit an iterative algorithm to solve (12) since $\mathbf{Z}$ relies on $\mathbf{W}$. The details of the proposed algorithm are shown in Algorithm 1. Next, we fix the subject similarity matrix $\mathbf{S}$ and then update $\mathbf{Q}$. Equation (7) is transformed into:

$$\min tr(\mathbf{Q}^T\mathbf{L_S}\mathbf{Q}), \quad s.t.\,\mathbf{Q} \in \mathbb{R}^{n\times r}, \mathbf{Q}^T\mathbf{Q} = \mathbf{E} \tag{13}$$

where the optimal solution $\mathbf{Q}$ of (13) is the r eigenvectors corresponding to the r minimum eigenvalues of $\mathbf{L_S}$. Next, we fix $\mathbf{W}$ and $\mathbf{Q}$, and then update $\mathbf{S}$. Equation (7) is transformed into:

$$\min \textstyle\sum_{i,j}\left(\left\|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}\right\|_2^2 \mathrm{s}_{ij} + \mu s_{ij}^2\right) + \sigma tr(\mathbf{Q^T L_S Q}), \quad s.t.\,\mathbf{s}^i\mathbf{1} = 1, 0 \le \mathrm{s}_{ij} \le 1, \tag{14}$$

We can transform (14) into the following form:

$$\min \textstyle\sum_{i,j}\left(\left\|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}\right\|_2^2 \mathrm{s}_{ij} + \mu s_{ij}^2\right) + \sigma\sum_{i,j}\left\|\mathbf{q}^i - \mathbf{q}^j\right\|_2^2 \mathrm{s}_{ij}, \quad s.t.\,\mathbf{s}^i\mathbf{1} = 1, 0 \le \mathrm{s}_{ij} \le 1 \tag{15}$$

The similarity vector of every subject is independent. Thus we can solve the above problem for the i-th subject, and we have:

$$\min \textstyle\sum_j\left(\left\|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}\right\|_2^2 \mathrm{s}_{ij} + \mu s_{ij}^2\right) + \sigma\sum_j\left\|\mathbf{q}^i - \mathbf{q}^j\right\|_2^2 \mathrm{s}_{ij}, \quad s.t.\,\mathbf{s}^i\mathbf{1} = 1, 0 \le \mathrm{s}_{ij} \le 1 \tag{16}$$

where we indicate $\mathbf{a}^i \in \mathbf{R}^{1\times n}$ with $\mathrm{a}_{ij} = \left\|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}\right\|_2^2 + \sigma\left\|\mathbf{q}^i - \mathbf{q}^j\right\|_2^2$. We can rewrite (16):

$$\min \left\|\mathbf{s}^i + \frac{1}{2\mu}\mathbf{a}^i\right\|_2^2, s.t.\,\mathbf{s}^i\mathbf{1} = 1, 0 \le \mathrm{s}_{ij} \le 1 \tag{17}$$

We consider two extreme cases of μ in (1). When $\mu = 0$, it makes only a single element of $\mathbf{s}^i$ not equal to zero. When $\mu = \infty$, it makes each element of $\mathbf{s}^i$ equal to $\frac{1}{n}$. The value of μ controls the number of neighbors of a subject. Thus the optimal value of μ needs to make most $\mathbf{s}^i$ have k non-zero elements, where k indicates the number of neighbors connected to $\mathbf{x}^i$. To achieve this goal, we consider the Lagrangian function of (17) as follows:

$$\mathcal{F}(\mathbf{s}^i,\beta,\boldsymbol{\varphi}^i,\boldsymbol{\tau}^i) = \frac{1}{2}\left\|\mathbf{s}^i + \frac{1}{2\mu}\mathbf{a}^i\right\|_2^2 + \beta(\mathbf{s}^i\mathbf{1} - 1) + \boldsymbol{\varphi}^i\left(\mathbf{s}^{i\mathbf{T}} - \mathbf{1}\right) + \boldsymbol{\tau}^i\left(-\mathbf{s}^{i\mathbf{T}}\right), \tag{18}$$

where β, $\boldsymbol{\varphi}^i$, and $\boldsymbol{\tau}^i$ indicate Lagrangian multipliers. According to the Karush–Kuhn–Tucker (KKT) condition, the optimal solution of $\mathbf{s}^i$ is:

$$\mathrm{s}_{ij} = -\frac{1}{2\mu}\mathrm{a}_{ij} - \beta, \tag{19}$$

Considering the convenience of expression, we suppose that $\mathrm{a}'_{i1}, \mathrm{a}'_{i2}, \ldots, \mathrm{a}'_{in}$ are sorted from small to large corresponding to $\mathrm{s}'_{i1}, \mathrm{s}'_{i2}, \ldots, \mathrm{s}'_{in}$. When the optimal $\mathbf{s}^i$ contains only k neighbors, we know $\mathrm{s}'_{ij} > 0\ (j = 1, 2, \ldots, k)$ and $\mathrm{s}'_{ij} = 0\ (j = k+1, k+2, \ldots, n)$. Therefore, we have:

$$\begin{cases} -\frac{1}{2\mu}\mathrm{a}'_{ij} - \beta > 0\ (j = 1, 2, \ldots, k) \\ -\frac{1}{2\mu}\mathrm{a}'_{ij} - \beta \le 0\ (j = k+1, k+2, \ldots, n) \end{cases}, \tag{20}$$

According to (19) and the constraint $\mathbf{s}^i\mathbf{1} = 1$, we get:

$$\textstyle\sum_{j=1}^k\left(-\frac{1}{2\mu}\mathrm{a}'_{ij} - \beta\right) = 1 \rightarrow \beta = -\frac{1}{k} - \sum_{j=1}^k\frac{1}{2k\mu}\mathrm{a}'_{ij}, \quad s.t.\,j = 1, 2, \ldots, k, \tag{21}$$

where the inequality for μ is given by (20) and (21):

$$\frac{k}{2}\mathrm{a}'_{ik} - \frac{1}{2}\textstyle\sum_{j=1}^k\mathrm{a}'_{ij} < \mu \le \frac{k}{2}\mathrm{a}'_{ik+1} - \frac{1}{2}\sum_{j=1}^k\mathrm{a}'_{ij}, \tag{22}$$

To obtain a good μ that can make the most $\mathbf{s}^i$ has k neighbors or non-zeros elements, we can calculate μ as:

$$\mu = \frac{1}{n}\textstyle\sum_{i=1}^n\left(\frac{k}{2}\mathrm{a}'_{ik+1} - \frac{1}{2}\sum_{j=1}^k\mathrm{a}'_{ij}\right), \tag{23}$$

Finally, we fix $\mathbf{W}$ and then update $\mathbf{G}$. Equation (7) is transformed into:

$$\min \textstyle\sum_{i,j}\left(\left\|\mathbf{w}^i - \mathbf{w}^j\right\|_2^2 \mathrm{g}_{ij} + \theta\mathrm{g}_{ij}^2\right), \quad s.t.\ \mathbf{g}^i\mathbf{1} = 1, 0 \le \mathrm{g}_{ij} \le 1 \tag{24}$$

The similarity vector of every feature is independent. Thus we can solve the above problem for the i-th feature, and we get:

$$\min \textstyle\sum_j\left(\left\|\mathbf{w}^i - \mathbf{w}^j\right\|_2^2 \mathrm{g}_{ij} + \theta\mathrm{g}_{ij}^2\right), \quad s.t.\ \mathbf{g}^i\mathbf{1} = 1, 0 \le \mathrm{g}_{ij} \le 1 \tag{25}$$

where we indicate $\mathbf{b}^i \in \mathbf{R}^{1\times d}$ with $\mathrm{b}_{ij} = \left\|\mathbf{w}^i - \mathbf{w}^j\right\|_2^2$. Therefore, we can rewrite (24) as:

Algorithm 2: Solution to (7)

Input: $\mathbf{X} \in \mathbb{R}^{d\times n}$, r, m, k, λ;
Output: $\mathbf{W} \in \mathbb{R}^{d\times m}$;
Initialize $\mathbf{S}$ by solving (1);
Initialize $\mathbf{G}$ by solving the following formula:
$\min \sum_{i,j}\left(\left\|\mathbf{x_i}-\mathbf{x_j}\right\|_2^2 \mathrm{g}_{ij}+\theta \mathrm{g}_{ij}{}^2\right)$,
$s.t.\, \mathbf{g}^i\mathbf{1}=1, 0\le \mathrm{g}_{ij}\le 1$;
Repeat
  Update $\mathbf{W}$ by Algorithm 1;
  Calculate the Laplacian matrix $\mathbf{L_S}$ of the similarity matrix $\mathbf{S}$;
  Update $\mathbf{Q}$ by solving (13);
  Update $\mathbf{S}$ by solving (17);
  Update $\mathbf{G}$ by solving (25);
Until the convergence or stop condition is satisfied.

$$\min \left\|\mathbf{g}^i+\frac{1}{2\theta}\mathbf{b}^i\right\|_2^2, s.t.\, \mathbf{g}^i\mathbf{1}=1, 0\le \mathrm{g}_{ij}\le 1 \tag{26}$$

The solution of (26) is the same as solving (17). We show the details of the total optimization in Algorithm 2.

E. Convergence Analysis of Algorithm 1

To prove that algorithm 1 converges, we first need to verify the following inequality.

Theorem 1: For any positive real numbers u and v, with p a constant $(0<p<2)$, it holds:

$$u^{\frac{p}{2}}-\frac{pu}{2v^{\frac{2-p}{2}}}\le v^{\frac{p}{2}}-\frac{pv}{2v^{\frac{2-p}{2}}} \tag{27}$$

Proof: We move the left side of the inequality (27) to its right-hand side, and then we need to prove the following inequality:

$$\mathfrak{I}(u,v)=v^{\frac{p}{2}}-\frac{pv}{2v^{\frac{2-p}{2}}}-u^{\frac{p}{2}}+\frac{pu}{2v^{\frac{2-p}{2}}}\ge 0$$
$$\rightarrow \mathfrak{I}(u,v)=(2-p)v+pu-2u^{\frac{p}{2}}v^{\frac{2-p}{2}}\ge 0 s.t. u>0, v>0 \tag{28}$$

To prove the inequality (28), we consider its Lagrangian function as:

$$\mathfrak{I}(u,v,\gamma_1,\gamma_2)=(2-p)v+pu-2u^{\frac{p}{2}}v^{\frac{2-p}{2}}-\gamma_1 u-\gamma_2 v, \tag{29}$$

where γ_1 and γ_2 indicate Lagrangian multipliers. According to the KKT condition, when u equals to v, (29) has the minimum value equal to zero. Thus, we can get:

$$\mathfrak{I}(u,v)\ge 0 \rightarrow u^{\frac{p}{2}}-\frac{pu}{2v^{\frac{2-p}{2}}}\le v^{\frac{p}{2}}-\frac{pv}{2v^{\frac{2-p}{2}}} \tag{30}$$

which concludes the proof.

