Information Fusion 66 (2021) 54–63

Contents lists available at ScienceDirect

Information Fusion

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

Heterogeneous data fusion for predicting mild cognitive impairment conversion[☆]

Heng Tao Shen[a], Xiaofeng Zhu[a,*], Zheng Zhang[b,f,*], Shui-Hua Wang[c,*], Yi Chen[d], Xing Xu[a], Jie Shao[a,e]

[a] School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518055, China
[c] School of Mathematics and Actuarial Science, University of Leicester, LE1 7RH, UK
[d] School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China
[e] Sichuan Artificial Intelligence Research Institute, Yibin, 644000, China
[f] Department of Computer and Information Science, University of Macau, Macau, 999078, China

ARTICLE INFO

Keywords:
Mild cognitive impairment
Transfer learning
Feature selection
Sparse learning
Alzheimer's disease

ABSTRACT

In the clinical study of Alzheimer's Disease (AD) with neuroimaging data, it is challenging to identify the progressive Mild Cognitive Impairment (pMCI) subjects from the stableMCI (sMCI) subjects (*i.e.,* the pMCI/sMCI classification) in an individual level because of small inter-group differences between two groups (*i.e.,* pMCIs and sMCIs) as well as high intra-group variations within each group. Moreover, there are a very limited number of subjects available, which cannot guarantee to find informative and discriminative patterns for achieving high diagnostic accuracy. In this paper, we propose a novel sparse regression method to fuse the auxiliary data into the predictor data for the pMCI/sMCI classification, where the predictor data is structural Magnetic Resonance Imaging (MRI) information of both pMCI and sMCI subjects and the auxiliary data includes the ages of the subjects, the Positron Emission Tomography (PET) information of the predictor data, and the structural MRI information of AD and Normal Controls (NC). Specifically, we incorporate the auxiliary data and the predictor data into a unified framework to jointly achieve the following objectives: i) *jointly selecting* informative features from both the auxiliary data and the predictor data; ii) *robust to outliers* from both the auxiliary data and the predictor data; and iii) *reducing the aging effect* due to the possible cause of brain atrophy induced by both the normal aging and the disease progression. As a result, our proposed method jointly selects the useful features from the auxiliary data and the predictor data by taking into account the influence of outliers and the age of the two kinds of data, *i.e.,* the pMCI and sMCI subjects as well as the AD and NC subjects. We further employ the linear Support Vector Machine (SVM) with the selected features of the predictor data to conduct the pMCI/sMCI classification. Experimental results on the public data of Alzheimer's Disease Neuroimaging Initiative (ADNI) show the proposed method achieved the best classification performance, compared to the best comparison method, in terms of four evaluation metrics.

1. Introduction and background

With the dramatic development and prevalence of Alzheimer's Disease (AD), it becomes vital to identify AD pathology at its early stage or even before its onset because the neuropathological progression in AD may begin many years before the appearance of clinical symptoms [1–3]. Many studies employed neuroimaging data such as structural Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) to conduct early disease diagnosis of AD [4–6]. For example, Khedher et al. presented a computer aided system which uses partial least squares and principle component analysis to conduct classification among the data such as AD, Mild Cognitive Impairment (MCI) and Normal Control (NC) [7], while Ortiz et al. investigated the ensemble of two deep learning architectures to discriminate AD from NC [8]. Zhu et al. proposed a joint regression and classification model using structural MRI and PET data [9].

[☆] This work was partially supported by the National Key Research and Development Program of China (Grant No. 2018AAA0102200), the National Natural Science Foundation of China (Grants No: 61632007 and 61876046), Key Project of Shenzhen Municipal Technology Research (Grant No. JSGG20200103103401723), and the Sichuan Science and Technology Program (Grants No: 2018GZDZX0032 and 2019YFG0535).

[*] Corresponding authors.
E-mail addresses: seanzhuxf@gmail.com (X. Zhu), darrenzz219@gmail.com (Z. Zhang), shuihuawang@ieee.org (S.-H. Wang).

https://doi.org/10.1016/j.inffus.2020.08.023
Received 21 April 2020; Received in revised form 19 July 2020; Accepted 29 August 2020
Available online 4 September 2020

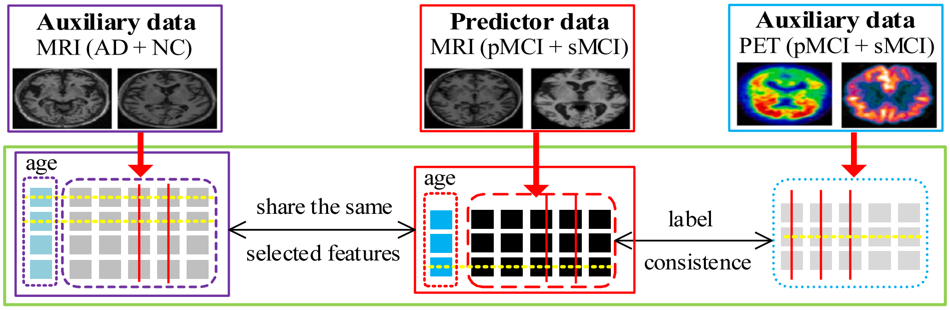


Fig. 1. An illustration of the proposed method. The boxes with purple solid lines, red solid lines, and blue solid lines, respectively, represent the original MRI images of ADs and NCs, the original MRI images of pMCIs and sMCIs, and the original PET images of pMCIs and sMCI. The boxes with purple dot lines, red dot lines, and blue dot lines, respectively, represent the MRI feature of ADs and NCs, the MRI features of pMCIs and sMCIs, and the PET features of pMCIs and sMCI. The green box indicates the combination process of the auxiliary data and the predictor data using our proposed method. Moreover, the vertical red solid line denotes the removal of uninformative features and the horizontal yellow dot line indicates robustness against outlier subjects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Mild Cognitive Impairment (MCI) is the symptomatic predementia stage of AD, characterized by cognitive and functional impairment not severe enough to fulfill the criteria for dementia [10,11]. The MCI possibly progressing to AD over times is called the progressive MCI (pMCI) and the MCI remaining stable for a long time is called the stable MCI (sMCI). In the early diagnosis of AD, it perhaps is the most interesting problem to distinguish pMCI subjects from sMCI subjects, *i.e.,* the pMCI/sMCI classification, due to that (1) i n the pathological spectrum of AD, MCI may be the optimal stage that clinical treatments or interventions can be effectively administered to prevent or delay decline to severe dementia; (2) 10% to 15% of MCI population will progress to AD annually [12,13]; and (3) the intervention treatment is more effective before the patients progress to AD [14–16].

