This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

Subspace Regularized Sparse Multi-Task Learning for Multi-Class Neurodegenerative Disease Identification

XiaofengvZhu, Heung-IlvSuk, *Member, IEEE*, Seong-Whan Lee*, *Fellow, IEEE*, Dinggang Shen*, *Senior Member, IEEE*, and the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—The high feature-dimension and low sample-size problem is one of the major challenges in the study of computer-aided Alzheimer's Disease (AD) diagnosis. To circumvent this problem, feature selection and subspace learning have been playing core roles in literature. Generally, feature selection methods are preferable in clinical applications due to their ease for interpretation, but subspace learning methods can usually achieve more promising results. In this paper, we combine two different methodological approaches to discriminative feature selection in a unified framework. Specifically, we utilize two subspace learning methods, namely, Linear Discriminant Analysis (LDA) and Locality Preserving Projection (LPP), which have proven their effectiveness in a variety of fields, to select class-discriminative and noise-resistant features. Unlike previous methods in neuroimaging studies that mostly focused on a binary classification, the proposed feature selection method is further applicable for multi-class classification in AD diagnosis. Extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset show the effectiveness of the proposed method over other state-of-the-art methods.**

***Index Terms*—Alzheimer's disease, feature selection, sparse coding, subspace learning, multi-class classification, mild cognitive impairment, neuroimaging data analysis**

I. Introduction

RECENTLY, neurodegenerative diseases, such as Alzheimer's Disease (AD), Parkinson's disease and Huntington's disease, have become highly prevalent within societies. Among these neurodegenerative diseases, AD is the most prevalent and was reported to be the sixth leading cause of death in the United States [1]. Hence, many research groups have devoted their efforts to understand underlying biological or physiological mechanisms behind AD.

Since neuroimaging tools, such as Magnetic Resonance Imaging (MRI) and Positron Emission Topography (PET), have been successfully applied to investigate neurophysiological characteristics of AD, machine learning techniques have also been greatly devised for analyzing neuroimaging data for AD diagnosis [2], [3], [4], [5], [6], [7], [8]. For example, Cuingnet *et al.* devised a general Supprot Vector Machine (SVM) framework for the study of AD [9], and Wang *et al.* proposed a sparse Bayesian multi-task learning algorithm for improving the prediction performance of AD diagnosis [10].

In AD studies, the feature dimensionality is high in nature [11], [12], [4], [13]. Thus, dimensionality reduction (such as subspace learning [14], [15] and feature selection [16], [17], [18], [19], [20]) has become one of the core steps in the field of machine learning. For example, Salas-Gonzalez *et al.* employed the statistical *t*-test method to select voxels of interest for AD diagnosis [17], while Zhou *et al.* combined Least Absolute Shrinkage and Selection Operator (LASSO) [21] and group sparse LASSO [22] to predict AD status [23], [24]. Feature selection methods, such as statistical *t*-test and sparse linear regression, find the informative feature subset from the original feature set [5], [6], [23], [25], [26], while subspace learning methods, such as Fisher's Linear Discriminant Analysis (LDA) [27] and Locality Preserving Projection (LPP) [28], transform original features into a low-dimensional space [29]. In regards to the interpretability of the results, feature selection methods are preferable compared to subspace learning methods, particularly in neuroimaging studies, as selected features directly link anatomical structures and thus provide an intuitive understanding. Meanwhile, subspace learning methods have recently presented promising performances in various applications [30], [15], [31], [32], [33]. For example, Sui *et al.* applied a number of subspace learning methods, such as Independent Component Analysis (ICA) [34], Canonical Correlation Analysis (CCA) [35], [36], and Partial Least Squares (PLS) [37], [38] for medical image analysis [33]. Liu *et al.* employed Local Linear Embedding (LLE) [39] to reduce feature dimensionality of multivariate MRI data to show that subspace learning methods are superior to feature selection methods, such as *t*-test and Chi-squared, in AD classification [15].

From a clinical standpoint, a model for AD/MCI diagnosis should be interpretable and able to accurately identify the disease status of a subject; therefore, it is reasonable to combine feature selection and subspace learning in a systematic manner. One intuitive way to do this is to design a two-stage method, *i.e.*, subspace learning before feature selection or subspace learning preceded by feature selection. However, because these approaches perform the methods individually, the results are likely to be suboptimal. It may be interesting to integrate them in a unified framework, where we can complement the limitations of each method.

* Corresponding authors: dgshen@med.unc.edu, sw.lee@korea.ac.kr.

Xiaofeng Zhu and Dinggang Shen are with the Department of Radiology and Biomedical Research Imaging Center (BRIC), The University of North Carolina at Chapel Hill, NC, USA. Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen are with the Department of Brain and Cognitive Engineering, Korea University, Republic of Korea.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

In this paper, we propose a novel feature selection method[1] to select class-discriminative and noise-resistant features from the original feature set by utilizing characteristics of subspace learning methods. Specifically, we inject two subspace learning methods (such as LDA [27] and LPP [28]) into a sparse least square regression framework. The rationale of using both LDA and LPP in our formulation is that LDA considers both the global information inherent in the observations and the class label information, with the goal of selecting class-discriminative features [27], [34], [41], while LPP preserves the neighborhood structure of each sample to reduce the adverse effect of noises or outliers [28], [36]. Mathematically, it is very similar to conduct feature selection by the sparse feature selection framework, except that the original data gets "adjusted" by the incorporation of the global information (*i.e.,* LDA) and local information (*i.e.,* LPP). Both LDA and LPP enable the proposed framework (with an intuitive and easy way) to select class-discriminative and noise-resistant features.

II. Materials and Image Preprocessing

In this work, we used the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset for performance evaluation. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, with a $60 million 5-year publicprivate partnership. The primary goal of ADNI was to demonstrate whether MRI, PET, other biological markers, and clinical and neuropsychological assessment could be combined to measure the progression of MCI and early AD. As a result, approximately 800 adults, aged 55 to 90, participated in this research.

A. Subjects

We describe the general inclusion/exclusion criteria of the subjects as follows: First, the MMSE (Mini-Mental State Examination) score of each NC subject is between 24 and 30 with Clinical Dementia Rating (CDR) of 0. Moreover, the NC subject is non-depressed, non MCI, and non-demented. Second, the MMSE score of each MCI subject is between 24 and 30 with CDR of 0.5. Moreover, each MCI subject is an absence of significant level of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. Last, the MMSE score of each Mild AD subject is between 20 and 26 with the CDR of 0.5 or 1.0.

We used baseline MRI and PET images obtained from 202 subjects, which included 51 AD subjects, 52 Normal Control (NC) subjects, and 99 MCI subjects. Moreover, 99 MCI subjects included 43 MCI Converters (MCI-C) and 56 MCI Non-Converters (MCI-NC). The detailed demographic information is summarized in Table I.

[1]This work focuses on multi-class classification of AD diagnosis with *either* single-modality data *or* multi-modality data, different from our previous work [40], which focused on joint regression and classification with only multi-modality data.

B. Image Preprocessing

We conducted image preprocessing separately for MRI and PET images of the selected 202 subjects. We downloaded raw Digital Imaging and COmmunications in Medicine (DICOM) MRI scans from the ADNI website[2]. All structural MR images used in this paper were acquired from 1.5T scanners. These MR images were already reviewed for quality, and automatically corrected for spatial distortion caused by gradient nonlinearity and B1 field inhomogeneity. All PET images were collected across a variety of scanners with protocols individualized for each scanner. We used 18-Fluoro-DeoxyGlucose (FDG) PET images. Also, we removed cerebellum in our preprocessing pipeline, as we mainly focused on brain regions in cerebrum for this study. These PET images were first acquired 30-60 minutes post-injection, and were then averaged, spatially aligned, interpolated to a standard voxel size, intensity normalized, followed by smoothing to a common resolution of $8mm$ full width at half maximum. Specifically, the image processing was conducted by the following steps: First, we performed anterior commissure-posterior commissure correction using MIPAV software[3] for all images, and then used the N3 algorithm [42] to correct the intensity inhomogeneity. Second, we extracted a brain on all structural MR images using a robust skull-stripping method [43], and then conducted manual edition and intensity inhomogeneity correction (if necessary). Third, we removed cerebellum based on registration and intensity inhomogeneity correction by repeating N3 for three times, and then we used the FAST algorithm in the FSL package [44] to segment structural MR images into three different tissues: Gray Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF). Next, we used HAMMER [45] for registration and then dissected images into 93 Regions-Of-Interest (ROIs) by labeling them based on the Jacob template [46]. After that, for each of all 93 ROIs in the labeled image of a subject, we computed the GM tissue volumes as features. For each subject, we aligned the PET images to their respective MR T1 images using affine registration and then computed the average intensity of each ROI as a feature. So, we extracted 93 features from MRI and 93 features from PET for each subject.

