frontiers in
AGING NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 07 August 2014
doi: 10.3389/fnagi.2014.00168

Subclass-based multi-task learning for Alzheimer's disease diagnosis

Heung-Il Suk[1], Seong-Whan Lee[2], Dinggang Shen[1,2]* and The Alzheimers Disease Neuroimaging Initiative

[1] *Department of Radiology, Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
[2] *Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea*

Edited by:
Rodrigo Orlando Kuljiš, Zdrav Mozak Limitada, Chile

Reviewed by:
Jieping Ye, Arizona State University, USA
Heng Huang, University of Texas at Arlington, USA

****Correspondence:***
Dinggang Shen, Department of Radiology, Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, 130 Mason Farm Road, Chapel Hill, NC 27599, USA
e-mail: dgshen@med.unc.edu

In this work, we propose a novel subclass-based multi-task learning method for feature selection in computer-aided Alzheimer's Disease (AD) or Mild Cognitive Impairment (MCI) diagnosis. Unlike the previous methods that often assumed a unimodal data distribution, we take into account the underlying multipeak[1] distribution of classes. The rationale for our approach is that it is highly likely for neuroimaging data to have multiple peaks or modes in distribution, e.g., mixture of Gaussians, due to the inter-subject variability. In this regard, we use a clustering method to discover the multipeak distributional characteristics and define subclasses based on the clustering results, in which each cluster covers a peak in the underlying multipeak distribution. Specifically, after performing clustering for each class, we encode the respective subclasses, i.e., clusters, with their unique codes. In encoding, we impose the subclasses of the same original class close to each other and those of different original classes distinct from each other. By setting the codes as new label vectors of our training samples, we formulate a multi-task learning problem in a $\ell_{2,1}$-penalized regression framework, through which we finally select features for classification. In our experimental results on the ADNI dataset, we validated the effectiveness of the proposed method by improving the classification accuracies by 1% (AD vs. Normal Control: NC), 3.25% (MCI vs. NC), 5.34% (AD vs. MCI), and 7.4% (MCI Converter: MCI-C vs. MCI Non-Converter: MCI-NC) compared to the competing single-task learning method. It is remarkable for the performance improvement in MCI-C vs. MCI-NC classification, which is the most important for early diagnosis and treatment. It is also noteworthy that with the strategy of modality-adaptive weights by means of a multi-kernel support vector machine, we maximally achieved the classification accuracies of 96.18% (AD vs. NC), 81.45% (MCI vs. NC), 73.21% (AD vs. MCI), and 74.04% (MCI-C vs. MCI-NC), respectively.

Keywords: Alzheimer's disease, mild cognitive impairment, neuroimaging analysis, feature selection, *K*-means clustering

1. INTRODUCTION

As the population is aging, the brain disorders under the broad category of dementia such as Alzheimer's Disease (AD), Parkinson's disease, etc. have been becoming great concerns around the world. In particular, AD, characterized by progressive impairment of cognitive and memory functions, is the most prevalent cause of dementia in elderly people. According to a recent report by Alzheimer's Association, the number of AD patients is significantly increasing every year, and 10–20 percent of people aged 65 or older have Mild Cognitive Impairment (MCI), a prodromal stage of AD (Alzheimer's Association, 2012). While there is no cure for AD to halt or reverse its progression, it has been of great importance for early diagnosis and prognosis of AD/MCI in the clinic, due to the symptomatic treatments available for a limited period in the spectrum of AD.

To this end, there have been a lot of studies to discover biomarkers and to develop a computer-aided diagnosis system with the help of neuroimaging such as Magnetic Resonance Imaging (MRI) (Cuingnet et al., 2011; Davatzikos et al., 2011; Wee et al., 2011; Zhou et al., 2011; Li et al., 2012; Zhang et al., 2012), Positron Emission Tomography (PET) (Nordberg et al., 2010), functional MRI (fMRI) (Greicius et al., 2004; Suk et al., 2013b). It has been also shown that fusing the complementary information from multiple modalities, e.g., MRI+PET, helps enhance the diagnostic accuracy (Fan et al., 2007; Perrin et al., 2009; Kohannim et al., 2010; Walhovd et al., 2010; Cui et al., 2011; Hinrichs et al., 2011; Zhang et al., 2011; Wee et al., 2012; Westman et al., 2012; Yuan et al., 2012; Zhang and Shen, 2012; Suk and Shen, 2013).

However, from a computational modeling perspective, while the feature dimension of those neuroimaging is high in nature,

[1]Even though the term of "multimodal distribution" is generally used in the literature, in order to avoid the confusion with the "multimodal" neuroimaging, we use the term of "multipeak distribution" throughout the paper.

we have a very limited number of observations/samples available. This so-called "small-*n*-large-*p*" problem (Fort and Lambert-Lacroix, 2005) has been of a great challenge in the field to build a robust model that can correctly identify a clinical label of a subject, e.g., AD, MCI, Normal Control (NC). For this reason, reducing the feature dimensionality, by which we can mitigate the overfitting problem and improve a model's generalizability, has been considered as a prevalent step in building a computer-aided AD diagnosis system as well as neuroimaging analysis (Mwangi et al., 2013).

In general, we can broadly categorize the approaches in the literature that aimed at lowering the feature dimensionality into feature-dimension reduction and feature selection. The methods of feature-dimension reduction find a mapping function that transforms the original feature space into a new low-dimensional space. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) (Martinez and Kak, 2001) are the representative methods of this category and to date, thanks to their computational efficiency, they have been the most widely used in various fields. The PCA finds a mapping function through which it still includes a large portion of the information in samples. Meanwhile, the LDA finds a transformation function that maps the original high-dimensional samples into the dimension-reduced ones by jointly maximizing the variance between classes and minimizing the variance within classes using a Fisher's criterion. However, since the learned projective functions in PCA or LDA are linear combinations of all the original features, it is often difficult to interpret the transformed features (Qiao et al., 2010). Clinically, it is unfavorable for the interpretational difficulty in neuroimaging analysis or classification.

Meanwhile, the feature selection approach that includes filter, wrapper, and embedded methods selects target-related features in the original feature space based on some criteria (Guyon and Elisseeff, 2003). Among these, the embedded methods, e.g., a ℓ_1-penalized linear regression model (Tibshirani, 1994) and its variants (Roth, 2004), have recently attracted researchers due to their theoretical strengths and effectiveness in neuroimage analysis (Varoquaux et al., 2010; Fazli et al., 2011; de Brecht and Yamagishi, 2012; Suk et al., 2013a). In the ℓ_1-penalized regression model, with a sparsity constraint using ℓ_1-norm, many elements in the weighting coefficient vector become zero, thus the corresponding features can be removed. From a machine learning point of view, since the ℓ_1-penalized linear regression model finds one weight coefficient vector that best regresses a target response vector, it is considered as a single-task learning. Hereafter, we use the terms of a ℓ_1-penalized regression model and a single-task learning interchangeably.

The main limitation of the previous methods of PCA, LDA, and ℓ_1-penalized regression model is that they consider a single mapping or a single weight coefficient vector in reducing the dimensionality. Here, if the underlying data distribution is not unimodal, e.g., mixture of Gaussians, then these methods would fail to find the proper mapping or weighting functions, and thus result in performance degradation. In this regard, Zhu and Martinez proposed a Subclass Discriminant Analysis (SDA) method (Zhu and Martinez, 2006) that first clustered samples of each class and then reformulated the conventional LDA by regarding clusters as subclasses. Recently, Liao and Shen applied the SDA method to segment prostate MR images and showed the effectiveness of the subclasses-based approach (Liao et al., 2013).

With respect to neuroimaging data, it is highly likely for the underlying data distribution to have multiple peaks due to the inter-subject variability (Fotenos et al., 2005; Noppeney et al., 2006; DiFrancesco et al., 2008). Here, it should be noted that although SDA was successfully applied to computer vision (Zhu and Martinez, 2006; Kim, 2010; Gkalelis et al., 2013) or medical image segmentation (Liao et al., 2013), as a variant of LDA, it still has an interpretational limitation. In this paper, we propose a novel method of feature selection for AD/MCI diagnosis by integrating the embedded method with the subclass-based approach. Specifically, we first divide each class into multiple subclasses by means of clustering, with which we can approximate the inherent multipeak data distribution of a class. Note that we regard each cluster as a subclass following Zhu and Martinez's work (Zhu and Martinez, 2006). Based on the clustering results, we encode the respective subclasses with their unique codes, for which we impose the subclasses of the same original class close to each other and those of different original classes distinct from each other. By setting the codes as new labels of our training samples, we finally formulate a multi-task learning problem in a $\ell_{2,1}$-penalized regression framework that takes into account the multipeak data distributions, and thus help enhance the diagnostic performances.