However, due to the subtle pathological changes (*i.e.,* brain atrophy) which may be masked by the normal aging effect and/or the inter-subject variations [17,18], it is very challenging to discriminate pMCI from sMCI. First, the pMCI has minor inter-group difference, compared to the sMCI so that many previous studies integrated the pMCI with the sMCI as a single category, *i.e.,* MCI [19]. Second, there is high intra-group variations for either the pMCI subjects or the sMCI subjects [20]. That is, different subjects in the same group (*i.e.,* either the pMCIs or the sMCIs) have high intra-group variations, making difficult construct classification models. Third, the number of MCI subjects is small, but the feature number is usually high. As a result, the pMCI/sMCI classification often suffers from the issues of small-sized sample and high-dimensional data [21–23] to easily result in the problem of curse of dimensionality [20,24]. Moreover, previous classification models are affected by redundant features and subject-level noise [21,25]. Hence, it is very vital to investigate informative and discriminative patterns to address above issues.

Recent studies have proposed a number of pattern recognition methods to improve the performance of the pMCI/sMCI classification by addressing one or more of the above issues [26]. For example, Zhu et al. proposed a Bayesian method to detect the association between structural MRI and genetic data by taking into account the ages of the subjects [27], assuming that brain atrophy is influenced by both the normal aging and the disease [17]. Wang et al. first demonstrated that the AD/NC classification is similar to the pMCI/sMCI classification, considering that the pMCI is like to the AD and the sMCI is like to the NC, and then employed the AD data and NC data for the pMCI/sMCI classification [28]. With the above assumption, the studies [25,29,30] designed transfer learning techniques for the pMCI/sMCI classification. Previous techniques share the common characteristics as follows, *i.e.,* using auxiliary information to improve the classification performance of the predictor data. By regarding the MRI information of both pMCI and sMCI subjects as the predictor data, the auxiliary data can be the information, such as the ages of the subjects, the MRI information of the AD and NC subjects, the PET data, and the genetic information. In practice, each kind of auxiliary information could be heterogeneous to others [31]. Hence, using these heterogeneous information together for the pMCI/sMCI classification is complex and challengeable.

In this paper, we extend our conference version in [20] to use MRI features as the predictor data to identify whether a MCI subject progresses to AD (*i.e.,* pMCI) or not (*i.e.,* sMCI) within 24 months. However, unlike the existing approaches that mostly utilized the predictor data only, we further exploit source information of a subject's age and available PET features. Specifically, we formulate a novel sparse regression model that jointly uses the auxiliary data and the predictor data for feature selection, so that the useful knowledge of the auxiliary data can be transferred to the predictor data. There are three key factors strengthening our method, (i) (*feature selection*) we use the auxiliary data to select informative features in the predictor data; (ii) (*outlier robustness*) our formulation is robust to outliers from both the auxiliary data and the predictor data; and (iii) (*aging effect removal*) we also include a subject's age as one of the features in each subject to learn its relationship with the neuroimaging features in the prediction model.

Different from previous studies of the pMCI/sMCI classification, the contributions of our method are twofold. First, we consider three kinds of auxiliary information, *i.e.,* the age, the MRI data of the AD and NC subjects, and the PET data of the pMCI and sMCI subjects. The auxiliary information is jointly involved in selecting features, which discriminates our method from the existing methods [17,29,32,33] that only used a part of them. We argue that the auxiliary information can be complementary and related to each other in some ways, and thus utilizing such information separately and sequentially is inevitably limited. Our method, in contrast, integrates all auxiliary information in a unified framework. Second, we conduct the pMCI/sMCI classification with the consideration of outlier influence on both the auxiliary data and the predictor data, to select useful features jointly and robustly. In contrast, Moradi et al. used the auxiliary data for feature selection only [17].

2. Method

We denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively, followed by denoting the Frobenius norm and the $\ell_{2,1}$-norm, respectively, of a matrix $\mathbf{X}$, as $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|_{2,1}$. We also denote the transpose operator, the trace operator, the rank, and the inverse of $\mathbf{X}$ as $\mathbf{X}^\top$, $tr(\mathbf{X})$, $rank(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively. We further denote α, β, and λ_i $(i = 1, \dots, 5)$ as non-negative parameters.

We denote the MRI feature matrix, the PET feature matrix, and the label matrix, respectively, for n_t subjects of the pMCIs and the sMCIs, as $\mathbf{X}_t \in \mathbb{R}^{n_t \times d}$, $\mathbf{X}_p \in \mathbb{R}^{n_t \times d}$, and $\mathbf{Y}_t \in \{0,1\}^{n_t \times c_t}$, where d denotes the feature dimension and c_t is the number of classes of the predictor data. We also denote the MRI feature matrix and the label matrix, of n_a subjects of the ADs and the NCs, respectively, as $\mathbf{X}_a \in \mathbb{R}^{n_a \times d}$ and $\mathbf{Y}_a \in \{0,1\}^{n_a \times c_a}$, where c_a is the number of classes of the auxiliary data. We further denote the age factors of the predictor data and the auxiliary data, as $\mathbf{x}_{tg} \in \mathbb{R}^{n_t}$ and $\mathbf{x}_{ag} \in \mathbb{R}^{n_a}$, respectively. In this paper, we extract Region-of-Interests (ROIs) based features for all neuroimaging data to conduct binary classification, *i.e.,* $c_t = c_a = 2$, which is straightforward to be extended to the multi-class classification problem.

2.1. Framework overview

In this paper, we propose to predict a subject to be either pMCI or sMCI based on MRI features, whose schematic illustration is shown in Fig. 1. Specifically, we devise a novel sparse regression method by taking MRI features of predictor group subjects (*i.e.,* pMCIs and sMCIs) as the predictors for prediction. Moreover,we further utilize PET features of the non-predictor group subjects and MRI features of the non-predictor group subjects (*i.e.,* ADs and NCs), and use subject's age to regress out the normal aging effect. Since the MRI features are used as predictors in our model, we regard other information including the ages of the subjects, PET features, and MRI features of the non-predictor group as 'auxiliary' data. Specifically, we use the auxiliary MRI data of AD and NC subjects in selecting useful features with the proposed joint feature selection formulation, while using the auxiliary PET data of pMCI and sMCI subjects to enhance the feature weight learning of the predictor data with the constraint of the consistency of label prediction. We further regard the ages of subjects as an additional feature to lessen the normal aging effect in the diagnosis.