Theorem 2: The iterative updating rules in Algorithm 1 will gradually reduce the objective value of (8) until convergence.

Proof: Supposing the updated $\mathbf{W}$ is $\widehat{\mathbf{W}}$, we can easily get:

$$tr\left(\widehat{\mathbf{W}}^T(\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{L_G})\widehat{\mathbf{W}}\right)+\lambda tr\left(\widehat{\mathbf{W}}^T\mathbf{Z}\widehat{\mathbf{W}}\right)$$
$$\le tr(\mathbf{W}^T(\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{L_G})\mathbf{W})+\lambda tr(\mathbf{W}^T\mathbf{Z}\mathbf{W}), \tag{31}$$

where we add $\lambda\sum_i \frac{p*eps}{2\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)^{\frac{2-p}{2}}}$ to both sides of the inequality (31) and replace the definition of $\mathbf{Z}$ with (11), and then we can rewrite the inequality (31):

$$tr\left(\widehat{\mathbf{W}}^T(\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{L_G})\widehat{\mathbf{W}}\right)+\lambda\sum_i\frac{p\left(\widehat{\mathbf{w}}^i\widehat{\mathbf{w}}^{i^T}+eps\right)}{2\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)^{\frac{2-p}{2}}}$$
$$\le tr(\mathbf{W}^T(\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{L_G})\mathbf{W})+\lambda\sum_i\frac{p\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)}{2\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)^{\frac{2-p}{2}}}, \tag{32}$$

Based on Theorem 1, we have:

$$\lambda\sum_i\left(\widehat{\mathbf{w}}^i\widehat{\mathbf{w}}^{i^T}+eps\right)^{\frac{p}{2}}-\lambda\sum_i\frac{p\left(\widehat{\mathbf{w}}^i\widehat{\mathbf{w}}^{i^T}+eps\right)}{2\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)^{\frac{2-p}{2}}}$$
$$\le \lambda\sum_i\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)^{\frac{p}{2}}-\lambda\sum_i\frac{p\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)}{2\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)^{\frac{2-p}{2}}} \tag{33}$$

We add inequalities (31) and (32) together, and then we can get:

$$tr\left(\widehat{\mathbf{W}}^T(\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{L_G})\widehat{\mathbf{W}}\right)+\lambda\sum_i\left(\widehat{\mathbf{w}}^i\widehat{\mathbf{w}}^{i^T}+eps\right)^{\frac{p}{2}}$$
$$\le tr(\mathbf{W}^T(\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{L_G})\mathbf{W})+\sum_i\left(\mathbf{w}^i\mathbf{w}^{i^T}+eps\right)^{\frac{p}{2}} \tag{34}$$

which concludes the proof.

IV. Experiment

A. Data Acquisition

The multimodal data used in this paper is obtained from the PPMI database. It is the first comprehensive, large-scale, and international database to study PD. In this study, we employ MRI and DTI data acquired by the Siemens MAGNETOM Trio 3.0 T MRI scanner. We select MRI data using these parameters: field strength= 3 tesla, flip angle= 9º, slice thickness =1 mm, TR = 2300 ms, TE= 2.98 ms, pulse sequence = GR/IR, and acquisition plane = SAGITTAL. For DTI data, the data acquisition parameters are: field strength = 3 tesla, flip angle = 90º, slice thickness = 2 mm, gradient directions = 64, TR = 600-1000 ms, and TE = 88 ms, pulse sequence = EP.

B. Subjects

In this paper, we collect baseline MRI and DTI data acquired from 238 subjects, including 62 normal control (NC) subjects, 142 PD subjects, and 34 SWEDD subjects. We collect 12-month data acquired from 186 subjects, including 54 NC subjects, 123 PD subjects, and 9 SWEDD subjects. We collect 24-month data acquired from 127 subjects, including 7 NC subjects, 98 PD subjects, and 22 SWEDD subjects. Geriatric Depression Scale (GDS), Epworth Sleepiness Scale (ESS), University of Pennsylvania Smell Identification Test (UPSIT), and Montreal Cognitive Assessment (MoCA) are used to estimate the depression, sleep, olfaction, and cognition scores, respectively. The GDS score is estimated according to a yes/no survey's answer. The GDS scores of normal condition, mild, moderate, and severe depression are between 0 and 4, 5 and 7, 8 and 11, 12 and 15, respectively. The ESS score is estimated according to a weighted sum of responses to several questions. The ESS scores of normal and sleepy subjects are between 0 and 9, 10 and 24, respectively. The UPSIT score is between 0 and 40. A low UPSIT score means that the subject has lost a lot of smell sense. The MoCA score is between 0 and 30. A low MoCA score means that the subject has lost a lot of cognitive ability. Table 1 summarizes the clinical information of the subjects.

TABLE I
CLINICAL DETAILS OF ALL SUBJECTS IN LONGITUDINAL TIME POINT

Time	Information	NC	PD	SWEDD
Baseline	Number	62	142	34
	GDS scores	5.1±1.2	5.3±1.5	5.6±1.3
	ESS scores	6.5±3.9	5.9±3.4	8.7±4.2
	UPSIT scores	33.4±4.7	22.5±8.5	29.9±8.4
	MoCA scores	28.2±1.1	27.5±2.1	26.9±2.9
12m	Number	54	123	9
	GDS scores	4.9±1.4	5.3±1.3	5.6±1.0
	ESS scores	6.1±3.7	6.6±4.3	7.2±3.7
	UPSIT scores	-	-	-
	MoCA scores	27.5±1.9	26.8±2.9	26.4±2.9
24m	Number	7	98	22
	GDS scores	4.9±0.4	5.7±1.6	5.4±1.5
	ESS scores	7.1±3.6	7.9±4.4	7.2±3.6
	UPSIT scores	-	-	-
	MoCA scores	28.3±1.0	26.6±2.9	26.0±2.7

C. Data Preprocessing

For MRI data, we conduct anterior commissure-posterior commissure reorientation exploiting center-of-mass method [37] for all images and use statistical parametric mapping (SPM8) (http://www.fil.ion.ucl.ac.uk/spm) tool to perform the preprocessing procedures based on a well-accepted pipeline. First, we correct the head movement and geometric distortion and then use the graph-cut method [38] to conduct skull-stripping. Then, all MRI images are registered with the international consortium for brain mapping template that provides coordinates of the relevant anatomical labels. After that, we segment the corresponding anatomical regions into GM, white matter, and cerebrospinal fluid (CSF). Meanwhile, these images are resampled to an isotropic resolution of 1.5mm. We spatially smooth the surface of these images using a 60-mm full width at half-maximum Gaussian kernel. The purpose of smoothing is to suppress the interference of noise. Finally, we use a toolbox for data processing and analysis for brain imaging (http://rfmri.org/dpabi) to register automated anatomical labeling (AAL) atlas [39] with GM and extract 116-dimensional features from GM exploiting the registered AAL atlas.

For DTI data, each subject contains 65 original format images where the b0 image does not activate the diffusion gradient, while the other 64 images have different gradient directions. First, we use the FMRIB Software Library (FSL) [40] to correct the b0 image distortion. Second, we use *bet* command of FSL tool to generate a mask image corresponding to the corrected b0 image. Third, we use dcm2nii tool (https://www.nitrc.org/frs/?group_id=152) to convert the 65 images into a 4D image and generate a b-vector file and a b-value file indicating each gradient direction and its scalar value, respectively. Fourth, we use *eddy_correct* command of FSL tool to correct the eddy current distortion on the 4D image. Fifth, we import the b-value file, the b-vector file, the mask image, and the corrected 4D image to *dtifit* command of FSL tool to calculate L1 and V1 images. Finally, we use AAL atlas to calculate the mean tissue density of each region of L1 and V1 and then obtain their 116-dimensional features, respectively.

D. Experimental Setting

We use the 10-fold cross-validation approach to verify the proposed approach in baseline multimodal data (GM, L1, and V1). Specifically, we randomly separate the baseline dataset in ten groups, where one group is used for testing, and the rest is used for training. We duplicate this process ten times to avert the probable bias during data partition. The final result is calculated by averaging the above results. We perform experiments on three binary classifications (i.e., NC vs. PD, NC vs. SWEDD, and PD vs. SWEDD) and prediction of four scores (i.e., depression, sleep, olfaction, and cognition scores) in baseline multimodal data. Due to the existence of data loss problem on longitudinal time points, we use the 5-fold cross-validation approach to verify the proposed approach on the 12-month and 24-month data. In addition, to enhance the generalization ability, we use baseline data as part of the training data to help 12-month data to learn the classification and regression prediction models and exploit baseline and 12-month data as part of the training data to assist 24-month data in determining the classification and regression prediction models. We also conduct three binary classification experiments on the 12-month and 24-month data. Due to the lack of olfaction scores on the 12-month and 24-month data, we only predict depression, sleep, and cognition scores. We determine the optimal SVC/SVR parameters of the support vector machine from $\mathrm{S\_C} \in \{2^{-10}, \ldots, 2^{10}\}$, and $\mathrm{S\_G} \in \{2^{-5}, \ldots, 2^{5}\}$, by performing grid search on the hyper-parameters of our objective function with the spaces of $\lambda \in \{10^0, \ldots, 10^8\}$, $m \in \{100,120 \ldots,240\}$, $k \in \{1,2 \ldots,10\}$, and $p \in \{0.1,0.4, \ldots,1.9\}$. Other parameters of the objective function can be adaptively determined during the model optimization.

E. Algorithm Comparison

The proposed approach is compared with state-of-the-art approaches including (1) principal component analysis (PCA) [18], which is added into the MATLAB software as an unsupervised dimensionality reduction method; (2) Laplacian score (Lscore) [17] (http://www.cad.zju.edu.cn/home/dengcai/), which is an unsupervised feature selection approach (http://www.cad.zju.edu.cn/home/dengcai/); the core idea of Lscore is to estimate the features based on their locality preserving ability; (3) RSFS [19] (https://github.com/LeiShiCS/RSFS), which concurrently exploits FME and ℓ_1 norm to robustly select the discriminative features; (4) the MMSL approach obtained from Lei *et al.* [15], which concurrently conducts classification and regression prediction of PD based on a united multi-task feature selection function that considers the similarity of difference among rows and columns in response matrix; (5) the joint multi-task learning (JMTL) approach from Lei *et al.* [5], which performs classification and prediction of PD based on a united multi-task feature selection function that explores multiple relationships in the response matrix, and (6) the M3T [30] approach based on $\ell_{2,1}$ norm, which learns a feature selection model to gain common relevant features of multiple tasks from every modality; the M3T approach is a particular case of MMSL approach when its two regularization terms set to zero.