2. MATERIALS AND IMAGE PROCESSING

2.1. SUBJECTS

In this work, we use the ADNI dataset publicly available on the web[2]. Specifically, we consider only the baseline MRI, 18-Fluoro-DeoxyGlucose (FDG) PET, and CerebroSpinal Fluid (CSF) data acquired from 51 AD, 99 MCI, and 52 NC subjects[3]. For the MCI subjects, they were further clinically subdivided into 43 MCI Converters (MCI-C), who progressed to AD in 18 months, and 56 MCI Non-Converters (MCI-NC), who did not progress to AD in 18 months. The demographics of the subjects are summarized in **Table 1**.

With regard to the general eligibility criteria in ADNI, subjects were in the age of between 55 and 90 with a study partner, who could provide an independent evaluation of functioning. General inclusion/exclusion criteria[4] are as follows: (1) healthy normal subjects: Mini Mental State Examination (MMSE) scores between 24 and 30 (inclusive), a Clinical Dementia Rating (CDR) of 0, non-depressed, non-MCI, and non-demented; (2) MCI subjects: MMSE scores between 24 and 30 (inclusive), a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia; and (3) mild AD: MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and meets the National Institute of Neurological and Communicative Disorders

[2]Available online at "http://adni.loni.usc.edu/"

[3]Although there exist in total more than 800 subjects in ADNI database, only 202 subjects have the baseline data including all the modalities of MRI, FDG-PET, and CSF.

[4]Refer to "http://www.adni-info.org/Home.aspx" for more details.

Table 1 | Demographic and clinical information of the subjects.

	AD	MCI converter	MCI non-converter	NC
	(*N* = 51)	(*N* = 43)	(*N* = 56)	(*N* = 52)
Female/male	18/33	15/28	17/39	18/34
Age (Mean ± *SD*)	75.2 ± 7.4 [59–88]	75.7 ± 6.9 [58–88]	75.0 ± 7.1 [55–89]	75.3 ± 5.2 [62–85]
Education (Mean ± *SD*)	14.7 ± 3.6 [4–20]	15.4 ± 2.7 [10–20]	14.9 ± 3.3 [8–20]	15.8 ± 3.2 [8–20]
MMSE (Mean ± *SD*)	23.8 ± 2.0 [20–26]	26.9 ± 2.7 [20–30]	27.0 ± 3.2 [18–30]	29 ± 1.2 [25–30]
CDR (Mean ± *SD*)	0.7 ± 0.3 [0.5–1]	0.5 ± 0 [0.5–0.5]	0.5 ± 0 [0.5–0.5]	0 ± 0 [0–0]

(MMSE, Mini Mental State Examination, CDR, Clinical Dementia Rating, N, number of subjects, SD, Standard Deviation, [min-max]).

and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD.

2.2. MRI AND PET SCANNING

The structural MR images were acquired from 1.5T scanners. We downloaded data in Neuroimaging Informatics Technology Initiative (NIfTI) format, which had been pre-processed for spatial distortion correction caused by gradient non-linearity and B1 field inhomogeneity. The FDG-PET images were acquired 30–60 min post-injection, averaged, spatially aligned, interpolated to a standard voxel size, normalized in intensity, and smoothed to a common resolution of 8 *mm* full width at half maximum. CSF data were collected in the morning after an overnight fast using a 20- or 24-gauge spinal needle, frozen within 1 h of collection, and transported on dry ice to the ADNI Biomarker Core laboratory at the University of Pennsylvania Medical Center.

2.3. IMAGE PROCESSING AND FEATURE EXTRACTION

The MR images were preprocessed by applying the typical procedures of Anterior Commissure (AC)-Posterior Commissure (PC) correction, skull-stripping, and cerebellum removal. Specifically, we used MIPAV software[5] for AC-PC correction, resampled images to 256 × 256 × 256, and applied N3 algorithm (Sled et al., 1998) to correct intensity inhomogeneity. An accurate and robust skull stripping (Wang et al., 2013) was performed, followed by cerebellum removal. We further manually reviewed the skull-stripped images to ensure clean removal. Then, FAST in FSL package[6] (Zhang et al., 2001) was used for structural MR image segmentation into three tissue types of Gray Matter (GM), White Matter (WM) and CSF. We finally pacellated them into 93 Regions Of Interests (ROIs) by warping Kabani et al.'s atlas (Kabani et al., 1998) to each subject's space via HAMMER (Shen and Davatzikos, 2002), although other advanced registration methods can also be applied for this process (Friston et al., 1995; Xue et al., 2006; Yang et al., 2008; Tang et al., 2009; Jia et al., 2010). In this work, we considered only GM for classification, because of its relatively high relatedness to AD/MCI compared to WM and CSF (Liu et al., 2012). Regarding FDG-PET images, they were rigidly aligned to the respective MR images, and then applied parcellation propagated from the atlas by registration.

For each ROI, we used the GM tissue volume from MRI, and the mean intensity from FDG-PET as features[7], which are most widely used in the field for AD/MCI diagnosis (Davatzikos et al., 2011; Hinrichs et al., 2011; Zhang and Shen, 2012; Suk et al., 2013a). Therefore, we have 93 features from a MR image and the same dimensional features from a FDG-PET image. Here, we should note that although it is known that the regions of medial temporal and superior parietal lobes are mainly affected by the disease, we assume that other brain regions, although their relatedness to AD is not clearly investigated yet, may also contribute to the diagnosis of AD/MCI and thus we consider 93 ROIs in our study. In addition, we have three CSF biomarkers of Aβ_{42}, t-tau, and p-tau as features.

3. METHODS

In this section, we first briefly introduce the mathematical background of single-task and multi-task learning, and then describe a novel subclass-based multi-task learning method for feature selection in AD/MCI diagnosis.

3.1. NOTATIONS

Throughout the 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. We further 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, and the ℓ_1-norm of a vector as $\|\mathbf{w}\|_1 = \sum_i |w_i|$.

3.2. BACKGROUND

Let $\mathbf{X} \in R^{N \times D}$ and $\mathbf{y} \in R^N$ denote, respectively, the D neuroimaging features and a clinical label of N samples[8]. Assuming that the clinical label can be represented by a linear combination of the neuroimaging features, many research groups have utilized a least square regression model with various regularization terms, which can be mathematically simplified as follows:

$$\min_{\mathbf{w}} \ \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_F^2 + \mathbb{R}(\mathbf{w}) \tag{1}$$

where $\mathbf{w} \in R^D$ is a weight coefficient vector and $\mathbb{R}(\mathbf{w})$ denotes a set of regularization terms. Regarding feature selection, despite its

[5]Available online at "http://mipav.cit.nih.gov/clickwrap.php"
[6]Available online at "http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/"

[7]While the most intuitive feature should be the voxel in MRI and FDG-PET, due to their extremely high dimensionality, in this paper, we take a ROI-based approach and consider the GM tissue volumes and the mean intensity for each ROI from MRI and FDG-PET, respectively, as the features. Furthermore, by using the ROI-based features for our classification, the performances can be less affected by the partial volume effect in PET imaging (Aston et al., 2002).
[8]In this work, we have one sample per subject and consider a binary classification.

simple form, the ℓ_1-penalized linear regression model has been widely and successfully used in the literature (Varoquaux et al., 2010; Fazli et al., 2011; de Brecht and Yamagishi, 2012; Suk et al., 2013a), formulated as follows:

$$\min_{\mathbf{w}} \ \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_F^2 + \lambda_1 \|\mathbf{w}\|_1 \tag{2}$$

where λ_1 denotes a sparsity control parameter. Since the method finds a single optimal weight coefficient vector $\mathbf{w}$ that regresses the target response vector $\mathbf{y}$, it is classified into a single-task learning **Figure 1A** in machine learning. In this framework, after finding an optimal weight coefficient vector of $\mathbf{w}$ by means of convex optimization, the features corresponding to zero (or close to zero) weight coefficients are discarded and the remaining ones are considered for the following steps.

If there exists additional class-related information, then we can further extend the ℓ_1-penalized linear regression model into a more generalized $\ell_{2,1}$-penalized one **Figure 1B** (Nie et al., 2010; Cai et al., 2011; Wang et al., 2011) as follows:

$$\min_{\mathbf{W}} \ \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_2 \|\mathbf{W}\|_{2,1} \tag{3}$$

where $\mathbf{Y} \in R^{N \times S}$ is a target response matrix, $\mathbf{W} \in R^{D \times S}$ is a weight coefficient matrix, S is the number of response variables, and λ_2 denotes a group sparsity control parameter. In machine learning, this framework is classified into a multi-task learning since it needs to find a set of weight coefficient vectors $\{\mathbf{w}_1, \cdots, \mathbf{w}_S\}$ by regressing multiple response values of $\mathbf{y}_1, \cdots, \mathbf{y}_S$, simultaneously[9].