III. Method

A. Notations

Throughout this paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively. For a matrix $\mathbf{X} = [x_{ij}]$, its *i*-th row and *j*-th column are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. Also, we denote the Frobenius norm and $\ell_{2,1}$-norm of a matrix $\mathbf{X}$ as $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}^i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}_j\|_2^2}$ and $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}^i\|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$, respectively. We further denote the transpose operator, the trace operator, and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

[2]http://www.loni.usc.edu/ADNI

[3]http://mipav.cit.nih.gov/clickwrap.php.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

TABLE I: Demographic information of the subjects. (MMSE: Mini-Mental State Examination; ADAS-Cog: Alzheimer's Disease Assessment Scale-Cognitive subscale; MCI-C: MCI Converters; MCI-NC: MCI Non-Converters)

	AD	NC	MCI-C	MCI-NC
Female/male	18/33	18/34	15/28	17/39
Age	75.2 ± 7.4	75.3 ± 5.2	75.8 ± 6.8	74.8 ± 7.1
Education	14.7 ± 3.6	15.8 ± 3.2	16.1 ± 2.6	15.8 ± 3.2
MMSE	23.8 ± 2.0	29.0 ± 1.2	26.6 ± 1.7	28.4 ± 1.7
ADAS-Cog	18.3 ± 6.0	7.3 ± 3.2	12.9 ± 3.9	10.2 ± 4.3

B. Sparse Multi-Task Learning with Subspace Regularization

Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ denote a feature matrix, where *d* and *n* are, respectively, the numbers of feature variables and subjects, and $\mathbf{Y} \in \mathbb{R}^{c\times n}$ denote a class indicator matrix with 0-1 encoding, where *c* is the number of classes. As for the feature selection, we use a sparse regression model, which has been successfully used in various applications [47], [48], [5], [10]. However, since the class indicator matrix $\mathbf{Y}$ includes multiple response variables, a regression model would find a regression coefficient vector for each response variable individually. In this regard, we regularize a least square regression model with an $\ell_{2,1}$-norm to find the features commonly used across the regression tasks as follows:

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

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is a regression coefficient matrix and λ is a sparsity control parameter. The $\ell_{2,1}$-norm $\|\mathbf{W}\|_{2,1}$ penalizes the coefficients in the same row of $\mathbf{W}$ together for joint selection or unselection in regressing the response variables in $\mathbf{Y}$. In Eq. (1), the optimal solution assigns a relatively large weight to the informative features and zero or a small weight to uninformative or less informative features [47], [49]. By viewing the regression of each response variable as one task, we call Eq. (1) as *multi-task learning*, and Argyriou *et al.* have shown that Eq. (1) successfully utilizes the correlation of different classes [47].

It is shown that LDA exploits the distributional characteristics that help find a generalized solution (*i.e.*, small bias), whereas LPP alleviates the sensitivity of the solution to noises or outliers in the training samples (*i.e.*, small variance) [27], [50]. However, in its current form, *i.e.*, Eq. (1), we cannot guarantee the class-discriminative power of selected features and the preservation of the neighborhood structure of data points, which are important characteristics to enhance classification performance. To resolve this drawback, we propose a novel sparse multi-task learning method by combining the methods of discriminant analysis and topological structure preservation jointly in a sparse regression framework. Specifically, we utilize a Fisher's LDA [34] that considers the global sample distributions by means of the ratio between within-class-variance and between-class-variance in a supervised manner. We also use an LPP [28] by constructing a Laplacian matrix to efficiently use the local topological relation among samples in an unsupervised manner.

In regards to Fisher's criterion for discriminative feature selection, a straightforward approach can penalize the objective function of Eq. (1) with the Fisher's ratio defined as follows:

$$R_G = \frac{\mathbf{W}^T\mathbf{\Sigma}_w\mathbf{W}}{\mathbf{W}^T\mathbf{\Sigma}_b\mathbf{W}} \tag{2}$$

where $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_b$ denote, respectively, the within-class covariance and the between-class covariance matrices. However, due to the non-convexity of Eq. (2), it is not trivial to find an optimal solution of the corresponding objective function. Interestingly, we can reformulate this multi-class LDA in a linear regression framework by replacing the original label indicator matrix $\mathbf{Y}$ with a specific class indicator matrix $\hat{\mathbf{Y}} = [\hat{y}_{ik}]$ defined as follows:

$$\hat{y}_{ik} = \begin{cases} \sqrt{\frac{n}{n_k}} - \sqrt{\frac{n_k}{n}}, & \text{if } \ l(\mathbf{x}_i) = k \\ -\sqrt{\frac{n_k}{n}}, & \text{otherwise} \end{cases} \tag{3}$$

where $l(\mathbf{x}_i)$ denotes the class label of $\mathbf{x}_i$ and n_k is the number of training samples of the class k[4]. That is, using a class indicator matrix $\hat{\mathbf{Y}}$ defined in Eq. (3), we can naturally incorporate the multivariate discriminant analysis of an embedding method to the sparse regression framework [51]. It is noteworthy that unlike the conventional LDA that projects features into an embedding space, in which it is generally difficult to interpret or investigate the results, we still work in the the original input space.

With respect to topological relation among samples, *i.e.*, local structural information, we use a graph Laplacian by defining a similarity matrix $\mathbf{S} = [s_{ij}] \in \mathbb{R}^{n\times n}$ between every pair of sample points $\mathbf{x}_i$ and $\mathbf{x}_j$ with a heat kernel[5] and define a regularization term as follows:

$$\begin{aligned} R_L &= \sum_{i,j}^{n} s_{ij}\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 \\ &= tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \end{aligned} \tag{4}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ and $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with its diagonal elements defined as $d_{ii} = \sum_j s_{ij}$.

By using the newly defined class indicator matrix $\hat{\mathbf{Y}}$ in Eq. (3) as the target response values and the locality preserving constraint in Eq. (4), we formulate our objective function as follows:

$$\begin{aligned} \min_{\mathbf{W}} \frac{1}{2}\|\hat{\mathbf{Y}} - \mathbf{W}^T\mathbf{X}\|_F^2 \ &+ \ \lambda_1 tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \\ &+ \ \lambda_2\|\mathbf{W}\|_{2,1} \end{aligned} \tag{5}$$

where λ_1 and λ_2 are the regularization tuning parameters. Here, we should note that Eq. (5) efficiently combines the subspace learning methods, *i.e.*, LDA and LPP, and a sparse regression-based feature selection method in a unified framework. Concretely, LDA utilizes class label information for

[4] $k \geq 3$. For the case of $k = 2$, it follows that $\hat{y}_i \in \{-2n_2/n, 2n_1/n\}$ and $\sum_{i=1}^{n} \hat{y}_i = 0$, where n_1 and n_2 denote the numbers of subjects from the negative and positive subjects, respectively [27], [34], [51].

[5] $H(\mathbf{x}_i, \mathbf{x}_j) = exp\left[-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}\right]$, where $\sigma \in \mathbb{R}^+$ defines a kernel width. For simplicity, we set $\sigma = 1$ in our experiments.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

discriminative feature selection, while LPP preserves the relationship between a sample and its neighborhood, which helps increase the robustness to noise.

Our method can be discriminated from the previous methods: (1) Unlike the previous sparse linear regression-based feature selection methods [48], [6], the proposed method finds the class-discriminative (via Fisher's criterion) and noise-resistant regression (via graph Laplacian), based on which we select informative features. (2) Compared to subspace learning methods, such as Principal Component Analysis (PCA) [52], LDA [34], and LPP [28], which all have an interpretational limitation, the proposed method selects features in the original space and thus allows intuitive investigation of the results. (3) Furthermore, while the conventional LDA finds at most $(c-1)$-dimension features for a c-class classification task, *e.g.*, 2-dimension features in a 3-class classification task, Eq. (5) can theoretically select at most d (in general, $d \gg c$ in the AD study) number of features.