F. Model Training

For PCA, we first learn the principal component coefficient matrix, and then we multiply the original data and the coefficient matrix to conduct the feature dimension reduction. Meanwhile, for a fair comparison, other approaches select features: we first calculate the ℓ_2 norm of each row of feature weight matrix $\mathbf{W}$ to obtain the column vector $\mathbf{b}$ and then get its

average value c. Based on the empirical results, we select features corresponding to element values of b greater than or equal to $0.2 \times c$. We import the selected features into the support vector machine (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) to learn SVC models for the classification and SVR models for the clinical score prediction.

G. Evaluation Criteria

To estimate the performance of the competing approaches, we use the quantitative measurements of accuracy (ACC), sensitivity (SEN), specificity (SPEC), precision (PREC), unweighted average recall (UAR), F1-score, and area under the receiver operating characteristic (ROC) curve (AUC) in classification tasks, and also the quantitative measurements of Pearson's correlation coefficient (CC), root mean squared error (RMSE), and mean absolute error (MAE) in regression tasks.

H. Classification Performance

Tables II, III, and IV show the classification performance of the competing approaches on longitudinal multimodal data. Meanwhile, the ROC curves of the related approaches are compared on three binary classification tasks in Fig. 2. According to these results, we can obtain the following findings.

First, unsupervised embedding approaches, such as RSFS and the proposed approach, have better performance than filter approaches such as Lscore and PCA. For example, regarding NC vs. SWEDD classification on the baseline data, RSFS reaches ACC of 85.56%, F1-score of 90.23%, and AUC of 0.75, while our approach reaches ACC of 89.78%, F1-score of 93.28%, and AUC of 0.84. However, PCA achieves ACC only of 80.22%, F1-score of 86.07%, and AUC of 0.75 and Lscore merely achieves ACC of 83.67%, F1-score of 88.68%, and AUC of 0.75. We see that PCA performs best with 24-month data in NC vs. SWEDD and obtains ACC of 82.48%, F1-score of 76%, and AUC of 0.95, and the other approaches perform relatively worse. The main reason may be that the sample size of 24-month data is too small in NC vs. SWEDD. Compared with Lscore, RSFS, M3T, MMSL, JMTL, and the proposed approaches that learn feature weight to select the discriminative features, PCA exploits feature coding to achieve feature dimensionality reduction, and thus it would be more efficient to perform better under limited data.

Second, unsupervised approaches are harder than supervised approaches because the label information is missing. However, the proposed approach has better classification performance than M3T, MMSL, and JMTL. For example, our approach achieves accuracies higher than M3T, MMSL, and JMTL on the baseline data, i.e., our approach and the other three approaches achieve accuracies of 83.33% vs. 78.90% vs. 81.33% vs. 81.81% for NC and PD, 89.78% vs. 82.44% vs. 87.44% vs. 88.67 for NC and SWEDD, and 88.69% vs. 85.85% vs. 87.52%

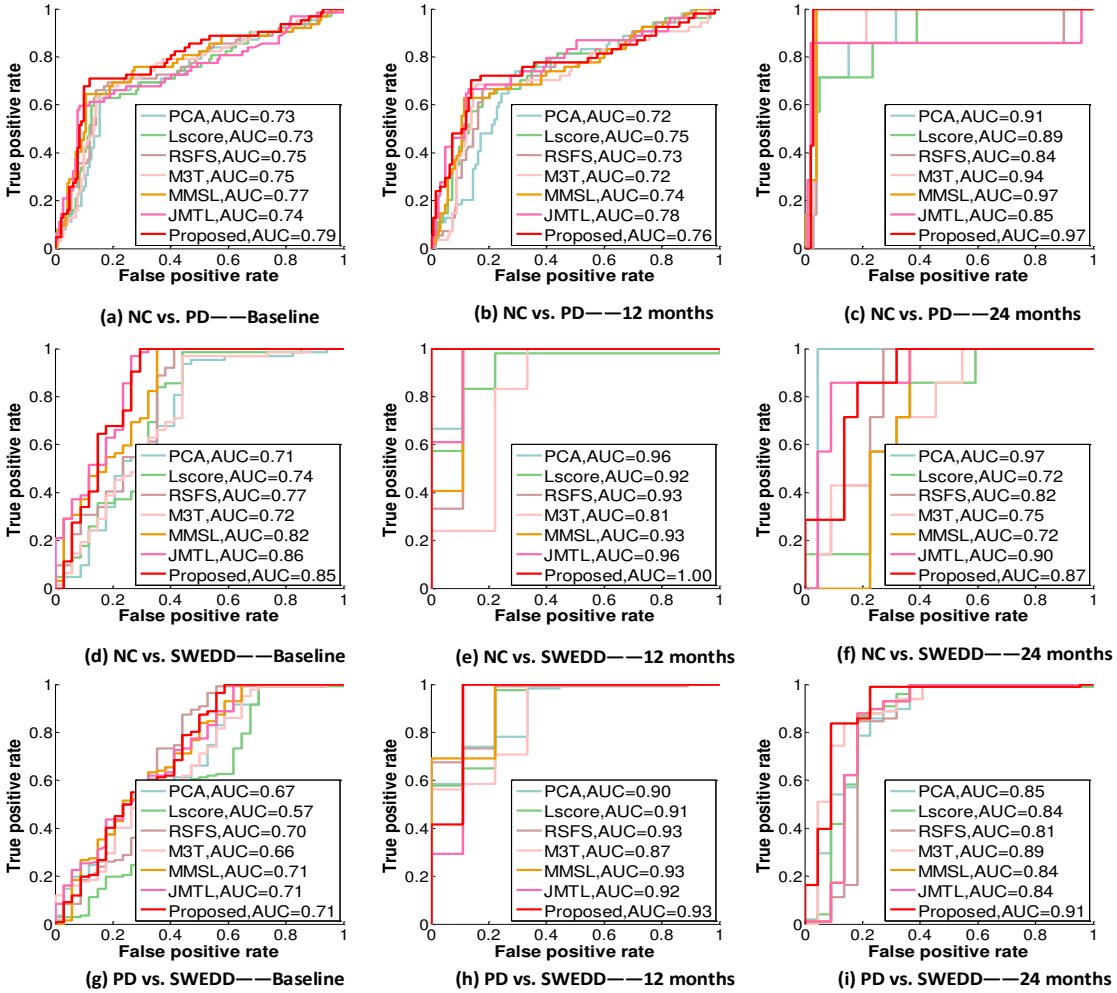


Fig. 2. Comparison of ROC curves for the competing approaches on longitudinal multimodal data.

TABLE II
CLASSIFICATION PERFORMANCE OF THE COMPETING APPROACHES IN NC VS. PD. BOLD INDICATES THE BEST PERFORMANCE (MEAN± STANDARD DEVIATION).

Time	Method	NC vs. PD						
		ACC(%)	SEN(%)	SPEC(%)	PREC(%)	UAR(%)	F1-score (%)	AUC
Baseline	PCA	78.45±6.89	64.05±17.69	84.43±11.12	67.71±15.32	74.24±7.79	63.99±10.54	0.73±0.11
	Lscore	79.45±9.20	63.57±18.20	86.71±10.60	70.60±21.06	75.14±10.43	65.05±15.98	0.74±0.09
	RSFS	79.43±7.25	66.67±19.54	85.33±11.57	69.94±17.75	76.00±8.60	65.80±12.13	0.74±0.11
	M3T	78.90±5.82	62.62±23.60	85.90±11.61	68.88±15.42	74.26±8.69	62.57±15.14	0.73±0.12
	MMSL.	81.33±6.58	64.52±18.63	88.62±10.24	76.06±18.56	76.57±8.20	67.09±11.99	0.75±0.09
	JMTL	81.81±4.90	59.76±15.65	**91.52±6.48**	**77.52±13.95**	75.64±7.11	65.90±10.66	0.75±0.07
	Proposed	**83.33±7.92**	**68.33±16.91**	90.10±6.95	76.52±16.11	**79.21±9.77**	**71.17±13.00**	**0.76±0.16**
12m	PCA	72.29±6.95	64.91±7.03	75.67±11.17	55.42±9.34	70.29±4.74	59.11±6.13	0.72±0.08
	Lscore	74.59±8.73	64.55±13.02	79.07±12.31	60.53±18.46	71.81±8.37	61.13±11.61	0.74±0.07
	RSFS	77.35±8.68	66.55±10.80	82.23±11.04	64.31±14.66	74.39±8.23	64.55±10.87	0.76±0.06
	M3T	79.65±8.45	66.36±9.96	85.47±8.68	67.99±17.01	75.92±8.89	66.88±12.97	0.73±0.11
	MMSL.	80.78±3.13	62.91±6.63	**88.60±5.32**	71.89±9.70	75.75±2.98	66.64±4.78	0.75±0.06
	JMTL	80.84±5.20	65.09±9.22	87.97±9.35	**73.84±18.17**	76.53±4.22	67.70±6.51	**0.79±0.08**
	Proposed	**81.44±8.35**	**70.18±8.59**	86.33±10.39	71.89±19.10	**78.26±7.78**	**70.30±11.83**	0.76±0.12
24m	PCA	94.23±4.12	70.00±44.72	95.89±4.37	60.00±43.46	82.95±22.14	59.33±37.89	0.89±0.11
	Lscore	93.28±2.78	70.00±44.72	94.89±3.72	50.00±37.27	82.45±21.69	52.67±31.30	0.89±0.11
	RSFS	94.33±7.83	90.00±22.36	95.00±8.66	78.33±33.12	92.50±10.61	77.33±25.21	0.88±0.18
	M3T	95.23±3.37	90.00±22.36	95.95±4.20	70.00±29.81	92.97±10.20	72.67±18.62	0.96±0.04
	MMSL	96.23±3.97	100.00±0.00	95.89±4.37	73.33±25.28	97.95±2.19	82.67±16.73	0.97±0.03
	JMTL	97.19±2.57	90.00±22.36	**98.00±2.74**	**83.33±23.57**	94.00±10.69	82.67±16.73	0.90±0.18
	Proposed	**97.19±4.24**	**100.00±0.00**	97.00±4.47	80.00±29.81	**98.50±2.24**	**86.00±21.91**	**0.98±0.03**

TABLE III
CLASSIFICATION PERFORMANCE OF THE COMPETING APPROACHES IN NC VS. SWEDD. BOLD INDICATES THE BEST PERFORMANCE (MEAN± STANDARD DEVIATION).