3.3. SUBCLASS-BASED MULTI-TASK LEARNING

We illustrate the proposed framework in **Figure 2**. In our framework, we first concatenate the multi-modal features into a long vector and then divide each class into a number of subclasses by means of clustering. Based on the clustering results, we encode new class-labels for subclasses and assign them to our training samples. Utilizing the new encoding, a multi-task learning is performed for feature selection. Finally, we train a linear Support Vector Machine (SVM) for classification.

As stated in section 1, it is likely for neuroimaging data to have multiple peaks in distribution due to the inter-subject variability (Fotenos et al., 2005; Noppeney et al., 2006; DiFrancesco et al., 2008). In this paper, we argue that it is necessary to consider the underlying multipeak data distribution in feature selection. To this end, we propose to divide classes into subclasses and to utilize the resulting subclass information in feature selection by means of a multi-task learning.

To divide the training samples in each class to subclasses, we use a clustering technique. Specifically, thanks to its simplicity and computational efficiency, especially in a high dimensional space, we apply a K-mean algorithm (Duda et al., 2001). Let $C = \{c_k\}_{k=1}^K$ denote a set of K clusters and $\{\boldsymbol{\mu}_k\}_{k=1}^K$ be the centers of the clusters (represented by row vectors). Given a set of training samples, the goal of K-means algorithm is to minimize the sum of the squared error over all K clusters:

$$J(C) = \sum_{k=1}^{K} \sum_{\mathbf{x}^i \in c_k} \|\mathbf{x}^i - \boldsymbol{\mu}_k\|^2. \tag{4}$$

The main steps of K-means algorithm can be summarized as follows (Jain and Dubes, 1988):

1. Initialize a set of K cluster means $\boldsymbol{\mu}_1^{(0)}, \cdots, \boldsymbol{\mu}_K^{(0)}$.
2. Assignment step: for each of the training samples $\{\mathbf{x}^i\}_{i=1}^N$, find a cluster $\gamma_i^{(t)}$ whose mean yields the least Euclidean distance to the sample as follows:

$$\gamma_i^{(t)} = \min_{c_k} \ \|\mathbf{x}^i - \boldsymbol{\mu}_k^{(t-1)}\|^2 \tag{5}$$

   where t denotes an index of iteration.
3. Update step: for every clusters $\{c_k\}_{k=1}^K$, compute the new mean with the samples assigned to the cluster as follows:

$$\boldsymbol{\mu}_k^{(t)} = \frac{1}{|c_k|} \sum_{i, \gamma_i^{(t)} = c_k} \mathbf{x}^i \tag{6}$$

   where $|c_k|$ denotes the number of samples assigned to the cluster c_k at the iteration t.
4. Repeat (2) and (3) until convergence.

After clustering the samples in each class independently, we divide the original classes into their respective subclasses by regarding each cluster as a subclass. We then encode the subclasses with their unique labels, for which we use "discriminative" sparse codes to

[9]To regress each response value is considered as a task.

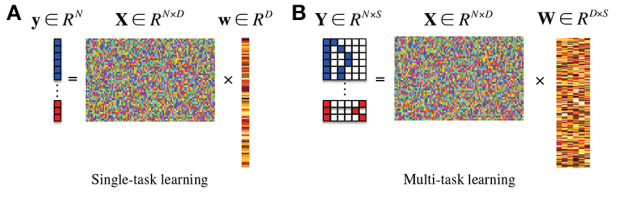

FIGURE 1 | In the response vector/matrix, the colors of blue, red, and white represent 1, −1, and 0, respectively. In multi-task learning, each row of the response matrix represents a newly defined sparse code for each sample by the proposed method. **(A)** Single-task learning, **(B)** multi-task learning.

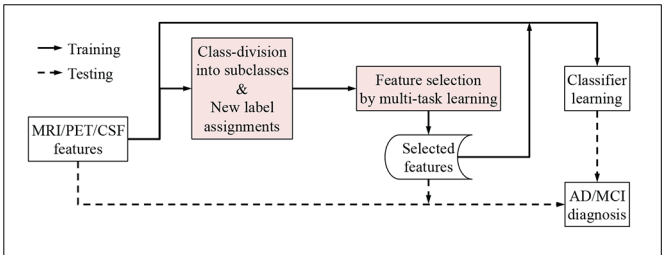

FIGURE 2 | A framework for AD/MCI diagnosis with the proposed subclass-based multi-task learning.

enhance classification performance. Let $K_{(+)}$ and $K_{(-)}$ denote, respectively, the number of clusters/subclasses for the original classes of "+" and "−." Without loss of generality, we define sparse codes for the subclasses of the original classes of "+" and "−" as follows:

$$\mathbf{s}_l^{(+)} = \begin{bmatrix} +1 & \mathbf{z}_l^{(+)} & \mathbf{0}_{K_{(-)}} \end{bmatrix} \quad (7)$$

$$\mathbf{s}_m^{(-)} = \begin{bmatrix} -1 & \mathbf{0}_{K_{(+)}} & \mathbf{z}_m^{(-)} \end{bmatrix} \quad (8)$$

where $l \in \{1, \cdots, K_{(+)}\}$, $m \in \{1, \cdots, K_{(-)}\}$, $\mathbf{0}_{K_{(+)}}$ and $\mathbf{0}_{K_{(-)}}$ denote, respectively, zero row vectors with $K_{(+)}$ and $K_{(-)}$ elements, and $\mathbf{z}_l^{(+)} \in \{0, 1\}^{K_{(+)}}$ and $\mathbf{z}_m^{(-)} \in \{0, -1\}^{K_{(-)}}$ denote, respectively, indicator row vectors in which only the l/m-th element is set to $1/-1$ and the others are 0. Thus, the full code set becomes:

$$\mathbb{S} = \{\mathbf{s}_1^{(+)}, \cdots, \mathbf{s}_l^{(+)}, \cdots, \mathbf{s}_{K_{(+)}}^{(+)}, \mathbf{s}_1^{(-)}, \cdots, \mathbf{s}_m^{(-)}, \cdots, \mathbf{s}_{K_{(-)}}^{(-)}\}. \quad (9)$$

For example, assume that we have three and two clusters for "+" and "−" classes, respectively. Then the code set is defined as follows:

$$\mathbb{S} = \left\{ \begin{array}{l} \mathbf{s}_1^{(+)} = [+1\ +1\ 0\ 0\ 0\ 0], \\ \mathbf{s}_2^{(+)} = [+1\ 0\ +1\ 0\ 0\ 0], \\ \mathbf{s}_3^{(+)} = [+1\ 0\ 0\ +1\ 0\ 0], \\ \mathbf{s}_1^{(-)} = [-1\ 0\ 0\ 0\ -1\ 0], \\ \mathbf{s}_2^{(-)} = [-1\ 0\ 0\ 0\ 0\ -1] \end{array} \right\}. \quad (10)$$

It is noteworthy that in our sparse code set, we reflect the original label information to our new codes by setting the first element of the sparse codes with their original label. Furthermore, by setting the indicator vectors $\{\mathbf{z}_m^{(-)}\}_{m=1}^{K_{(-)}}$ to be negative, the distances become close among the subclasses of the same original class and distant among the subclasses of the different original classes. That is, in the code set of Equation (10), the squared Euclidean distance between subclasses of the same original class is 2, but that between subclasses of different original classes is 6.

Using the newly defined sparse codes, we assign a new label vector $\mathbf{y}^i$ to a training sample $\mathbf{x}^i$ as follows:

$$\mathbf{y}^i = \mathbf{s}_{\gamma_i}^{(y_i)} \quad (11)$$

where $y_i \in \{+, -\}$ is the original label of the sample $\mathbf{x}^i$, and γ_i denotes the cluster to which the sample $\mathbf{x}^i$ was assigned in the K-means algorithm. In this way, we extend the original scalar labels of $+1$ or -1 into sparse code vectors in $\mathbb{S}$.