C. Optimization

Eq. (5) is a convex but non-smooth function. In this work, we solve it by designing a new accelerated proximal gradient method [53]. We first conduct the proximal gradient method on Eq. (5) by defining

$$\begin{aligned} f(\mathbf{W}) &= \tfrac{1}{2}\|\hat{\mathbf{Y}} - \mathbf{W}^T\mathbf{X}\|_F^2 \\ &\quad + \lambda_1 tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}), \\ \mathcal{L}(\mathbf{W}) &= f(\mathbf{W}) + \lambda_2\|\mathbf{W}\|_{2,1}. \end{aligned} \tag{6}$$

$f(\mathbf{W})$ is convex and differentiable, while $\lambda_2\|\mathbf{W}\|_{2,1}$ is convex but non-smooth [53]. To optimize $\mathbf{W}$ with the proximal gradient method, we iteratively update it with the following rule:

$$\mathbf{W}(t+1) = \arg\min_{\mathbf{W}} G_{\eta(t)}(\mathbf{W}, \mathbf{W}(t)), \tag{7}$$

where $G_{\eta(t)}(\mathbf{W}, \mathbf{W}(t)) = f(\mathbf{W}(t)) + \langle \nabla f(\mathbf{W}(t)), \mathbf{W} - \mathbf{W}(t)\rangle + \frac{\eta(t)}{2}\|\mathbf{W} - \mathbf{W}(t)\|_F^2 + \lambda_2\|\mathbf{W}\|_{2,1}$, $\nabla f(\mathbf{W}(t)) = (\mathbf{X}\mathbf{X}^T + \lambda_1\mathbf{X}\mathbf{L}\mathbf{X}^T)\mathbf{W}(t) - \mathbf{X}\hat{\mathbf{Y}}^T$, $\langle\cdot,\cdot\rangle$ is an inner product operator, $\eta(t)$ is determined by the line search (refer to [49] for detailed description), and $\mathbf{W}(t)$ is the value of $\mathbf{W}$ obtained at the t-iteration.

By ignoring the terms independent of $\mathbf{W}$ in Eq. (7), we can rewrite it as

$$\begin{aligned} \mathbf{W}(t+1) &= \pi_{\eta(t)}(\mathbf{W}(t)) \\ &= \arg\min_{\mathbf{W}} \ \tfrac{1}{2}\|\mathbf{W} - \mathbf{U}(t)\|_2^2 \\ &\quad + \tfrac{\lambda_2}{\eta(t)}\|\mathbf{W}\|_{2,1} \end{aligned} \tag{8}$$

where $\mathbf{U}(t) = \mathbf{W}(t) - \frac{1}{\eta(t)}\nabla f(\mathbf{W}(t))$ and $\pi_{\eta(t)}(\mathbf{W}(t))$ is the Euclidean projection of $\mathbf{W}(t)$ onto the convex set $\eta(t)$, and $\frac{1}{\eta(t)}$ denotes a stepsize at the *t*-iteration. Thanks to the separability of $\mathbf{W}(t+1)$ on each row, *i.e.*, $\mathbf{w}^i(t+1)$, we can update the weights for each row individually:

$$\mathbf{w}^i(t+1) = \arg\min_{\mathbf{w}^i} \ \frac{1}{2}\|\mathbf{w}^i - \mathbf{u}^i(t)\|_2^2 + \frac{\lambda_2}{\eta(t)}\|\mathbf{w}^i\|_2, \tag{9}$$

Algorithm 1: Pseudo code of solving Eq. (5).

Input: $\eta(0) = 1, \alpha(1) = 1$, $\gamma = 0.2$, λ_1, λ_2;
Output: $\mathbf{W}$;

```
Initialize t = 1;
Initialize W(1) as a random diagonal matrix;
repeat
    while L(W(t)) > G_{η(t-1)}(π_{η(t-1)}(W(t)), W(t)) do
        Set η(t-1) = γη(t-1)        /* γ is a predefined constant (For details, refer to Appendix A) */;
    end
    Set η(t) = η(t-1);
    Compute W(t+1) = arg min_W G_{η(t)}(W, V(t));
    Compute α(t+1) = (1+sqrt(1+4α(t)^2))/2;
    Compute Eq. (11);
until Eq. (5) converges;
```

where $\mathbf{u}^i(t) = \mathbf{w}^i(t) - \frac{1}{\eta(t)}\nabla f(\mathbf{w}^i(t))$. In Eq. (9), $\mathbf{w}^i(t+1)$ takes a closed form solution [49] as follows:

$$\mathbf{w}^{i*} = \begin{cases} (1 - \frac{\lambda_2}{\eta(t)\|\mathbf{u}(t)^i\|_2^2})\mathbf{u}^i(t), \\ \qquad \text{if } \|\mathbf{u}^i(t)\|_2^2 > \frac{\lambda_2}{\eta(t)} \\ 0, \qquad \text{otherwise.} \end{cases} \tag{10}$$

Meanwhile, in order to accelerate the proximal gradient method in Eq. (7), we further introduce an auxiliary variable $\mathbf{V}(t+1)$ as follows:

$$\mathbf{V}(t+1) = \mathbf{W}(t) + \frac{\alpha(t)-1}{\alpha(t+1)}(\mathbf{W}(t+1) - \mathbf{W}(t)), \tag{11}$$

where the coefficient $\alpha(t+1)$ is usually set as $\alpha(t+1) = \frac{1+\sqrt{1+4\alpha(t)^2}}{2}$ [53].

We summarize the pseudo code for the proposed sparse multi-task learning with subspace regularization in Algorithm 1 and prove the convergence of Algorithm 1 in Appendix A.

D. Feature Selection and Multi-Class Classification

Because we use an $\ell_{2,1}$-norm regularizer in our objective function, after finding the optimal solution with Algorithm 1, we have some zero row vectors in $\mathbf{W}$. Thus, we discard the features, whose regression coefficient vectors are zero, by regarding them as being uninformative in representing the target response variables, *i.e.*, class labels.

After conducting feature selection, we build a multi-class classifier with a Support Vector Machines (SVM) [54]. There are two approaches for multi-class classification [55], [6], such as *one-against-rest* and *one-against-one*. The one-against-rest method builds *c* binary classifiers (here *c* is the number of classes) with each binary classifier κ_i $(i = 1, ..., c)$ built between the *i*-th class and the other $(c-1)$ classes, while the one-against-one method builds $\frac{c(c-1)}{2}$ binary classifiers, with each binary classifier $\kappa_{i,j}$ $(j = 1, ..., c)$ built between the *i*-th class and the *j*-th class $(i \neq j)$. In terms of computational efficiency and the training cost, we choose to use the one-against-one approach, which classifies a test sample $\mathbf{x}_{te}$ with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

the following rule:

$$\kappa(\mathbf{x}_{te}) = \arg\max_i(\Sigma_j \kappa_{i,j}(\mathbf{x}_{te})). \quad (12)$$

IV. Experimental Results

A. Experimental Settings

We conducted performance evaluation on a subset of the ADNI dataset by including 51 AD, 43 MCI-C, 56 MCI-NC[6], and 52 NC subjects. We considered two multi-class classification problems: (1) AD vs. MCI vs. NC (3-class) and (2) AD vs. MCI-C vs. MCI-NC vs. NC (4-class). In the 3-class classification, we included both MCI-C and MCI-NC as MCI. For the modality fusion of MRI and PET (MRI+PET), we concatenated their features into a long vector of 186 features. We employed the metrics of classification ACCuracy (ACC) to evaluate the performance of all competing methods.

We compared the proposed method with Fisher Score (FS) [27], LPP [28], standard LDA [34], and PCA [52]. FS is a feature selection method that selects features based on the score ranking in the original feature space. Meanwhile, LPP, LDA, and PCA are the subspace learning methods, which are used to consider local topological structures, global structures, and maximal variance of the samples, respectively. For these four methods, we solved them with a generalized eigen-decomposition method and determined dimensions based on their respective eigenvalues. We also compared the proposed method with other state-of-the-art feature selection methods, namely, Sparse Joint Classification and Regression (SJCR) [5] and Multi-Modal Multi-Task (M3T) [6]. SJCR uses a logistic loss function and a least square loss function simultaneously, along with an $\ell_{2,1}$-norm for multi-task feature selection. It has been used to conduct multi-class feature selection. M3T uses multi-task learning with an $\ell_{2,1}$-norm to select a common set of features for tasks of regression and binary classification. In order to show the validity of feature selection strategies, we also conducted a classification task without feature selection, *i.e.*, using all features (denoted as 'Original').

We used a 10-fold cross-validation technique because of the limited number of samples. Specifically, we first randomly partitioned the whole dataset into 10 subsets and then selected one subset for testing and used the remaining 9 subsets for training. We repeated the whole process 10 times to avoid any possible bias during dataset partitioning for cross-validation. The final result was computed by averaging the results from all of the experiments. We used an LIBSVM toolbox [56] for SVM training. For the model selection, *i.e.*, tuning parameters[7] in Eq. (5) and the soft margin parameter[8] in SVM, we further split the training dataset into 5 subsets for 5-fold inner cross-validation. The parameters that showed the best performance in the inner cross-validation were used in testing[9].

B. Classification Accuracy

Table II summarizes the classification accuracy of all competing methods for two multi-class classification problems. The proposed method outperformed all competing methods in all experiments. For example, in the 3-class classification problem, our method improved the classification accuracy by 4.29% (MRI), 4.01% (PET), and 5.44% (MRI+PET), respectively, compared to the best performances among the competing methods with the respective modality. Meanwhile, in the 4-class classification problem, the classification improvements were even higher than the best with as much as 7.61% (MRI), 4.44% (PET), and 5.08% (MRI+PET), respectively. Based on these results, we argue that the proposed discriminative and noise-resistant feature selection method helped enhance classification performances.