2.2. Feature selection on predictor data

Given the predictor data $\mathbf{X}_t$ and its corresponding label matrix $\mathbf{Y}_t$, a robust sparse regression method linearly estimates a coefficient matrix $\mathbf{W}_t \in \mathbb{R}^{d\times c_t}$ by optimizing the following objective function:

$$\min_{\mathbf{W}_t} \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \lambda_1\|\mathbf{W}_t\|_{2,1} \tag{1}$$

The $\ell_{2,1}$-norm loss function, *i.e.,* a robust loss function in first term of Eq. (1), makes Eq. (1) robust against the subject-level outliers [34–36]. Specifically, each row of $(\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t)$ in Eq. (1) corresponds to the prediction residual of one subject. Under the $\ell_{2,1}$-norm operation, the residual value of each row (*i.e.,* subject) is combined via ℓ_2-norm, *i.e.,* the square root of the sum of the squares, and thus are less affected by the outliers, compared to the least square loss function [37,38]. The $\ell_{2,1}$-norm regularization term on $\mathbf{W}_t$ penalizes $\mathbf{W}_t$ by encouraging the row sparsity, *i.e.,* all elements of some rows of $\mathbf{W}_t$ are all zeros, to select the corresponding features in $\mathbf{X}_t$ [39,40].

2.3. Feature selection on predictor and auxiliary data

Using Eq. (1) directly on the predictor data (*i.e.,* the MRI features of the pMCI and sMCI subjects) for the MCI conversion classification could still be ineffective due to the limited training data. To circumvent the lack of training samples, recent studies [17,33,41,42] exploited auxiliary information from non-predictor groups, *e.g.,* AD and NC subjects. The rationale of using such auxiliary data is that in terms of the AD pathological spectrum, *i.e.,* the sMCI is closer to the NC while the pMCI is closer to the AD. Thus, the features that are informative for the AD/NC classification could be also useful for the pMCI/sMCI classification [41,42]. In this paper, we also utilize such auxiliary data for feature selection. However, unlike previous methods [33,42] that mostly first learned a classification model over only the auxiliary data and then transferred the learned model to build a target-oriented model, we devise a novel sparse feature selection model that jointly exploits both the predictor data and the auxiliary data.

With the assumption that MRI features selected for the AD/NC classification could be also informative for the pMCI/sMCI classification, we propose to use MRI features of AD and NC subjects (*i.e.,* auxiliary data $\mathbf{X}_a$), to help in selecting MRI features of pMCI and sMCI subjects as follows:

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{W}_a} \quad & \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \lambda_1\|\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a\|_{2,1} \\ & + \lambda_2\|[\mathbf{W}_t,\mathbf{W}_a]\|_{2,1} + \lambda_3\|[\mathbf{W}_t,\mathbf{W}_a]\|_F^2\end{aligned} \tag{2}$$

where $\mathbf{W}_a \in \mathbb{R}^{d\times c_a}$ is a coefficient matrix for the auxiliary data. The reason to use $\ell_{2,1}$-norm on the loss function of auxiliary data (*i.e.,* the second term in Eq. (2)) is similar to Eq. (1), *i.e.,* for robustness to outliers. The Frobenius norm on $[\mathbf{W}_t\ \mathbf{W}_a]$ in the fourth term of Eq. (2) is used to provide a group effect, which tends to select highly correlated features together, by countering for some weaknesses of the sparsity constraint [43,44]. The $\ell_{2,1}$-norm regularizer on $[\mathbf{W}_t, \mathbf{W}_a] \in \mathbb{R}^{d\times(c_t+c_a)}$ encourages the row-wise joint sparsity [37,38]. This sparsity constraint encourages the same set of features to be selected for both $\mathbf{X}_t$ and $\mathbf{X}_a$ (recall that $\mathbf{X}_t$ and $\mathbf{X}_a$ denote the feature matrix for the predictor and auxiliary data, respectively). With the sparsity regularization term $\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1}$, the useful features are kept by satisfying the AD/NC classification constraint (via $\mathbf{W}_a$) and the AD/NC classification (via $\mathbf{W}_t$), simultaneously. The jointly learned model is more robust than the individual models of either only satisfying the pMCI/sMCI classification constraint (via $\mathbf{W}_a$) [17,41] which does not consider the pathological difference in the pMCI and sMCI subjects, or only satisfying the pMCI/sMCI classification constraint (via $\mathbf{W}_t$) in [33,42] which has been reported to have limited performance due to the small number of subjects.

In the multiple-modality AD study, it has shown that the PET data and the MRI data could provide complementary information to each other [17,29,42]. In this paper, we use the PET data of the pMCI and sMCI subjects, *i.e.,* $\mathbf{X}_p$, as other kind of auxiliary data, to help learn the coefficient matrix $\mathbf{W}_t$ of the predictor data. More specifically, we constrain the predicted values from the PET data and the MRI data to be close to each other, as both modalities share the same label information. As a result, we have the following objective function

$$\min_{\mathbf{W}_t,\mathbf{W}_p} \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_4\|\mathbf{W}_p\|_{2,1} \tag{3}$$

where $\mathbf{W}_p \in \mathbb{R}^{d\times c_t}$ is a coefficient matrix to the PET data. $\mathbf{X}_t\mathbf{W}_t$ and $\mathbf{X}_p\mathbf{W}_p$ are the predictions of the label matrix using the MRI and PET data, respectively. Thus, their difference, measured by the summation of element-wise similarity, should be as small as possible. Combining Eq. (2) with Eq. (3), we obtain the following objective function, which learns $\mathbf{W}_t$ with the MRI data of AD/NC subjects and the PET data of pMCI/sMCI subjects,

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{W}_a,\mathbf{W}_p} \quad & \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \lambda_1\|\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a\|_{2,1} \\ & + \lambda_2\|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_3\|[\mathbf{W}_t,\mathbf{W}_a]\|_{2,1} \\ & + \lambda_4\|[\mathbf{W}_t,\mathbf{W}_a]\|_F^2 + \lambda_5\|\mathbf{W}_p\|_{2,1}.\end{aligned} \tag{4}$$

2.4. Aging effect removal

The studies (*e.g.,* [17,45]) showed that both the normal aging and the AD pathology contribute to brain atrophy and it is necessary to remove the aging effect to the brain atrophy before analysis. The first method designed for the aging effect removal fits a linear regression model between the features and the age of NC subjects to obtain a coefficient matrix [17,45]. This coefficient matrix denotes how the age affects the feature values. The second method directly fits the model by using both features and the age as covariates [17,27]. Actually, both of them assume that there is linear relationship among the labels, features and ages. Hence, we use the ages of the subjects as one feature in both the predictor data and the auxiliary data to have our final objective function as follows.