Thanks to our new sparse codes, it becomes natural to convert a single-task learning in Equation (2) into a multi-task learning in Equation (3) by replacing the original label vector $\mathbf{y}$ in Equation (2) with a matrix $\mathbf{Y} = [\mathbf{y}^i]_{i=1}^N \in \{-1, 0, 1\}^{N \times (1+K_{(+)}+K_{(-)})}$ where $K_{(+)}$ and $K_{(-)}$ denote the number of clusters in the original classes of "+" and "−," respectively. **Figure 1B** illustrates the conceptual meaning of our subclass-based multi-task learning, in which the regression of each column vector of $\mathbf{y}$ is considered as a task. Therefore, we have now $(1 + K_{(+)} + K_{(-)})$ tasks. Note that the task of regressing the first column response vector $\mathbf{y}_1$ corresponds to our binary classification problem between the original classes of "+" and "−." Meanwhile, the tasks of regressing the remaining column vectors $\{\mathbf{y}_i\}_{i=2}^{1+K_{(+)}+K_{(-)}}$ formulate new binary classification problems between one subclass and all the other subclasses. It should be noted that unlike the single-task learning that finds a single mapping $\mathbf{w}$ between regressors $\mathbf{X}$ and the response $\mathbf{y}$, the subclass-based multi-task learning finds multiple mappings $\{\mathbf{w}_1, \cdots, \mathbf{w}_{(1+K_{(+)}+K_{(-)})}\}$, and thus allows us to efficiently use the underlying multipeak data distribution in feature selection.

3.4. FEATURE SELECTION AND CLASSIFIER LEARNING

Because of the $\ell_{2,1}$-norm regularizer in our objective function of Equation (3), after finding the optimal solution, we have some zero row-vectors in $\mathbf{W}$. In terms of the linear regression, the corresponding features are not informative in regressing the response values. In this regard, we finally select the features whose weight coefficient vector is non-zero, i.e., $\|\mathbf{w}^j\|_2 > 0$. With the selected features, we then train a linear SVM, which have been successfully used in many applications (Zhang and Shen, 2012; Suk and Lee, 2013).

4. EXPERIMENTAL RESULTS

4.1. EXPERIMENTAL SETTING

We considered four binary classification problems: AD vs. NC, MCI vs. NC, AD vs. MCI, and MCI-C vs. MCI-NC. In the classifications of MCI vs. NC and AD vs. MCI, we labeled both MCI-C and MCI-NC as MCI. Due to the limited number of samples, we applied a 10-fold cross-validation technique in each binary classification problem. Specifically, we randomly partitioned the samples of each class into 10 subsets with approximately equal size without replacement. We then used 9 out of 10 subsets for training and the remaining one for testing. We reported the performances by averaging the results of 10 cross-validations.

For model selection, i.e., number of clusters K in Equation (4), sparsity control parameters of λ_1 in Equation (2) and λ_2 in Equation (3), and the soft margin parameter C in SVM, we further split the training samples into 5 subsets for nested cross-validation. To be more specific, we defined the spaces of the model parameters as follows: $K \in \{1, 2, 3, 4, 5\}$, $C \in \{2^{-10}, \ldots, 2^5\}$, $\lambda_1 \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5\}$, and $\lambda_2 \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5\}$. The parameters that achieved the best classification accuracy in the inner cross-validation were finally used in testing. In our implementation, we used a SLEP toolbox[10] for feature selection and a LIBSVM toolbox[11] for SVM classifier learning.

To validate the effectiveness of the proposed Subclass-based Multi-Task Learning (SMTL) method, we compared it to the Single-Task Learning (STL) method that used only the original class label as the target response vector in Equation (2). For each set of experiments, we used 93 MRI features, 93 PET features, and/or 3 CSF features as regressors in the respective least

[10] Available online at "http://www.public.asu.edu/~jye02/Software/SLEP/index.htm."

[11] Available online at "http://www.csie.ntu.edu.tw/~cjlin/libsvm/."

square regression models. Regarding the multimodal neuroimaging fusion, e.g., MRI+PET (MP) and MRI+PET+CSF (MPC), we constructed a long feature vector by concatenating features of the modalities. It should be noted that the only difference between the proposed SMTL method and the competing STL method lies in the way of selecting features.

4.2. DATA DISTRIBUTIONS

We visualized the data distributions of our dataset in **Figure 3**. Due to the high dimensionality of the original feature vectors, we first transformed them into their respective 2D eigenspace, whose bases were obtained via principal component analysis (Duda et al., 2001). From the scatter plots, we can see that most of the data distributions look more like having multiple peaks rather than a single peak. For a quantitative evaluation, we also performed Henze-Zirkler's multivariate normality test (Henze and Zirkler, 1990) and summarized the results in **Table 2**. In our test, the null hypothesis was that the samples could come from a multivariate normal distribution. Regarding MRI, the null hypothesis was rejected for both AD and MCI. With respect to PET, the test rejected the hypothesis for MCI. In the meantime, it turned out that the CSF samples of all the disease labels didn't follow a multivariate Gaussian distribution. Based on these qualitative and quantitative evaluations, we could confirm the multipeak data distributions and justify the necessity of the subclass-based approach, which can sufficiently handle such multipeak distribution problem.

4.3. PERFORMANCE MEASUREMENTS

Let TP, TN, FP, and FN denote, respectively, True Positive, True Negative, False Positive, and False Negative.

Table 2 | A summary of Henze-Zirkler's multivariate normality test on our dataset.

Modality	AD	MCI	NC
MRI	0.0005 (R)	0.0004 (*R*)	0.6967 (*A*)
PET	0.4273 (A)	0.0239 (*R*)	0.3150 (*A*)
CSF	0.0049 (R)	<0.0001 (*R*)	<0.0001 (*R*)

"R" or "A" in parentheses denotes whether the null hypothesis (that the samples could come from a multivariate normal distribution) is rejected or accepted at the 5% significance level.

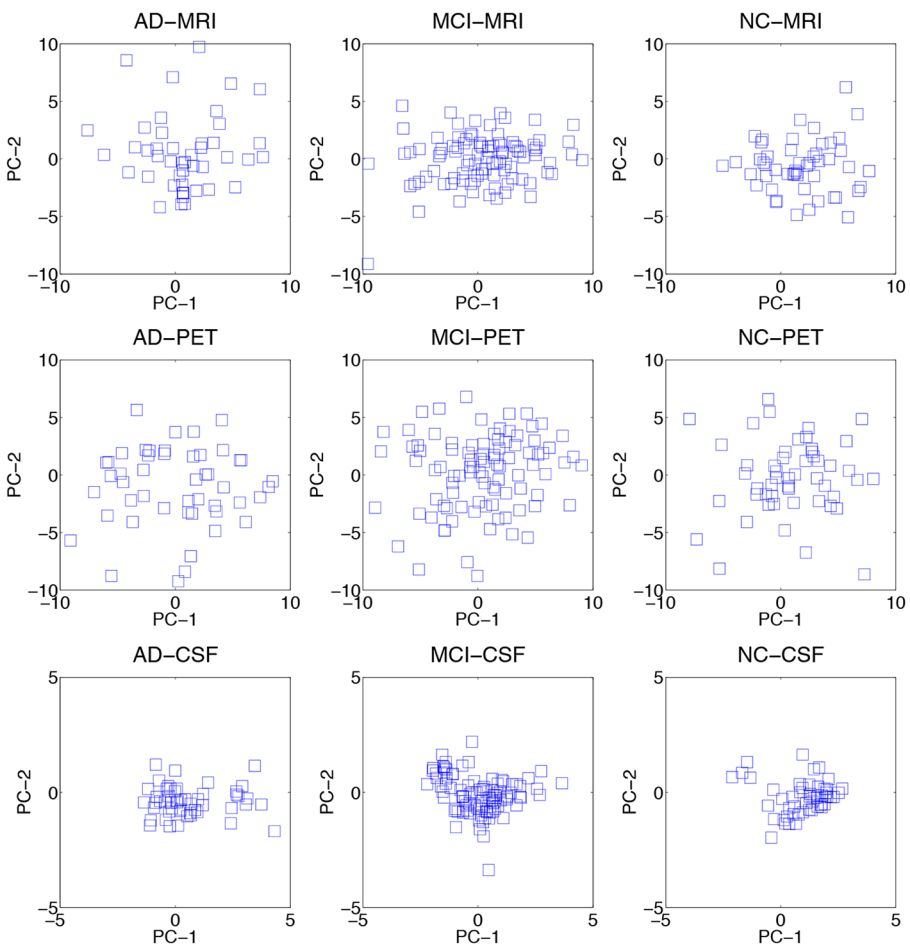

FIGURE 3 | Data distributions of three modalities over different disease labels. For visualization, we transformed the original features in an ambient space into their respective 2D eigenspace, whose bases (PC-1 and PC-2) were obtained via principal component analysis.

In our experiments, we considered the following five metrics:

- ACCuracy (ACC) = (TP+TN) / (TP+TN+FP+FN).
- SENsitivity (SEN) = TP / (TP+FN).
- SPECificity (SPEC) = TN / (TN+FP).
- Balanced ACCuracy (BAC) = (SEN+SPEC) / 2.
- Area Under the receiver operating characteristic Curve (AUC).