It is noticeable from Table II that all feature selection methods (except for LDA) outperformed the method of exploiting full features (*i.e.*, Original), which implies the effectiveness of feature selection in solving the *high-dimension and small sample size* problem in classification. We found that LDA achieved the lowest classification accuracies among the competing methods. The main reason was that LDA projected the original high dimensional feature space into only two or three dimensional subspace, respectively. In such low-dimensional space, the performance was very limited. On the other hand, the subspace learning methods, except for LDA, outperformed the feature selection method of FS. This verified the conclusion that subspace learning methods outperform feature selection methods [36]. Thus, it is reasonable to integrate subspace learning into the feature selection framework, which aims at enhancing the classification power of the proposed feature selection model in the multi-class AD diagnosis. Moreover, the proposed method was able to outperform both the conventional feature selection and subspace learning methods by combining the two approaches.

Fig. 1 presents the parameters' sensitivity by changing values of C in SVM and (λ_1, λ_2) in Eq. (5). The results show that our method was sensitive to the parameters within only a small range, and the best parameter combination was always found in our experiments, such as $\lambda_1 = 10^3$, $\lambda_2 = 10$, and $C = 3$ for the 3-class classification task with MRI+PET data in Fig. 1.(c).

Finally, we also conducted three binary classification tasks by following the definition of response variables in [27], [34], [51] (Please see the detail in Footnote 4) and reported respective results in Table III. Similarly, the proposed method

[6]In this paper, MCI-C and MCI-NC denote the conversion status from MCI to AD in 18 months of follow-up. Specially, MCI-C indicated the subjects converted from MCI to AD in 36 months, while MCI-NC subjects were not converted to AD in both 18 months and 36 months. The remaining MCI subjects were partitioned into a group not converted in 18 months but converted in 36 months and another group with observation information in baseline but missing information in 18 months.

[7]$\lambda_1 \in \{10^{-5}, ..., 10^2\}$ and $\lambda_2 \in \{10^{-5}, ..., 10^2\}$

[8]$C \in \{2^{-5}, ..., 2^5\}$

[9]We also conducted 10-fold cross-validation technique ten times on all competing methods and then reported the averaging results of all experiments. It is worth noting that, for fair comparison, we optimize parameter values for each competing method. Specifically, for all subspace methods such as FS, LPP, PCA and LDA, we determine their optimal dimensionality based on their respective eignevalues computed by the generalized eigen-decomposition method, according to [13], [27], [28], [34], [52]. For sparse learning methods such as SJCR and M3T, we optimize their sparsity parameter by cross-validating its value in the ranges of $\{10^{-5}, ..., 1, ..., 10^5\}$ (as in [5]) and $\{10^{-5}, ..., 10^2\}$, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

achieved the best results, outperforming all the competing methods.

V. Discussion

A. Role of LDA and LPP in the Proposed Method

In this section, we justify the rationale of applying both LPP and LDA in the proposed framework. To this end, we further consider the LDA Sparse Regression (LDA-SR) as Eq. (5) without the LPP regularization term and also the LPP Sparse Regerssion (LPP-SR) as Eq. (5) replacing $\hat{\mathbf{Y}}$ with the 0-1 encoding method for representing class labels. Table IV summarizes the classification performance of both LDA-SR and LPP-SR on two classification tasks. Obviously, LDA-SR utilizes the discriminative information of the data compared to M3T [6] but does not have the graph Laplacian regularization term compared to our method, while LPP-SR exploits the graph Laplacian regularization term compared to M3T but does not have the LDA parts compared to our methods.

When comparing the performances summarized in Table II and Table IV, we find that LDA-SR, on average, improved by 0.99% more than M3T. The results support the efficacy of applying discriminant analysis in the sparse linear regression model. We also observe that LPP-SR improved by 2.89% more than M3T. This indicates the effectiveness in adding local information into the sparse linear regression model, while also verifying that the LPP regularization term could successfully characterize local topological structures of the data in the least square regression [57]. Furthermore, LDA-SR and LPP-SR, on average, improved by 1.38% and 2.37%, respectively, compared to SJCR.

Recent studies have indicated that LDA was able to capture the global distributional characteristics of the training samples, while LPP was able to preserve the local topological structures of the data [27], [57], [50]. In real applications, since the inherent structure of data is often complex and a single characterization (either global or local) may not be able to sufficiently represent underlying patterns. Lastly, we have found that LDA-SR and LPP-SR were worse than our method as much as 4.76% and 2.86%, respectively. This indicates that combining both LDA and LPP in a unified framework can help find a more generalized solution (*i.e.,* small bias) via LDA and alleviate the sensitivity of the classifier to noises or outliers (*i.e.,* small variance) via LPP.

B. Effects of Dimensionality on Classification Accuracy

We investigated the performance changes of the four competing feature selection methods, *i.e.,* FS, SJCR, M3T, and the proposed method. We plotted the performance changes in Fig. 2 by varying the dimensionality from 10 to 90 with an increment of 10 for MRI and PET, and from 20 to 180 with an increment of 20 for MRI+PET, respectively. It is noteworthy that the proposed method consistently showed the best performance over the varying dimensions. For the 3-class classification problem, the proposed method reported performance improvements on average of 4.92% (MRI), 4.58% (PET), and 5.35% (MRI+PET) compared to FS, by 4.04% (MRI), 3.19% (PET), and 3.24% (MRI+PET) compared to SJCR, and by 5.01% (MRI), 4.18% (PET), and 5.34% (MRI+PET) compared to M3T. For the 4-class classification problem, the proposed method improved on average by 4.61% (MRI), 3.03% (PET), and 8.27% (MRI+PET) compared to FS, by 4.17% (MRI), 2.04% (PET), and 4.42% (MRI+PET) compared to SJCR, and by 7.85% (MRI), 5.38% (PET), and 6.59% (MRI+PET) compared to M3T.

Interestingly, the classification accuracies of the feature selection methods began to decrease after a certain dimensionality, from which we believe that the intrinsic class-discriminative feature dimensionality for the classification is low [58].

C. Most Discriminative Brain Regions

We also investigated the potential of brain regions as biomarkers in AD diagnosis based on the selected frequency of the ROIs and also compared the results among the feature selection methods[10] with MRI+PET. Fig. 3 shows the frequency of the ROIs selected by the proposed method in two multi-class classification problems. We also visualized the 10 most frequently selected ROIs by the proposed method in Fig. 4 and Fig. 5. We compared the 10 most frequently selected ROIs by different feature selection methods in Table V and Table VI.

From Fig. 3, Table V and Table VI, we can see that the commonly selected regions in two multi-class classification tasks were uncus right (22)[11], hippocampal formation right (30), uncus left (46), middle temporal gyrus left (48), hippocampal formation left (69), amygdala left (76), middle temporal gyrus right (80), and amygdala right (83) from MRI; precuneus right (26), precuneus left (41), and angular gyrus left (87) from PET. These regions were also selected by the proposed method and the competing methods with MRI+PET. Moreover, these discriminative brain regions have been pointed out in the previous literatures on binary classification [6] and have been also shown to be highly related to AD and MCI in clinical diagnosis [59], [60], [61], [62]. In this regard, we can say that these regions can be the potential biomarkers for AD/MCI diagnosis.