Time	Method	NC vs. SWEDD						
		ACC(%)	SEN(%)	SPEC(%)	PREC(%)	UAR(%)	F1-score (%)	AUC
Baseline	PCA	80.22±9.07	95.00±8.05	51.67±23.83	79.03±7.53	73.33±11.82	86.07±6.53	0.75±0.12
	Lscore	83.67±9.44	98.33±5.27	57.50±24.04	81.41±10.17	77.92±11.63	88.68±6.49	0.75±0.15
	RSFS	85.56±10.82	100.00±0.00	60.00±27.16	82.79±10.62	80.00±13.58	90.23±6.71	0.75±0.20
	M3T	82.44±9.32	96.67±7.03	55.83±25.17	80.98±8.36	76.25±11.63	87.74±5.96	0.69±0.11
	MMSL	87.44±10.72	100.00±0.00	64.17±28.88	85.24±12.13	82.08±14.44	91.61±7.12	0.77±0.19
	JMTL	88.67±7.35	100.00±0.00	68.33±21.08	85.63±8.82	84.17±10.54	92.05±5.03	**0.85±0.13**
	Proposed	**89.78±12.48**	**100.00±0.00**	**70.83±33.16**	**88.24±12.82**	**85.42±16.58**	**93.28±7.82**	0.84±0.19
12m	PCA	96.92±4.21	98.18±4.07	90.00±22.36	98.33±3.73	94.09±10.85	98.18±2.50	0.95±0.12
	Lscore	95.26±4.34	98.18±4.07	70.00±44.72	96.67±4.56	84.09±21.68	97.31±2.46	0.94±0.10
	RSFS	98.33±3.73	100.00±0.00	80.00±44.72	98.33±3.73	90.00±22.36	99.13±1.94	0.84±0.37
	M3T	95.38±6.88	100.00±0.00	70.00±44.72	95.26±6.96	85.00±22.36	97.46±3.75	0.83±0.32
	MMSL	98.46±3.44	100.00±0.00	90.00±22.36	98.33±3.73	95.00±11.18	99.13±1.94	0.94±0.14
	JMTL	98.46±3.44	100.00±0.00	90.00±22.36	98.33±3.73	95.00±11.18	99.13±1.94	0.93±0.16
	Proposed	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**1.00±0.00**
24m	PCA	**82.48±12.04**	**100.00±0.00**	**76.00±17.82**	**63.33±21.73**	**88.00±8.91**	**76.00±14.61**	0.95±0.11
	Lscore	49.71±17.42	100.00±0.00	34.00±23.02	33.05±10.93	67.00±11.51	48.89±12.04	0.75±0.25
	RSFS	72.95±13.57	100.00±0.00	65.00±15.41	48.33±14.91	82.50±7.71	64.00±14.61	0.87±0.19
	M3T	60.57±17.97	100.00±0.00	47.00±26.60	39.67±10.83	73.50±13.30	56.10±11.41	0.80±0.20
	MMSL	72.10±23.56	100.00±0.00	64.00±30.70	53.00±28.20	82.00±15.35	66.10±21.86	0.73±0.28
	JMTL	72.29±27.73	100.00±0.00	65.00±34.28	60.00±37.91	82.50±17.14	69.33±30.04	**0.96±0.09**
	Proposed	67.90±23.05	100.00±0.00	57.00±31.14	51.67±29.11	78.50±15.57	64.67±22.80	0.84±0.11

TABLE IV
CLASSIFICATION PERFORMANCE OF THE COMPETING APPROACHES IN PD VS. SWEDD. BOLD INDICATES THE BEST PERFORMANCE (MEAN± STANDARD DEVIATION).

Time	Method	PD vs. SWEDD						
		ACC(%)	SEN(%)	SPEC(%)	PREC(%)	UAR(%)	F1-score (%)	AUC
Baseline	PCA	85.82±4.76	99.29±2.26	28.33±28.65	85.76±4.99	63.81±13.71	91.92±2.54	0.62±0.18
	Lscore	85.78±4.78	99.29±2.26	28.33±29.19	85.87±6.07	63.81±13.63	91.93±2.57	0.51±0.18
	RSFS	88.04±3.31	99.29±2.26	39.17±22.58	87.74±3.59	69.23±10.70	93.09±1.78	0.64±0.21
	M3T	85.85±4.65	99.29±2.26	29.17±26.13	85.76±4.99	64.23±12.40	91.92±2.54	0.63±0.19
	MMSL	87.52±5.69	100.00±0.00	34.17±27.90	86.92±5.69	67.08±13.95	92.91±3.19	0.65±0.22
	JMTL	88.07±5.58	100.00±0.00	37.50±27.29	87.40±5.28	68.75±13.64	93.20±3.04	0.67±0.18
	Proposed	**88.69±5.79**	**100.00±0.00**	**40.83±28.72**	**88.00±5.50**	**70.42±14.36**	**93.54±3.17**	**0.69±0.19**
12m	PCA	96.95±3.13	100.00±0.00	50.00±50.00	96.92±3.14	75.00±25.00	98.41±1.62	0.89±0.21
	Lscore	97.75±3.33	100.00±0.00	70.00±44.72	97.72±3.35	85.00±22.36	98.82±1.73	0.90±0.22
	RSFS	97.78±3.31	99.20±1.79	80.00±44.72	98.52±3.31	89.60±22.15	98.82±1.73	0.90±0.20
	M3T	97.72±3.35	100.00±0.00	60.00±54.77	97.72±3.35	80.00±27.39	98.82±1.73	0.79±0.29
	MMSL	98.52±3.31	100.00±0.00	80.00±44.72	98.52±3.31	90.00±22.36	99.23±1.72	0.89±0.25
	JMTL	98.52±2.03	100.00±0.00	80.00±27.39	99.23±1.72	90.00±13.69	99.61±0.88	**0.94±0.13**

	Proposed	**99.26±1.66**	**100.00±0.00**	**90.00±22.36**	**99.23±1.72**	**95.00±11.18**	**99.61±0.88**	0.92±0.18
	PCA	92.49±5.46	100.00±0.00	60.00±29.37	91.81±5.78	80.00±14.68	95.65±3.11	0.88±0.11
	Lscore	92.49±5.46	98.95±2.35	65.00±23.45	92.53±5.12	81.97±12.20	95.58±3.20	0.83±0.18
	RSFS	93.32±5.61	100.00±0.00	64.00±30.70	92.71±5.88	82.00±15.35	96.14±3.17	0.82±0.24
24m	M3T	92.49±5.46	100.00±0.00	60.00±29.37	91.81±5.78	80.00±14.68	95.65±3.11	0.88±0.11
	MMSL	93.36±4.64	100.00±0.00	65.00±23.45	92.63±5.06	82.50±11.73	96.12±2.70	0.82±0.20
	JMTL	93.36±4.64	**100.00±0.00**	65.00±23.45	92.63±5.06	82.50±11.73	96.12±2.70	0.82±0.20
	Proposed	**95.03±3.40**	98.95±2.35	**77.00±22.80**	**95.36±4.55**	**87.97±10.78**	**97.04±1.97**	**0.89±0.10**

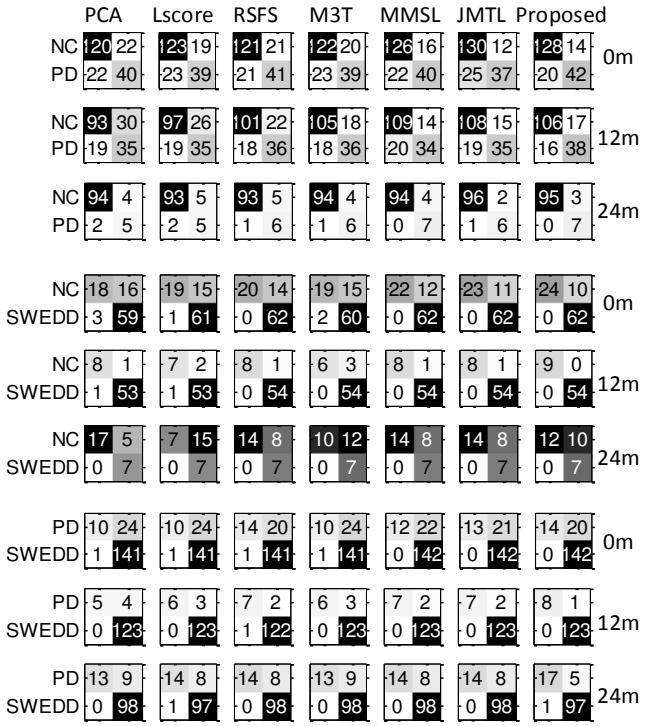


Fig. 3. Confusion matrices for the competing approaches on longitudinal data.

vs. 88.07% for PD vs. SWEDD. The results demonstrate that the proposed approach can better learn the structural property intrinsic in data.

Third, on the baseline data, specificity values of the competing approaches are low for discriminating PD and SWEDD. For example, PCA, Lscore, RSFS, M3T, MMSL, JMTL, and the proposed approaches have specificity values of 28.33%, 28.33%, 39.17%, 29.17%, 34.17%, 37.50%, and 40.83%. The reasons may be that it is difficult to distinguish SWEDD patients from PD since both PD and SWEDD have asymmetric rest tremor, and the size of data is limited. When we exploit baseline data to help 12-month data to learn the classification model and exploit baseline and 12-month data to assist 24-month data in learning the classification and model, the classification performance improves. For instance, on the 12-month data, our approach achieves ACC of 99.26%, SEN of 100%, SPEC of 90%, PREC of 99.23%, UAR of 95%, F1-score of 99.61%, and AUC of 0.92. On the 24-month data, the proposed approach has ACC of 95.03%, SEN of 98.95%, SPEC of 77%, PREC of 95.36%, UAR of 87.97%, F1-score of 97.04%, and AUC of 0.89.

Finally, the proposed approach achieves the best performance on most longitudinal time points. Taking NC vs. PD as an example, on the baseline data, our approach has ACC of 83.33%, SEN of 68.33%, SPEC of 90.10%, PREC of 76.52%, UAR of 79.21%, F1-score of 71.17%, and AUC of 0.76. On the 12-month data, the proposed approach has ACC of 81.44%, SEN of 70.18%, SPEC of 86.33%, PREC of 71.89%, UAR of 78.26%, F1-score of 70.30%, and AUC of 0.76. On the 24-month data, our approach has ACC of 97.19%, SEN of 100%, SPEC of 97%, PREC of 80%, UAR of 98.5%, F1-score of 86%, and AUC of 0.98. Meanwhile, we see that our approach obtains the highest UAR overall on the longitudinal data in Tables II, III, and IV, also demonstrating the superiority of our approach. Further, we obtain the overall/average confusion matrices through linearly connecting the predicted labels and the actual labels of the test data in the entire cross-validation process. Fig.3 shows the confusion matrices of the competing approaches on longitudinal data. We can observe that our approach obtains the optimum performance.

I. Regression Performance

Tables V, VI, and VII summarize the regression performances of the competing approaches on longitudinal multimodal data. According to CC results, our approach and JMTL approach obtain the optimum regression performance overall.