The accuracy that counts the number of correctly classified samples in a test set is the most direct metric for comparison between methods. Regarding the sensitivity and specificity, the higher the values of these metrics, the lower the chance of mis-diagnosing. Note that in our dataset, in terms of the number of samples available for each class, they are highly imbalanced, i.e., AD(51), MCI(99), and NC(52). Therefore, it is likely to have an inflated performance estimates for the classifications of MCI vs. NC and AD vs. MCI. For this reason, we also consider a balanced accuracy that considers the imbalance of a test set. Lastly, one of the most effective measurements of evaluating the performance of diagnostic tests in brain disease as well as other medical areas is the Area Under the receiver operating characteristic Curve[12] (AUC). The AUC can be thought as a measure of the overall performance of a diagnostic test. The larger the AUC, the better the overall performance of the diagnostic test.

4.4. CLASSIFICATION RESULTS

We summarized the performances of the competing methods with various modalities for AD and NC classification in **Table 3**. The proposed method showed the mean ACCs of 93.27% (MRI), 89.27% (PET), 95.18% (MP), and 95.27% (MPC). Compared to the STL method that showed the ACCs of 90.45% (MRI), 86.27% (PET), 92.27% (MP), and 94.27% (MPC), the proposed method improved by 2.82% (MRI), 3% (PET), 2.91% (MP), and 1% (MPC) in accuracy. The proposed SMTL method achieved higher AUC values than the STL method for all the cases. It is also remarkable that, except for the metric of specificity with PET, 90.33% (STL) vs. 88.33% (SMTL), the proposed method consistently outperformed the competing STL method over all the metrics and modalities.

In the discrimination of MCI from NC, as reported in **Table 4**, the proposed method showed the ACCs of 76.82% (MRI), 74.18% (PET), 79.52% (MP), and 80.07% (MPC). Meanwhile, the STL method showed the ACCs of 74.85% (MRI), 69.51% (PET), 74.85% (MP), and 76.82% (MPC). Again, the proposed method outperformed the STL method by improving ACCs of 1.97% (MRI), 4.67% (PET), 4.67% (MP), and 3.25% (MPC), respectively. It is believed that the high sensitivities and the low specificities for both competing methods resulted from the imbalanced data between MCI and NC. In the metrics of BAC and AUC that somehow reflect the imbalance of the test samples, the proposed method achieved the best BAC of 77.06% and the best AUC of 81.82% with MPC.

From a clinical point of view, establishing the boundaries between preclinical AD and mild AD, i.e., MCI, has practical and economical implications. To this end, we also performed experiments on AD vs. MCI classification and summarized the results in **Table 5**. Similar to the MCI vs. NC classification, because of the imbalanced data, we had a large gap between sensitivities and specificities. Nevertheless, the proposed method still showed the best ACC of 74.60%, the best BAC of 67.83%, and the best AUC of 72.85% with MP.

[12]The receiver operating characteristic curve is defined as a plot of test true positive rate vs. its false positive rate.

Table 3 | A summary of the performances for AD vs. NC classification.

Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	AUC (%)
STL	MRI	90.45 ± 6.08	82.67	**98.33**	90.50	93.55
	PET	86.27 ± 8.59	82.00	90.33	86.17	90.12
	MP	92.27 ± 5.93	90.00	94.67	92.33	94.91
	MPC	94.27 ± 6.54	**94.00**	94.33	94.17	95.74
SMTL	MRI	93.27 ± 6.33	88.33	**98.33**	93.33	94.19
	PET	89.27 ± 7.43	90.00	88.33	89.17	91.67
	MP	95.18 ± 6.65	**94.00**	96.33	**95.17**	96.15
	MPC	**95.27 ± 6.58**	**94.00**	96.33	**95.17**	**97.13**

(STL, Single-Task Learning; SMTL, Subclass-based Multi-Task Learning). The boldface denotes the best performance in each metric.

Table 4 | A summary of the performances for MCI vs. NC classification.

Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	AUC (%)
STL	MRI	74.85 ± 5.92	80.67	64.00	72.33	76.55
	PET	69.51 ± 10.11	74.78	59.67	67.22	73.54
	MP	74.85 ± 3.91	84.78	56.00	70.39	78.79
	MPC	76.82 ± 7.15	85.89	59.33	72.61	79.25
SMTL	MRI	76.82 ± 7.15	85.78	59.67	72.72	77.84
	PET	74.18 ± 7.18	81.89	59.67	70.78	72.73
	MP	79.52 ± 5.39	**88.89**	62.00	75.44	77.91
	MPC	**80.07 ± 8.42**	86.78	**67.33**	**77.06**	**81.82**

(STL, Single-Task Learning; SMTL, Subclass-based Multi-Task Learning). The boldface denotes the best performance in each metric.

Table 5 | A summary of the performances for AD vs. MCI classification.

Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	AUC (%)
STL	MRI	62.68 ± 7.01	4.00	93.00	48.50	59.16
	PET	72.02 ± 6.73	31.33	93.00	62.17	69.50
	MP	69.26 ± 8.66	51.00	78.56	64.78	71.40
	MPC	68.40 ± 14.48	41.33	82.44	61.89	70.19
SMTL	MRI	70.60 ± 5.97	39.00	86.67	62.83	66.90
	PET	73.31 ± 3.25	33.00	**94.00**	63.50	67.78
	MP	**74.60 ± 9.57**	**46.67**	89.00	**67.83**	**72.85**
	MPC	72.60 ± 9.88	37.33	**91.00**	64.17	71.74

(STL, Single-Task Learning, SMTL, Subclass-based Multi-Task Learning). The boldface denotes the best performance in each metric.

Lastly, we conducted experiments of MCI-C and MCI-NC classification, and compared the results in **Table 6**. The proposed SMTL method achieved the best ACC of 72.02%, the best BAC of 70.33%, and the best AUC of 69.64% with MP. In line with the fact that the classification between MCI-C and MCI-NC is the most important for early diagnosis and treatment, it is remarkable that compared to the STL method, the ACC improvements by the proposed method were 4.62% (MRI), 5.15% (PET), 7.4% (MP), and 7.22% (MPC), respectively.

In order to further verify the superiority of the proposed SMTL method compared to the STL method, we also performed a statistical significance test to assess whether the differences in classification ACCs between the methods are at a significant level on the dataset by means of a paired t-test. Here, the null hypothesis in our work was that the proposed SMTL method produced the same mean ACCs as the STL method. The p-values were 8.884e-04 (AD vs. NC), 4.85e-05 (MCI vs. NC), 1.11e-03 (AD vs. MCI), 7.48e-03 (MCI-C vs. MCI-NC), respectively. That is, the proposed SMTL method statistically outperformed the STL method for all the cases, rejecting the null hypothesis beyond the 95% confidence level.

Table 6 | A summary of the performances for MCI-C vs. MCI-NC classification.

Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	AUC (%)
STL	MRI	56.98 ± 20.61	51.00	60.67	55.83	58.85
	PET	61.58 ± 17.79	55.00	66.00	60.50	60.63
	MP	64.62 ± 14.04	62.50	66.00	64.25	63.87
	MPC	62.89 ± 12.29	58.50	66.00	62.25	58.31
SMTL	MRI	61.60 ± 13.12	44.00	75.67	59.83	60.76
	PET	66.73 ± 11.32	39.00	**88.00**	63.50	65.57
	MP	**72.02 ± 13.80**	58.00	82.67	**70.33**	**69.64**
	MPC	70.11 ± 14.21	**59.00**	78.67	68.83	67.36

(STL, Single-Task Learning, SMTL, Subclass-based Multi-Task Learning). The boldface denotes the best performance in each metric.

4.5. DISCUSSION

In the classifications of AD vs. MCI and MCI-C vs. MCI-NC, the proposed SMTL method with MP, rather than with MCP, achieved the best performances. That is, although we used richer information with MPC, i.e., additional CSF features, the performances with MPC were lower than with MP in those classification problems. Based on the results, fusing the CSF features with the other modalities turned out to be a confounding factor in the classifications of AD vs. MCI and MCI-C vs. MCI-NC. Furthermore, in our experiments above, the selected features were fed into a SVM classifier and in this stage, the features of different modalities have equal weights in decision, which can be a potential problem degrading the performances. To this end, we additionally performed experiments by replacing a Single-Kernel linear SVM (SK-SVM) with a Multi-Kernel linear SVM (MK-SVM) (Gönen and Alpaydin, 2011), with which we could find optimal weights for the modalities. The modality weights were determined by nested cross-validation similarly for model parameters selection described in section 4.1. Specifically, we applied a grid search with an interval of 0.1 with the constraint of the sum of the modality weights to be one. In **Figure 4**, we compared the best performances of SK-SVM, i.e., equal weights for modalities, with those of MK-SVM. It should be noted that for both methods of

Table 7 | Comparison of classification accuracies with the state-of-the-art methods that used multimodal neuroimaing for AD/MCI vs. NC.