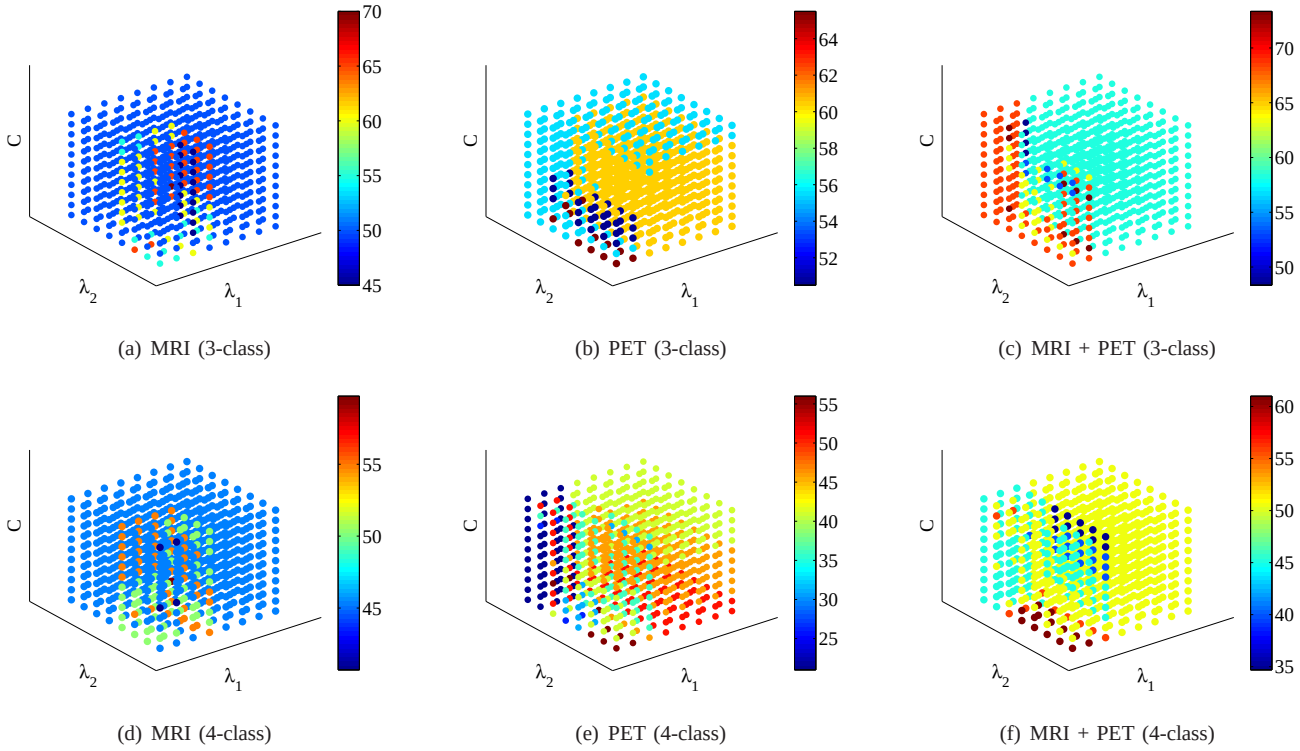
Our method selected, on average, 50.5 and 34.3 features for MRI+PET (186 dimensional features) for the 3-class classification task and the 4-class classification task, respectively. It is interesting that the smaller number of features was selected in a 4-class classification tasks rather than in a 3-class classification task, whereas the larger number of features was selected from MRI rather than from PET in both 3-class and 4-class classification problems. Furthermore, from Table II, we can see that MRI-based methods achieved better performance than the PET-based methods. Based on these observations, it is likely that the structural MR image provides more discriminative information in identifying the clinical status related to AD, compared to the functional PET image.

Here, we should mention that most of the methods selected similar features from the top 10 brain regions, but our method

[10]Note that the methods (such as PCA, LPP, and LDA) do not conduct feature selection, so they cannot output the selected regions.

[11]The number in the parentheses represents an index of an ROI. Please refer to Table IX for the full name of the respective ROI.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering



(a) MRI (3-class) (b) PET (3-class) (c) MRI + PET (3-class)

(d) MRI (4-class) (e) PET (4-class) (f) MRI + PET (4-class)

Fig. 1: Classification accuracy on different parameters' setting, *i.e.*, $C \in [-5:5]$ (upward), $\lambda_1 \in \{10^{-5}, ..., 10^{-2}\}$ (rightward), and $\lambda_2 \in \{10^{-5}, ..., 10^{-2}\}$ (leftward).

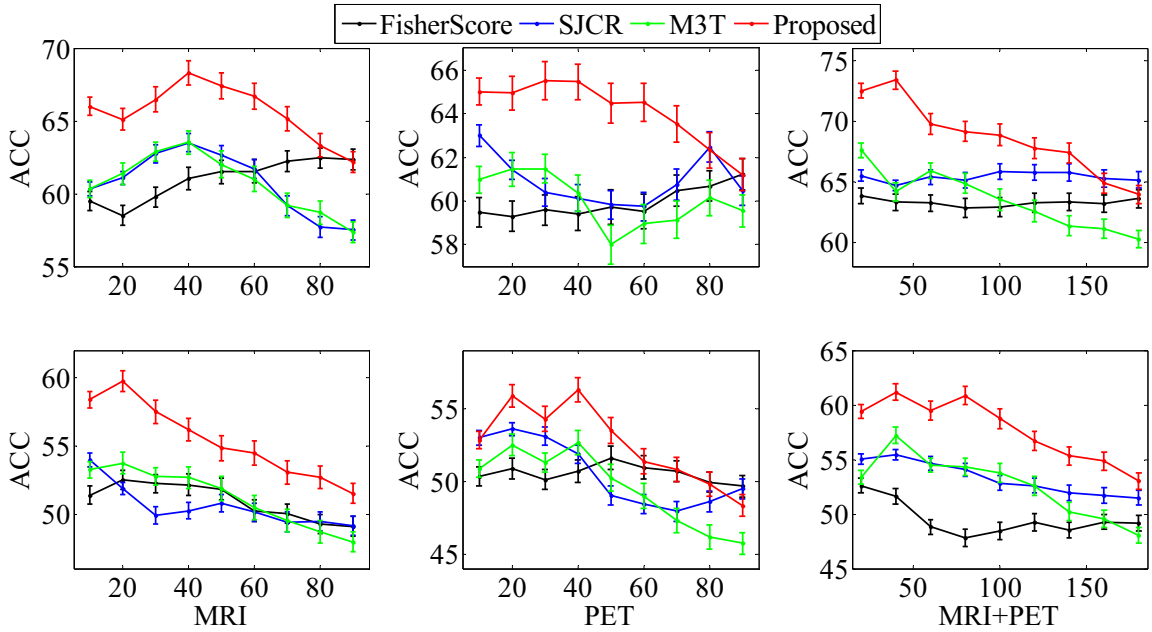


Fig. 2: Classification ACCuracy (ACC) of using different number of features in four feature selection methods, on a 3-class classification task (top) and a 4-class classification task (bottom), respectively. Note that the horizontal axis represents different number of features selected by various feature selection methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

TABLE II: Comparison of classification accuracy ((mean±standard deviation)%) for two multi-class classification tasks. The boldface denotes the best performance for each modality or combined modalities in each classification task. The values in the parentheses indicated the average number of selected features by all the methods in total 100 runs.

Method	AD/MCI/NC			AD/MCI-C/MCI-NC/NC		
	MRI	PET	MRI+PET	MRI	PET	MRI+PET
Orignal	61.96±1.46 (93.0)	57.99±1.75 (93.0)	62.59±1.77 (186)	49.13±1.62 (93.0)	47.98±1.54 (93.0)	49.89 ±1.56 (186)
FS	62.33±1.56 (46.2)	60.11±1.54 (42.3)	62.88±1.31 (72.3)	50.87±1.73 (38.7)	50.44±1.49 (37.1)	51.76 ±1.58 (59.0)
PCA	63.71±1.30 (35.2)	61.49±1.58 (38.5)	64.61±1.60 (62.8)	51.05±1.64 (36.2)	51.51±1.62 (35.0)	52.20±1.60 (61.3)
LPP	63.21±1.91 (39.3)	61.03±1.22 (32.8)	64.35±1.29 (65.2)	51.72±1.42 (33.2)	51.39±1.58 (26.3)	52.60±1.37 (53.2)
LDA	49.01±1.71 (2.00)	39.02±1.23 (2.00)	51.85±1.66 (2.00)	35.25±1.65 (3.00)	31.82±1.40 (3.00)	36.32±1.64 (3.00)
SJCR	64.02±1.36 (38.2)	61.31±1.73 (29.2)	67.66±1.63 (58.2)	52.13±1.73 (28.1)	51.85±1.68 (27.4)	55.98±1.65 (49.4)
M3T	63.30±1.66 (36.1)	61.32±1.90 (28.4)	67.91±1.91 (55.5)	51.89±1.61 (25.7)	50.91±1.83 (26.6)	54.47±1.67 (47.9)
Proposed	**68.31±1.23 (32.7)**	**65.50±1.50 (28.8)**	**73.35±1.53 (50.5)**	**59.74±1.52 (20.1)**	**56.29±1.53 (19.7)**	**61.06±1.40 (34.3)**

TABLE III: Comparison of classification accuracy ((mean±standard deviation)%) for three binary classification tasks. The boldface denotes the best performance for each modality or combined modalities in each classification task.

