NeuroImage 100 (2014) 91–105

Contents lists available at ScienceDirect

NeuroImage

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

A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis

Xiaofeng Zhu [a], Heung-Il Suk [a], Dinggang Shen [a,b,*]

[a] Department of Radiology and BRIC, The University of North Carolina at Chapel Hill, USA
[b] Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Accepted 31 May 2014
Available online 7 June 2014

Keywords:
Alzheimer's disease (AD)
Feature selection
Joint sparse learning
Manifold learning
Mild Cognitive Impairment (MCI) conversion

ABSTRACT

Recent studies on AD/MCI diagnosis have shown that the tasks of identifying brain disease and predicting clinical scores are highly related to each other. Furthermore, it has been shown that feature selection with a manifold learning or a sparse model can handle the problems of high feature dimensionality and small sample size. However, the tasks of clinical score regression and clinical label classification were often conducted separately in the previous studies. Regarding the feature selection, to our best knowledge, most of the previous work considered a loss function defined as an element-wise difference between the target values and the predicted ones. In this paper, we consider the problems of joint regression and classification for AD/MCI diagnosis and propose a novel matrix-similarity based loss function that uses high-level information inherent in the target response matrix and imposes the information to be preserved in the predicted response matrix. The newly devised loss function is combined with a group lasso method for joint feature selection across tasks, i.e., predictions of clinical scores and a class label. In order to validate the effectiveness of the proposed method, we conducted experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and showed that the newly devised loss function helped enhance the performances of both clinical score prediction and disease status identification, outperforming the state-of-the-art methods.

© 2014 Elsevier Inc. All rights reserved.

Introduction

Alzheimer's disease (AD) is the most common form of dementia that often appears in the persons over 65 years old. Brookmeyer et al. showed that there are 26.6 million AD patients worldwide and 1 out of 85 people will be affected by AD by 2050 (Brookmeyer et al., 2007; Fan et al., 2007; Wee et al., 2011). Thus, for timely treatment that might be effective to slow the progression, it's highly important for early diagnosis of AD and its early stage, Mild Cognitive Impairment (MCI). Studies have shown that AD may significantly affect both structures and functions of the brain (Greicius et al., 2004; Guo et al., 2010; Wang et al., 2011; Zhang and Shen, 2012). Greicius et al. demonstrated that the disrupted connectivity between posterior cingulate and hippocampus led to the posterior cingulate hypometabolism (Greicius et al., 2004). Guo et al. reported that AD patients exhibited significant decrease of gray matter volume in the hippocampus, parahippocampal gyrus, and insula and superior temporal gyrus (Guo et al., 2010). However, previous imaging studies for the diagnosis of AD employed either univariate methods or group-comparison methods, thus limiting their application to disease diagnosis on an individual level (Chu et al., 2012; Lemoine et al., 2010; Li et al., 2012; Liu et al., 2012; Salas-Gonzalez et al., 2010; Wee et al., 2012; Zhang et al., 2012; Zhou et al., 2011).

For the last decades, neuroimaging has been successfully used to investigate the characters of neurodegenerative progression in the spectrum between cognitive normal and AD. Particularly, different modalities provide different kinds of information for helping monitoring AD, e.g., structural brain atrophy by Magnetic Resonance Imaging (MRI) (De Leon et al., 2007; Du et al., 2007; Fjell et al., 2010; McEvoy et al., 2009), metabolic alterations in the brain by Positron Emission Tomography (PET) (Morris et al., 2001; Santi et al., 2001), and pathological amyloid depositions through CerebroSpinal Fluid (CSF) (Buchhave et al., 2009; Fjell et al., 2010; Hansson et al., 2006; Seppälä et al., 2011). It has been shown that the analysis of patterns in neuroimaging data for AD/MCI diagnosis can be efficiently handled by machine learning and pattern recognition methods. However, the previous studies mostly focused on developing classification models for predicting categorical class labels such as AD, MCI, and healthy Normal Control (NC). Recently, regression models have also been investigated to predict clinical scores such as Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) and Mini-Mental State Examination (MMSE) from individual MRI and/or PET scans (Cheng et al.,

* Corresponding author.
E-mail address: dgshen@med.unc.edu (D. Shen).

http://dx.doi.org/10.1016/j.neuroimage.2014.05.078
1053-8119/© 2014 Elsevier Inc. All rights reserved.

2013; Franke et al., 2010; Stonnington et al., 2010; Walhovd et al., 2010). For example, Cheng et al. presented a novel semi-supervised multi-modal relevance vector regression method for predicting clinical scores of neurological diseases (Cheng et al., 2013); Duchesne et al. employed linear regression models to estimate one-year MMSE changes from structural MRI (Duchesne et al., 2009); Fan et al. and Wang et al. designed, independently, high-dimensional kernel-based regression methods to estimate ADAS-Cog and MMSE (Wang et al., 2010).

Unlike those previous studies that focused on only one of the tasks (Jie et al., 2013; Liu et al., 2014; Suk and Shen, 2013), there have been also efforts to tackle both tasks simultaneously in a unified framework. For example, Zhang and Shen proposed a method of joint feature selection for both disease diagnosis and clinical score prediction, and showed that the features used for these tasks were highly correlated (Zhang and Shen, 2012). For better understanding of the underlying mechanism of AD, our interest in this paper is to predict both clinical scores and disease status jointly, and here we call it as a Joint Regression and Classification (JRC) problem.

For a robust model construction, it has been a long issue in the field of medical image analysis to filter out uninformative features and to overcome the small sample size problem. Wang et al. showed that only a few brain areas (such as medial temporal lobe structures, medial and lateral parietal, as well as prefrontal cortical areas) may predict memory scores and thus can be used to discriminate AD from NC (Wang et al., 2011). Regarding the small sample size problem, in the diagnosis of AD, the available sample size is usually small, while the feature dimensionality is high. For example, the sample size used in (Jie et al., 2013; Liu et al., 2014) was as small as 103 (i.e., 51 AD and 52 NC), while the dimensionality of features (including MRI features and PET features) was hundreds or even thousands. The small sample size makes it difficult to build a generalized model, and the high-dimensional data could lead to the over-fitting issue (Zhu et al., 2012) although the number of intrinsic features may be low (Weinberger et al., 2004).

In order to tackle these problems, feature selection has been commonly used in the literature. Zhang and Shen embedded an $\ell_{2,1}$-norm regularizer into a sparse learning model for multi-task learning (Zhang and Shen, 2012). Recent studies on neuroimage-based AD/MCI diagnosis demonstrated that the consideration of the manifold of the data can further improve the performance of the feature selection model (Zhu et al., 2013a, 2013b). Moreover, manifold learning techniques have been used in the feature selection models for either regression or classification (Cho et al., 2012; Cuingnet et al., 2011; Jie et al., 2013; Liu et al., 2013, 2014). Cho et al. adopted a manifold harmonic transformation method on the cortical thickness data (Cho et al., 2012). Liu et al. conducted the manifold learning between a predicted graph and a target graph for AD classification (Liu et al., 2013), while Jie et al. proposed a manifold regularized multi-task learning framework to jointly select features from multi-modal data for AD diagnosis (Jie et al., 2013). To our best knowledge, previous methods usually first conducted feature selection and then built regression or classification models for the diagnosis of AD. From a mathematical standpoint, the previous methods used a loss function defined as sum of the element-wise difference between target values and predicted ones, and considered only the manifold of feature observations, not the manifold of the target variables. Furthermore, none of the previous methods considered a manifold-based feature selection method for the JRC problem.

In this paper, we propose a novel loss function that considers a high-level information inherent in the observations, and combine it with a group lasso (Yuan and Lin, 2006) for joint sparse feature selection in the JRC problem. The rationale for our approach is that, compared to the low-level neuroimaging features, it is less likely for the high-level clinical label and clinical scores to be contaminated by noises (Zhang and Shen, 2012). For this reason, we build a more robust model by taking into account the relational information between high-level clinical label and clinical scores as well as the relation among samples in feature selection. This discriminates our method from the previous methods that considered only the relation among feature samples. Specifically, we define a loss function as a matrix similarity and impose the high-level information in the target response matrix to be preserved in the predicted response matrix. For the high-level information, we use the relations between response samples and the relations between response variables in a response matrix, each of which we call as 'sample-sample relation' and 'variable–variable relation'. Hereafter, each column and each row of a matrix denote, respectively, one sample and one response variable. In our work, a sample in a response matrix consists of clinical scores and a class label, and each of the clinical scores or a class label is considered as a response variable. By utilizing these high-level information inherent in the target response matrix and imposing them to be preserved in the predicted response matrix, we define a more sophisticated loss function, which affects feature selection, and thus enhances the performances of the regression and classification in AD/MCI diagnosis.

Materials and image preprocessing

For performance evaluation, we use the ADNI dataset publicly available on the web. 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. The main goal of ADNI was designed to test if the serial of MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. To this end, ADNI recruited over 800 adults (aged 55 to 90) to participate in the research. More specifically, approximately 200 cognitively normal older individuals were followed for 3 years, 400 people with MCI were followed for 3 years, and 200 people with early AD were followed for 2 years.[1] The research protocol was approved by each local institutional review board and the written informed consent was obtained from each participant.

Subjects

The general inclusion/exclusion criteria of the subjects are briefly described as follows:

1. The MMSE score of each healthy subject (a.k.a., Normal Control (NC)) is between 24 and 30. Their Clinical Dementia Rating (CDR) is of 0. Moreover, the healthy subject is non-depressed, non MCI, and non-demented.
2. The MMSE score of each MCI subject is between 24 and 30. Their CDR is 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.
3. The MMSE score of each Mild AD subject is between 20 and 26, with the CDR of 0.5 or 1.0.

In this paper, we use baseline MRI, PET, and CSF data obtained from 202 subjects including 51 AD subjects, 52 NC subjects, and 99 MCI subjects[2]. The detailed demographic information is summarized in Table 1.

MRI, PET, and CSF

We downloaded raw Digital Imaging and Communications in Medicine (DICOM) MRI scans from the public ADNI website. These MRI scans were already reviewed for quality, and automatically corrected for spatial distortion caused by gradient nonlinearity and B1

[1] Please refer to 'www.adni-info.org' for up-to-date information.
[2] Including 43 MCI converters and 56 MCI non-converters.

Table 1
Demographic information of the subjects. The numbers in parentheses denote the number of subjects in each clinical category. (MCI-C: MCI Converters, MCI-NC: MCI Non-converters).

	AD	NC	MCI-C	MCI-C
	(51)	(52)	(43)	(56)
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	12.1 ± 3.8	12.9 ± 3.9	8.03 ± 3.8

field inhomogeneity. PET images were acquired 30–60 min post-injection. They were then averaged, spatially aligned, interpolated to a standard voxel size, intensity normalized, 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. In this study, CSF Aβ42, CSF *t*-tau, and CSF *p*-tau are used as features.

Image analysis

The image processing for all MR and PET images was conducted following the same procedures in Zhang and Shen (2012). Specifically, we first performed anterior commissure–posterior commissure correction using MIPAV software[3] on all images, and used the N3 algorithm (Sled et al., 1998) to correct the intensity inhomogeneity. Second, we extracted a brain on all structural MR images using a robust skull-stripping method, followed by manual edition and intensity inhomogeneity correction. After removal of cerebellum based on registration and intensity inhomogeneity correction by repeating N3 for three times, we used FAST algorithm in the FSL package (Zhang et al., 2001) to segment the structural MR images into three different tissues: Gray Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF). Next, we used HAMMER[4] (Shen and Davatzikos, 2002) (although other methods (Jia et al., 2010; Qiao et al., 2009; Shen et al., 1999; Shen and Davatzikos, 2004; Yang et al., 2008; Zacharaki et al., 2008) can be used) to conduct registration and obtained the Region-Of-Interest (ROI)-labeled image based on the Jacob template, which dissects a brain into 93 ROIs (Kabani, 1998). For each of all 93 ROI regions in the labeled image of one subject, we computed the GM tissue volumes in the ROI region by integrating the GM segmentation result of this subject. And, for each subject, we first aligned the PET image to its respective MR T1 image using affine registration and then computed the average intensity of each ROI in the PET image. Finally, for each subject, we obtained totally 93 features from MRI, 93 features from PET, and 3 features from CSF. In order for multi-modality fusion, we simply concatenated the features of modalities into a long feature vector.