In the regression performance for NC vs. PD, our approach achieves the optimal performance for olfaction and cognition scores on the baseline data. The corresponding CC, RMSE, and MAE are 0.553, 8.243, and 6.820 for olfaction score, and 0.608, 3.355, and 2.744 for cognition score. JMTL shows the optimal performance for depression score on the baseline data. The corresponding CC, RMSE, and MAE are 0.603, 4.320, and 3.466 for depression score. MMSL shows the optimal performance for sleep score on the baseline data. The corresponding CC, RMSE, and MAE are 0.569, 5.743, and 4.683. On the 12-month data, our approach achieves the optimal performance for depression and sleep scores. The corresponding CC, RMSE, MAE are 0.543, 2.083, and 1.643 for depression score, 0.514, 6.019, and 4.800 for sleep score. JMTL shows the optimal performance for cognition score. The corresponding CC, RMSE, and MAE are 0.590, 2.681, and 2.006 for cognition score. On the 24-month data, our approach has the optimal performance for depression and cognition scores. The corresponding CC, RMSE, and MAE are 0.655, 1.546, and 1.064 for depression score and 0.726, 2.113, and 1.641 for cognition score. JMTL has the optimal performance for sleep score on the 24-month data. The corresponding CC, RMSE, and MAE are 0.587, 5.425, and 4.293.

In the regression performance for NC vs. SWEDD, our approach obtains the optimal performance for sleep, olfaction, and cognition scores on the baseline data. The corresponding results are CC of 0.763, RMSE of 5.147, and MAE of 4.427 for sleep score, CC of 0.775, RMSE of 5.752, and MAE of 4.374 for olfaction score, and CC of 0.790, RMSE of 3.632, and MAE of 3.062 for cognition score. JMTL obtains the optimal performance for depression score on the baseline data. The corresponding results are CC of 0.777, RMSE of 2.473, and MAE

TABLE V
REGRESSION PERFORMANCE OF THE COMPETING APPROACHES IN NC VS. PD. BOLD INDICATES THE BEST PERFORMANCE (MEAN).

Time	Method	Depression score			Sleep score			Olfaction score			Cognition score		
		CC	RMSE	MAE	CC	RMSE	MAE	CC	RMSE	MAE	CC	RMSE	MAE
Baseline	PCA	0.446	4.648	3.812	0.461	4.573	3.684	0.451	8.744	7.286	0.523	3.693	2.949
	Lscore	0.516	3.852	3.026	0.498	4.755	3.780	0.489	8.384	6.976	0.535	**2.968**	**2.348**
	RSFS	0.560	2.605	1.931	0.561	4.463	**3.536**	0.525	8.477	7.091	0.582	5.372	4.343
	M3T	0.527	3.612	2.680	0.492	4.864	3.943	0.532	8.952	7.571	0.567	3.185	2.505
	MMSL	0.570	4.404	3.557	**0.569**	5.743	4.683	0.525	8.447	7.081	0.581	4.138	3.440
	JMTL	**0.603**	4.320	3.466	0.545	**4.355**	3.539	0.533	8.535	7.151	0.587	3.36	2.617
	Proposed	0.601	**2.410**	**1.829**	0.545	4.879	3.981	**0.553**	**8.243**	**6.820**	**0.608**	3.355	2.744
12m	PCA	0.430	2.652	2.099	0.421	6.086	4.712	--	--	--	0.491	3.723	2.892
	Lscore	0.440	2.955	2.358	0.400	5.543	4.305	--	--	--	0.462	**2.605**	**1.943**
	RSFS	0.502	3.639	2.972	0.414	6.516	5.409	--	--	--	0.540	3.722	2.769
	M3T	0.489	4.058	3.236	0.407	5.340	**4.079**	--	--	--	0.531	2.639	1.989
	MMSL	0.525	3.271	2.615	0.498	**5.168**	4.092	--	--	--	0.557	3.167	2.435
	JMTL	0.535	3.029	2.373	0.469	6.164	4.784	--	--	--	**0.590**	2.681	2.006
	Proposed	**0.543**	**2.083**	**1.643**	**0.514**	6.019	4.800	--	--	--	0.574	3.489	2.433
24m	PCA	0.522	2.136	1.507	0.444	5.139	4.147	--	--	--	0.635	4.336	3.506
	Lscore	0.548	2.067	1.555	0.434	**4.542**	**3.514**	--	--	--	0.621	2.669	2.022
	RSFS	0.619	1.579	1.157	0.547	5.184	3.800	--	--	--	0.695	3.645	2.948
	M3T	0.562	1.644	1.154	0.476	4.971	3.949	--	--	--	0.655	3.811	2.952
	MMSL	0.637	1.792	1.271	0.557	5.919	4.688	--	--	--	0.676	2.411	1.824
	JMTL	0.621	1.654	1.150	**0.587**	5.425	4.293	--	--	--	0.705	2.158	1.682
	Proposed	**0.655**	**1.546**	**1.064**	0.550	4.878	3.814	--	--	--	**0.726**	**2.113**	**1.641**

TABLE VI
REGRESSION PERFORMANCE OF THE COMPETING APPROACHES IN NC VS. SWEDD. BOLD INDICATES THE BEST PERFORMANCE (MEAN).

Time	Method	Depression score			Sleep score			Olfaction score			Cognition score		
		CC	RMSE	MAE	CC	RMSE	MAE	CC	RMSE	MAE	CC	RMSE	MAE
Baseline	PCA	0.714	3.032	2.312	0.600	4.443	3.602	0.673	7.284	5.811	0.654	4.328	3.414
	Lscore	0.639	4.545	3.532	0.686	5.282	4.392	0.693	7.003	5.263	0.656	2.932	**2.050**
	RSFS	0.720	2.971	2.377	0.739	4.951	4.060	0.684	**5.716**	**4.224**	0.713	3.687	2.917
	M3T	0.693	2.902	2.291	0.708	4.969	4.094	0.711	6.740	5.234	0.675	3.007	2.299
	MMSL	0.764	4.703	3.829	0.728	5.199	4.317	0.748	6.069	4.576	0.759	**2.835**	2.298
	JMTL	**0.777**	2.473	1.877	0.731	**3.720**	**3.233**	0.727	7.198	5.273	0.782	2.922	2.238
	Proposed	0.775	**2.274**	**1.736**	**0.763**	5.147	4.427	**0.775**	5.752	4.374	**0.790**	3.632	3.062
12m	PCA	0.551	3.066	2.554	0.588	4.074	3.313	--	--	--	0.640	**2.443**	**2.121**
	Lscore	0.623	1.757	1.389	0.507	4.062	3.353	--	--	--	0.602	2.485	2.125
	RSFS	0.708	**1.108**	**0.811**	0.633	3.889	3.112	--	--	--	0.709	3.445	2.776
	M3T	0.749	2.833	2.004	0.564	4.580	3.584	--	--	--	0.622	2.491	2.133
	MMSL	0.735	2.717	2.032	0.694	**3.309**	**2.707**	--	--	--	0.699	2.922	2.439
	JMTL	**0.809**	1.289	0.887	**0.763**	3.828	3.122	--	--	--	**0.788**	3.605	2.871
	Proposed	0.733	2.872	2.263	0.683	5.866	4.961	--	--	--	0.720	3.365	2.912
24m	PCA	0.745	1.640	1.240	0.818	5.760	4.518	--	--	--	0.787	4.138	3.410
	Lscore	0.758	**1.303**	**0.946**	0.756	3.988	3.171	--	--	--	0.818	4.427	3.606
	RSFS	0.875	1.609	1.209	0.916	**3.097**	**2.492**	--	--	--	0.893	3.962	3.280
	M3T	0.846	1.318	1.021	0.836	5.402	4.382	--	--	--	0.909	4.761	3.938
	MMSL	0.873	2.586	2.185	0.900	6.230	5.364	--	--	--	0.980	4.433	3.460
	JMTL	0.877	3.869	3.203	**0.938**	7.763	6.491	--	--	--	**0.990**	**3.128**	**2.446**
	Proposed	**0.888**	1.542	1.185	0.924	3.791	3.127	--	--	--	0.944	3.864	3.339

TABLE VII
REGRESSION PERFORMANCE OF THE COMPETING APPROACHES IN PD VS. SWEDD. BOLD INDICATES THE BEST PERFORMANCE (MEAN).

Time	Method	Depression score			Sleep score			Olfaction score			Cognition score		
		CC	RMSE	MAE	CC	RMSE	MAE	CC	RMSE	MAE	CC	RMSE	MAE
Baseline	PCA	0.471	**2.128**	**1.558**	0.501	**3.918**	**3.084**	0.500	8.552	6.979	0.589	3.185	2.393
	Lscore	0.487	3.824	2.952	0.403	4.757	3.705	0.439	8.875	7.431	0.587	**2.869**	**2.186**
	RSFS	0.558	3.277	2.559	0.523	4.477	3.642	0.474	8.934	7.374	0.590	3.624	2.660
	M3T	0.545	4.107	3.038	0.532	4.464	3.536	0.507	8.398	**6.927**	0.594	3.272	2.522
	MMSL	**0.601**	4.237	3.344	0.554	5.065	3.981	0.525	8.439	7.218	0.627	3.916	3.132
	JMTL	0.599	4.966	3.994	**0.595**	5.431	4.323	0.522	8.663	7.258	**0.644**	2.890	2.216
	Proposed	0.593	2.482	1.931	0.558	4.802	3.841	**0.575**	**8.277**	6.945	0.632	3.402	2.649
12m	PCA	0.437	**1.748**	**1.311**	0.350	5.537	4.525	--	--	--	0.489	3.760	2.925
	Lscore	0.394	2.178	1.706	0.453	4.890	3.948	--	--	--	0.517	4.311	3.269
	RSFS	0.533	3.269	2.614	0.494	4.782	3.853	--	--	--	0.553	**2.790**	**2.104**
	M3T	0.489	4.523	3.678	0.445	5.313	3.996	--	--	--	0.526	3.744	2.974
	MMSL	0.529	4.498	3.637	0.527	5.225	4.091	--	--	--	0.609	3.215	2.342
	JMTL	**0.588**	2.908	2.384	**0.556**	**4.546**	**3.652**	--	--	--	**0.641**	3.266	2.501

	Proposed	0.561	2.123	1.719	0.535	4.862	3.885	--	--	--	0.580	3.716	2.811
	PCA	0.511	1.452	1.064	0.441	4.814	3.750	--	--	--	0.686	3.678	3.003
	Lscore	0.491	1.563	1.123	0.392	5.568	4.382	--	--	--	0.673	3.082	2.503
	RSFS	0.576	2.713	2.193	0.501	**4.420**	**3.373**	--	--	--	0.726	3.119	2.520
24m	M3T	0.533	3.147	2.554	0.442	4.781	3.936	--	--	--	0.670	3.086	2.503
	MMSL	0.558	2.158	1.644	0.495	5.080	3.962	--	--	--	0.703	3.113	2.521
	JMTL	0.558	2.158	1.644	0.496	6.480	5.120	--	--	--	0.698	3.400	2.699
	Proposed	**0.636**	**1.335**	**0.982**	**0.540**	4.688	3.535	--	--	--	**0.726**	**2.995**	**2.491**