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag},\mathbf{W}_p} \quad & \|\mathbf{Y}_t - [\mathbf{X}_t,\mathbf{x}_{tg}][\mathbf{W}_t^\top,\mathbf{w}_{tg}^\top]^\top\|_{2,1} \\ & + \lambda_1\|\mathbf{Y}_a - [\mathbf{X}_a,\mathbf{x}_{ag}][\mathbf{W}_a^\top,\mathbf{w}_{ag}^\top]^\top\|_{2,1} \\ & + \lambda_2\|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} \\ & + \lambda_3\|[\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag}]\|_F^2 \\ & + \alpha\|[\mathbf{W}_t,\mathbf{W}_a]\|_{2,1} + \beta\|\mathbf{W}_p\|_{2,1}\end{aligned} \tag{5}$$

where $\mathbf{w}_{tg} \in \mathbb{R}^{1\times c_t}$ and $\mathbf{w}_{ag} \in \mathbb{R}^{1\times c_a}$ are coefficient matrices. In Eq. (5), the last two terms help select common useful features for the first two data fitting terms, while the third term imposes label prediction consistency between $\mathbf{X}_t$ and $\mathbf{X}_p$. In addition, the use of the $\ell_{2,1}$-norm loss function helps to learn $\mathbf{X}_t$, $\mathbf{X}_p$, and $\mathbf{X}_a$ by reducing the influence of outliers.

It is time-consuming to tune 5 parameters (*i.e.,* λ_1, λ_2, λ_3, α, and β) for the optimization of Eq. (5). To address this issue, we add a square root operator in the terms, *i.e.,* the 2nd term, the 3rd term, and the 4th term, in Eq. (5) and obtain our final objective function as follows

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag},\mathbf{W}_p} & \|\mathbf{Y}_t - [\mathbf{X}_t,\mathbf{x}_{tg}][\mathbf{W}_t^\top,\mathbf{w}_{tg}^\top]^\top\|_{2,1} \\ & + \sqrt{\|\mathbf{Y}_a - [\mathbf{X}_a,\mathbf{x}_{ag}][\mathbf{W}_a^\top,\mathbf{w}_{ag}^\top]^\top\|_{2,1}} \\ & + \sqrt{\|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1}} \\ & + \sqrt{\|[\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag}]\|_F^2} \\ & + \alpha\|[\mathbf{W}_t,\mathbf{W}_a]\|_{2,1} + \beta\|\mathbf{W}_p\|_{2,1}\end{aligned} \tag{6}$$

There is no explicit weight between the loss function and the regularization terms. The square root operators in Eq. (6) facilitate to learn implicit weights to balance the loss function and each regularization term. Specifically, the Lagrange function of Eq. (6) is

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag},\mathbf{W}_p} & \|\mathbf{Y}_t - [\mathbf{X}_t,\mathbf{x}_{tg}][\mathbf{W}_t^\top,\mathbf{w}_{tg}^\top]^\top\|_{2,1} \\ & + \sqrt{\|\mathbf{Y}_a - [\mathbf{X}_a,\mathbf{x}_{ag}][\mathbf{W}_a^\top,\mathbf{w}_{ag}^\top]^\top\|_{2,1}} \\ & + \sqrt{\|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1}} \\ & + \sqrt{\|[\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag}]\|_F^2} \\ & + \alpha\|[\mathbf{W}_t,\mathbf{W}_a]\|_{2,1} + \beta\|\mathbf{W}_p\|_{2,1} \\ & + \mathbb{G}(\boldsymbol{\Lambda},\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag},\mathbf{W}_p)\end{aligned} \tag{7}$$

where $\boldsymbol{\Lambda}$ is the Lagrange multiplier and $\mathbb{G}(\boldsymbol{\Lambda},\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag},\mathbf{W}_p)$ is the formalized term derived from the constraints. Taking the derivative of Eq. (7) *w.r.t.* $\mathbf{W}_p$ and setting the derivative to zero, we have

$$\lambda_2\frac{\partial\|\mathbf{X}_t\mathbf{W}_t-\mathbf{X}_p\mathbf{W}_p\|_{2,1}}{\partial\mathbf{W}_p} + \beta_2\frac{\partial\|\mathbf{W}_p\|_{2,1}}{\partial\mathbf{W}_p} + \frac{\partial\mathbb{G}(\boldsymbol{\Lambda},\mathbf{B},\mathbf{F})}{\partial\mathbf{F}} = 0, \tag{8}$$

where

$$\lambda_2 = \frac{1}{2\sqrt{\|\mathbf{X}_t\mathbf{W}_t-\mathbf{X}_p\mathbf{W}_p\|_{2,1}}}. \tag{9}$$

It is noteworthy that λ_2 in Eq. (9) is dependent on the variable of $\mathbf{W}_p$ and Eq. (8) cannot be directly solved without knowing the value of λ_2. Once the value of λ_2 is available, Eq. (8) can be considered as the solution of the following problem

$$\min_{\mathbf{W}_p} \lambda_2\|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \beta\|\mathbf{W}_p\|_{2,1}. \tag{10}$$

The optimization of $\mathbf{W}_p$ in Eq. (10) is straightforward shown in next section. Once the value of $\mathbf{W}_p$ is obtained, the value of λ_2 in Eq. (9) can be obtained as well. This motivates us to optimize $\mathbf{W}_p$ and λ_2 using the alternating optimization strategy [46]. As a result, the solution of $\mathbf{W}_p$ is a local optimal solution of Eq. (6) and the value of λ_2 is the weight between the loss function and the second term in Eq. (6). Moreover, if the alternating optimization strategy converges (it will be proved later), the solution of $\mathbf{W}_p$ in Eq. (10) satisfies the Karush–Kuhn–Tucker (KKT) conditions of Eq. (6).