Methods	Subjects (AD/MCI/NC)	Modality	AD vs. NC (%)	MCI vs. NC (%)
Kohannim et al., 2010	40/83/43	MRI+PET+CSF	90.7	75.8
Hinrichs et al., 2011	48/119/66	MRI+PET	92.4	n/a
Zhang et al., 2011	51/99/52	MRI+PET+CSF	93.2	76.4
Westman et al., 2012	96/162/111	MRI+CSF	91.8	77.6
Liu et al., 2013	51/99/52	MRI+PET	94.37	78.80
Proposed method	51/99/52	MRI+PET+CSF	**96.18**	**81.45**

The boldface denotes the best performance in each classification task.

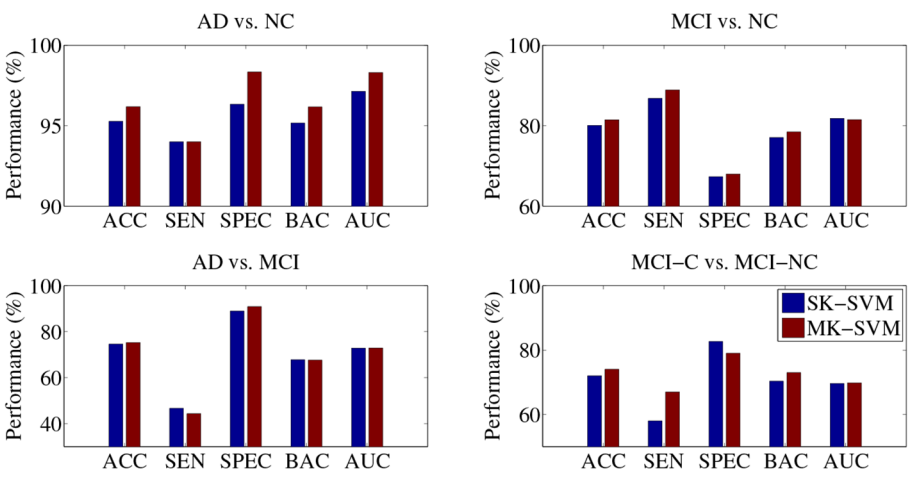

FIGURE 4 | Performance comparison between SK-SVM and MK-SVM in four binary classifications. For both methods, the feature selection was performed on the concatenated feature vectors with the proposed subclass-based multi-task learning.

SK-SVM and MK-SVM, we applied the proposed STML method for feature selection. By means of a modality-adaptive weighting strategy with MK-SVM, we obtained the maximal ACCs of 96.18% (AD vs. NC), 81.45% (MCI vs. NC), 73.21% (AD vs. MCI), and 74.04% (MCI-C vs. MCI-NC). That is, MK-SVM clearly outperformed the SK-SVM by improving the ACCs of 0.91% (AD vs. NC), 1.41% (MCI vs. NC), 0.67% (AD vs. MCI), and 2.02% (MCI-C vs. MCI-NC), respectively.

In **Table 7**, we also compared the classification accuracies of the proposed method with those of the state-of-the-art methods that fused multimodal neuroimaing for the classifications of AD vs. NC and MCI vs. NC. Note that, due to different datasets and different approaches of extracting features and building classifiers, it may not be fair to directly compare the performances among the methods. Nevertheless, the proposed method showed the highest accuracies among the methods in both classification problems. In particular, it is noteworthy that compared to Zhang and Shen's work (Zhang et al., 2011) in which they used the same dataset with ours, the proposed method enhanced the accuracies by 2.98 and 5.05% for the classifications of AD vs. NC and MCI vs. NC, respectively. Furthermore, in comparison with Liu et al.'s work (Liu et al., 2013), where they used the same types of features from MRI and PET and the same number of subjects with ours, our method improved the accuracies by 1.81% (AD vs. NC) and 2.65% (MCI vs. NC), respectively.

Regarding the interpretation of the selected ROIs, due to the involvement of cross-validation, multimodal neuroimaging fusion, and multiple binary classifications in our experiments, it was not straightforward to analyze the selected ROIs. In this work, we first built a histogram of the frequency of the selected ROIs of MRI and PET over cross-validations per binary classification, and normalized it by considering only the ROIs whose frequency was larger than the mean frequency and set the frequency of the disregarded ROIs to zero. **Figure 5** presents the normalized frequency of the selected ROIs in each binary classification. We then added the four normalized histograms in **Figure 5** to find the relative frequency of the selected ROIs over four classification problems. We finally selected ROIs whose frequency was larger than the mean normalized frequency and visualized them in **Figure 6**. Those ROIs include amygdala, hippocampus, parahippocampal gyrus (Braak and Braak, 1991; Visser et al., 2002; Mosconi, 2005; Lee et al., 2006; Devanand et al., 2007; Burton et al., 2009; Desikan et al., 2009; Walhovd et al., 2010; Ewers et al., 2012), superior frontal gyrus, insula, anterior/posterior cingulate gyrus, inferior occipital gyrus, post central gyrus, supramarginal gyrus (Buckner et al., 2005; Desikan et al., 2009; Dickerson et al., 2009; Schroeter et al., 2009), precuneus, paracentral lobule (Bokde et al., 2006; Singh et al., 2006; Davatzikos et al., 2011), heschl gyrus (Supekar et al., 2008), superior/middle temporal gyrus, temporal pole, inferior temporal (Chan et al., 2001; Visser et al., 2002; Burton et al., 2009).

5. CONCLUSIONS

In this paper, we proposed a novel method that formulates a subclass-based multi-task learning. Specifically, to take into account the underlying multipeak data distribution of the

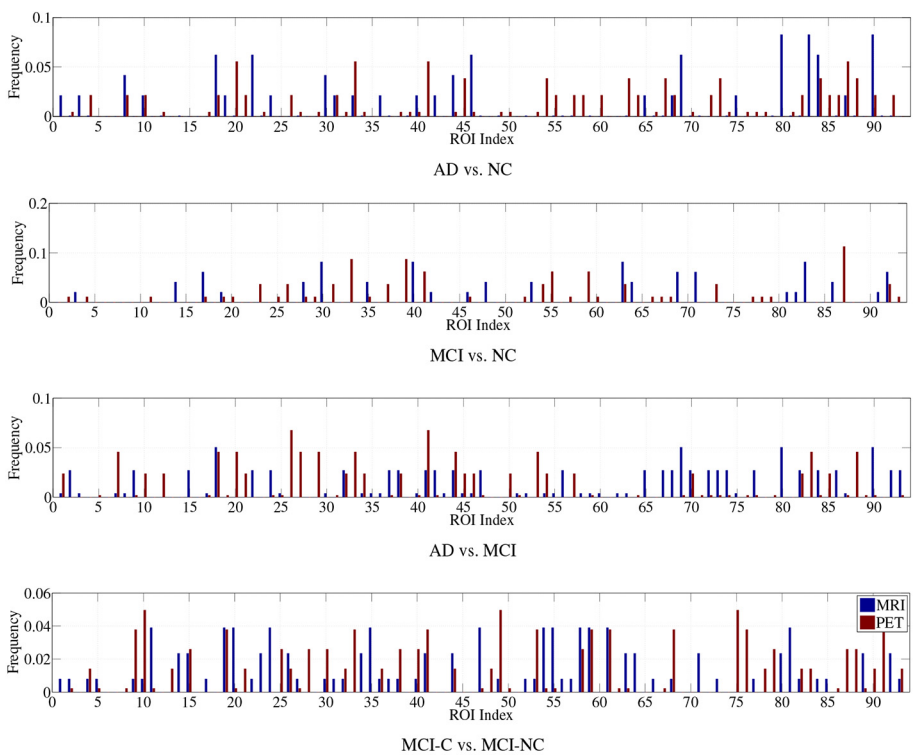

FIGURE 5 | Normalized histograms of the selected features in four binary classification problems.