Method	AD vs. NC			MCI vs. NC			MCI-C vs. MCI-NC		
	MRI	PET	MRI+PET	MRI	PET	MRI+PET	MRI	PET	MRI+PET
Orignal	89.5±1.34	86.2±1.85	89.7±1.48	68.3±1.72	69.0±1.12	71.6±0.95	60.3±1.23	62.2±1.54	62.7±1.56
FS	90.2±1.24	88.5±0.48	91.5±1.48	75.9±1.44	74.9±1.04	75.9±0.48	64.5±1.47	63.4±0.48	65.1±1.10
PCA	91.2±0.89	89.2±0.68	92.0±0.95	76.2±1.06	75.1±0.96	77.2±0.21	65.3±1.11	64.9±0.75	66.2±1.81
LPP	92.0±1.91	90.2±0.92	93.2±1.01	77.1±1.81	75.9±1.58	78.0±0.10	66.2±1.15	65.3±0.65	66.8±1.50
LDA	80.2±1.71	80.1±0.94	86.2±1.11	65.3±1.01	66.5±1.40	68.2±0.14	59.3±1.01	58.3±0.59	59.1±0.90
SJCR	92.9±1.36	92.6±0.95	94.2±1.22	78.2±1.51	77.1±0.85	78.6±0.95	68.0±0.93	67.0±0.65	68.6±0.86
M3T	92.6±1.12	92.3±1.48	94.0±2.14	78.1±1.15	77.2±1.47	78.4±0.15	67.1±0.62	67.0±0.54	67.9±1.00
Proposed	**94.3±0.95**	**93.3±0.79**	**95.5±1.05**	**79.3±1.10**	**79.1±0.99**	**79.7±0.21**	**70.1±1.00**	**69.9±0.52**	**71.2±1.22**

TABLE IV: Classification accuracy ((mean±standard deviation)%) of the LDA-SR and the LPP-SR method.The values in the parentheses indicated the average number of selected features by all the methods in total 100 runs.

Method	AD/MCI/NC			AD/MCI-C/MCI-NC/NC		
	MRI	PET	MRI+PET	MRI	PET	MRI+PET
LDA-SR	64.27±2.02 (36.3)	62.02±2.45 (32.1)	69.45±3.06 (52.3)	52.53±1.80 (20.5)	51.45±2.36 (22.9)	56.02±1.86 (39.2)
LPP-SR	65.04±1.17 (30.2)	63.96±1.56 (33.2)	71.31±1.47 (49.8)	55.45±1.48 (22.9)	53.85±1.74 (23.2)	57.54±1.37 (43.5)

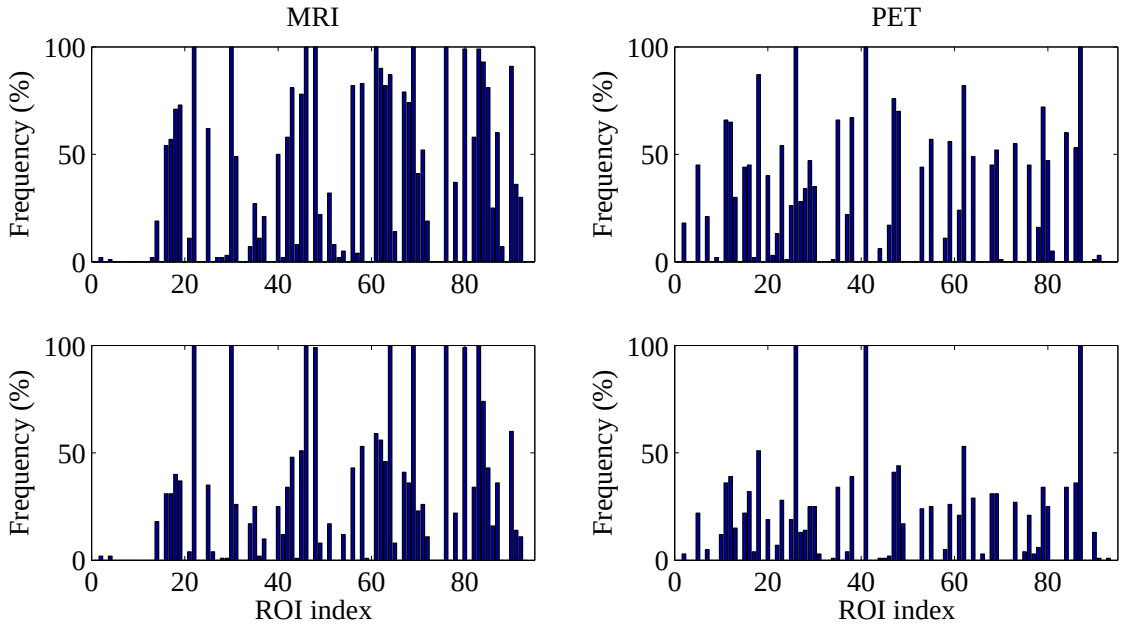


Fig. 3: Frequency of the selected ROIs by the proposed method with MRI+PET in a 3-class classification task (top) and a 4-class classification task (bottom), respectively. For example, $Frenquency_{22} = 100$ in the upper left sub-figure means that the 22nd ROI was selected 100 times over 100 repeats by the proposed method.

selected them with the highest frequency[12]. For example, in the 3-class classification task with MRI+PET, M3T selected the brain regions of middle temporal gyrus right (80) and amygdala right (83) from MRI (see the last column of Table V), which are ranked top 6 and top 8 with the frequency of 95% and 92%, respectively, while our method selected them

[12]In our experiments, we conducted 10-fold cross-validation ten times to obtain 100 groups of reduced feature sets, we define the term 'Frequency' as $Frenquency_i = \frac{\text{the times of the } i\text{−th feature appeared in 100 groups}}{100} \times 100\%$.

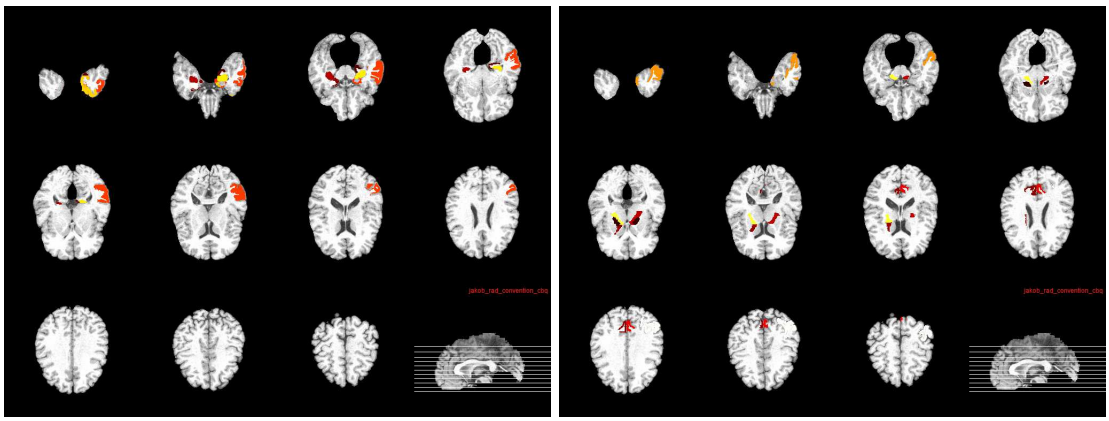
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

TABLE V: Top 10 selected ROIs by feature selection methods on the 3-class classification task. Note that in the last column, the values on the left-side of the semicolon denote the regions selected from MRI, while the values next to the semicolon indicate the regions selected from PET. Please refer to Table IX for the full names of the ROIs.

Method	MRI	PET	MRI+PET
FS	17,30,46,48,63,69,76,80,83,84	11,12,18,26,41,48,62,79,83,90	30,46,48,69,76,80,83; 26,41,87
SJCR	22,30,46,63,64,69,76,79,80,83	11,12,16,18,26,29,62,64,79,87	22,30,46,48,62,76,83; 16,41,87
M3T	17,22,30,46,48,61,64,69,76,83	11,18,26,29,35,41,48,64,79,87	25,30,46,62,76,80,83; 16,26,87
Proposed	17,22,30,46,48,61,63,64,69,83	11,12,26,29,35,41,62,64,79,87	22,30,46,48,61,69,76; 26,41,87

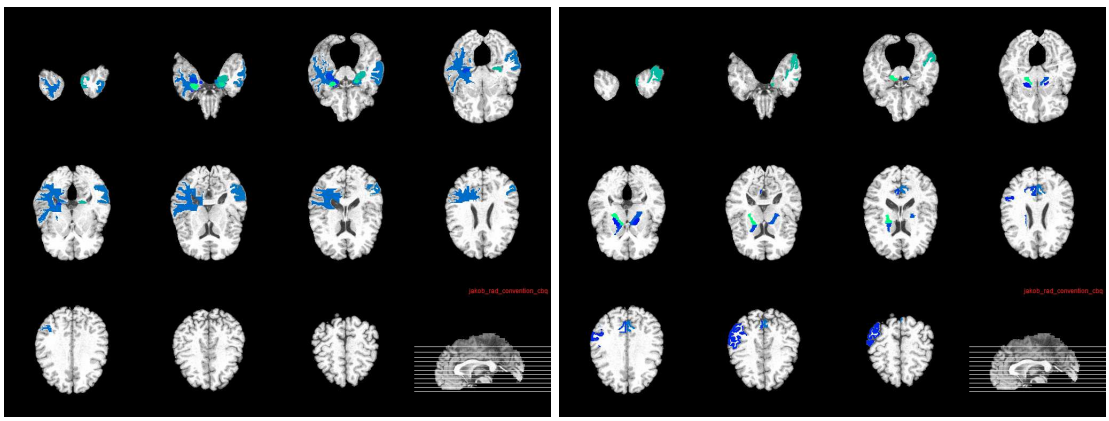
TABLE VI: Top 10 selected ROIs by feature selection methods on the 4-class classification task. Note that in the last column, the values on the left-side of the semicolon denote the regions selected from MRI, while the values next to the semicolon indicate the regions selected from PET. Please refer to Table IX for the full names of the ROIs.