The value of λ_2 is also related to the update of $\mathbf{W}_t$. Moreover, the value of λ_1 is related to the updates of $\mathbf{W}_a$ and $\mathbf{w}_{ag}$, and the value of λ_3 is related to the optimization of $\mathbf{W}_t$ and $\mathbf{w}_{tg}$. Based on the similar steps from Eq. (7) to Eq. (10), we have the following result while individually optimizing $\mathbf{W}_p$, $\mathbf{W}_a$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{w}_{tg}$

$$\begin{cases}\begin{aligned}\min_{\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag},\mathbf{W}_p} & \|\mathbf{Y}_t - [\mathbf{X}_t,\mathbf{x}_{tg}][\mathbf{W}_t^\top,\mathbf{w}_{tg}^\top]^\top\|_{2,1} \\ & + \lambda_1\|\mathbf{Y}_a - [\mathbf{X}_a,\mathbf{x}_{ag}][\mathbf{W}_a^\top,\mathbf{w}_{ag}^\top]^\top\|_{2,1} \\ & + \lambda_2\|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} \\ & + \lambda_3\|[\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag}]\|_F^2 \\ & + \alpha\|[\mathbf{W}_t,\mathbf{W}_a]\|_{2,1} + \beta\|\mathbf{W}_p\|_{2,1} \qquad \text{(a)}\end{aligned} \\ \lambda_1 = \frac{1}{2\sqrt{\|\mathbf{Y}_a-[\mathbf{X}_a,\mathbf{x}_{ag}][\mathbf{W}_a^\top,\mathbf{w}_{ag}^\top]^\top\|_{2,1}}} \qquad \text{(b)} \\ \lambda_2 = \frac{1}{2\sqrt{\|\mathbf{X}_t\mathbf{W}_t-\mathbf{X}_p\mathbf{W}_p\|_{2,1}}} \qquad \text{(c)} \\ \lambda_3 = \frac{1}{2\sqrt{\|[\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag}]\|_F^2}}. \qquad \text{(d)}\end{cases} \tag{11}$$

Algorithm 1: The pseudo of solving Eq. (6).

```
Input: $\mathbf{X}_t$, $\mathbf{X}_p$, $\mathbf{Y}_t$, $\mathbf{X}_a$, $\mathbf{Y}_a$, $\mathbf{x}_{tg}$, $\mathbf{x}_{ag}$, $\alpha$, and $\beta$;
Output: $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$;
Randomly initialize $\mathbf{W}_t$, $\mathbf{w}_{tg}$, $\mathbf{W}_p$, $\mathbf{W}_a$, and $\mathbf{w}_{ag}$;
Initialize $\lambda_2$ via $\lambda_2 = \frac{1}{2\|[\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag}]\|_F}$;
repeat
    Calculate $\mathbf{A}$ and $\mathbf{B}$ by Eq. (13);
    Calculate $\mathbf{C}$ and $\mathbf{U}$ by Eq. (16);
    Calculate $\mathbf{V}$ by Eq. (21);
    Update $\mathbf{w}_{tg}$ and $\mathbf{w}_{ag}$ by Eq. (12);
    Update $\mathbf{W}_a$ by Eq. (18);
    Update $\mathbf{W}_t$ by Eq. (15);
    Update $\mathbf{W}_p$ by Eq. (20);
    Update $\lambda_1$ by Eq. (11b);
    Update $\lambda_2$ by Eq. (11c);
    Update $\lambda_3$ by Eq. (11d);
until Eq. (6) converges;
```

The values of λ_1, λ_2, and λ_3 are automatically obtained without the tuning process and can be regarded as the weight of the corresponding term. For example, if the prediction error is small, the value of λ_1 is large, *i.e.,* the second term in Eq. (6) is more important. Hence, the optimization of the value of λ_1 automatically balances the contribution of the second term in Eq. (6). As a result, the optimization of Eq. (6) is changed to optimize Eq. (11a). It is noteworthy that we keep two parameters (*i.e.,* α and β) in Eq. (6) to be tuned because they control the sparsity of the terms $\|[\mathbf{W}_t,\mathbf{W}_a]\|_{2,1}$ and $\|\mathbf{W}_p\|_{2,1}$. Moreover, their sparsity will change based on the data distribution [23,37,39].

2.5. Optimization

Eq. (11a) is not convex to all variables, *i.e.,* $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$, but is convex to each variable while fixing the others to achieve a local minima. Hence, in this paper, we employ the alternating optimization strategy [46] to solve Eq. (11a). Specifically, our method iteratively optimizes each parameter by fixing the others until the objective function value is stable. We list the pseudo of our method in Algorithm 1.

2.5.1. Update $\mathbf{w}_{tg}$ and $\mathbf{w}_{ag}$ by fixing $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$

With the fixed $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$, the optimization with respect to the variables $\mathbf{w}_{tg}$ and $\mathbf{w}_{ag}$ are independent to each other. Thus we individually set the derivative of Eq. (11a) with respect to $\mathbf{w}_{tg}$ and $\mathbf{w}_{ag}$ to zero to obtain

$$\begin{aligned}\hat{\mathbf{w}}_{tg} &= (\mathbf{x}_{tg}^\top\mathbf{A}\mathbf{x}_{tg} + \lambda_2)^{-1}\mathbf{x}_{tg}^\top\mathbf{A}(\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t) \\ \hat{\mathbf{w}}_{ag} &= (\mathbf{x}_{ag}^\top\mathbf{B}\mathbf{x}_{ag} + \lambda_2)^{-1}\mathbf{x}_{ag}^\top\mathbf{B}(\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a)\end{aligned} \tag{12}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$, respectively, are diagonal matrices where the diagonal elements are defined as

$$\begin{cases} a_{jj} = \frac{1}{2\|(\mathbf{Y}_t-[\mathbf{X}_t,\mathbf{x}_{tg}][\mathbf{W}_t^\top,\mathbf{w}_{tg}^\top]^\top)^j\|_2^2}, j = 1,\dots,n. \\ b_{jj} = \frac{1}{2\|(\mathbf{Y}_a-[\mathbf{X}_a,\mathbf{x}_{ag}][\mathbf{W}_a^\top,\mathbf{w}_{ag}^\top]^\top)^j\|_2^2}, j = 1,\dots,n. \end{cases} \tag{13}$$