Method

In this section, we describe our framework for joint regression and classification in AD/MCI diagnosis and propose a novel matrix similarity-based loss function and feature selection. Fig. 1 presents a schematic diagram of our method for predictions of clinical scores and a class label. Given MRI, PET, and CSF data, we first extract features from MRI and PET, while we use the CSF data itself as CSF features. We then construct a feature matrix $\mathbf{X}$ with a concatenation of multi-modal features at each column, and a corresponding response matrix $\mathbf{Y}$ with a concatenation of clinical scores (e.g., ADAS-Cog, MMSE) and a class label at each column. With our new loss function and a group lasso method, we select features that are jointly used to represent the clinical scores and the class label. By using the dimension-reduced data, we build clinical score regression models and a clinical label identification model with Support Vector Regression (SVR) and Support Vector Classification (SVC), respectively.

Notations

In 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.

Matrix-similarity based loss function

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{c \times n}$, where n, d, and c denote the numbers of samples (or subjects),[5] feature variables, and response variables, respectively. In our work, the response variables correspond to ADAS-Cog, MMSE, and a class label. We assume that the response variables can be predicted by a weighted linear combination of the features as follows:

$$\mathbf{Y} \approx \mathbf{W}^T\mathbf{X} = \hat{\mathbf{Y}} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times c}$ is a regression matrix. By regarding the prediction of each response variable as a task and constraining the same features to be used across tasks, we can use a group lasso method (Yuan and Lin, 2006) formulated as follows:

$$\min_{\mathbf{W}} f(\mathbf{W}) + \lambda \|\mathbf{W}\|_{2,1} \tag{2}$$

where $f(\mathbf{W})$ is a loss function depending on $\mathbf{W}$ and λ is a sparsity control parameter. Note that each element in a column $\mathbf{w}_k$ of $\mathbf{W}$ assigns a weight to each of the observed features in predicting the k-th response variable. The $\ell_{2,1}$-norm regularizer $\|\mathbf{W}\|_{2,1}$ penalizes all coefficients in the same row of $\mathbf{W}$ together for joint selection or un-selection in predicting the response variables. Specifically, the ℓ_2-norm regularizer enforces the selection of the same features across all tasks, and the ℓ_1-norm imposes the feature sparseness in the linear combination. In our JRC problem, this $\ell_{2,1}$-norm selects the ROIs that are highly relevant to the estimation of both clinical scores and a class label.

With regard to the loss function in Eq. (2), the most commonly used metric in the literature is the element-wise distance between the target response matrix $\mathbf{Y}$ and the predicted response matrix $\hat{\mathbf{Y}}$ as follows:

$$\begin{aligned} f(\mathbf{W}) &= \left\|\mathbf{Y} - \mathbf{W}^T\mathbf{X}\right\|_F^2 \\ &= \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_F^2 \\ &= \sum_{i=1}^{c}\sum_{j=1}^{n}\left(y_{ij} - \hat{y}_{ij}\right)^2. \end{aligned} \tag{3}$$

This element-wise loss function has been successfully used in many objective functions in the literature (Suk et al., 2013; Yuan and Lin, 2006; Zhang and Shen, 2012). From a matrix similarity point of view, Eq. (3) measures the matrix similarity between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ with the sum

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

[4] Although there exist many recent methods for registration, HAMMER has already been validated on many datasets including the ADNI dataset and continuously improved for the last decade.

[5] In this work, we have one sample per subject.

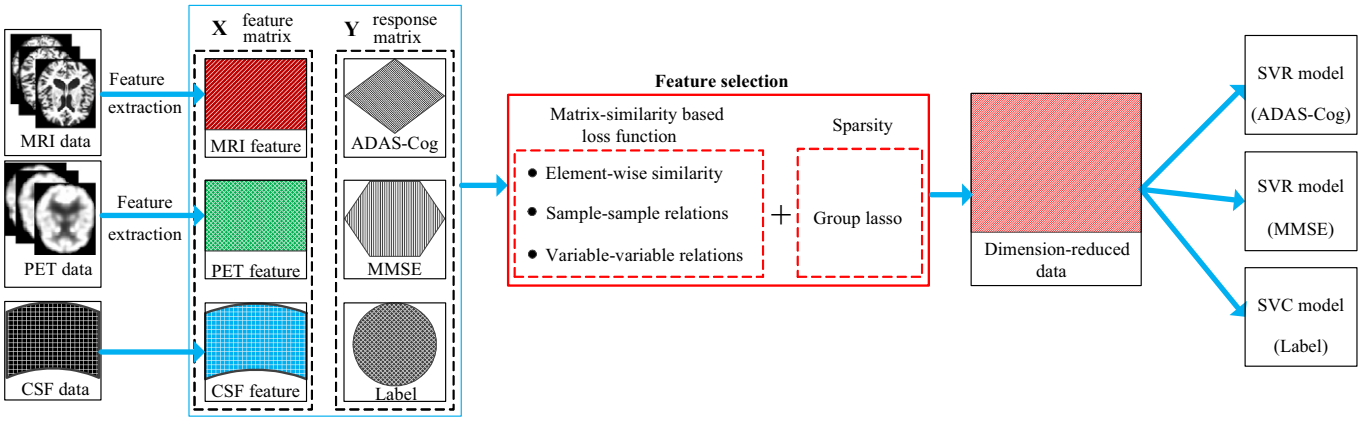


Fig. 1. The framework of the proposed method.

of the element-wise differences between matrices. Note that, in this case, the lower the score is, the more similar they are. However, we believe that there exists additional information inherent in the matrices, which we can use in measuring the similarity, such as the relations between any pair of columns and the relations between any pair of rows in a matrix. In our case, the columns and the rows correspond, respectively, to samples and response variables. Ideally, besides the element-wise values, those relations in the target response matrix $\mathbf{Y}$ should be preserved in the predicted response matrix $\hat{\mathbf{Y}}$. Concretely, the row-wise relations find the correlations of a clinical label and ADAS-Cog, a clinical label and MMSE, and ADAS-Cog and MMSE over samples, and the column-wise relations represent the correlation between any pair of samples over response variables. By enriching the loss function with the higher-level information and imposing the information to be matched between two matrices, we can find an optimal regression matrix $\mathbf{W}$ that helps accurately predict the target response values, and thus select useful features. The selected features can be finally used for more accurate prediction of testing samples in both the clinical scores and a class label.

To better characterize the newly devised loss function, we explain them in the context of a graph matching. We illustrate the sample–sample (a pair of columns) relations, e.g., $(\mathbf{y}_i-\mathbf{y}_j)$ or $(\hat{\mathbf{y}}_i-\hat{\mathbf{y}}_j)$, and the variable–variable (a pair of rows) relations, e.g., $(\mathbf{y}^k-\mathbf{y}^l)$ or $(\hat{\mathbf{y}}^k-\hat{\mathbf{y}}^l)$, by means of a graph in Figs. 2(a) and (b), respectively. In Fig. 2(a), a node represents one sample, i.e., a column vector $\mathbf{y}_i$ or $\hat{\mathbf{y}}_i$ in the respective matrices, an edge in a graph denotes the relation between the connected nodes, and different colors denote different class labels. In the graph, the samples of the same class would have a small distance, whereas the samples of different classes would have a large distance. In Fig. 2(b), a node represents a set of observations for a response variable, i.e., a row vector in the respective matrices, and an edge denotes the relation between nodes.

As explained above, we impose these relational properties in a target response matrix, now represented by graphs, to be preserved in the respective graphs for the predicted response matrix as follows:

$$G_{\mathbf{Y}}^{S} \approx G_{\hat{\mathbf{Y}}}^{S} \tag{4}$$

$$G_{\mathbf{Y}}^{V} \approx G_{\hat{\mathbf{Y}}}^{V} \tag{5}$$

where $G_{\mathbf{Y}}^{S}$ and $G_{\hat{\mathbf{Y}}}^{S}$ denote, respectively, graphs representing the sample–sample relations for the target response matrix $\mathbf{Y}$ and the predicted response matrix $\hat{\mathbf{Y}}$, and $G_{\mathbf{Y}}^{V}$ and $G_{\hat{\mathbf{Y}}}^{V}$ denote, respectively, graphs representing the variable–variable relations for the target response matrix $\mathbf{Y}$ and the predicted response matrix $\hat{\mathbf{Y}}$. Hereafter, we call the graphs representing the sample–sample relations and the variable–variable relations as '*S*-graph' and '*V*-graph', respectively. We formulate the problem of matching two sets of graphs, i.e., *S*-graph and *V*-graph, as follows:

$$\begin{aligned} M_S &= \sum_{i,j=1}^{n} \left\| \left(\mathbf{y}_i-\mathbf{y}_j\right) - \left(\hat{\mathbf{y}}_i-\hat{\mathbf{y}}_j\right) \right\|_2^2 \\ &= \sum_{i,j=1}^{n} \left\| \left(\mathbf{y}_i-\mathbf{y}_j\right) - \left(\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j\right) \right\|_2^2 \end{aligned} \tag{6}$$

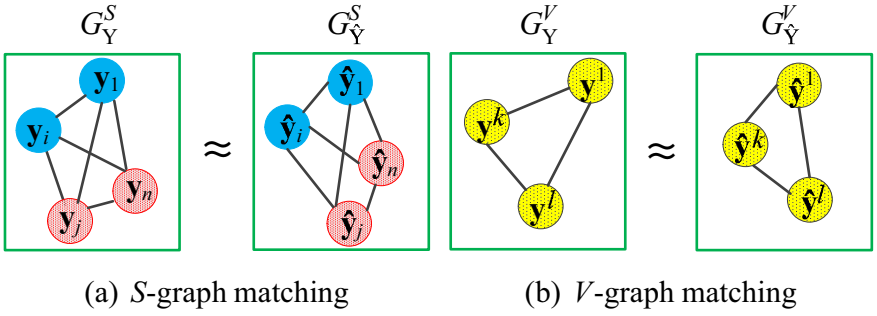


Fig. 2. An illustration of measuring matrix similarity by means of a graph matching. For simplicity, we showed only a small number of nodes. (a) Each node represents a column vector of the target or the predicted response matrix, edges represent the distance between nodes, and colors represent class labels. (b) Each node represents a row vector of the target or the predicted response matrix and edges denote the distance between nodes.

$$M_V = \sum_{k,l=1}^{c} \left\| \left(\mathbf{y}^k - \mathbf{y}^l\right) - \left(\hat{\mathbf{y}}^k - \hat{\mathbf{y}}^l\right) \right\|_2^2 = \sum_{k,l=1}^{c} \left\| \left(\mathbf{y}^k - \mathbf{y}^l\right) - \left((\mathbf{w}_k)^T\mathbf{X} - (\mathbf{w}_l)^T\mathbf{X}\right) \right\|_2^2 \quad (7)$$

where M_S and M_V denote, respectively, the graph matching scores between $G_{\mathbf{Y}}^S$ and $G_{\hat{\mathbf{Y}}}^S$, and between $G_{\mathbf{Y}}^V$ and $G_{\hat{\mathbf{Y}}}^V$, and n and c denote, respectively, the numbers of samples and response variables in the matrices as mentioned above. By introducing these newly devised graph matching terms into the loss function of Eq. (3), our new loss function becomes as follows:

$$f(\mathbf{W}) = \left\| \mathbf{Y} - \mathbf{W}^T\mathbf{X} \right\|_F^2 + \alpha_1 M_S + \alpha_2 M_V \quad (8)$$

where α_1 and α_2 denote, respectively, the control parameters for the respective terms. Compared to the conventional element-wise loss function in Eq. (3), the proposed function additionally considers two graph matching regularization terms.

Finally, our objective function for feature selection can be written as follows:

$$\min_{\mathbf{W}} \left\| \mathbf{Y} - \mathbf{W}^T\mathbf{X} \right\|_F^2 + \alpha_1 \sum_{i,j=1}^{n} \left\| \left(\mathbf{y}_i - \mathbf{y}_j\right) - \left(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right) \right\|_2^2 + \alpha_2 \sum_{k,l=1}^{c} \left\| \left(\mathbf{y}^k - \mathbf{y}^l\right) - \left((\mathbf{w}_k)^T\mathbf{X} - (\mathbf{w}_l)^T\mathbf{X}\right) \right\|_2^2 + \lambda \|\mathbf{W}\|_{2,1}. \quad (9)$$

It is worth noting that unlike the previous manifold learning methods, i.e., local linear embedding (Roweis and Saul, 2000), locality preserving projection (He et al., 2005), and high-order graph matching (Liu et al., 2013), that focused on the sample similarities by imposing nearby samples to be still nearby in the transformed space, the proposed method imposes more strict constraints, i.e., sample–sample relations and variable–variable relations, in finding the optimal regression matrix **W**.