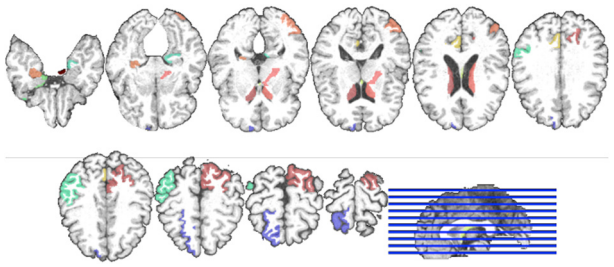

FIGURE 6 | Visualization of the selected ROIs by the proposed method. Different colors denote different brain areas.

original classes, we applied a clustering method to partition each class into multiple clusters, which further considered as subclasses. Here, we can think that one cluster, i.e., subclass, represents one peak in distribution. The respective subclasses were encoded with their unique codes, for which we imposed the subclasses of the same original class close to each other and those of different original classes distinct from each other. We assigned the newly defined codes to our training samples as new label vectors and applied a $\ell_{2,1}$-norm regularizer in a linear regression framework, thus formulated a multi-task learning problem. We finally selected features based on the optimal weight coefficients. It is noteworthy that unlike the previous methods of PCA, LDA, and other embed methods for dimensionality reduction, the proposed method considered multiple mapping functions to reflect the underlying multipeak data distributions, and thus to enhance performances in AD/MCI diagnosis. In our experimental results on the publicly available ADNI dataset, we proved the validity of the proposed method by outperforming the competing methods in four binary classifications of AD vs. NC, MCI vs. NC, AD vs. NC, and MCI-C vs. MCI-NC.

In the context of the practical application of the proposed method, it should be considered for how to determine the optimal number of clusters, i.e., K, for each class, although, in this paper, we applied a cross-validation technique for dealing with this issue. One potential solution for this issue is to use affinity propagation algorithm (Frey and Dueck, 2007) that does not require the number of clusters to be determined. The other potential limitation of our work is that outliers or contaminated features could affect our clustering results, thus causing performance degradation by selecting uninformative features or unselecting informative features. All these limitations will be considered in our future research.

ACKNOWLEDGEMENT

This work was supported in part by NIH grants EB006733, EB008374, EB009634, AG041721, MH100217, and AG042599, and also by the National Research Foundation grant (No. 2012-005741) funded by the Korean government.

REFERENCES

Alzheimer's Association. (2012). 2012 Alzheimer's disease facts and figures. *Alzheimer's Dement.* 8, 131–168. doi: 10.1016/j.jalz.2012.02.001

Aston, J. A. D., Cunningham, V. J., Asselin, M.-C., Hammers, A., Evans, A. C., and Gunn, R. N. (2002). Positron emission tomography partial volume correction: estimation and algorithms. *J. Cereb. Blood Flow Metab.* 22, 1019–1034. doi: 10.1097/00004647-200208000-00014

Bokde, A. L. W., Lopez-Bayo, P., Meindl, T., Pechler, S., Born, C., Faltraco, F., et al. (2006). Functional connectivity of the fusiform gyrus during a face-matching task in subjects with mild cognitive impairment. *Brain* 129, 1113–1124. doi: 10.1093/brain/awl051

Braak, H., and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82, 239–259. doi: 10.1007/BF00308809

Buckner, R. L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, A. F., et al. (2005). Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *J. Neurosci.* 25, 7709–7717. doi: 10.1523/JNEUROSCI.2177-05.2005

Burton, E. J., Barber, R., Mukaetova-Ladinska, E. B., Robson, J., Perry, R. H., Jaros, E., et al. (2009). Medial temporal lobe atrophy on MRI differentiates Alzheimer's disease from dementia with lewy bodies and vascular cognitive impairment: a prospective study with pathological verification of diagnosis. *Brain* 132, 195–203. doi: 10.1093/brain/awn298

Cai, X., Nie, F., Huang, H., and Ding, C. (2011). "Multi-class l2,1-norm support vector machine," in *2011 IEEE 11th International Conference on Data Mining*, 91–100. doi: 10.1109/ICDM.2011.105

Chan, D., Fox, N., Scahill, R., Crum, W., Whitwell, J., Leschziner, G., et al. (2001). Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Ann. Neurol.* 49, 433–442. doi: 10.1002/ana.92

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., et al. (2011). Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PLoS ONE* 6:e21896. doi: 10.1371/journal.pone.0021896

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using ADNI database. *Neuroimage* 56, 766–781. doi: 10.1016/j.neuroimage.2010.06.013

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., and Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol. Aging* 32, 2322.e19–2322.e27. doi: 10.1016/j.neurobiolaging.2010.05.023

de Brecht, M., and Yamagishi, N. (2012). Combining sparseness and smoothness improves classification accuracy and interpretability. *Neuroimage* 60, 1550–1561. doi: 10.1016/j.neuroimage.2011.12.085

Desikan, R., Cabral, H., Hess, C., Dillon, W., Salat, D., Buckner, R., et al. (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain* 132, 2048–2057. doi: 10.1093/brain/awp123

Devanand, D. P., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., et al. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment. *Neurology* 68, 828–836. doi: 10.1212/01.wnl.0000256697.20968.d7

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., et al. (2009). The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cereb. Cortex* 19, 828–836. doi: 10.1093/cercor/bhn113

DiFrancesco, M., Hollandm, S., and Szaflarski, J. (2008). Simultaneous EEG/functional magnetic resonance imaging at 4 tesla: correlates of brain activity to spontaneous alpha rhythm during relaxation. *J. Clin. Neurophysiol.* 25, 255–264. doi: 10.1097/WNP.0b013e3181879d56

Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification*. Wiley.

Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack, C. R. Jr., et al. (2012). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiol. Aging* 33, 1203–1214. doi: 10.1016/j.neurobiolaging.2010.10.019

Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., et al. (2007). Multivariate examination of brain abnormality using both structural and functional MRI. *Neuroimage* 36, 1189–1199. doi: 10.1016/j.neuroimage.2007.04.009

Fazli, S., Danóczy, M., Schelldorfer, J., and Müller, K.-R. (2011). ℓ1-penalized linear mixed-effects models for high dimensional data with application to BCI. *Neuroimage* 56, 2100–2108. doi: 10.1016/j.neuroimage.2011.03.061

Fort, G., and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics* 21, 1104–1111. doi: 10.1093/bioinformatics/bti114

Fotenos, A., Snyder, A., Girton, L., Morris, J., and Buckner, R. (2005). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology* 64, 1032–1039. doi: 10.1212/01.WNL.0000154530.72969.11

Frey, B. J., and Dueck, D. (2007). Clustering by passing messages between data points. *Science* 315, 972–976. doi: 10.1126/science.1136800

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J. B., Heather, J. D., and Frackowiak, R. S. J. (1995). Spatial registration and normalization of images. *Hum. Brain Mapp.* 3, 165–189. doi: 10.1002/hbm.460030303

Gkalelis, N., Mezaris, V., Kompatsiaris, I., and Stathaki, T. (2013). Mixture subclass discriminant analysis link to restricted gaussian model and other generalizations. *Neural Netw. Learn. Syst. IEEE Trans.* 24, 8–21. doi: 10.1109/TNNLS.2012.2216545

Gönen, M., and Alpaydin, E. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* 12, 2211–2268.

Greicius, M. D., Srivastava, G., Reiss, A. L., and Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4637–4642. doi: 10.1073/pnas.0308627101

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Henze, N., and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Commun. Stat. Theory Methods* 19, 3595–3617. doi: 10.1080/03610929008830400

Hinrichs, C., Singh, V., Xu, G., and Johnson, S. C. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* 55, 574–589. doi: 10.1016/j.neuroimage.2010.10.081

Jain, A. K., and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc.

Jia, H., Wu, G., Wang, Q., and Shen, D. (2010). ABSORB: Atlas building by self-organized registration and bundling. *Neuroimage* 51, 1057–1070. doi: 10.1016/j.neuroimage.2010.03.010

Kabani, N., MacDonald, D., Holmes, C., and Evans, A. (1998). A 3D atlas of the human brain. *Neuroimage* 7:S717.

Kim, S.-W. (2010). A pre-clustering technique for optimizing subclass discriminant analysis. *Patt. Recogn. Lett.* 31, 462–468. doi: 10.1016/j.patrec.2009.07.007

Kohannim, O., Hua, X., Hibar, D. P., Lee, S., Chou, Y.-Y., Toga, A. W., Jr., et al. (2010). Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiol. Aging* 31, 1429–1442. doi: 10.1016/j.neurobiolaging.2010.04.022

Lee, A. C. H., Buckley, M. J., Gaffan, D., Emery, T., Hodges, J. R., and Graham, K. S. (2006). Differentiating the roles of the hippocampus and perirhinal cortex in processes beyond long-term declarative memory: a double dissociation in dementia. *J. Neurosci.* 26, 5198–5203. doi: 10.1523/JNEUROSCI.3157-05.2006

Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., et al. (2012). Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiol. Aging* 33, 427.e15–427.e30. doi: 10.1016/j.neurobiolaging.2010.11.008

Liao, S., Gao, Y., Oto, A., and Shen, D. (2013). "Representation learning: a unified deep learning framework for automatic prostate MR segmentation." in *Medical Image Computing and Computer-Assisted Intervention*, Vol. 8150, Lecture Notes in Computer Science (Nagoya), 254–261.