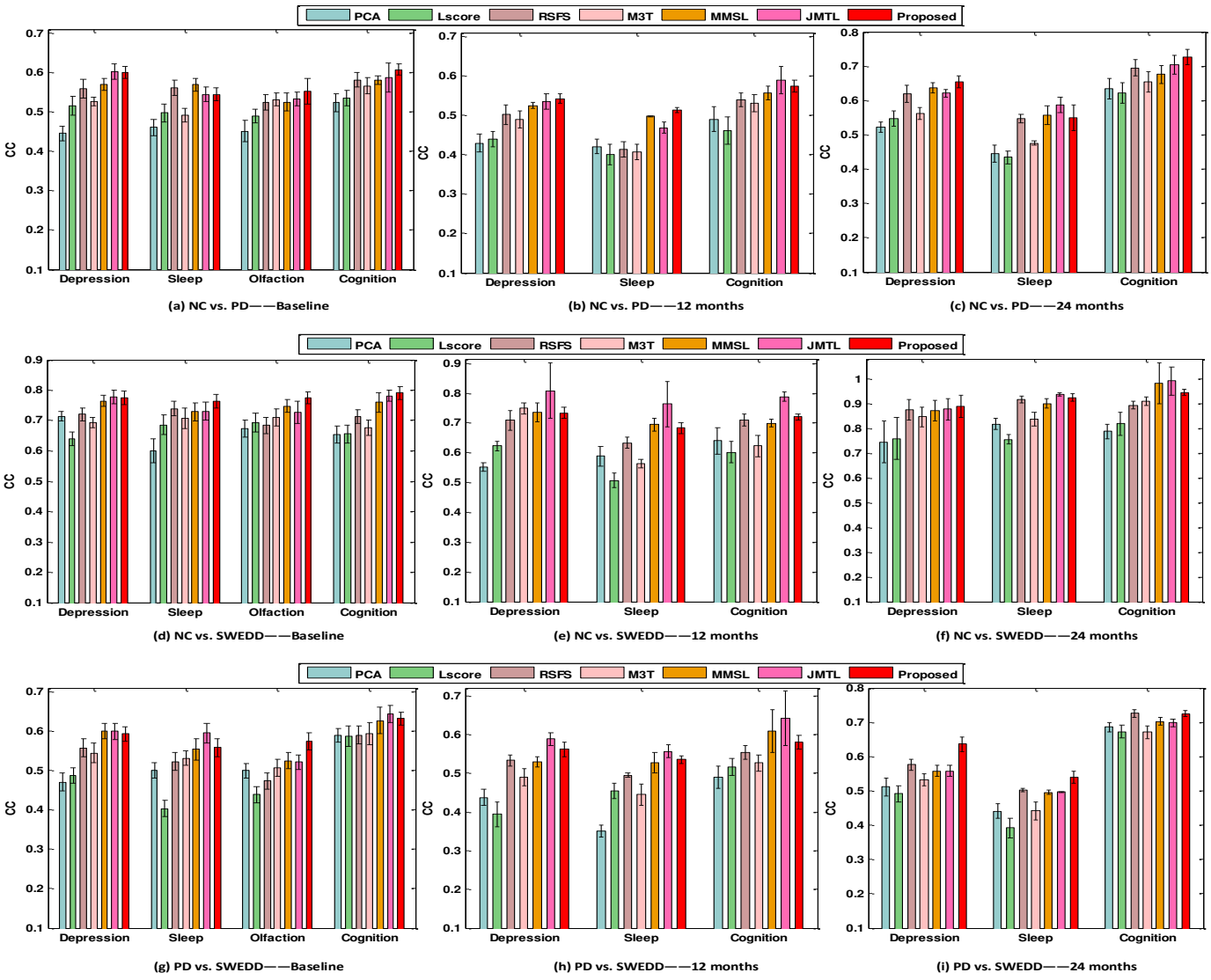


Fig. 4. Regression performance for the competing approaches on longitudinal multimodal data.

of 1.877 for depression score. On the 12-month data, JMTL shows the optimal performance for depression, sleep, and cognition scores. The corresponding results are CC of 0.809, RMSE of 1.289, and MAE of 0.887 for depression, CC of 0.763, RMSE of 3.828, and MAE of 3.122 for sleep score, and CC of 0.788, RMSE of 3.605, and MAE of 2.871 for cognition score. On the 24-month data, our approach achieves the optimal performance for depression score. The corresponding results are CC of 0.888, RMSE of 1.542, and MAE of 1.185 for depression score. JMTL shows the optimal performance for sleep and cognition scores. The corresponding results are CC of 0.938, RMSE of 7.763, and MAE of 6.491 for sleep score and CC of 0.990, RMSE of 3.128, and MAE of 2.446 for cognition score.

In the regression performance for PD vs. SWEDD, our approach shows the optimal performance for olfaction score in baseline data. The corresponding CC, RMSE, and MAE are 0.575, 8.277, and 6.945. MMSL shows the optimal performance for depression score. The corresponding CC, RMSE, and MAE are 0.601, 4.237, and 3.344. JMTL shows the optimal performance for sleep and cognition scores. The corresponding CC, RMSE, and MAE are 0.595, 5.431, and 4.323 for sleep score and 0.644, 2.890, and 2.216 for cognition score. On the 12-month data, JMTL shows the optimal performance for depression, sleep, and cognition scores. The corresponding CC, RMSE, and MAE are 0.588, 2.908, and 2.384 for depression score, 0.556, 4.546, and 3.652 for sleep score, and 0.641, 3.266,

and 2.501 for cognition score. On the 24-month data, our approach shows the optimal performance for depression, sleep, and cognition scores. The corresponding CC, RMSE, and MAE are 0.636, 1.335, and 0.982 for depression score, 0.540, 4.688, and 3.535 for sleep score, and 0.726, 2.995, and 2.491 for cognition score. Fig. 4 also illustrates regression performance for the competing approaches on longitudinal data, and our approach has good performance.

J. *Results Summary*

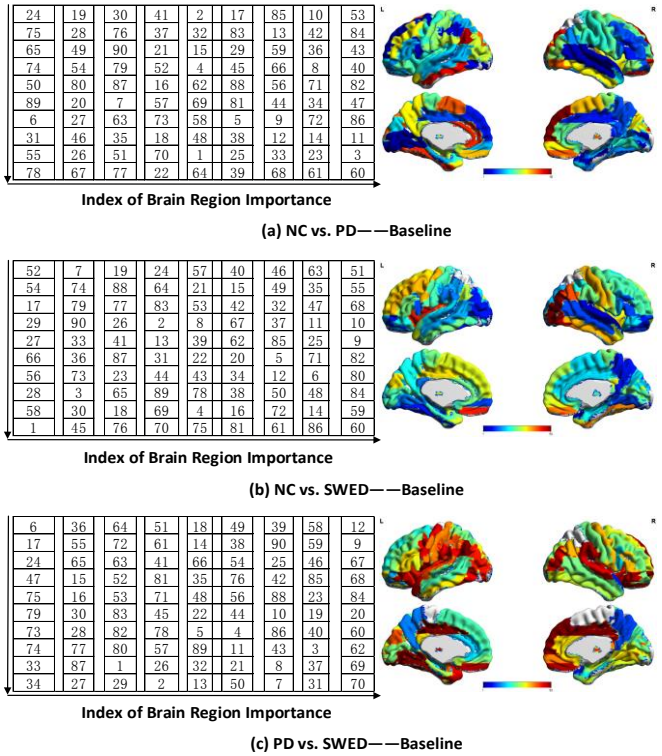


Fig. 5. Index of brain region importance and corresponding visual representation on the baseline data (Note that the importance of brain regions is sorted from top to bottom and left to right in the left table).

According to the experimental results, we observe that our approach has excellent classification performance and regression performance in NC vs. PD, NC vs. SWEDD, and PD vs. SWEDD. The proposed approach is far superior to all other approaches for the classification tasks on longitudinal data. Meanwhile, we sum up the weights of the same brain regions from three modalities to obtain the discriminative brain region index. Fig.5. shows the index of brain region importance of baseline time point and its corresponding visual representation exploiting the BrainNet Viewer tool [41]. For NC vs. PD on the baseline data, the top ten brain regions are: right superior frontal gyrus, left pallidum, left angular gyrus, right putamen, right superior occipital gyrus, left inferior temporal gyrus, right superior frontal gyrus (orbital part), left anterior cingulate and paracingulate gyri, left fusiform gyrus, and right thalamus. For NC vs. SWEDD, the top ten brain regions are right middle occipital gyrus, right inferior occipital gyrus, left Rolandic operculum, left insula, left gyrus rectus, right angular gyrus, right fusiform gyrus, right gyrus rectus, right postcentral gyrus, and left precentral gyrus. For PD vs. SWEDD, the top ten brain regions are right superior frontal gyrus (orbital part), left Rolandic operculum, right superior frontal gyrus, left lingual gyrus, left pallidum, left Heschl gyrus, bilateral putamen, and bilateral median cingulate and paracingulate gyri. It is noteworthy these brain regions follow the previous PD studies such as superior frontal gyrus, pallidum, angular, and putamen in [42], superior frontal gyrus (orbital part), inferior temporal gyrus, and fusiform gyrus in [43]. The top regions strongly correlate with PD, which may be the potential factors causing the disease.

To further discover the connection relationships between the top ten brain regions and other brain regions, we use the feature weights obtained by ten-fold cross-validation on our approach to calculate the Pearson correlation coefficient matrix. We then exploit the matrix to obtain five other brain regions strongly connected to each top brain region. In Fig 6, we exploit the *circularGraph* function of MATLAB software to generate the correlation maps. In NC vs. PD, we can see that some other brain regions strongly connected to multiple top brain regions follow the previous PD studies such as Heschl gyrus and superior parietal gyrus [44, 45]. We can also see that the top ten brain regions are not only strongly related to each other but also connected to other brain regions in NC vs. PD and PD vs. SWEDD. However, in [5], most of the top brain regions are only related to each other. Compared with the results in [5], our correlation maps can show richer information and thus are more suitable as a complement for PD diagnosis. The possible reason is that the multi-modal data (i.e., GM, L1, and V1) used in this paper is more able to present brain change information more effectively. For example, GM of MRI is widely used to obtain information about changes in nerve cells. L1 and V1 of DTI indicate the largest diffusion coefficient and its direction vector, respectively. Therefore, L1 and V1 may be more sensitive to neurodegeneration in the brain. Meanwhile, many current approaches [46, 47] use resting-state functional MRI data to draw correlation maps, which can effectively show the functional connection differences between different groups or stages. In this paper, we use MRI and DTI data to draw correlation maps, which can effectively show the connection relationships between the top brain regions and other related brain regions at the structural level, which is complementary to these functional studies for PD diagnosis. In addition, we can also see that the brain region importance and the correlation maps obtained by our approach have many differences from those obtained in [5]. The main reason is that we use data from different modalities in the two works. Specifically, we use GM, L1, and V1 data in this paper while using GM, CSF of MRI, and mean diffusion (MD) coefficient of DTI in [5].