Objective function optimization

After some mathematical transformations, we can simplify M_S and M_V as follows:

$$M_S = tr\left(2\mathbf{W}^T\mathbf{X}\mathbf{H}_n\mathbf{X}^T\mathbf{W} - 4\mathbf{Y}\mathbf{H}_n\mathbf{X}^T\mathbf{W}\right) \quad (10)$$

$$M_V = tr\left(2\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{W}^T\mathbf{X} - 4\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{Y}\right) \quad (11)$$

where $\mathbf{H}_n = n\mathbf{I}_n - 1_n(1_n)^T$ and $\mathbf{H}_c = c\mathbf{I}_c - 1_c(1_c)^T$, $\mathbf{I}_n$ (or $\mathbf{I}_c$) is an identity matrix of size n (or c), and 1_n (or 1_c) is a column vector of n (or c) ones. By replacing the graph matching terms M_S and M_V with Eqs. (10) and (11), our objective function in Eq. (9) can be rewritten as follows:

$$\min_{\mathbf{W}} \left\| \mathbf{Y} - \mathbf{W}^T\mathbf{X} \right\|_F^2 + \alpha_1 tr\left(2\mathbf{W}^T\mathbf{X}\mathbf{H}_n\mathbf{X}^T\mathbf{W} - 4\mathbf{Y}\mathbf{H}_n\mathbf{X}^T\mathbf{W}\right) + \alpha_2 tr\left(2\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{W}^T\mathbf{X} - 4\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{Y}\right) + \lambda\|\mathbf{W}\|_{2,1}. \quad (12)$$

By setting the derivative of the objective function in Eq. (12) with respect to **W** as zero, we can obtain an equation formed as follows:

$$\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{B} = \mathbf{C} \quad (13)$$

where $\mathbf{A} = -(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{X}^T + 2\alpha_1\mathbf{X}\mathbf{H}_n\mathbf{X}^T + \lambda\mathbf{Q})$, $\mathbf{B} = 2\alpha_2\mathbf{H}_c$, $\mathbf{C} = -(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{Y}^T + 2\alpha_1\mathbf{X}\mathbf{H}_n\mathbf{Y}^T + 2\alpha_2\mathbf{X}\mathbf{Y}^T\mathbf{H}_c)$, and $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with the i-th diagonal element set to

$$q_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}. \quad (14)$$

Although the objective function in Eq. (12) is convex, due to the non-smooth term of $\|\mathbf{W}\|_{2,1}$, it is not straightforward to find the global optimum. Furthermore, due to the inter-dependence in computing matrices of **W** and **Q**, it's not trivial to solve Eq. (13). To this end, in this work, we apply an iterative approach to optimize Eq. (13) by alternatively computing **Q** and **W**. That is, at the t-th iteration, we first update the matrix $\mathbf{W}(t)$ with the matrix $\mathbf{Q}(t-1)$ fixed and then update the matrix $\mathbf{Q}(t)$ with the updated matrix $\mathbf{W}(t)$. Refer to Algorithm 1[6] and Appendix A, respectively, for implementation details and the proof of convergence of our algorithm.

Algorithm 1. Pseudo code of solving Eq. (12).

Input: $\mathbf{X} \in \mathbb{R}^{d\times n}$, $\mathbf{Y} \in \mathbb{R}^{c\times n}$, $\alpha_1, \alpha_2, \lambda$;
Output: $\mathbf{W}$;
Initialize $t = 0$, $\mathbf{Q}(t)$ as a random diagonal matrix;
repeat
  Update $\mathbf{W}(t+1)$ by solving Eq. (13)[6];
  Update $\mathbf{Q}(t+1)$ via Eq. (14);
  $t = t+1$;
until *Eq. (12) converges*;

Feature selection and model training

Due to the use of an $\ell_{2,1}$-norm regularizer in our objective function, after finding the optimal solution with Algorithm 1, we have some zero (or close to zero) row vectors in **W**, whose corresponding features are not useful in joint predictions of clinical scores and a class label. Furthermore, following the literatures (Zhu et al., 2013b, 2013c), we believe that the lower the ℓ_2-norm value of a row vector, the less informative the respective feature in our observation. To this end, we first sort rows in **W** in a descending order based on their ℓ_2-norm values, i.e., $\|\mathbf{w}^j\|_2$, $j \in \{1, \ldots, d\}$, to find K top-ranked rows,[7] and then select the respective features. Note that the selected features are jointly used to predict clinical scores and a class label.

With the selected features, we then train support vector machines, which have been successfully used in many fields (Suk and Lee, 2013; Zhang and Shen, 2012). Specifically, we build two SVR (Smola and Schölkopf, 2004) models for predicting ADAS-Cog and MMSE scores, respectively, and a SVC (Burges, 1998) model for identifying a class label.[8]

Experimental results

We conducted various experiments to compare the proposed method with the state-of-the-art methods, as detailed below.

Experimental settings

We considered three binary classification problems: AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC. For MCI vs. NC, both MCI-C and MCI-NC were labeled as MCI. For each set of experiments, we used features from MRI, PET, MRI + PET (MP for short), or MRI + PET + CSF (MPC for short) for training our feature selection model with the

[6] In our work, we used the built-in function 'lyap' in MATLAB, i.e., vec(W(t)) = (I ⊗ A(W(t − 1)) + BT ⊗ I) − 1vec(C), where A is a function of W.

[7] Following the previous work (Zhu et al., 2013a, 2013b, 2013c), we set K as the number of non-zero row vectors, i.e., $K = \sum_i \delta\left(\|\mathbf{w}^i\|_2 > \theta\right)$, where $\delta(\cdot)$ is a Kronecker delta function and θ is a threshold. In our experiments, we set $\theta = 10^{-5}$ empirically.

[8] We used the LIBSVM toolbox available at 'http://www.csie.ntu.edu.tw/~cjlin/libsvm/'.

same target responses, i.e., 2 clinical scores and 1 class label. Then, with the respectively selected features, we trained two regression models, each of which was for a clinical score of ADAS-Cog and MMSE, respectively, and one classification model for a class label.

To evaluate the performance of all competing methods, we employed the metrics of Correlation Coefficient (CC) and Root Mean Squared Error (RMSE) between the predicted clinical scores and the target clinical scores in regression, and also the metrics of classification ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), and Area Under Curve (AUC) in classification.

We used 10-fold cross-validation to compare all methods. Specifically, we first randomly partitioned the whole dataset into 10 subsets. We then selected one subset for testing and used the remaining 9 subsets for training. We repeated the whole process 10 times to avoid the possible bias during dataset partitioning for cross-validation. The final result was computed by averaging results from all experiments. For the model selection, i.e., tuning parameters[9] in Eq. (12) and in the LIBSVM toolbox,[10] we further split the training dataset into 5 subsets for 5-fold inner cross-validation. The parameters that resulted in the best performance in the inner cross-validation were used in testing.

Competing methods

We particularly selected the following methods/ways for comparison.

- Original features based method: We conducted the tasks of regression and classification using the original features with no feature selection step, and used them as baseline method. In the following, we denote this method with the suffix "N".
- Single-task based method: We conducted each of regression or classification tasks separately by using the objective function in Eq. (12). In particular, although here we used the same original features as the proposed method, we performed the task of regression or classification separately at each time for selecting their own sets of features. In the following, we use the suffix "S" to represent the type of single-task based method. For example, MP-S denotes a single-task based feature selection method on the MP data.
- M3T (Zhang and Shen, 2012): This Multi-Modal Multi-Task method includes two key steps: (1) using multi-task feature selection for determining a common subset of relevant features for multiple response variables (or multiple tasks) from each modality, and (2) a multi-kernel decision fusion for integrating the selected features from all modalities for prediction. It is worth noting that M3T is a special case of our method, i.e., by setting $\alpha_1 = 0$ and $\alpha_2 = 0$ in Eq. (12).
- HOGM (Liu et al., 2013): High-Order Graph Matching method uses a sample–sample relation in a matrix and applies an ℓ_1-norm regularization term with a single response or a single task.
- M2TFS (Jie et al., 2013): Manifold regularized Multi-Task Feature Selection (M2TFS) conducts feature selection by combining the least square loss function with an $\ell_{2,1}$-norm regularizer and a graph regularizer, and then perform multi-modality classification as a multi-task learning framework with each task focusing on each modality. This method is designed only for conducting classification. In our experiments, M2TFS included two versions, i.e., (1) M2TFS-C, denoting the use of simple concatenation of multi-modality features for classification, and (2) M2TFS-K, denoting the use of multiple kernels for fusing information from multi-modality data. Since M2TFS was designed for multi-modality data, requiring each modality with the same feature dimensionality, we applied it to only MP in our experiments.

[9] $\alpha_1 \in \{10^{-5}, \ldots, 10^{2}\}$, $\alpha_2 \in \{10^{-5}, \ldots, 10^{2}\}$, and $\lambda \in \{10^{2}, \ldots, 10^{8}\}$ in our experiments.
[10] $C \in \{2^{-5}, \ldots, 2^{5}\}$ in our experiments.

Simulation study

In this section, we justify the validity of the proposed method on simulation data and compare with the competing methods. For the simulation study, we generated data using a linear regression model of $\mathbf{Y} = \mathbf{W}^T\mathbf{X} + \mathbf{E}$, where $\mathbf{X} \in \mathbb{R}^{d \times n}$ is a regressor matrix, $\mathbf{W} \in \mathbb{R}^{d \times 3}$ is a coefficient matrix, $\mathbf{E} \in \mathbb{R}^{3 \times n}$ is a noise matrix, and $\mathbf{Y} = [\mathbf{y}_1^T, \mathbf{y}_2^T, \mathbf{y}_3^T]^T \in \mathbb{R}^{3 \times n}$ is a response matrix. Specifically, we generated two datasets to consider the cases of single-modality and multi-modality. (1) Single-modality: For each class, we generated n_i $(i = 1,2)$ samples by setting the first d_0 rows relevant to the classes and the remaining d–d_0 rows irrelevant for discrimination. The samples of each class were generated from multivariate normal distribution. The class labels of all samples were set in $\mathbf{y}_3$. We constructed $\mathbf{W}$ by setting the first d_0 rows with the values drawn from $\mathcal{N}(0, 1)$ and the rest d–d_0 rows zero. We then obtained the noise $\mathbf{E}$ from $\mathcal{N}\left(0, 10^{-3}\Sigma(0.1)\right)$, where $\Sigma(0.1)$ was a covariance matrix with the diagonal elements of 1 and the off-diagonal elements of 0.1. After obtaining $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{E}$ as described above, we obtained the observation $[\mathbf{y}_2^T, \mathbf{y}_3^T]^T$ via the linear regression model and then centered and standardized it. We generated data sets of 'Data1' by setting $n_1 = 50$, $n_2 = 60$, $d = 80$, and $d_0 = 30$, and 'Data2' by setting $n_1 = 50$, $n_2 = 50$, $d = 120$, and $d_0 = 60$. (2) Multi-modality: Applying the same setting with the single-modality, we generated $\mathbf{W}$, and $\mathbf{E}$, and two regression matrices with the same dimensionality. $\mathbf{X}$ includes these two regression matrices to form multi-modality data. Finally, we obtained $\mathbf{Y}$ and then centered and standardized it. We generated data sets of 'Data3' $n_1 = 50$, $n_2 = 50$, $d = 140$, and $d_0 = 50$, and 'Data4' by setting $n_1 = 50$, $n_2 = 40$, $d_1 = 120$, and $d_0 = 50$.

We applied the proposed method and the competing methods on these simulated data according to the experimental setting in Section Experimental settings, and evaluated the performances using the metrics of Correlation Coefficient (CC) and ACCuracy (ACC) for regression and classification, respectively. Table 2 shows the results on the four simulation datasets. The proposed method obtained the best performance in both classification and regression. Specifically, first, the method without feature selection obtained the worst performance for both classification and regression in the four simulated dataset. This shows the importance of conducting feature selection on the

Table 2
Performance comparison on simulated data. The number in parentheses is a standard deviation. Note that 'Data1-N' means the original features of 'Data1' and 'Data2-S' means the single-task based feature selection method on 'Data2'. The boldface denotes the best performance in each metric and each dataset.