Liu, F., Wee, C.-Y., Chen, H., and Shen, D. (2013). "Inter-modality relationship constrained multi-task feature selection for ad/mci classification," in *Medical Image Computing and Computer-Assisted Intervention, Volume 8149 of Lecture Notes in Computer Science*, eds K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab (Berlin; Heidelberg: Springer), 308–315.

Liu, M., Zhang, D., and Shen, D. (2012). Ensemble sparse classification of Alzheimer's disease. *Neuroimage* 60, 1106–1116. doi: 10.1016/j.neuroimage.2012.01.055

Martinez, A. M., and Kak, A. (2001). PCA versus LDA. *IEEE Trans. Patt. Anal. Mach. Intell.* 23, 228–233. doi: 10.1109/34.908974

Mosconi, L. (2005). Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. *Eur. J. Nucl. Med. Mol. Imaging* 32, 486–510. doi: 10.1007/s00259-005-1762-7

Mwangi, B., Tian, T., and Soares, J. (2013). A review of feature reduction techniques in neuroimaging. *Neuroinformatics* 12, 229–244. doi: 10.1007/s12021-013-9204-3

Nie, F., Huang, H., Cai, X., and Ding, C. H. (2010). "Efficient and robust feature selection via joint l2,1-norms minimization," in *Advances in Neural Information Processing Systems 23*, eds J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Vancouver, BC) 1813–1821.

Noppeney, U., Penny, W. D., Price, C. J., Flandin, G., and Friston, K. J. (2006). Identification of degenerate neuronal systems based on intersubject variability. *Neuroimage* 30, 885–890. doi: 10.1016/j.neuroimage.2005.10.010

Nordberg, A., Rinne, J. O., Kadir, A., and Langstrom, B. (2010). The use of PET in Alzheimer disease. *Nat. Rev. Neurol.* 6, 78–87. doi: 10.1038/nrneurol.2009.217

Perrin, R. J., Fagan, A. M., and Holtzman, D. M. (2009). Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature* 461, 916–922. doi: 10.1038/nature08538

Qiao, L., Chen, S., and Tan, X. (2010). Sparsity preserving projections with applications to face recognition. *Patt. Recogn.* 43, 331–341. doi: 10.1016/j.patcog.2009.05.005

Roth, V. (2004). The generalized LASSO. *IEEE Trans. Neural Netw.* 15, 16–28. doi: 10.1109/TNN.2003.809398

Schroeter, M. L., Stein, T., Maslowski, N., and Neumann, J. (2009). Neural correlates of alzheimer's disease and mild cognitive impairment: a systematic and quantitative meta-analysis involving 1351 patients. *Neuroimage* 47, 1196–1206. doi: 10.1016/j.neuroimage.2009.05.037

Shen, D., and Davatzikos, C. (2002). HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Trans. Med. Imaging* 21, 1421–1439. doi: 10.1109/TMI.2002.803111

Singh, V., Chertkow, H., Lerch, J. P., Evans, A. C., Dorr, A. E., and Kabani, N. J. (2006). Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. *Brain* 129, 2885–2893. doi: 10.1093/brain/awl256

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* 17, 87–97. doi: 10.1109/42.668698

Suk, H.-I., and Lee, S.-W. (2013). A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces. *IEEE Trans. Patt. Anal. Mach. Intell.* 35, 286–299. doi: 10.1109/TPAMI.2012.69

Suk, H.-I., Lee, S.-W., and Shen, D. (2013a). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* doi: 10.1007/s00429-013-0687-3. [Epub ahead of print].

Suk, H.-I., and Shen, D. (2013). "Deep learning-based feature representation for AD/MCI classification," in *Medical Image Computing and Computer-Assisted Intervention*, Volume 8150, Lecture Notes in Computer Science (Nagoya), 583–590.

Suk, H.-I., Wee, C.-Y., and Shen, D. (2013b). "Discriminative group sparse representation for mild cognitive impairment classification," in *Machine Learning in Medical Imaging*, Volume 8184, Lecture Notes in Computer Science (Nagoya), 131–138.

Supekar, K., Menon, V., Rubin, D., Musen, M., and Greicius, M. D. (2008). Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Comput. Biol.* 4:e1000100. doi: 10.1371/journal.pcbi.1000100

Tang, S., Fan, Y., Wu, G., Kim, M., and Shen, D. (2009). RABBIT: Rapid alignment of brains by building intermediate templates. *Neuroimage* 47, 1277–1287. doi: 10.1016/j.neuroimage.2009.02.043

Tibshirani, R. (1994). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B* 58, 267–288.

Varoquaux, G., Gramfort, A., Poline, J.-B., and Thirion, B. (2010). "Brain covariance selection: better individual functional connectivity models using population prior," in *Advanced in Neural Information Processing Systems 23*, eds J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Vancouver, BC) 2334–2342.

Visser, P. J., Verhey, F. R. J., Hofman, P. A. M., Scheltens, P., and Jolles, J. (2002). Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. *J. Neurol. Neurosurg. Psychiatry* 72, 491–497. doi: 10.1136/jnnp.72.4.491

Walhovd, K., Fjell, A., Brewer, J., McEvoy, L., Fennema-Notestine, C., Hagler, D. J. Jr., et al. (2010). Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer disease. *Am. J. Neuroradiol.* 31, 347–354. doi: 10.3174/ajnr.A1809

Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A., et al. (2011). "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance," in *2011 IEEE International Conference on Computer Vision* (Barcelona) 557–562. doi: 10.1109/ICCV.2011.6126288

Wang, Y., Nie, J., Yap, P.-T., Li, G., Shi, F., Geng, X., et al. (2013). Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. *PLoS ONE* 9:e77810. doi: 10.1371/journal.pone.0077810

Wee, C.-Y., Yap, P.-T., Li, W., Denny, K., Browndyke, J. N., Potter, G. G., et al. (2011). Enriched white matter connectivity networks for accurate identification of MCI patients. *Neuroimage* 54, 1812–1822. doi: 10.1016/j.neuroimage.2010.10.026

Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., et al. (2012). Identification of MCI individuals using structural and functional connectivity networks. *Neuroimage* 59, 2045–2056. doi: 10.1016/j.neuroimage.2011.10.015

Westman, E., Muehlboeck, J.-S., and Simmons, A. (2012). Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *Neuroimage* 62, 229–238. doi: 10.1016/j.neuroimage.2012.04.056

Xue, Z., Shen, D., and Davatzikos, C. (2006). Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping. *Med. Image Anal.* 10, 740–751. doi: 10.1016/j.media.2006.06.007

Yang, J., Shen, D., Davatzikos, C., and Verma, R. (2008). "Diffusion tensor image registration using tensor geometry and orientation features," in *Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Vol. 5242, Lecture Notes in Computer Science, eds D. Metaxas, L. Axel, G. Fichtinger, and G. Szkely (New York, NY), 905–913.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., and Ye, J. (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *Neuroimage* 61, 622–632. doi: 10.1016/j.neuroimage.2012.03.059

Zhang, D., and Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 59, 895–907. doi: 10.1016/j.neuroimage.2011.09.069

Zhang, D., Shen, D., and ADNI. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS ONE* 7:e33182. doi: 10.1371/journal/pone.0033182

Zhang, D., Wang, Y., Zhou, L., Yuan, H., and Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 55, 856–867. doi: 10.1016/j.neuroimage.2011.01.008

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

Zhou, L., Wang, Y., Li, Y., Yap, P.-T., Shen, D., and ADNI. (2011). Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. *PLoS ONE* 6:e21935. doi: 10.1371/journal.pone.0021935

Zhu, M., and Martinez, A. M. (2006). Subclass discriminant analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* 28, 1274–1286. doi: 10.1109/TPAMI.2006.172

Conflict of Interest Statement: The Reviewer Dr. Heng Huang declares that, despite having collaborated with the authors, the review process was handled objectively and no conflict of interest exists. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 January 2014; accepted: 30 June 2014; published online: 07 August 2014.
*Citation: Suk H-I, Lee S-W, Shen D and The Alzheimers Disease Neuroimaging Initiative (2014) Subclass-based multi-task learning for Alzheimer's disease diagnosis. Front. Aging Neurosci. **6**:168. doi: 10.3389/fnagi.2014.00168*
This article was submitted to the journal Frontiers in Aging Neuroscience.
Copyright © 2014 Suk, Lee, Shen and The Alzheimers Disease Neuroimaging Initiative. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.