2.5.2. Update $\mathbf{W}_t$ by fixing $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_a$, and $\mathbf{W}_p$

Given $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_a$, and $\mathbf{W}_p$, Eq. (11a) is change to

$$\begin{aligned}\min_{\mathbf{W}_t} & \|\mathbf{Y}_t - [\mathbf{X}_t, \mathbf{x}_{tg}][\mathbf{W}_t^\top, \mathbf{w}_{tg}^\top]^\top\|_{2,1} \\ & + \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} \\ & + \|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F.\end{aligned} \tag{14}$$

By setting the derivative of Eq. (14) with respect to $\mathbf{W}_t$ to zero and solving the resulting equations, we can obtain

$$\hat{\mathbf{W}}_t = \mathbf{G}^{-1}\mathbf{H} \tag{15}$$

where $\mathbf{G} = (\mathbf{X}_t^\top(\mathbf{A} + \mathbf{C})\mathbf{X}_t + \lambda_1\mathbf{U} + \lambda_2\mathbf{I}_d)$ and $\mathbf{H} = (\mathbf{X}_t^\top\mathbf{A}(\mathbf{Y}_t - \mathbf{x}_{tg}\mathbf{w}_{tg}) + \mathbf{X}_t^\top\mathbf{C}\mathbf{X}_p\mathbf{W}_p)$, $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is an identity matrix, and $\mathbf{C} \in \mathbb{R}^{n\times n}$ and $\mathbf{U} \in \mathbb{R}^{d\times d}$ are diagonal matrices and their respective diagonal elements are

$$\begin{cases} c_{jj} = \frac{1}{2\|(\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p)^j\|_2^2}, j = 1, \dots, n. \\ u_{kk} = \frac{1}{2\|(\mathbf{W}_t, \mathbf{W}_a)^k\|_2^2}, j = 1, \dots, n. \end{cases} \tag{16}$$

2.5.3. Update $\mathbf{W}_a$ by fixing $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_p$

With fixed $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_p$, Eq. (11a) becomes:

$$\begin{aligned}\min_{\mathbf{W}_a} & \|\mathbf{Y}_a - [\mathbf{X}_a, \mathbf{x}_{ag}][\mathbf{W}_a^\top, \mathbf{w}_{ag}^\top]^\top\|_{2,1} \\ & + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} + \|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F.\end{aligned} \tag{17}$$

By setting the derivative of Eq. (17) with respect to $\mathbf{W}_t$ to zero and solving the equations, we obtain

$$\hat{\mathbf{W}}_a = (\mathbf{X}_a^\top\mathbf{B}\mathbf{X}_t + \lambda_1\mathbf{U} + \lambda_2\mathbf{I}_d)^{-1}\mathbf{X}_a^\top\mathbf{B}(\mathbf{Y}_t - \mathbf{x}_{tg}\mathbf{w}_{tg}). \tag{18}$$

2.5.4. Update $\mathbf{W}_p$ by fixing $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_a$

Given $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_a$, Eq. (11a) becomes:

$$\min_{\mathbf{W}_p} \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_3\|\mathbf{W}_p\|_{2,1}. \tag{19}$$

By setting the derivative of Eq. (19) with respect to $\mathbf{W}_t$ to zero and solving the equations, we obtain

$$\hat{\mathbf{W}}_p = (\mathbf{X}_p^\top\mathbf{C}\mathbf{X}_p + \lambda_3\mathbf{V})^{-1}\mathbf{X}_p^\top\mathbf{C}\mathbf{X}_t\mathbf{W}_t \tag{20}$$

where $\mathbf{V} \in \mathbb{R}^{d\times d}$ is a diagonal matrix whose diagonal elements are defined as

$$v_{kk} = \frac{1}{2\|(\mathbf{W}_p)^k\|_2^2}, k = 1, \dots, d. \tag{21}$$

2.5.5. Convergence, initialization, and complexity

Algorithm 1 iteratively optimizes the variables (*i.e.,* $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$) under the framework of the alternating optimization strategy in [46]. Moreover, each variable has closed-form solution. The convergence of the alternating optimization strategy framework has been theoretically proved, therefore Algorithm 1 converges to a local minimization.

Our proposed method randomly initializes the variables, *i.e.,* $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$ to result in that Algorithm 1 typically achieves convergence within 30 iterations.

The time complexity of optimizing $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$, respectively, is $O(n^2c)$, $O(n^2c)$, $O(d^3n)$, $O(d^3n)$, and $O(d^3n)$, where n, d, and c, respectively, represent the numbers of the samples, the features, and the classes. Hence, the time complexity of Algorithm 1 is $min\{O(d^3n), O(n^2)\}$.

3. Experiments

3.1. Data sets

In our experiments, we used the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set publicly available on the web (adni.loni.usc.edu).[1] We preprocessed the MRI and PET images by sequentially applying spatial distortion correction, skull-stripping, and cerebellum removal, followed by segmenting the MRI images into gray matter, white matter, and cerebrospinal fluid, and then warped them into the AAL template to obtain 90 regions. We further aligned the PET images to their respective MRI images. We finally obtained 90 gray matter volumes from a MRI image and 90 mean intensities from a PET image and used them for features. Besides, the age feature was the age of the subject in this paper.

We generated two data sets from the ADNI cohort: (1) 'Data1' included all subjects with complete MRI and PET data, consisting of 93 AD, 99 NC, 55 pMCI, and 59 sMCI subjects, and (2) 'Data2' includes all subjects with complete MRI, PET, and CerebroSpinal Fluid (CSF),[2] consisting of 50 AD, 51 NC, 31 pMCI, and 30 sMCI subjects. The pMCI subjects were those who converted to AD within 24 months, while sMCI subjects were those who did not convert to AD within 24 months.

3.2. Experimental settings

We defined a baseline model that utilized the original features for classification ('Original') and also considered other state-of-the-art feature selection methods, namely, General Sparsity Regularized feature selection (GSR) [47], Semi-Supervised Learning (SSL) [17], and Domain Transfer Learning (DTL) [29]. We list the detail of the comparison methods as follows.

- Original is the baseline method which uses all predictor data to perform classification without removing any features.
- GSR conducts feature selection by optimizing an $\ell_{2,r}$-norm ($0 < r \le 2$) loss function and an $\ell_{2,p}$-norm ($0 < p \le 1$) regularization term to reduce the influence of subject-level outliers. In our experiments, we considered to form its two variants: 'GSR-Pre' (using the predictor data alone) and 'GSR-Aux' (using the auxiliary data of the AD and NC subjects alone, *i.e.,* [33]).
- SSL sequentially performs the aging effect removal and feature selection using the AD and NC subjects.
- DTL conducts feature selection using both the predictor data and the auxiliary data, without taking into account the aging effect removal and the robustness against outliers in the data.