Dataset	Method	ACC	CC (ADAS-Cog)	CC (MMSE)
Data1	Data1-N	0.701 (0.093)	0.806 (0.118)	0.787 (0.142)
	Data1-S	0.701 (0.086)	0.923 (0.065)	0.898 (0.070)
	HOGM	0.712 (0.090)	0.949 (0.020)	0.938 (0.023)
	M3T	0.704 (0.093)	0.950 (0.118)	0.948 (0.032)
	Proposed	**0.720 (0.073)**	**0.984 (0.016)**	**0.980 (0.017)**
Data2	Data2-N	0.709 (0.102)	0.765 (0.132)	0.769 (0.131)
	Data2-S	0.725 (0.099)	0.799 (0.105)	0.800 (0.123)
	HOGM	0.720 (0.106)	0.832 (0.073)	0.827 (0.088)
	M3T	0.719 (0.105)	0.857 (0.169)	0.830 (0.161)
	Proposed	**0.747 (0.071)**	**0.896 (0.061)**	**0.879 (0.080)**
Data3	Data3-N	0.640 (0.115)	0.780 (0.196)	0.696 (0.189)
	Data3-S	0.650 (0.128)	0.783 (0.149)	0.718 (0.170)
	M2TFS-C	0.655 (0.138)	0.798 (0.132)	0.734 (0.166)
	M2TFS-K	0.668 (0.129)	0.812 (0.124)	0.746 (0.153)
	HOGM	0.678 (0.119)	0.802 (0.142)	0.748 (0.170)
	M3T	0.654 (0.141)	0.820 (0.159)	0.751 (0.202)
	Proposed	**0.698 (0.107)**	**0.850 (0.101)**	**0.780 (0.155)**
Data4	Data4-N	0.626 (0.111)	0.821 (0.117)	0.650 (0.205)
	Data4-S	0.641 (0.096)	0.848 (0.114)	0.695 (0.208)
	M2TFS-C	0.649 (0.084)	0.861 (0.077)	0.739 (0.173)
	M2TFS-K	0.658 (0.799)	0.875 (0.090)	0.745 (0.154)
	HOGM	0.664 (0.101)	0.868 (0.088)	0.754 (0.173)
	M3T	0.651 (0.100)	0.879 (0.155)	0.750 (0.217)
	Proposed	**0.684 (0.070)**	**0.922 (0.098)**	**0.788 (0.153)**

Table 3
Comparison of classification performances (%) of the competing methods. (ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), and Area Under Curve (AUC)). The boldface denotes the best performance in each metric and each feature.

Feature	Method	AD vs. NC				MCI vs. NC				MCI-C vs. MCI-NC			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
MRI	MRI-N	89.5	82.7	86.3	95.3	68.3	92.6	39.2	82.5	60.25	15.5	92.3	68.7
	MRI-S	91.2	85.9	92.5	96.7	76.7	93.3	37.6	83.7	64.5	24.9	**95.8**	70.6
	HOGM	93.4	89.5	92.5	97.1	77.7	**95.6**	51.4	84.4	66.8	36.7	95.0	72.2
	M3T	92.6	87.2	95.9	97.5	78.1	94.5	54.0	83.1	67.1	37.7	92.0	72.5
	Proposed	**93.8**	**89.7**	**96.7**	**97.9**	**79.7**	95.0	**56.1**	**85.2**	**70.8**	**40.7**	94.0	**75.6**
PET	PET-N	86.2	83.5	84.8	94.8	69.0	95.0	30.8	77.9	62.2	21.6	93.1	71.3
	PET-S	87.9	85.7	90.9	94.7	73.8	**96.5**	36.2	78.7	65.1	31.0	**95.5**	73.5
	HOGM	91.7	91.1	92.8	95.6	74.7	**96.5**	43.2	79.3	66.6	35.5	**95.5**	72.4
	M3T	90.9	90.5	93.1	96.4	77.2	94.5	44.3	80.5	67.0	39.1	93.2	73.1
	Proposed	**92.3**	**92.3**	**93.9**	**96.6**	**79.1**	96.1	**47.2**	**81.2**	**70.9**	**42.7**	94.1	**77.4**
MP	MP-N	89.7	92.2	85.9	96.1	71.6	96.1	43.9	82.7	62.7	22.6	93.5	73.2
	MP-S	90.8	92.6	93.8	96.7	76.3	**97.0**	39.9	83.4	66.9	33.9	96.0	75.7
	M2TFS-C	91.0	90.4	91.4	95	73.4	76.5	**67.1**	78.0	58.4	52.3	63.0	60.0
	M2TFS-K	95.0	**94.9**	95.0	97.0	79.3	85.9	66.6	82.0	68.9	**64.7**	71.8	70.0
	HOGM	95.2	92.8	95.4	97.8	79.5	96.6	58.6	84.6	67.6	45.5	**96.8**	75.1
	M3T	94.0	92.0	96.3	98.0	78.4	95.0	57.7	83.9	67.9	47.0	93.3	75.7
	Proposed	**95.3**	93.5	**98.1**	**98.3**	**80.2**	96.5	59.7	**85.5**	**72.0**	48.1	94.3	**78.7**
MPC	MPC-N	90.8	93.1	88.3	96.5	72.5	96.3	47.1	84.1	64.1	23.1	93.6	73.9
	MPC-S	92.5	94.1	93.8	97.6	77.1	97.1	47.5	83.9	67.8	34.1	96.8	75.8
	HOGM	95.6	94.5	96.9	98.5	80.6	96.7	**64.7**	86.2	68.8	47.5	**98.5**	75.3
	M3T	94.6	93.1	96.4	98.5	80.1	95.2	58.7	84.3	68.5	47.5	92.7	76.0
	Proposed	**95.9**	**95.7**	**98.6**	**98.8**	**82.0**	**98.0**	60.1	**87.0**	**72.6**	**48.5**	94.4	**78.8**

high-dimensional features before performing classification or regression. Second, our joint classification and regression framework outperform the single-task framework since the joint framework uses more information than the single-task framework. Third, all methods with multi-modality data improved performances compared to the methods with single-modality data.

Classification results

Table 3 shows the classification performance for all methods. Fig. 3 shows the classification accuracy of the proposed method using single-task or multi-task formulation. Fig. 4 shows the Receiver Operating Characteristic (ROC) curves of the proposed method using four different combinations of data, i.e., MRI, PET, MP, and MPC. From the results, it is clear that the proposed method outperforms the competing methods in all experiments. Specifically, we observe the following results.

- It is important to conduct feature selection on the high-dimensional features before performing classification. The worst results were obtained by the methods without feature selection, i.e., MRI-N, PET-N, MP-N, and MPC-N. For example, for MRI-based classification as shown in Table 3, even using a simple feature selection method, i.e., MRI-S, can still increase the classification accuracy by 1.7%, 8.4%, and 4.25% compared to MRI-N in AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC classifications, respectively. Our method with MPC improved the classification accuracy by 5.1%, 9.5%, and 8.5%, in AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC classifications, respectively.
- It is beneficial to use joint regression and classification framework for feature selection, even only for the task of classification. As shown in Table 3 and Fig. 3, the proposed method that performed feature selection for joint regression and classification achieved better classification performance than the single-task based classification methods (MRI-S, PET-S, MP-S, and MPC-S). For example, for MRI-based classification, our method improved the classification accuracy by 2.6%, 3.0%, and 6.3% compared to MRI-S based method in AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC classifications, respectively.
- Multi-modality data helps improve classification performance. As shown in Table 3, in all experiments, the classification performances of all methods with multi-modality data such as MP and MPC were better than the same methods with single-modality data such as MRI or PET. Also, the classification performance by MPC was generally better than MP. For example, in classifying AD from NC, the proposed method with MPC achieved the classification accuracy of 95.9%, sensitivity of 95.7%, specificity of 98.6%, and AUC of 98.8%, while the best performance among other competing methods with single-modality data was only 93.8% (accuracy), 92.3% (sensitivity), 96.7% (specificity), and 97.9% (AUC), respectively, and the best performance among other competing methods with MP data was 95.3% (accuracy), 94.9% (sensitivity), 98.1% (specificity), and 98.3% (AUC), respectively.

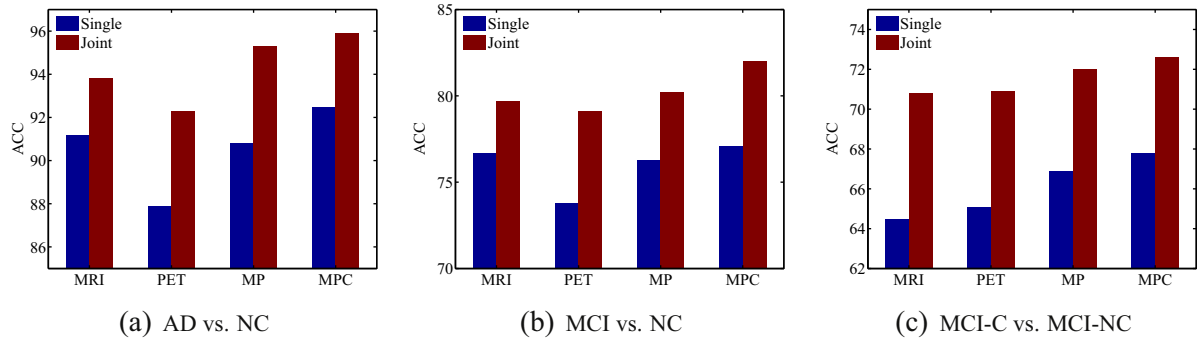


Fig. 3. Comparison of classification ACCuracy (ACC) of the proposed method with single-task ("Single") or multi-task ("Joint") learning.

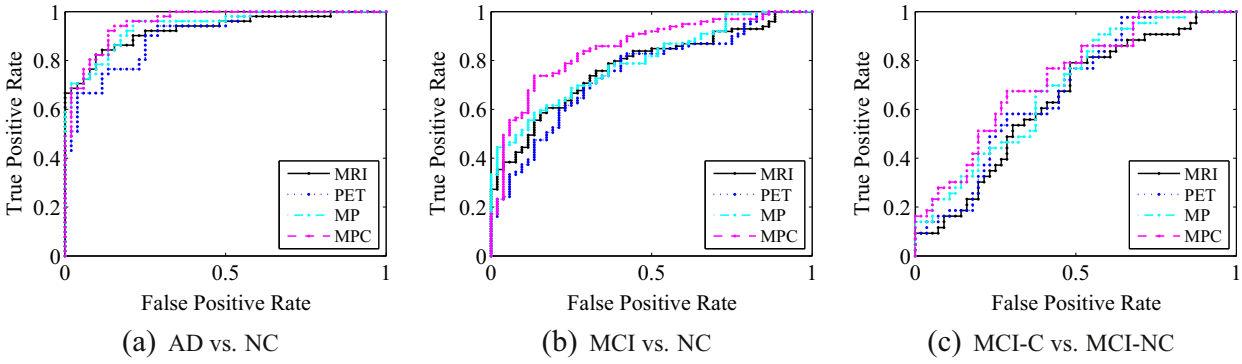


Fig. 4. Receiver Operating Characteristic (ROC) curves for the proposed method using 4 different types of data.

In classifying MCI from NC, the proposed method with MPC achieved the classification accuracy of 82.0%, sensitivity of 98.0%, specificity of 60.1%, and AUC of 87.0%, while the best performance among other competing methods with single-modality data was only 79.7% (accuracy), 96.5% (sensitivity), 56.1% (specificity), and 85.2% (AUC), respectively, and the best performance among other competing methods with MP data was 80.2% (accuracy), 97.0% (sensitivity), 67.1% (specificity), and 85.5% (AUC), respectively. In classifying MCI-C from MCI-NC, the proposed method with MPC achieved the classification accuracy of 72.6%, sensitivity of 48.5%, specificity of 94.4%, and AUC of 78.8%, while the best accuracy among other competing methods with single-modality data was only 70.9% (accuracy), 42.7% (sensitivity), 95.5% (specificity), and 77.4% (AUC), respectively, and the best performance among other competing methods with MP was 72.0% (accuracy), 64.7% (sensitivity), 96.8% (specificity), and 78.7% (AUC), respectively.

Regression results

We evaluated the regression performance through the estimation of clinical scores (i.e., ADAS-Cog and MMSE) for the cases of using MRI, PET, MP, and MPC, respectively. We presented the results of CCs and RMSEs of all competing methods in Table 4 and Figs. 5–9, respectively.

Table 4 shows that the proposed method outperforms all other competing methods, when using a combination of three multi-modality data. Fig. 5 shows the regression performance of our method with a single-task or a multi-task learning scheme. Figs. 6–9 further show the scatter plots of the target scores vs. the estimated scores of our method for ADAS-Cog and MMSE, respectively, when using 4 different types of data. In these figures, the horizontal axis represents the predicted values of ADAS-Cog (top in Figs. 6–9) or MMSE (bottom in Figs. 6–9), and the vertical axis represents the target values.