Method	MRI	PET	MRI+PET
FS	17,30,46,48,61,69,76,80,83,84	12,18,26,38,41,47,48,62,86,87	30,46,48,69,76 80,83 ; 26,41,87
SJCR	30,43,48,56,63,64,76,80,83,84	12,16,18,26,35,55,41,62,79,87	22,46,48,64,69,76,90 ; 26,41,87
M3T	22,30,46,56,58,64,69,76,83,90	11,16,18,26,29,35,41,55,64,79	30,46,48,61,64,69,83 ; 26,41,87
Proposed	17,30,43,46,48,63,64,69,76,83	11,12,18,26,29,35,41,62,64,79	22,30,46,64,69,76,83 ; 26,41,87



(a) MRI (b) PET

Fig. 4: Top 10 selected regions in the 3-class classification task with MRI/PET.



(a) MRI (b) PET

Fig. 5: Top 10 selected regions in the 4-class classification task with MRI/PET.

with the frequency of 99% and 99% for MRI, respectively, but ranked them in top 11 and top 12, due to the high frequency (100%) of all other top 10 regions (7 for MRI and 3 for PET). On the other hand, most of the methods also selected other brain regions (different from the aforementioned potential biomarkers) as the top ones in our experiments, such as parahippocampal gyrus left (17), temporal pole left (63), and entorhinal cortex left (64) from MRI, and globus palladus right (11) and anterior limb of internal capsule right (79) from PET. These regions may also be potential biomarkers for multi-class AD diagnosis.

D. Large MRI Dataset from ADNI

We further evaluate performance on a large MRI dataset from the ADNI cohort, including 186 AD, 118 MCI-C, 124 MCI-NC, and 226 NC. We used the same setting as in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

TABLE VII: Comparison of classification accuracy ((mean±standard deviation)%) for two multi-class classification tasks with MRI. The boldface denotes the best performance in each classification task. The values in the parentheses indicated the average number of selected features by all the methods in total 100 runs.

Method	AD/MCI/NC	AD/MCI-C/MCI-NC/NC
Orignal	61.98±2.51 (93.0)	48.01±1.73 (93.0)
FS	62.56±1.79 (43.2)	50.80±1.09 (36.6)
PCA	64.76±1.61 (36.5)	51.49±1.58 (32.1)
LPP	64.32±1.49 (31.5)	55.84±1.64 (29.3)
LDA	49.13±1.65 (2.00)	45.71±2.16 (3.00)
SJCR	64.87±1.78 (42.6)	53.98±1.57 (39.2)
M3T	64.75±1.16 (31.2)	52.32±1.34 (28.6)
LDA-SR	64.88±1.52 (35.8)	56.34±1.78 (24.8)
LPP-SR	65.13±0.76 (32.2)	57.19±1.67 (26.8)
Proposed	**68.49±0.89 (29.3)**	**61.86±1.22 (23.2)**

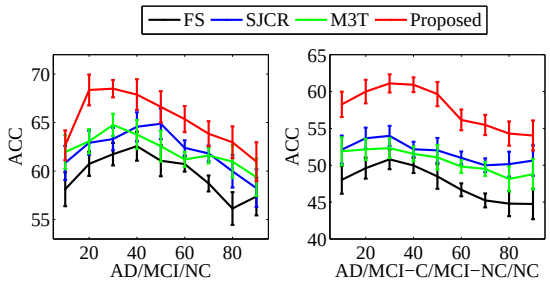


Fig. 6: Accuracy changes in four methods with MRI on a 3-class classification task (left) and a 4-class classification task (right), respectively.

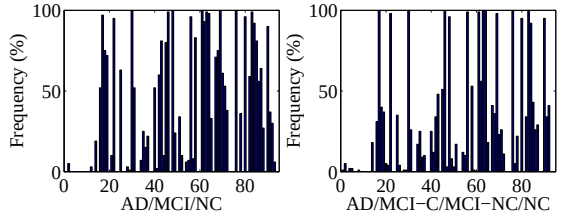
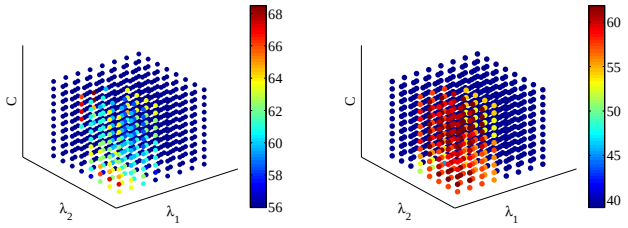


Fig. 7: Frequency of the selected ROIs by the proposed method on a large MRI dataset in a 3-class classification task (left) and a 4-class classification task (right), respectively.

Section IV-A. The experimental results are reported in Tables VII and VIII, as well as Figures 6, 7, and 8. Again, the proposed method achieved the best results, outperforming all the competing methods. The feature selection strategies were also helpful in enhancing classification accuracy, compared to the 'Original' method.

VI. Conclusion

In this paper, we focused on the *high feature-dimension* problem for multi-class classification in AD diagnosis. Specifically, we proposed a novel feature selection method by integrating subspace learning, which utilized both the global and the local topological information inherent in the data, in a sparse linear regression framework. In our experimental results on the ADNI dataset, we validated the efficacy of the proposed method by enhancing classification accuracies in multi-class



(a) 3-class (b) 4-class

Fig. 8: Classification accuracy on different parameters' setting, *i.e.*, $C \in [-5:5]$ (upward), $\lambda_1 \in \{10^{-5}, ..., 10^{-2}\}$ (rightward), and $\lambda_2 \in \{10^{-5}, ..., 10^{-2}\}$ (leftward).

classification problems. In our future works, we will extend the proposed linear feature selection model to the nonlinear model via kernel functions to capture complex patterns between brain images and the corresponding AD status.

VII. Acknowledgements

This work was supported in part by NIH grants (EB006733, EB008374, EB009634, MH100217, AG041721, AG042599), the ICT R&D program of MSIP/IITP [B0101-15-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Centre)], and the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2015R1A2A1A05001867). Xiaofeng Zhu was supported in part by the National Natural Science Foundation of China under grant 61263035.

Appendix

Regarding the convergence of the optimization, we can use the following theorem proved in [53]:

Theorem 1. *[53] Let $\{\mathbf{W}(t)\}$ be the sequence generated by Algorithm 1, then for $\forall\ t \geq 1$, the following holds*
$\mathcal{L}(\mathbf{W}(t)) - \mathcal{L}(\mathbf{W}^*) \leq \frac{2\gamma\vartheta\|\mathbf{W}(1)-\mathbf{W}^*\|_F^2}{(t+1)^2}$,
where $\gamma > 0$ is a predefined constant, ϑ is the Lipschitz constant of the gradient of $f(\mathbf{W})$ in Eq. (6), and $\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{W})$.

Theorem 1 shows that the convergence rate of the proposed accelerated proximal gradient method is $\mathcal{O}(\frac{1}{t^2})$, where t denotes an iteration number.

References

[1] Alzheimer's Association, "2012 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 8, no. 2, pp. 131–168, 2012.
[2] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon, "Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 13, pp. 4637–4642, 2004.
[3] X. Guo, Z. Wang, K. Li, Z. Li, Z. Qi, Z. Jin, L. Yao, and K. Chen, "Voxel-based assessment of gray and white matter volumes in Alzheimer's disease," *Neuroscience Letters*, vol. 468, no. 2, pp. 146–150, 2010.
[4] J. Taquet and C. Labit, "Hierarchical oriented predictions for resolution scalable lossless and near-lossless compression of CT and MRI biomedical images," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2641–2652, 2012.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2466616, IEEE Transactions on Biomedical Engineering

TABLE VIII: Top 10 selected ROIs by feature selection methods on a 3-class classification task (second column) and a 4-class classification task (third column), respectively, on a large MRI dataset. Refer to Table IX for the full names of the ROIs.

Method	AD/MCI/NC	AD/MCI-C/MCI-NC/NC
FS	17,30,46,48,63,69,76,79,83,84	17,22,46,48,61,69,76,80,83,90
SJCR	17,30,46,63,64,69,76,79,80,83	17,43,48,56,61,64,76,80,83,84
M3T	17,22,30,46,48,61,76,79,83,84	22,30,46,56,58,64,69,76,83,84
Proposed	17,30,46,48,61,63,64,69,76,83	17,30,46,56,61,63,64,69,76,83

TABLE IX: The names of the selected ROIs in this work.