We have two proposed methods, *i.e.,* Pro-Age in Eq. (5) tuning 5 parameters and ProAuto-Age in Eq. (6) only tuning 2 parameters. It is noteworthy that most of comparison methods did not take the aging effect removal into account. Hence, we used 'Pro-noAge' and 'ProAuto-noAge', to denote our two proposed methods without taking into account the aging effect removal.

We repeated the 10-fold cross-validation for 100 times on all methods, each of which conducted 5-fold nested cross-validations for model selection. We used grid search in the search range of $\{10^{-5}, \dots, 10^5\}$ to determine the related parameter values and in the search range of $C \in \{2^{-5}, 2^{-4}, \dots, 2^5\}$ in SVM, for all methods so that all methods outputted their best performance.

In this paper, we conducted the experiments using two sets of predictor data, *i.e.,* 'MRI', and 'MRI + PET', respectively, to indicate single modality predictors (only MRI features) and multi-modality predictors

[1] 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. The investigators in ADNI contributed to provide the ADNI data, but did not participate in either the analysis or the writing of this report.

[2] It is noteworthy that we did not use CSF features in our experiments and CSF features only help us to decide the used subjects.

Table 1
Classification performance (%) of all methods on two real data sets with different features. The value in parentheses is the standard deviation and 'Data 1 (MRI) indicates the data set 'Data 1' with the MRI feature.

Data set	Metrics	Original	GSR-Pre	GSR-Aux	SSL	DTL	Pro-noAge	ProAuto-noAge
Data 1 (MRI)	Accuracy	53.40 (1.3)	55.60 (0.6)	57.60 (1.0)	58.60 (0.8)	57.70 (2.1)	64.80 (0.7)	65.70 (1.0)
	Sensitivity	62.95 (2.2)	62.04 (1.3)	63.07 (1.5)	54.85 (1.1)	56.40 (2.4)	63.19 (2.2)	65.49 (1.6)
	Specificity	44.50 (0.9)	49.60 (1.6)	52.50 (0.5)	62.10 (0.7)	58.90 (0.2)	66.30 (0.3)	65.90 (0.2)
	AUC	59.60 (0.1)	64.20 (0.2)	65.10 (0.3)	63.90 (0.5)	63.80 (0.6)	65.90 (0.1)	69.60 (0.2)
Data 2 (MRI)	Accuracy	50.30 (0.8)	52.43 (0.7)	54.60 (1.0)	56.80 (1.4)	61.90 (1.1)	62.50 (2.1)	65.90 (1.1)
	Sensitivity	49.08 (0.8)	52.20 (0.7)	55.50 (1.0)	57.92 (1.4)	62.13 (1.1)	62.22 (2.1)	69.11 (1.2)
	Specificity	51.57 (0.9)	52.61 (0.6)	53.62 (1.1)	55.65 (1.0)	61.66 (1.8)	62.79 (1.4)	62.58 (0.6)
	AUC	60.50 (0.4)	61.90 (0.2)	63.01 (0.2)	63.85 (1.6)	65.76 (1.7)	66.93 (0.9)	68.43 (1.4)
Data 1 (MRI + PET)	Accuracy	65.50 (1.4)	67.20 (0.7)	68.90 (1.9)	68.88 (0.6)	70.90 (1.7)	76.60 (0.6)	78.70 (0.2)
	Sensitivity	67.90 (0.5)	56.26 (2.0)	58.60 (1.8)	60.86 (1.2)	62.96 (1.3)	74.56 (1.1)	77.30 (0.9)
	Specificity	63.20 (1.9)	77.40 (2.3)	78.50 (1.7)	76.20 (1.0)	78.30 (2.0)	78.50 (1.5)	80.00 (1.0)
	AUC	64.20 (1.4)	68.50 (1.8)	70.20 (2.0)	68.90 (0.9)	71.10 (1.3)	75.60 (1.3)	77.56 (0.8)
Data 2 (MRI + PET)	Accuracy	62.30 (1.5)	64.30 (1.2)	66.80 (1.7)	69.80 (1.8)	72.10 (1.8)	76.20 (1.3)	78.20 (1.0)
	Sensitivity	68.20 (2.3)	70.40 (1.9)	68.93 (2.0)	73.19 (0.8)	74.85 (1.2)	76.10 (1.0)	77.51 (1.7)
	Specificity	56.20 (0.6)	57.90 (1.9)	64.60 (1.7)	66.30 (2.2)	69.20 (2.0)	76.30 (2.2)	78.90 (2.1)
	AUC	70.20 (1.0)	73.50 (0.6)	76.80 (1.9)	75.90 (1.6)	77.20 (1.7)	78.20 (2.4)	79.00 (1.9)

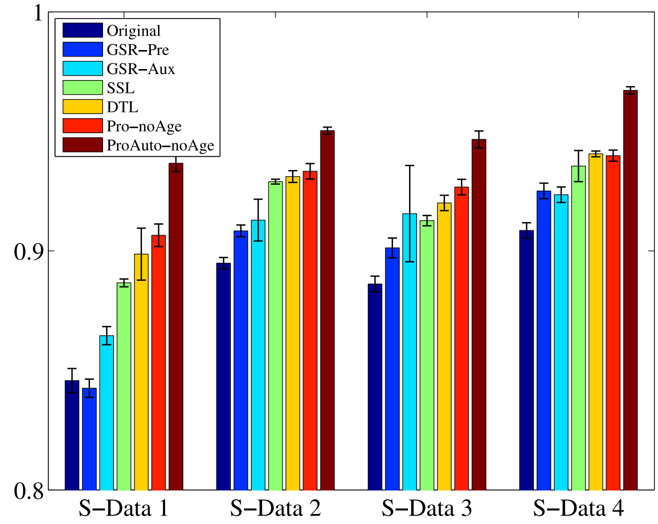


Fig. 2. Classification performance of all methods on four simulated data sets, where the tuple (n_t, n_a, d_t, d_p, d_a, $ratio$) was set as (1000, 200, 400, 400, 200, 0.2), (1000, 200, 400, 400, 200, 0.6), (1000, 2000, 400, 400, 200, 0.2), and (1000, 2000, 400, 400, 200, 0.6), respectively, for S-Data 1, S-Data 2, S-Data 3, and S-Data 4.