In Table 4, we can see that the regression performance of the methods without feature selection (MRI-N, PET-N, MP-N and MPC-N) was worse than methods with feature selection. Moreover, our method consistently achieved the best performance compared to other competing methods. Table 4 and Figs. 6–9 also indicate that our method with MPC consistently outperformed the same method with MP on each performance measure, although the method with MP already achieved a better performance than our method with a single modality such as MRI or PET. This scenario was observed for all other competing methods. In the prediction of ADAS-Cog and MMSE scores in AD vs. NC, our method with MPC obtained the CCs of 0.668 and 0.685, respectively, and the RMSEs of 4.47 and 1.78, respectively. The best performance among other competing methods with features of a single modality such as MRI or PET was 0.663 and 0.650 (CCs), and 4.58 and

Table 4
Comparison of regression performances of the competing methods in terms of Correlation Coefficient (CC) and Root Mean Square Error (RMSE). The boldface denotes the best performance in each metric and each feature.

Feature	Method	AD vs. NC				MCI vs. NC				MCI-C vs. MCI-NC			
		ADAS-Cog		MMSE		ADAS-Cog		MMSE		ADAS-Cog		MMSE	
		CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE
MRI	MRI-N	0.587	4.96	0.520	2.02	0.329	4.48	0.309	1.90	0.420	4.10	0.441	1.51
	MRI-S	0.591	4.85	0.566	1.95	0.347	4.27	0.367	1.64	0.426	4.01	0.482	1.44
	HOGM	0.625	**4.53**	0.598	1.91	0.352	4.26	0.371	1.63	0.435	**3.94**	0.521	1.41
	M3T	0.649	4.60	0.638	1.91	0.445	4.27	0.420	1.66	0.497	4.01	0.550	1.41
	Proposed	**0.661**	4.58	**0.650**	**1.89**	**0.461**	**4.21**	**0.441**	**1.62**	**0.543**	3.97	**0.573**	**1.39**
PET	PET-N	0.597	4.86	0.514	2.04	0.333	4.34	0.331	1.70	0.382	4.08	0.452	1.50
	PET-S	0.620	4.83	0.593	2.00	0.356	4.26	0.359	1.69	0.437	4.00	0.478	1.48
	HOGM	0.600	4.69	0.515	1.99	0.360	**4.21**	0.368	1.67	0.430	4.03	0.523	**1.41**
	M3T	0.647	4.67	0.593	1.92	0.447	4.24	0.432	1.68	0.520	3.91	0.569	1.45
	Proposed	**0.663**	**4.64**	**0.610**	**1.89**	**0.452**	**4.21**	**0.444**	**1.66**	**0.542**	**3.88**	**0.571**	1.43
MP	MP-N	0.626	4.80	0.587	1.99	0.365	4.29	0.335	1.69	0.431	4.09	0.455	1.47
	MP-S	0.634	4.83	0.585	1.92	0.359	4.25	0.371	1.67	0.449	4.00	0.496	1.41
	M2TFS-C	0.641	4.89	0.636	1.87	0.446	4.25	0.408	1.64	0.504	3.99	0.545	1.38
	M2TFS-K	0.645	4.59	0.648	1.82	0.458	4.21	0.415	1.63	0.517	3.99	0.557	1.37
	HOGM	0.633	4.64	0.602	1.83	0.364	**4.20**	0.365	1.65	0.450	3.93	0.531	1.40
	M3T	0.653	4.61	0.639	1.91	0.450	4.23	0.433	1.64	0.522	3.81	0.567	1.36
	Proposed	**0.666**	**4.53**	**0.651**	**1.80**	**0.463**	**4.20**	**0.448**	**1.62**	**0.542**	**3.76**	**0.579**	**1.35**
MPC	MPC-N	0.629	4.79	0.588	1.97	0.368	4.29	0.337	1.70	0.449	4.08	0.457	1.46
	MPC-S	0.638	4.81	0.599	1.92	0.366	4.25	0.394	1.66	0.461	4.00	0.517	1.40
	HOGM	0.639	4.63	0.611	1.81	0.365	4.20	0.368	1.65	0.454	3.92	0.534	1.40
	M3T	0.665	4.59	0.663	1.81	0.451	4.19	0.441	1.62	0.530	3.72	0.570	1.31
	Proposed	**0.668**	**4.47**	**0.685**	**1.78**	**0.470**	**4.16**	**0.456**	**1.59**	**0.556**	**3.62**	**0.584**	**1.29**

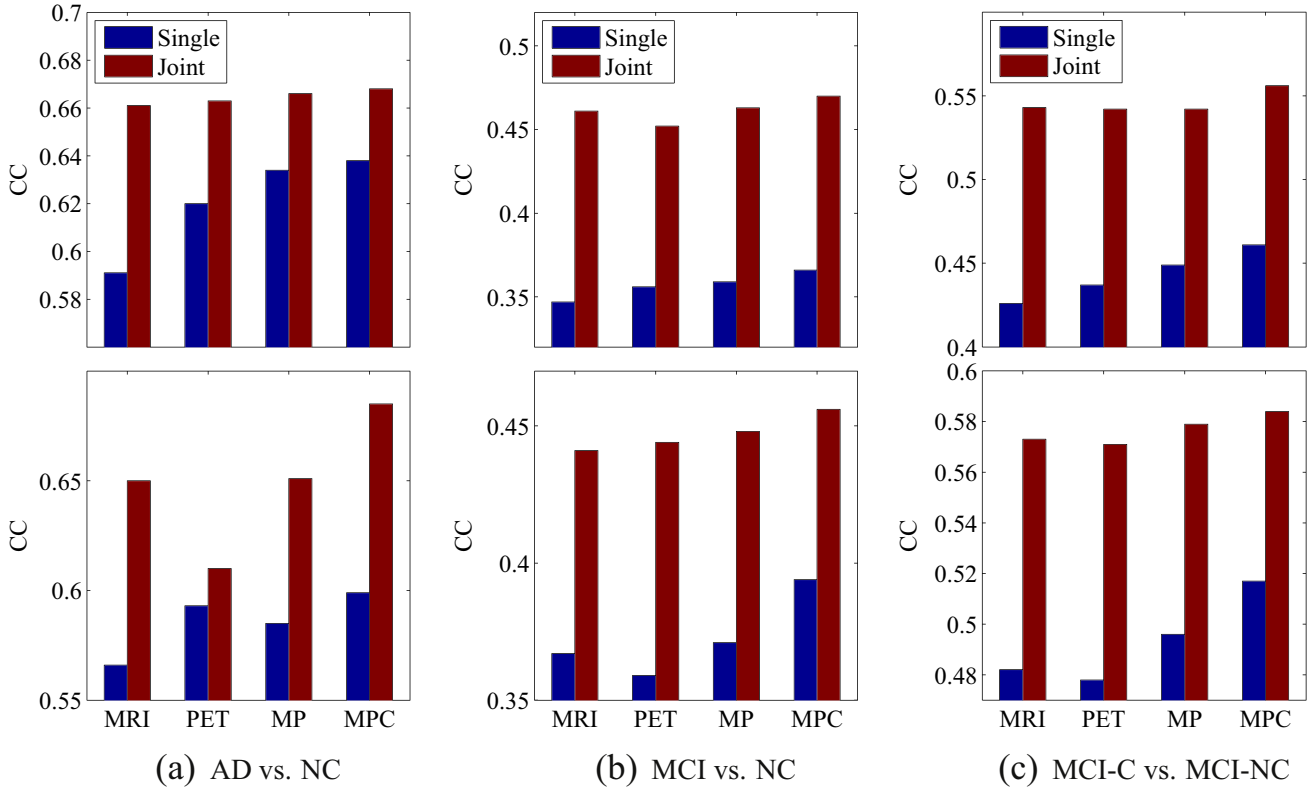


Fig. 5. Correlation Coefficients (CC) for ADAS-Cog (top) and MMSE (bottom) score prediction with our method formulated for single-task ("Single") or multi-task ("Joint") regression.

1.89 (RMSEs), respectively, and the best performance by other competing methods with MP features was 0.666 and 0.651 (CCs), and 4.53 and 1.80 (RMSEs), respectively. In MCI vs. NC, our method obtained CCs of 0.47 (ADAS-Cog) and 0.456 (MMSE), and RMSEs of 4.16 (ADAS-Cog) and 1.59 (MMSE), which were superior to those of single-modality or MP. The proposed method also obtained the best results for the predictions of ADAS-Cos and MMSE scores in MCI-C vs. MCI-NC.

We also compared the proposed method with its variants, i.e., the single response (or single task) based method in Fig. 5. From the figure, we can see that the joint formulation of registration and classification outperforms the single-task based regression, same as for the classification task above.

Results summary

From our experimental results, we found that (1) the proposed method formulated in a joint regression and classification framework outperformed its counterpart that was formulated separately for regression or classification; (2) the joint use of multiple modalities outperformed the case of using a single modality separately. Moreover, the paired-sample t-tests (at 95% significance level) between results of our method and all other competing methods (e.g., with the p-values of all cases less than 0.025 and most cases less than 0.001) showed that our method was significantly better than all other methods on the tasks of predicting clinical scores (i.e., ADAS-Cog and MMSE) and identifying class label.

We also compared our method with M3T (Zhang and Shen, 2012) that used the element-wise loss function in feature selection and also the methods that considered only either 'sample–sample relation' (*S*-graph) or 'variable–variable relation' (*V*-graph). In Fig. 10, we can see that (1) both *S*-graph and *V*-graph based methods showed better performances in regression and classification than M3T. The mean improvement by both *S*-graph and *V*-graph based methods was about 1% compared to M3T. (2) Although there was no significant difference between *S*-graph and *V*-graph based methods (at 95% significance level in the paired-sample *t*-tests), our method that considered both graphs simultaneously were statistically significant different from them and M3T.

Most discriminative brain regions

We also investigated the most discriminative regions that were selected by the proposed feature selection method. Since the feature selection in each fold was performed only based on the training set, the selected features could vary across different cross-validations. We thus defined the most discriminative regions based on the selected frequency of each region over the cross-validations. The top 10 selected regions in MCI vs. NC classification with MPC were marked in Fig. 11. They were amygdala right, hippocampal formation left, hippocampal formation right, entorhinal cortex left, temporal pole left, parahippocampal gyrus left, uncus left, perirhinal cortex left, cunecus left, and temporal pole right. It is noteworthy that the top six-ranked brain regions are known to be highly related to AD and MCI in many previous studies (Chételat et al., 2005; Convit et al., 2000; Fox and Schott, 2004; Liu et al., 2014; Misra et al., 2009; Zhang and Shen, 2012). Moreover, according to Table 5, almost all the competing methods[11] selected these six regions as the top selected regions. Even though most of the methods (including our methods and the competing methods) in our experiments selected these six features as a part of their final feature set, our proposed method outperforms the competing methods since it can select more useful features than the competing methods thanks to consideration of high-level information.

[11] The selected brain regions by both M2TFS-C and M2TFS-K were conducted on MP data.