Finally, Fig. 7 further illustrates the top twenty brain regions connectivity on longitudinal data, where the blue, green, and cyan lines indicate baseline and 12-month, and 24-month data, respectively. We can see that the condition worsens over time, and the brain lesion regions also change, which can verify that the proposed approach is effective. In NC vs. PD, we can also see that some top brain regions present at three-time points follow the previous PD studies such as angular gyrus, superior occipital gyrus, and inferior temporal gyrus [48, 49]. These brain regions are also related to brain cognition. At present, most studies mainly focus on PD, and there are few studies on SWEDD [5]. In Fig. 6 and Fig. 7, we also show the correlation and connection maps of NC vs. SWEDD and PD vs. SWEDD, which is helpful to guide doctors to further divide the disease.

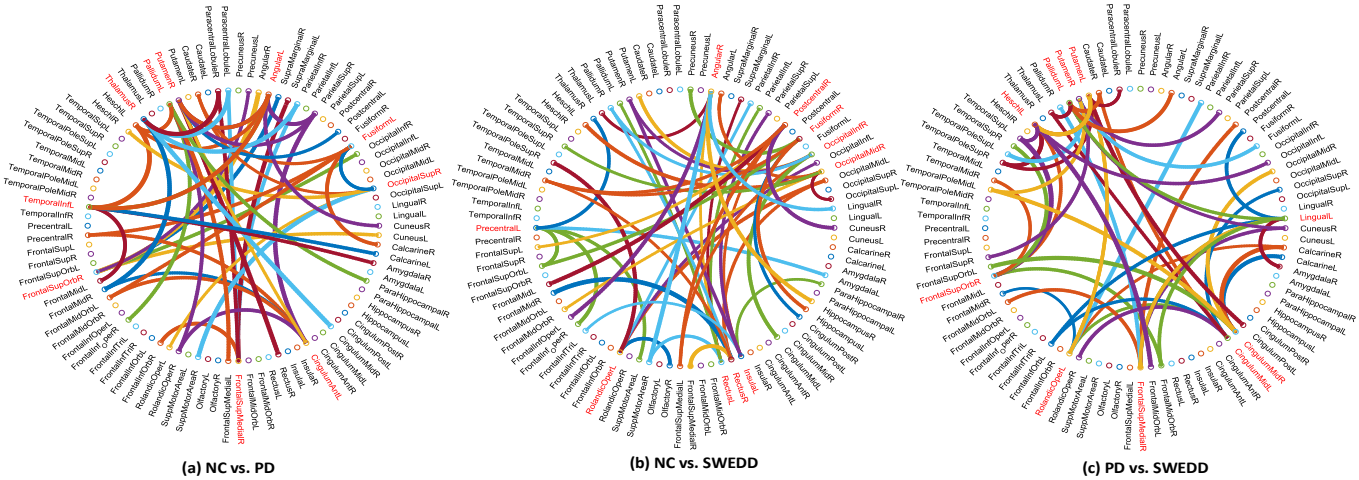


Fig. 6. Correlation maps between top ten brain regions with red fonts and other brain regions with black fonts on the baseline data.

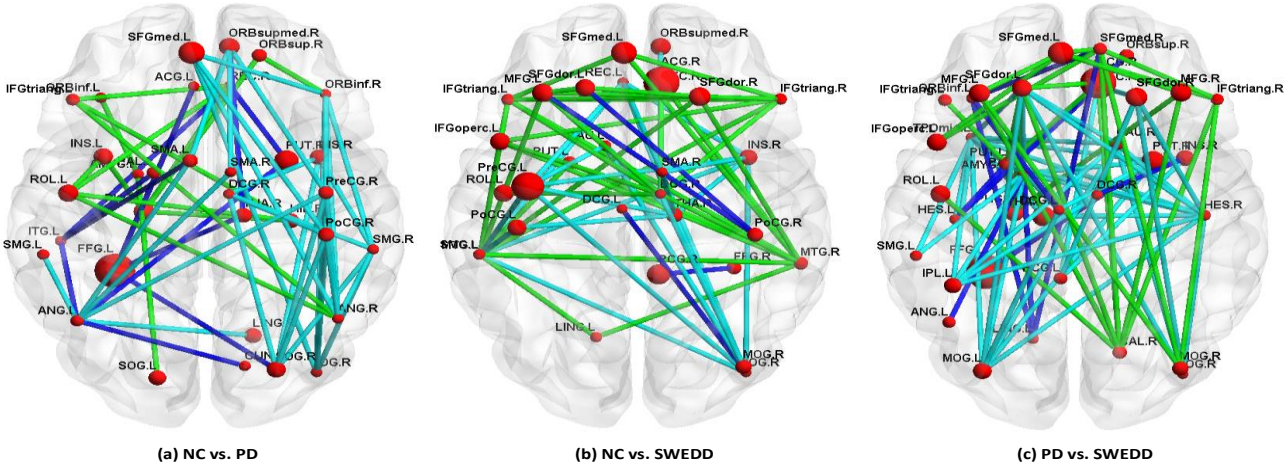


Fig. 7. Top twenty brain regions connectivity on longitudinal data.

V. Discussions

Our approach has achieved promising performance, but there are still several limitations. First, in the data preprocessing, MRI images are registered to a standard template for tissue segmentation, which may remove some pathological changes of PD. We can use deep learning techniques to segment neuroimaging data to eliminate this effect [50-52]. Second, in this paper, we conduct the three binary classification tasks instead of a multi-class classifier for disease classification, which is consistent with the doctor's clinical diagnosis. The doctor first diagnoses whether the subject is sick and then further tests the condition of the subject in detail. Many previous studies have learned multiple binary classifiers instead of a multi-classifier for disease classification [7]. Third, our approach does not consider the relationship among longitudinal data. We are considering to add the regularization term between longitudinal data to enhance the generalization ability of the proposed approach [53]. Fourth, our work only uses MRI and DTI data to perform united classification and regression. Since some genes are related to PD [54], we can combine gene and neuroimaging data for improving the classification and regression performance. Finally, we do not analyze the full extent DTI data in this paper. For example, we can use DTI data to generate fiber bundle imaging of top brain regions and quantify white matter fiber differences between NC and PD [55], which may clinically contribute to the classification and prediction of PD.

VI. Conclusions

In this paper, we propose a novel adaptive unsupervised feature selection approach through embedding learning using longitudinal multimodal data for the united classification and clinical score prediction of PD. Specifically, the proposed approach concurrently performs adaptive embedding learning and sparse regression; the similarity matrices and discriminative features thus can be determined adaptively. Meanwhile, we dynamically update the similarity matrices among subjects and features and have the connected number of the similarity matrix among subjects equal to the number of classes to gain the intrinsic structural property of the data. An effective iterative optimization algorithm is proposed to solve this problem. By constructing the united embedding and sparse regression framework, our approach can find the most disease-related biomarkers, which is helpful for PD monitoring. We perform abundant experiments on the PPMI dataset to verify the validity of the proposed approach. We use longitudinal data to boost the performance of regression and classification effectively. The proposed approach is shown to surpass other state-of-the-art methods.

REFERENCES

[1] L. V. Kalia and A. E. Lang, "Parkinson's disease," *The Lancet,* vol. 386, pp. 896-912, Aug 2015.

[2] R. B. Postuma, D. Berg, M. Stern, W. Poewe, C. W. Olanow, W. Oertel*, et al.*, "MDS clinical diagnostic criteria for Parkinson's disease," *Mov Disord,* vol. 30, pp. 1591-601, Oct 2015.

[3] J. Lotharius and P. Brundin, "Pathogenesis of parkinson's disease: dopamine, vesicles and [alpha]-synuclein," *Nat Rev Neurosci,* vol. 3, pp. 932-942, Jan 2002.

[4] B. S. Connolly and A. E. Lang, "Pharmacological Treatment of Parkinson Disease: A Review," *JAMA,* vol. 311, pp. 1670-1683, 2014.

[5] H. Lei, Z. Huang, F. Zhou, A. Elazab, E. Tan, H. Li*, et al.*, "Parkinson's Disease Diagnosis via Joint Learning From Multiple Modalities and Relations," *IEEE Journal of Biomedical and Health Informatics,* vol. 23, pp. 1437-1449, 2019.

[6] Y. Li, J. Liu, Z. Tang, and B. Lei, "Deep Spatial-Temporal Feature Fusion From Adaptive Dynamic Functional Connectivity for MCI Identification," *IEEE Transactions on Medical Imaging,* vol. 39, pp. 2818-2830, 2020.

[7] E. Adeli, K. Thung, L. An, G. Wu, F. Shi, T. Wang*, et al.*, "Semi-Supervised Discriminative Classification Robust to Sample-Outliers and Feature-Noises," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, pp. 515-522, 2019.

[8] H. I. Suk, S. W. Lee, and D. Shen, "Subclass-based multi-task learning for Alzheimer's disease diagnosis," *Front Aging Neurosci,* vol. 6, p. 168, 2014.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian*, et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis,* vol. 42, pp. 60-88, 2017.

[10] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *the Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1302-1308.

[11] X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen, "Subspace Regularized Sparse Multi-Task Learning for Multi-Class Neurodegenerative Disease Identification," *IEEE transactions on biomedical engineering,* vol. 63, pp. 607-618, 2016.

[12] H. Lei, Z. Huang, A. Elazab, H. Li, and B. Lei, "Longitudinal and Multi-modal Data Learning via Joint Embedding and Sparse Regression for Parkinson's Disease Diagnosis," in *Machine Learning in Medical Imaging*, 2018, pp. 310-318.

[13] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B-Statistical Methodology,* vol. 67, pp. 301-320, 2005.

[14] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B-Methodological,* vol. 58, pp. 267-288, 1996.

[15] H. Lei, Z. Huang, J. Zhang, Z. Yang, E.-L. Tan, F. Zhou*, et al.*, "Joint detection and clinical score prediction in Parkinson's disease via multi-modal sparse learning," *Expert Systems with Applications,* vol. 80, pp. 284-296, Sep 2017.