Index	ROI Name	Index	ROI Name
5	precentral gyrus right	10	superior frontal gyrus right
11	globus palladus right	12	globus palladus left
15	putamen right	16	frontal lobe WM right
17	parahippocampal gyrus left	18	angular gyrus right
19	temporal pole right	20	subthalamic nucleus right
22	uncus right	25	frontal lobe WM left
26	precuneus right	29	posterior limb of internal capsule right
30	hippocampal formation right	35	anterior limb of internal capsule left
36	occipital lobe WM right	41	precuneus left
42	parietal lobe WM left	43	temporal lobe WM right
46	uncus left	47	middle occipital gyrus right
48	middle temporal gyrus left	53	postcentral gyrus left
55	precentral gyrus left	56	temporal lobe WM left
57	medial front-orbital gyrus left	61	perirhinal cortex left
62	inferior temporal gyrus left	63	temporal pole left
64	entorhinal cortex left	69	hippocampal formation left
73	postcentral gyrus right	76	amygdala left
79	anterior limb of internal capsule right	80	middle temporal gyrus right
82	corpus callosum	83	amygdala right
84	inferior temporal gyrus right	87	angular gyrus left
90	lateral occipitotemporal gyrus left		

[5] H. Wang, F. Nie, H. Huang, S. Risacher, A. J. Saykin, and L. Shen, "Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression," in *MICCAI*, 2011, pp. 115–123.

[6] 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, no. 2, pp. 895–907, 2012.

[7] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[8] ——, "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis," *Brain Structure & Function*, vol. 220, no. 2, pp. 841–859, 2015.

[9] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, and O. Colliot, "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," *NeuroImage*, vol. 56, no. 2, pp. 766–781, 2011.

[10] J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen, "Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease," in *CVPR*, 2012, pp. 940–947.

[11] S. Liao and D. Shen, "A feature-based learning framework for accurate prostate localization in CT images," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3546–3559, 2012.

[12] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014.

[13] 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. 14, no. 0, pp. 1–30, 2014.

[14] R. Guerrero, C. Ledig, and D. Rueckert, "Manifold alignment and transfer learning for classification of Alzheimer's disease," in *MLMI*, 2014, vol. 8679, pp. 77–84.

[15] X. Liu, D. Tosun, M. W. Weiner, and N. Schuff, "Locally linear embedding for MRI based Alzheimer's disease classification," *NeuroImage*, vol. 83, no. 0, pp. 148 – 157, 2013.

[16] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin, "Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images," *NeuroImage*, vol. 60, no. 1, pp. 59–70, 2012.

[17] D. Salas-Gonzalez, J. M. Grriz, J. Ramrez, I. A. Illn, M. Lpez, F. Segovia, R. Chaves, P. Padilla, C. G. Puntonet, and T. A. D. N. Initiative, "Feature selection using factor analysis for Alzheimers diagnosis using F18-FDG PET images," *Medical Physics*, vol. 37, no. 11, pp. 6084–6095, 2010.

[18] J. Young, G. Ridgway, K. Leung, and S. Ourselin, "Classification of Alzheimer's disease patients with hippocampal shape, wrapper based feature selection and support vector machine," in *SPIE*, vol. 8314, 2012.

[19] H.-I. Suk, S.-W. Lee, and D. Shen, "Subclass-based multi-task learning for Alzheimer's disease diagnosis," *Frontiers in Aging Neuroscience*, vol. 6, no. 168, 2014.

[20] ——, "Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis," *Brain Structure & Function*, pp. 1–19, 2015.

[21] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[22] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[23] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, no. 0, pp. 233 – 248, 2013.

[24] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, "Supervised discriminative group sparse representation for mild cognitive impairment diagnosis," *Neuroinformatics*, vol. 13, no. 3, pp. 277–295, 2015.

[25] X. Zhu, L. Zhang, and Z. Huang, "A sparse embedding and least variance encoding approach to hashing," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3737–3750, 2014.

[26] J. Li, Y. Jin, Y. Shi, I. D. Dinov, D. J. J. Wang, A. W. Toga, and P. M. Thompson, "Voxelwise spectral diffusional connectivity and its applications to alzheimer's disease and intelligence prediction," in *MICCAI*, 2013, pp. 655–662.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[28] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *NIPS*, 2005, pp. 1–8.

[29] L. Zhang, L. Wang, and W. Lin, "Conjunctive patches subspace learning with side information for collaborative image retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3707–3720, 2012.

[30] X. Zhu, Z. Huang, H. Cheng, J. Cui, and H. T. Shen, "Sparse hashing for fast multimedia search," *ACM Transactions on Information Systems*, vol. 31, no. 2, p. 9, 2013.

[31] L. Zhan, J. Zhou, Y. Wang, Y. Jin, N. Jahanshad, G. Prasad, T. M. Nir, C. D. Leonardo, J. Ye, and P. M. Thompson, "Comparison of 9

tractography algorithms for detecting abnormal structural brain networks in alzheimers disease," *Frontiers in Aging Neuroscience*, vol. 7, no. 48, 2015.

[32] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 110–121, 2011.

[33] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun, "A review of multivariate methods for multimodal fusion of brain imaging data," *Journal of neuroscience methods*, vol. 204, no. 1, pp. 68–81, 2012.

[34] G. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2004.

[35] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[36] X. Zhu, Z. Huang, H. T. Shen, J. Cheng, and C. Xu, "Dimensionality reduction by mixed kernel canonical correlation analysis," *Pattern Recognition*, vol. 45, no. 8, pp. 3003–3016, 2012.

[37] H. Wold, "Partial least squares," *Encyclopedia of statistical sciences*, 1985.

[38] X. Zhu, X. Li, and S. Zhang, "Block-row sparse multiview multilabel learning for image classification," *IEEE Transactions Cybernetics*, vol. 0, no. 0, p. online, 2015.

[39] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[40] X. Zhu, H.-I. Suk, and D. Shen, "Multi-modality canonical feature selection for Alzheimer's disease diagnosis," in *MICCAI*, 2014, pp. 162–169.

[41] X. Zhu, Z. Huang, Y. Yang, H. T. Shen, C. Xu, and J. Luo, "Self-taught dimensionality reduction on the high-dimensional small-sized data," *Pattern Recognition*, vol. 46, no. 1, pp. 215–229, 2013.

[42] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998.

[43] Y. Wang, J. Nie, P.-T. Yap, G. Li, F. Shi, X. Geng, L. Guo, and D. Shen, "Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates," *PLoS One*, vol. 9, p. e77810, 2014.

[44] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[45] D. Shen and C. Davatzikos, "HAMMER: hierarchical attribute matching mechanism for elastic registration," *IEEE Transactions on Medical Imaging*, vol. 21, no. 11, pp. 1421–1439, 2002.

[46] N. J. Kabani, "3D anatomical atlas of the human brain," *NeuroImage*, vol. 7, pp. 0700–0717, 1998.

[47] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[48] Y. Cho, J.-K. Seong, Y. Jeong, and S. Y. Shin, "Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data," *NeuroImage*, vol. 59, no. 3, pp. 2217–2230, 2012.

[49] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization," in *UAI*, 2009, pp. 339–348.

[50] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[51] J. Ye, "Least squares linear discriminant analysis," in *ICML*, 2007, pp. 1087–1093.

[52] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.

[53] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer, 2004, vol. 87.

[54] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discover*, vol. 2, no. 2, pp. 121–167, 1998.

[55] H.-I. Suk and S.-W. Lee, "A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 286–299, 2013.

[56] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[57] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *IJCAI*, 2011, pp. 1294–1299.

[58] K. Q. Weinberger, F. Sha, and L. K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *ICML*, 2004, pp. 17–24.

[59] G. Chételat, F. Eustache, F. Viader, V. D. L. Sayette, A. Pélerin, F. Mézenge, D. Hannequin, B. Dupuy, J.-C. Baron, and B. Desgranges, "FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment," *Neurocase*, vol. 11, no. 1, pp. 14–25, 2005.

[60] A. Convit, J. De Asis, M. De Leon, C. Tarshish, S. De Santi, and H. Rusinek, "Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimers disease," *Neurobiology of Aging*, vol. 21, no. 1, pp. 19–26, 2000.

[61] N. C. Fox and J. M. Schott, "Imaging cerebral atrophy: normal ageing to Alzheimer's disease," *The Lancet*, vol. 363, no. 9406, pp. 392–394, 2004.

[62] C. Misra, Y. Fan, and C. Davatzikos, "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI," *NeuroImage*, vol. 44, no. 4, pp. 1415–1422, 2009.