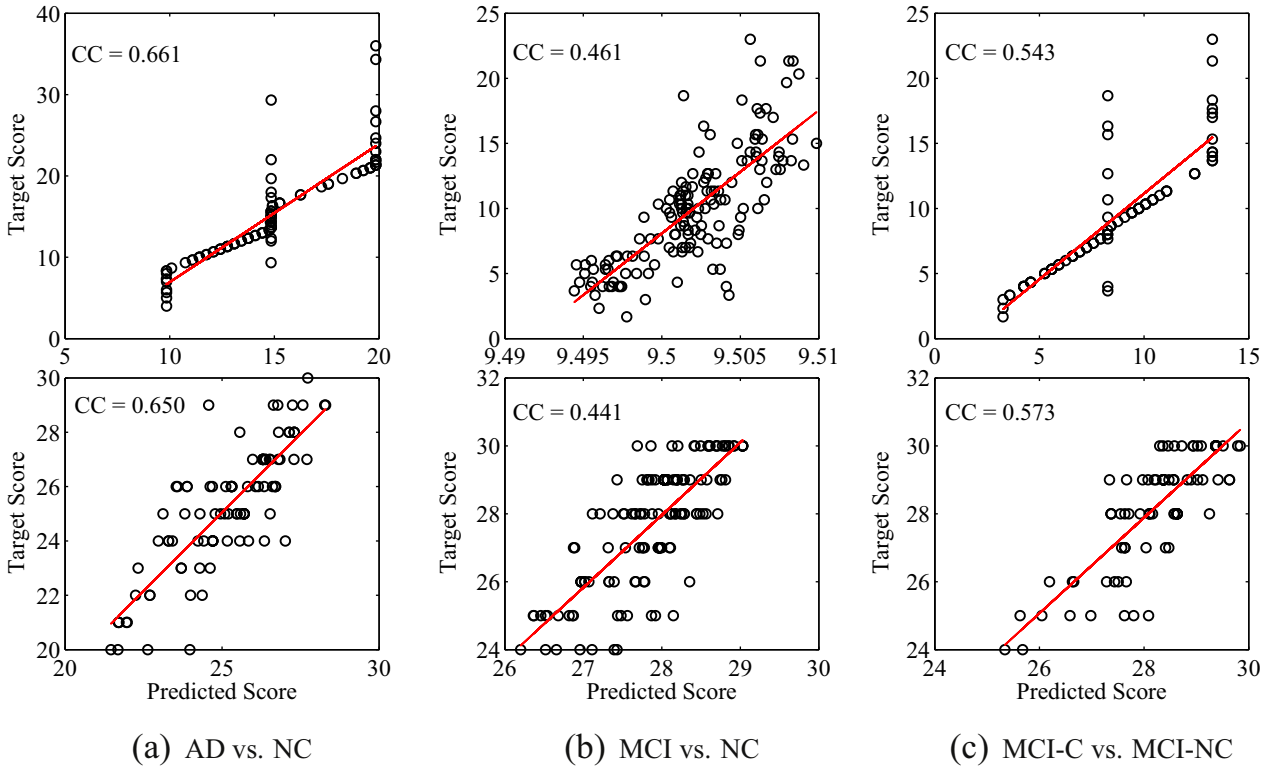


Fig. 6. Scatter plots and the respective Correlation Coefficients (CCs) obtained by the proposed method on MRI data (top: ADAS-Cog, bottom: MMSE).

Conclusion

In this work, we proposed a novel loss function in the context of a matrix similarity. Specifically, we used high-level information inherent in the target response matrix and imposed the information to be preserved in the predicted response matrix. Our objective function for joint feature selection was formulated by combining the newly devised loss function with a group lasso. In our extensive experiments on ADNI

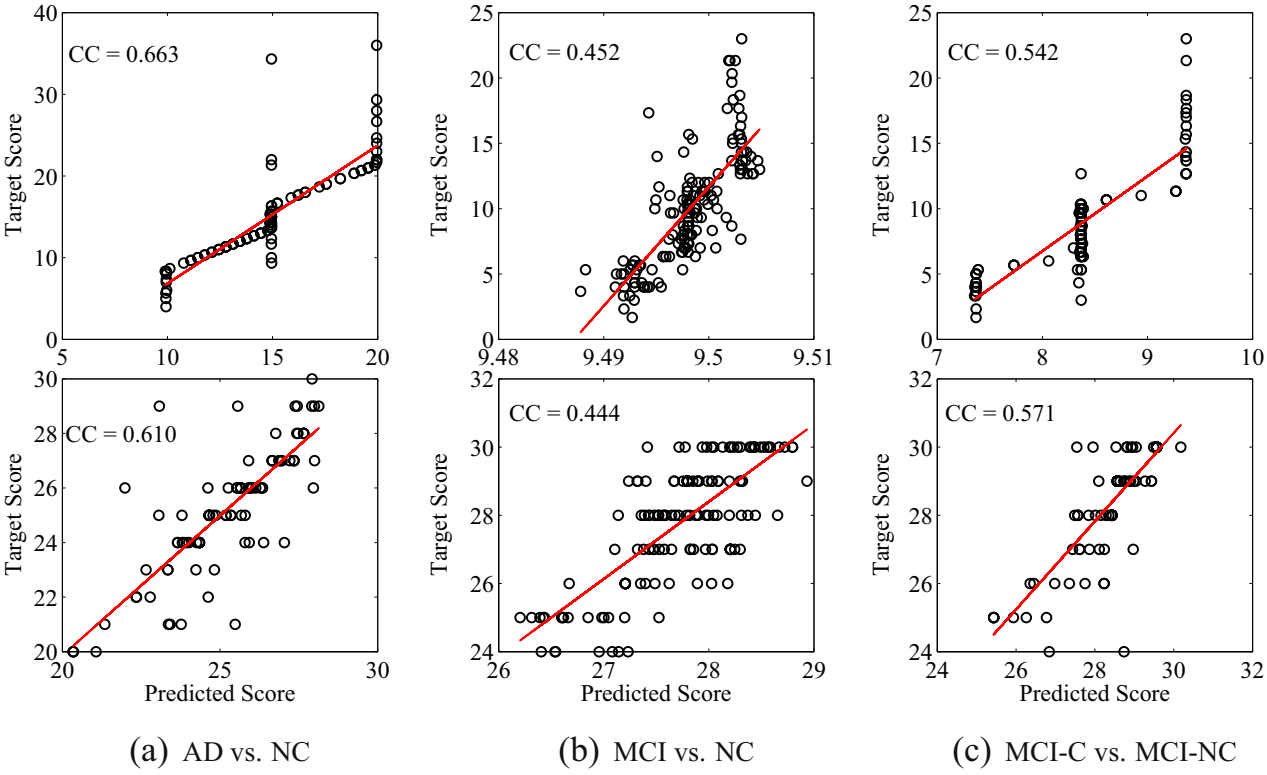


Fig. 7. Scatter plots and the respective Correlation Coefficients (CCs) obtained by the proposed method on PET data (top: ADAS-Cog, bottom: MMSE).

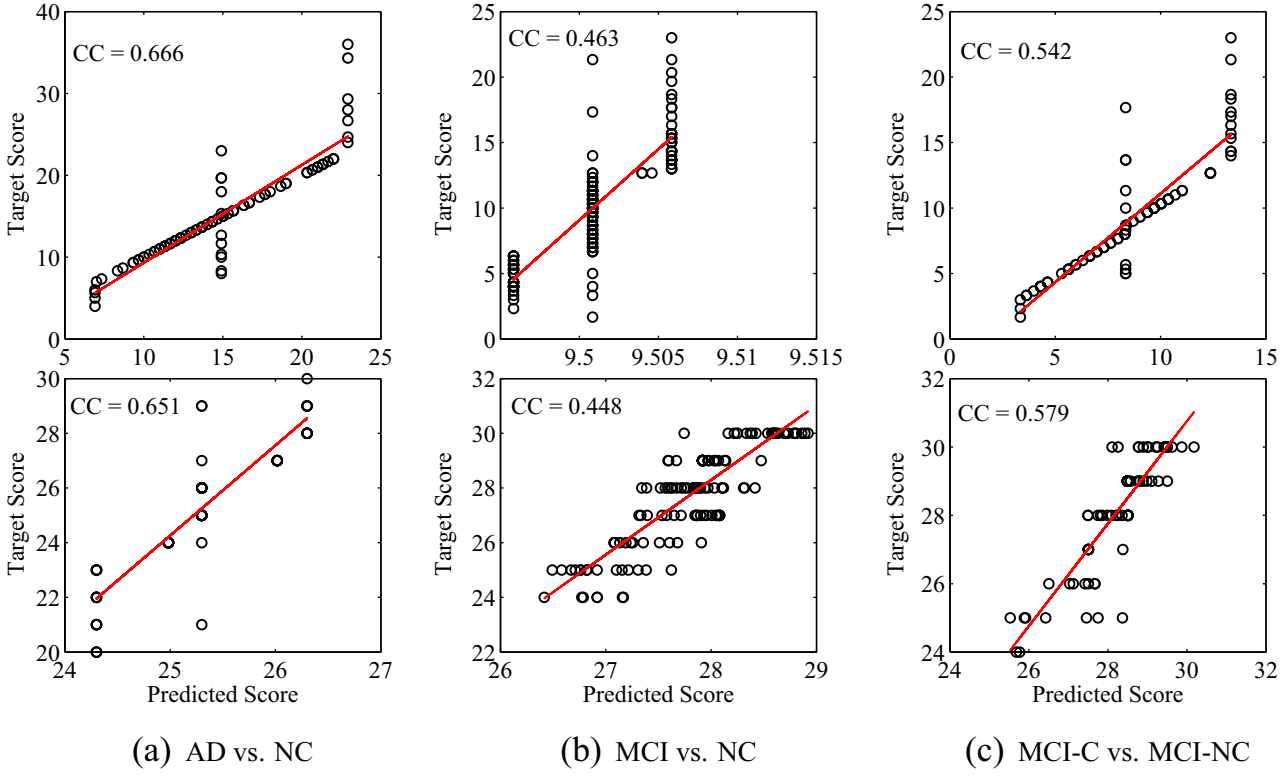


Fig. 8. Scatter plots and the respective Correlation Coefficients (CCs) obtained by the proposed method on the MP data (top: ADAS-Cog, bottom: MMSE).

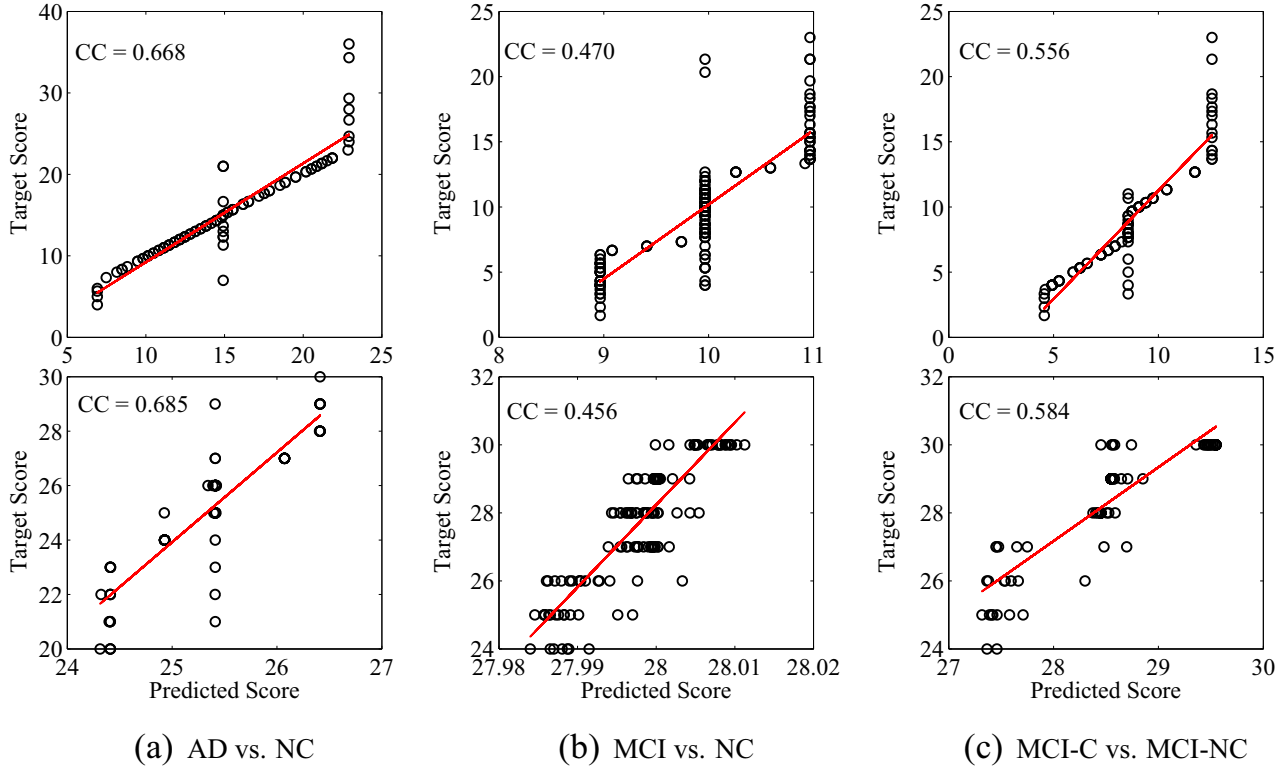


Fig. 9. Scatter plots and the respective Correlation Coefficients (CCs) obtained by the proposed method on the MPC data (top: ADAS-Cog, bottom: MMSE).