[16] F. Liu, C.-Y. Wee, H. Chen, and D. Shen, "Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification," *NeuroImage,* vol. 84, pp. 466-475, 2014.

[17] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *the Proceedings of the 18th International Conference on Neural Information Processing Systems*, 2005, pp. 507–514.

[18] I. Jolliffe, *Principal component analysis*: Springer, 2011.

[19] L. Shi, L. Du, and Y. D. Shen, "Robust Spectral Learning for Unsupervised Feature Selection," in *2014 IEEE International Conference on Data Mining*, 2014, pp. 977-982.

[20] S. Wang, J. Tang, and H. Liu, "Embedded Unsupervised Feature Selection," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 470–476.

[21] X. Zhu, H. I. Suk, and D. Shen, "A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis," *Neuroimage,* vol. 100, pp. 91-105, Oct 2014.

[22] X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, and D. Shen, "A novel relational regularization feature selection method for joint regression and classification in AD diagnosis," *Medical Image Analysis,* vol. 38, pp. 205-214, May 2017.

[23] L. Liu, Q. Wang, E. Adeli, L. Zhang, H. Zhang, and D. Shen, "Exploring diagnosis and imaging biomarkers of Parkinson's disease via iterative canonical correlation analysis based feature selection," *Computerized Medical Imaging and Graphics,* vol. 67, pp. 21-29, 2018.

[24] E. Adeli, G. Wu, B. Saghafi, L. An, F. Shi, and D. Shen, "Kernel-based Joint Feature Selection and Max-Margin Classification for Early Diagnosis of Parkinson's Disease," *Scientific Reports,* vol. 7, p. 41069, 2017.

[25] P. A. Pérez-Toro, J. C. Vásquez-Correa, T. Arias-Vergara, E. Nöth, and J. R. Orozco-Arroyave, "Nonlinear dynamics and Poincaré sections to model gait impairments in different stages of Parkinson's disease," *Nonlinear Dynamics,* vol. 100, pp. 3253-3276, 2020.

[26] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, R. Arora, E. Nöth, N. Dehak, H. Christensen*, et al.*, "Multi-view representation learning via gcca for multimodal analysis of Parkinson's disease," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2966-2970.

[27] E. Adeli, F. Shi, L. An, C.-Y. Wee, G. Wu, T. Wang*, et al.*, "Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data," *NeuroImage,* vol. 141, pp. 206-219, Nov 2016.

[28] J. Shi, Z. Xue, Y. Dai, B. Peng, Y. Dong, Q. Zhang*, et al.*, "Cascaded Multi-Column RVFL+ Classifier for Single-Modal Neuroimaging-Based Diagnosis of Parkinson's Disease," *IEEE Transactions on Biomedical Engineering,* vol. 66, pp. 2362-2371, 2019.

[29] P. Yang, F. Zhou, D. Ni, Y. Xu, S. Chen, T. Wang*, et al.*, "Fused Sparse Network Learning for Longitudinal Analysis of Mild Cognitive Impairment," *IEEE Transactions on Cybernetics,* pp. 1-14, 2019.

[30] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *Neuroimage,* vol. 59, pp. 895-907, Jan 2012.

[31] B. Lei, P. Yang, T. Wang, S. Chen, and D. Ni, "Relational-Regularized Discriminative Sparse Learning for Alzheimer's Disease Diagnosis," *IEEE Trans Cybern,* vol. 47, pp. 1102-1113, Apr 2017.

[32] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 333–342.

[33] F. Nie, D. Xu, I. W. H. Tsang, and C. Zhang, "Flexible Manifold Embedding: A Framework for Semi-Supervised and Unsupervised Dimension Reduction," *IEEE Transactions on Image Processing,* vol. 19, pp. 1921-1932, 2010.

[34] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection," *IEEE Transactions on Cybernetics,* vol. 44, pp. 793-804, 2014.

[35] B. Mohar, Y. Alavi, G. Chartrand, O. Oellermann, and A. Schwenk, "The Laplacian spectrum of graphs," *Graph theory, combinatorics, and applications*, vol. 2, pp. 871-898, 1991.

[36] K. Fan, "On a Theorem of Weyl Concerning Eigenvalues of Linear Transformations. I," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 35, pp. 652-655, 1949.

[37] Y. C. Pai and J. Patton, "Center of mass velocity-position predictions for balance control," *J Biomech,* vol. 30, pp. 347-54, Apr 1997.

[38] S. A. Sadananthan, W. Zheng, M. W. Chee, and V. Zagorodnov, "Skull stripping using graph cuts," *Neuroimage,* vol. 49, pp. 225-239, Jan 2010.

[39] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix*, et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage,* vol. 15, pp. 273-289, Jan 2002.

[40] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *Neuroimage,* vol. 62, pp. 782-790, Aug 2012.

[41] M. Xia, J. Wang, and Y. He, "BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics," *PLOS ONE,* vol. 8, p. e68910, 2013.

[42] R. Gonzalez-Redondo, D. Garcia-Garcia, P. Clavero, C. Gasca-Salas, R. Garcia-Eulate, J. L. Zubieta*, et al.*, "Grey matter hypometabolism and atrophy in Parkinson's disease with cognitive impairment: a two-step process," *Brain,* vol. 137, pp. 2356-67, Aug 2014.

[43] D. Zhang, J. Wang, X. Liu, J. Chen, and B. Liu, "Aberrant Brain Network Efficiency in Parkinson's Disease Patients with Tremor: A Multi-Modality Study," *Frontiers in aging neuroscience,* vol. 7, pp. 169-169, 2015.

[44] C. Nombela, L. E. Hughes, A. M. Owen, and J. A. Grahn, "Into the groove: can rhythm influence Parkinson's disease?," *Neurosci Biobehav Rev,* vol. 37, pp. 2564-2570, Dec 2013.

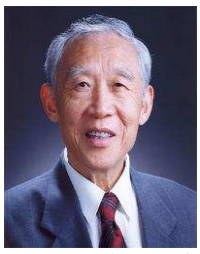
[45] T. Wu and M. Hallett, "A functional MRI study of automatic movements in patients with Parkinson's disease," *Brain,* vol. 128, pp. 2250-2259, 2005.
[46] J. Kim, M. Criaud, S. S. Cho, M. Díez-Cirarda, A. Mihaescu, S. Coakeley, *et al.*, "Abnormal intrinsic brain functional network dynamics in Parkinson's disease," *Brain,* vol. 140, pp. 2955-2967, 2017.
[47] N. Tuovinen, K. Seppi, F. de Pasquale, C. Müller, M. Nocker, M. Schocke, *et al.*, "The reorganization of functional architecture in the early-stages of Parkinson's disease," *Parkinsonism & Related Disorders,* vol. 50, pp. 61-68, 2018.
[48] R. González-Redondo, D. García-García, P. Clavero, C. Gasca-Salas, R. García-Eulate, J. L. Zubieta, *et al.*, "Grey matter hypometabolism and atrophy in Parkinson's disease with cognitive impairment: a two-step process," *Brain,* vol. 137, pp. 2356-2367, 2014.
[49] Y. Hosokai, Y. Nishio, K. Hirayama, A. Takeda, T. Ishioka, Y. Sawada, *et al.*, "Distinct patterns of regional cerebral glucose metabolism in Parkinson's disease with and without mild cognitive impairment," *Movement Disorders,* vol. 24, pp. 854-862, 2009.
[50] J. Xu, F. Jiao, Y. Huang, X. Luo, Q. Xu, L. Li, *et al.*, "A Fully Automatic Framework for Parkinson's Disease Diagnosis by Multi-Modality Images," *Frontiers in Neuroscience,* vol. 13, 2019.
[51] M. Ariz, R. C. Abad, G. Castellanos, M. Martínez, A. Muñoz-Barrutia, M. A. Fernández-Seara, *et al.*, "Dynamic Atlas-Based Segmentation and Quantification of Neuromelanin-Rich Brainstem Structures in Parkinson Disease," *IEEE Transactions on Medical Imaging,* vol. 38, pp. 813-823, 2019.
[52] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage,* vol. 170, pp. 446-455, 2018.
[53] X. Wang, D. Shen, and H. Huang, "Prediction of Memory Impairment with MRI Data: A Longitudinal Study of Alzheimer's Disease," in *Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 273-281.
[54] C. Klein and A. Westenberger, "Genetics of Parkinson's disease," *Cold Spring Harbor perspectives in medicine,* vol. 2, p. a008888, 2012.
[55] D. Zhu, K. Li, L. Guo, X. Jiang, T. Zhang, D. Zhang, *et al.*, "DICCCOL: dense individualized and common connectivity-based cortical landmarks," *Cereb Cortex,* vol. 23, pp. 786-800, Apr 2013.

Zhongwei Huang received the M.Eng. degree from Shenzhen University, Shenzhen, China, in 2019. He is also currently pursuing the Ph.D. degree at Shenzhen University.

His current research interests include medical image analysis, machine learning, and pattern recognition.

Haijun Lei received the Ph.D. degree from Huazhong University of Science and Technology, Wuhan, in 2003.

He is currently with the College of Computer Science and Software Engineering, Shenzhen University, China. His current research interests include medical image analysis, machine learning, and pattern recognition.



Guoliang Chen received the B.Sc. degree from Xi'an Jiaotong University, Xi'an, China, in 1961.

His research interests include parallel algorithm, computer architecture, computer network, and computational intelligence.

Prof. Chen is an Academician of Chinese Academy of Sciences.

Alejandro F. Frangi received Ph.D. degree from Utrecht University, Utrecht, in 2001.

His current research interests include medical image computing, medical imaging, and image-based computational physiology.

Yanwu Xu received the B.Eng. and PhD degrees from the University of Science and Technology of China, in 2004 and 2009, respectively.

His current research interests include ocular imaging, medical image analysis, computer vision, and machine learning.

Ahmed Elazab received the Ph.D. degree from University of Chinese Academy of Sciences, China, in 2017.

His current research interests include machine and deep learning, pattern recognition, medical image analysis, and computer-aided diagnosis.

Jing Qin received the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2009.

His current research interests include innovations for healthcare and medicine applications, medical image processing, deep learning, visualization and human-computer interaction, and health informatics.

Baiying Lei received her M.Eng. degree in electronics science and technology from Zhejiang University, China, in 2007, and Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2013.

She is currently with School of Biomedical Engineering, Health Science Center, Shenzhen University, China. Her current research interests include medical image analysis, machine learning, and pattern recognition.

Dr. Lei has coauthored more than 120 scientific articles, e.g., IEEE TCYB, IEEE TMI, IEEE TBME, IEEE JBHI. Pattern Recognition and Information Sciences. She is an IEEE senior member and serves as the editorial board member of Scientific Reports, Frontiers in Neuroinformatics, Frontiers in Aging Neuroscience, and Academic Editor of Plos One.