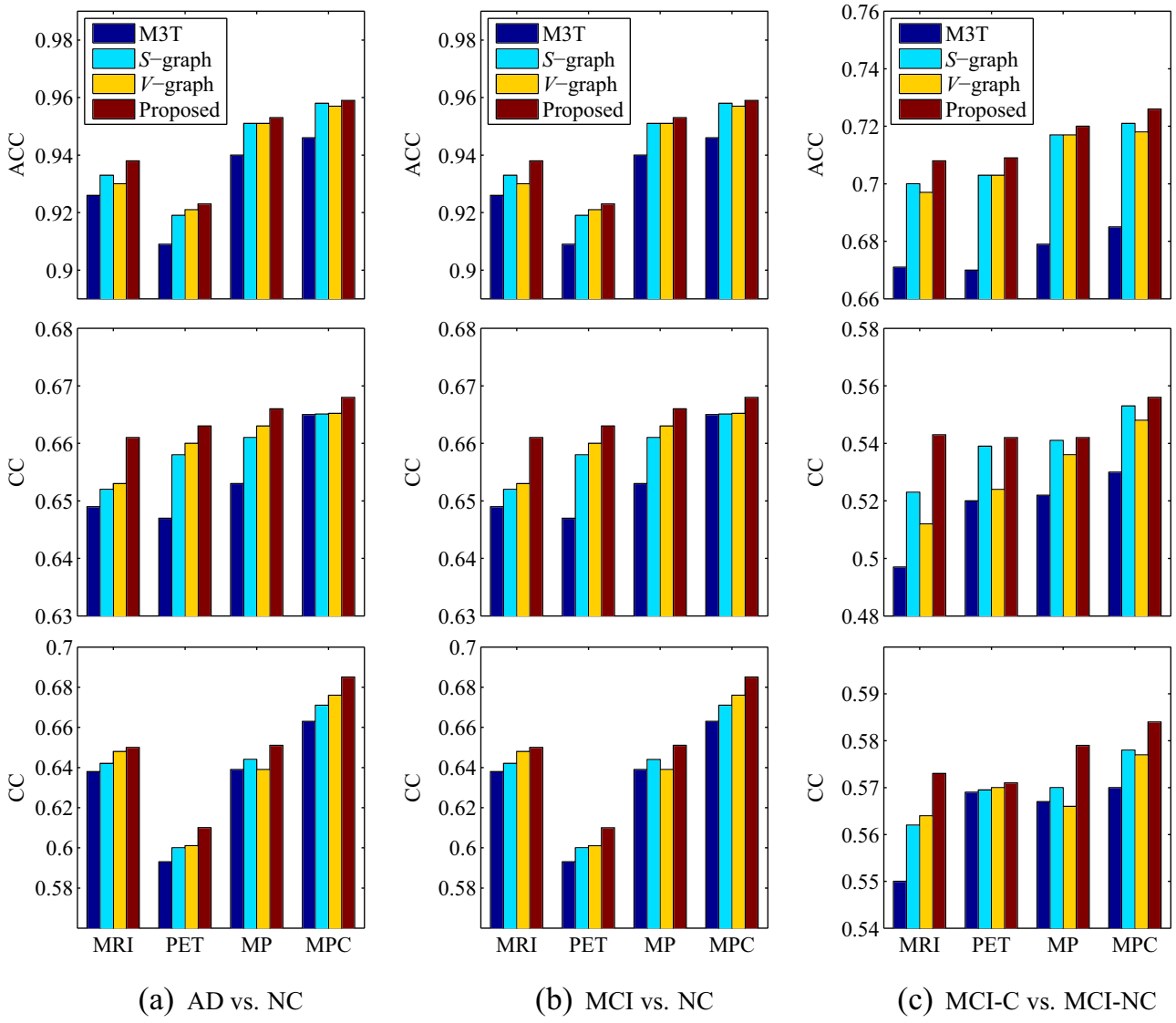


Fig. 10. Comparison of ACCuracy (ACC) (top), Correlation Coefficient (CC) of ADAS-Cog (middle), and CC of MMSE (bottom) among three graph based methods and also M3T.

dataset, we validated the effectiveness of the proposed method by showing the performance enhancements of both the clinical scores (ADAS-Cog and MMSE) prediction and the class label identification, outperforming the state-of-the-art methods. In the future work, we will extend the proposed framework to the problem of incomplete data, which often occurs in clinical trails and longitudinal follow-up studies.

Acknowledgments

Many thanks for the constructive advice from Feng Liu, Guan Yu, and Kim-Han Thung. This study was supported by National Institutes of Health (EB006733, EB008374, EB009634, AG041721, AG042599, and MH100217). Xiaofeng Zhu was partly supported by the National Natural Science Foundation of China under grant 61263035.

Appendix A

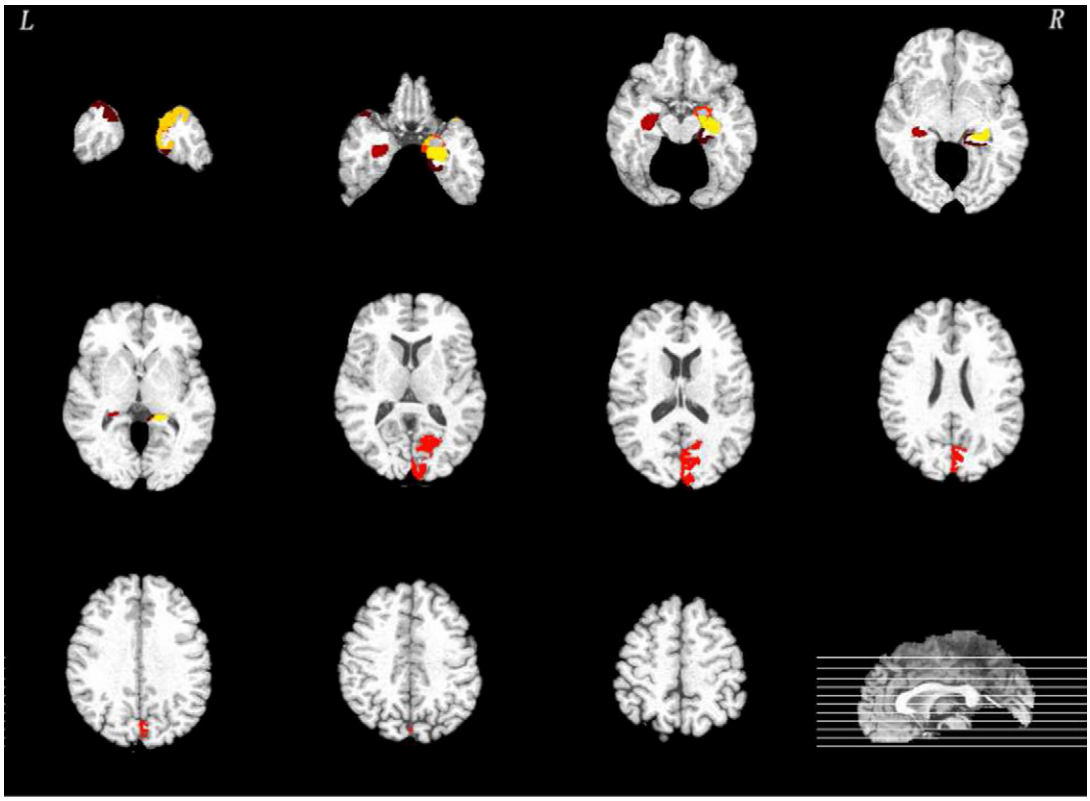
We prove that the proposed Algorithm 1 makes the value of the objective function in Eq. (12) monotonically decrease. We first give a Lemma from (Zhu et al., 2013a,2013b, 2013c) as follows, which will be used in our proof.

Lemma 1. *For any nonzero row vectors* $(\mathbf{w}(t))^i \in \mathbb{R}^c$ *and* $(\mathbf{w}(t+1))^i \in \mathbb{R}^c$, $i = 1, \ldots, d$, *and* t *denotes an iteration index, the following holds*:

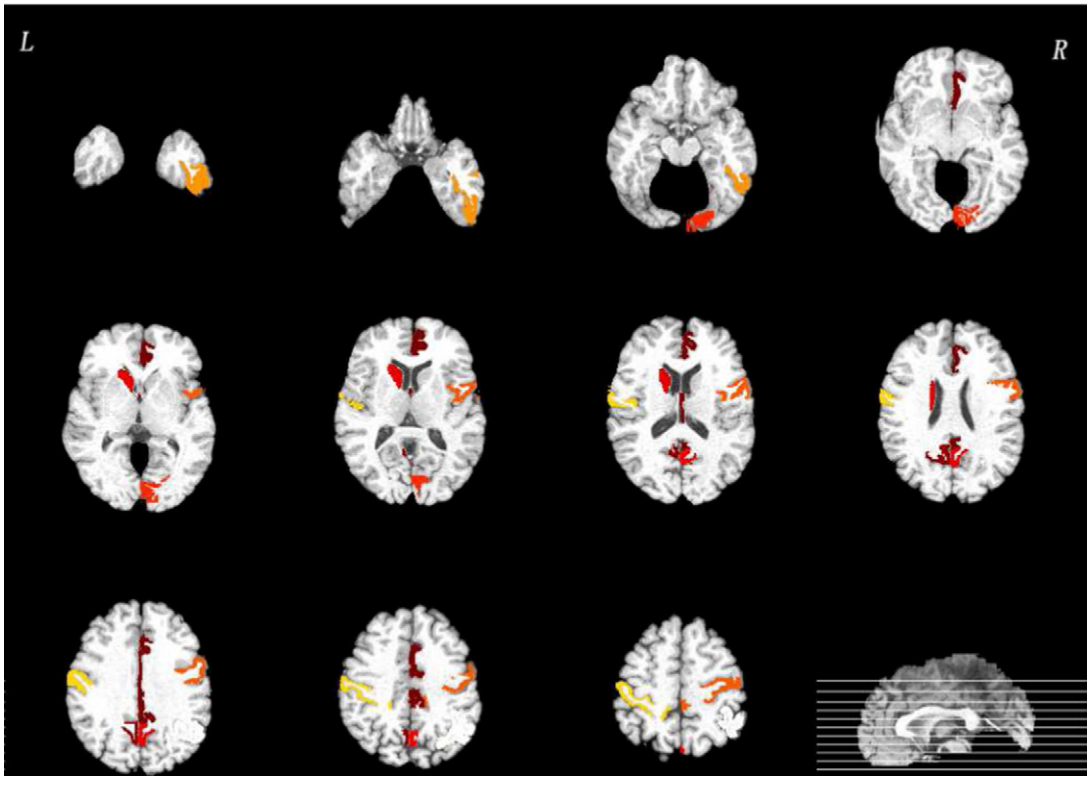
$$\sum_{i=1}^{d}\left(\left(\frac{\left\|(\mathbf{w}(t+1))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2}-\left\|(\mathbf{w}(t+1))^i\right\|_2\right)-\left(\frac{\left\|(\mathbf{w}(t))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2}-\|(\mathbf{w}(t))^i\|_2\right)\right)\geq 0. \tag{A.1}$$

Theorem 1. *In each iteration, Algorithm 1 monotonically decreases the objective function value in Eq. (12).*

Proof. In Algorithm 1, we denote that part of Eq. (12), i.e., without the last term $\lambda\|\mathbf{W}\|_{2,1}$, in the t-th iteration as $\mathscr{L}(t) = \|\mathbf{Y} - (\mathbf{W}(t))^T\mathbf{X}\|_F^2 + \alpha_1 tr(2(\mathbf{W}(t))^T\mathbf{X}\mathbf{H}_n\mathbf{X}^T\mathbf{W}(t) - 4\mathbf{Y}\mathbf{H}_n\mathbf{X}^T\mathbf{W}(t)) + \alpha_2 tr(2\mathbf{X}^T\mathbf{W}(t)\mathbf{H}_c(\mathbf{W}(t))^T\mathbf{X} - 4\mathbf{X}^T\mathbf{W}(t)\mathbf{H}_c\mathbf{Y})$. We also denote $\mathbf{Q}(t)$ as the optimal value in the t-th iteration for $\mathbf{Q}$. According to Zhu et al. (2013a, 2013b,2013c), optimizing



(a) MRI



(b) PET

Fig. 11. Top 10 selected MRI/PET regions in the MCI classification with MPC. The brain regions were color-coded. Moreover, different colors indicate different brain regions.

Table 5
The six brain regions selected by the competing methods. 'Y/N' denotes, respectively, whether a brain region was ranked within the top 10 or not; For the cases of 'N', we reported its ranking in the parentheses with boldface.

Regions	MPC-S	M2TFS-C	M2TFS-K	HOGM	M3T
Parahippocampal gyrus left	Y	Y	Y	**N(11)**	Y
Hippocampal formation right	**N(15)**	Y	Y	Y	Y
Temporal pole left	Y	Y	Y	Y	Y
Entorhinal cortex left	Y	Y	Y	Y	Y
Hippocampal formation left	**N(18)**	Y	Y	Y	Y
Amygdala right	Y	Y	Y	Y	Y

the non-smooth convex form $\|\mathbf{W}\|_{2,1}$ can be transferred to iteratively optimize $\mathbf{Q}$ and $\mathbf{W}$ in $tr(fW^T\mathbf{Q}\mathbf{W})$. Therefore, according to the 3-rd step of Algorithm 1, we have

$$\mathscr{S}(t+1)+\lambda tr\Big((\mathbf{W}(t+1))^T\mathbf{Q}(t)\mathbf{W}(t+1)\Big)\le\mathscr{S}(t)+\lambda tr\Big((\mathbf{W}(t))^T\mathbf{Q}(t)\mathbf{W}(t)\Big). \tag{A.2}$$

By changing the trace form into the form of summation, we have

$$\mathscr{S}(t+1)+\lambda\sum_{i=1}^{d}\frac{\left\|(\mathbf{w}(t+1))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2}\le\mathscr{S}(t)+\lambda\sum_{i=1}^{d}\frac{\left\|(\mathbf{w}(t))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2}. \tag{A.3}$$

By simple modification, we can have

$$\begin{aligned}\mathscr{S}(t+1)+\lambda\sum_{i=1}^{d}\left(\frac{\left\|(\mathbf{w}(t+1))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2}-\left\|(\mathbf{w}(t+1))^i\right\|_2+\left\|(\mathbf{w}(t+1))^i\right\|_2\right)\\ \le\mathscr{S}(t)+\lambda\sum_{i=1}^{d}\left(\frac{\left\|(\mathbf{w}(t))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2-\left\|(\mathbf{w}(t))^i\right\|_2+\left\|(\mathbf{w}(t))^i\right\|_2}\right).\end{aligned} \tag{A.4}$$

After reorganizing terms, we finally have

$$\begin{aligned}\mathscr{S}(t+1)+\lambda\sum_{i=1}^{d}\left\|(\mathbf{w}(t+1))^i\right\|_2+\lambda\sum_{i=1}^{d}\left(\left(\frac{\left\|(\mathbf{w}(t+1))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2}6-\left\|(\mathbf{w}(t+1))^i\right\|_2\right)\right.\\ \left.-\left(\frac{\left\|(\mathbf{w}(t))^i\right\|_2^2}{2\left\|(\mathbf{w}(t))^i\right\|_2-\left\|(\mathbf{w}(t))^i\right\|_2}\right)\right)\le\mathscr{S}(t)+\lambda\sum_{i=1}^{d}\left\|(\mathbf{w}(t))^i\right\|_2.\end{aligned} \tag{A.5}$$

According to Lemma 1, the third term of the left side in Eq. (A.5) is non-negative. Therefore, the following inequality holds

$$\mathscr{S}(t+1)+\lambda\sum_{i=1}^{d}\left\|(\mathbf{w}(t+1))^i\right\|_2\le\mathscr{S}(t)+\lambda\sum_{i=1}^{d}\left\|(\mathbf{w}(t))^i\right\|_2. \tag{A.6}$$

□

References

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, M.H., 2007. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement. 3 (3), 186–191.

Buchhave, P., Blennow, K., Zetterberg, H., Stomrud, E., Londos, E., Andreasen, N., Minthon, L., Hansson, O., 2009. Longitudinal study of CSF biomarkers in patients with Alzheimer's disease. PLoS One 4 (7), 62–94.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. 2 (2), 121–167.

Cheng, B., Zhang, D., Chen, S., Kaufer, D., Shen, D., 2013. Semi-supervised multimodal relevance vector regression improves cognitive performance estimation from imaging and biological biomarkers. Neuroinformatics 11 (3), 339–353.

Chételat, G., Eustache, F., Viader, F., Sayette, V.D.L., Pélerin, A., Mézenge, F., Hannequin, D., Dupuy, B., Baron, J.-C., Desgranges, B., 2005. FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. Neurocase 11 (1), 14–25.

Cho, Y., Seong, J.-K., Jeong, Y., Shin, S.Y., 2012. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. NeuroImage 59 (3), 2217–2230.

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., 2012. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. NeuroImage 60 (1), 59–70.

Convit, A., De Asis, J., De Leon, M., Tarshish, C., De Santi, S., Rusinek, H., 2000. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. Neurobiol. Aging 21 (1), 19–26.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56 (2), 766–781.

De Leon, M., Mosconi, L., Li, J., De Santi, S., Yao, Y., Tsui, W., Pirraglia, E., Rich, K., Javier, E., Brys, M., Glodzik, L., Switalski, R., Saint Louis, L., Pratico, D., 2007. Longitudinal CSF isoprostane and MRI atrophy in the progression to AD. J. Neurol. 254 (12), 1666–1675.

Du, A.-T.T., Schuff, N., Kramer, J.H., Rosen, H.J., Gorno-Tempini, M.L.L., Rankin, K., Miller, B.L., Weiner, M.W., 2007. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130, 1159–1166.

Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., Frisoni, G.B., 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. NeuroImage 47 (4), 1363–1370.

Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B.B., Gee, J.C., Wang, J., Shen, D., 2007. Multivariate examination of brain abnormality using both structural and functional MRI. NeuroImage 36 (4), 1189–1199.

Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., 2010. The Alzheimer's Disease Neuroimaging Initiative, 2010. CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. J. Neurosci. 30 (6), 2088–2101.

Fox, N.C., Schott, J.M., 2004. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. Lancet 363 (9406), 392–394.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. NeuroImage 50 (3), 883–892.

Greicius, M.D., Srivastava, G., Reiss, A.L., 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 (13), 4637–4642.

Guo, X., Wang, Z., Li, K., Li, Z., Qi, Z., Jin, Z., Yao, L., Chen, K., 2010. Voxel-based assessment of gray and white matter volumes in Alzheimer's disease. Neurosci. Lett. 468 (2), 146–150.

Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K., Minthon, L., 2006. Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. Lancet Neurol. 5 (3), 228–234.

He, X., Cai, D., Niyogi, P., 2005. Laplacian score for feature selection. NIPS, pp. 1–8.

Jia, H., Wu, G., Wang, Q., Shen, D., 2010. ABSORB: atlas building by self-organized registration and bundling. NeuroImage 51 (3), 1057–1070.

Jie, B., Zhang, D., Cheng, B., Shen, D., 2013. Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease. MICCAI, pp. 9–16.

Kabani, N.J., 1998. 3D anatomical atlas of the human brain. NeuroImage 7, 0700–0717.

Lemoine, B., Rayburn, S., Benton, R., 2010. Data fusion and feature selection for Alzheimer's diagnosis. Brain Informatics, pp. 320–327.

Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., Shen, D., 2012. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. Neurobiol. Aging 33 (2), 427.e15–427.e30.

Liu, M., Zhang, D., Shen, D., 2012. Ensemble sparse classification of Alzheimer's disease. NeuroImage 60 (2), 1106–1116.

Liu, F., Suk, H.-I., Wee, C.-Y., Chen, H., Shen, D., 2013. High-order graph matching based feature selection for Alzheimer's disease identification. MICCAI, pp. 311–318.

Liu, F., Wee, C.-Y., Chen, H., Shen, D., 2014. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. vol. 84, 466–475.

McEvoy, L.K., Fennema-Notestine, C., Roddey, J.C., Hagler, D.J., Holland, D., Karow, D.S., Pung, C.J., Brewer, J.B., Dale, A.M., 2009. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. Radiology 251 (5), 195–205.

Misra, C., Fan, Y., Davatzikos, C., 2009. 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 44 (4), 1415–1422.

Morris, J., Storandt, M., Miller, J., et al., 2001. Mild cognitive impairment represents early-stage Alzheimer disease. Arch. Neurol. 58 (3), 397–405.

Qiao, H., Zhang, H., Zheng, Y., Ponde, D.E., Shen, D., Gao, F., Bakken, A.B., Schmitz, A., Kung, H.F., Ferrari, V.A., et al., 2009. Embryonic stem cell grafting in normal and infarcted myocardium: serial assessment with MR imaging and PET dual detection1. Radiology 250 (3), 821–829.

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

Salas-Gonzalez, D., Górriz, J., Ramírez, J., Illán, I., López, M., Segovia, F., Chaves, R., Padilla, P., Puntonet, C., et al., 2010. Feature selection using factor analysis for Alzheimer's diagnosis using F18-FDG PET images. Med. Phys. 37 (11), 6084–6095.

Santi, S.D., de Leon, M.J., Rusinek, H., Convit, A., Tarshish, C.Y., Roche, A., Tsui, W.H., Kandil, E., Boppana, M., Daisley, K., Wang, G.J., Schlyer, D., Fowler, J., 2001. Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiol. Aging 22 (4), 529–539.

Seppälä, T.T., Koivisto, A.M., Hartikainen, P., Helisalmi, S., Soininen, H., Herukka, K., 2011. Longitudinal changes of CSF biomarkers in Alzheimer's disease. J. Alzheimers Dis. 25 (4), 583–594.

Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21 (11), 1421–1439.
Shen, D., Davatzikos, C., 2004. Measuring temporal morphological changes robustly in brain mr images via 4-dimensional template warping. NeuroImage 21 (4), 1508–1517.
Shen, D., Wong, W.-h., Ip, H.H., 1999. Affine-invariant image retrieval by correspondence matching of shapes. Image Vis. Comput. 17 (7), 489–499.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.
Stonnington, C.M., Chu, C., Klöppel, S., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S., 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage 51 (4), 1405–1413.
Suk, H.-I., Lee, S.-W., 2013. A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces. IEEE Trans. Pattern Anal. Mach. Intell. 35 (2), 286–299.
Suk, H.-I., Shen, D., 2013. Deep learning-based feature representation for AD/MCI classification. MICCAI, pp. 583–590.
Suk, H.-I., Wee, C.-Y., Shen, D., 2013. Discriminative group sparse representation for mild cognitive impairment classification. MLMI, pp. 131–138.
Walhovd, K., Fjell, A., Dale, A., McEvoy, L., Brewer, J., Karow, D., Salmon, D., Fennema-Notestine, C., 2010. Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobiol. Aging 31 (7), 1107–1121.
Wang, Y., Fan, Y., Bhatt, P., Davatzikos, C., 2010. High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. NeuroImage 50 (4), 1519–1535.
Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., Shen, L., 2011. Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. MICCAI, pp. 115–123.
Wee, C.-Y., Yap, P.-T., Li, W., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2011. Enriched white matter connectivity networks for accurate identification of MCI patients. NeuroImage 54 (3), 1812–1822.
Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K. A., Wang, L., Shen, D., 2012. Identification of MCI individuals using structural and functional connectivity networks. NeuroImage 59 (3), 2045–2056.
Weinberger, K.Q., Sha, F., Saul, L.K., 2004. Learning a kernel matrix for nonlinear dimensionality reduction. ICML, pp. 17–24.
Yang, J., Shen, D., Davatzikos, C., Verma, R., 2008. Diffusion tensor image registration using tensor geometry and orientation features. Medical Image Computing and Computer-Assisted Intervention–MICCAI, pp. 905–913.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B 68 (1), 49–67.
Zacharaki, E.I., Shen, D., koo Lee, S., Davatzikos, C., 2008. ORBIT: a multiresolution framework for deformable registration of brain tumor images. IEEE Trans. Med. Imaging 27 (8), 1003–1017.
Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59 (2), 895–907.
Zhang, Y., Brady, M., 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 (1), 45–57.
Zhang, D., Shen, D., et al., 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7 (3), e33182.
Zhou, L., Wang, Y., Li, Y., Yap, P.-T., Shen, D., et al., 2011. Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. PLoS One 6 (7), e21935.
Zhu, X., Huang, Z., Shen, H.T., Cheng, J., Xu, C., 2012. Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recogn. 45 (8), 3003–3016.
Zhu, X., Huang, Z., Cui, J., Shen, T., 2013a. Video-to-shot tag propagation by graph sparse group lasso. IEEE Trans. Multimedia 13 (3), 633–646.
Zhu, X., Huang, Z., Yang, Y., Shen, H.T., Xu, C., Luo, J., 2013b. Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recogn. 46 (1), 215–229.
Zhu, X., Wu, X., Ding, W., Zhang, S., 2013c. Feature selection by joint graph sparse coding. SDM, pp. 803–811.