(MRI and PET features). That is, we total conducted 4 experiments on two data sets with two different kinds of features.

We employed classification ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), and Area Under the receiver operating characteristic Curve (AUC) as the evaluation metrics.

3.3. Simulation study

In this section, we investigate the validity of our proposed method compared with all comparison methods on the simulation data. By setting the linear regression model as $\mathbf{Y} = \mathbf{XW} + \mathbf{E}$ where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a regressor matrix, $\mathbf{W} \in \mathbb{R}^{d\times 2}$ is a coefficient matrix, $\mathbf{E} \in \mathbb{R}^{n\times 2}$ is a noise matrix, and $\mathbf{Y} = \{0,1\}^{n\times 2}$ is a response matrix, we first generated the data pairs $(\mathbf{Y}_t, \mathbf{X}_t)$ by three steps as follows. (i) The samples of each class were generated from multivariate normal distribution. Moreover, the first d_0 rows of n_i $(i = 1, 2)$ samples were related to the classes and the remaining $d - d_0$ rows were unrelated to the classes. (ii) The first d_0 rows in $\mathbf{W}$ were drawn from $\mathcal{N}(0, 1)$ and the rest $d - d_0$ rows were set as zero. (iii) $\mathbf{E}$ was generated from $\mathcal{N}(0, 10^{-3}\Sigma(0.1))$, where $\Sigma(0.1)$ was a covariance matrix with the diagonal elements of 1 and the off-diagonal elements of 0.1. We followed the above steps to obtain data pairs $(\mathbf{Y}_a, \mathbf{X}_a)$ and $(\mathbf{Y}_t, \mathbf{X}_p)$. Moreover, the samples of each class in $\mathbf{X}_a$ came from multivariate normal distribution and the samples of each class in $\mathbf{X}_p$ were from multivariate t-distribution.

By setting different values to the tuple (n_t, n_a, d_t, d_p, d_a, $ratio$) where n_t, n_a, d_t, d_p, d_a, and $ratio$, respectively, represent the number of the samples of $\mathbf{X}_t$, the samples of $\mathbf{X}_a$, the dimensions of $\mathbf{X}_t$, the dimensions of $\mathbf{X}_p$, the dimensions of $\mathbf{X}_a$, and the percentage of kept features, we evaluated our proposed methods and all comparison methods on four simulated data sets with the experimental setting in Section 3.2 in terms of classification accuracy. We reported the results in Fig. 2, where our proposed method obtained the best performance on all four data sets. This verified the advantages of our proposed methods, compared to all comparison methods. Moreover, Original conducting classification with all features obtained the worst classification performance. This indicates that it is essential to conduct feature selection for high-dimensional data, shown in the literature [9,22,25].

3.4. Results analysis on ADNI data sets

Table 1 lists the classification performance of all methods at different scenarios. We listed our observations as follows.

First, our proposed methods (*i.e.,* ProAuto-noAge and Pro-noAge) achieved the best performance, followed by DTL, SSL, GSR-Aux, GRS-Pre, and Original. Specifically, ProAuto-noAge and Pro-noAge improved on average by 6.48% and 4.38%, compared to the best comparison method, *i.e.,* DTL, in terms of the classification accuracy across two data sets with two different kinds of features. In particular, ProAuto-noAge improved on averages by 13.11%, compared to the worst comparison method, *i.e.,* Original, in terms of all evaluation metrics across all 4 experiments. This verifies the effectiveness of our proposed method for the pMCI/sMCI classification.

Second, all methods achieved larger improvement (in comparison with Original) on Data 2, compared to their corresponding improvement on Data 1. For example, the difference between our ProAuto-noAge and Original was on average 7.88% and 5.86%, respectively, on Data 2 and Data 1 across two different kinds of features. This may imply that the auxiliary information can improve the prediction ability of the predictor data, especially when the sample size of the predictor data is small while the sample size of the auxiliary data are the same.

Third, by regarding the use of the auxiliary data from the AD and NC subjects, GSR-Aux consistently outperformed its counterpart GSR-Pre in all experiments. Specifically, GSR-Aux used the AD and NC subjects to construct the AD/NC classifier to classify the predictor data, *i.e.,* distinguishing pMCI subjects from sMCI subjects, while GSR-Pre employs the pMCI and sMCI subjects to classify the target data. In our experiments, the AD/NC classifier (*i.e.,* GSR-Aux) improves the classification performance by 2.10% in terms of all evaluation metrics across

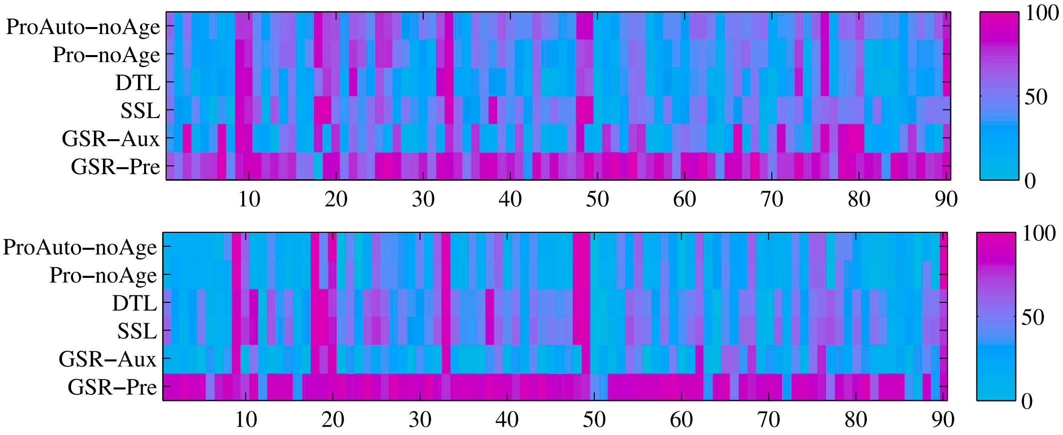


Fig. 3. The feature selection frequency map (*i.e.,* how frequent a feature out of 90 ROIs is selected in 1000 experiments) of MRI features in the pMCI/sMCI classification. In this map, only MRI features are used as the predictor data on Data1 (upper) and Data2 (bottom). The horizontal axis indicates the indices of ROIs..

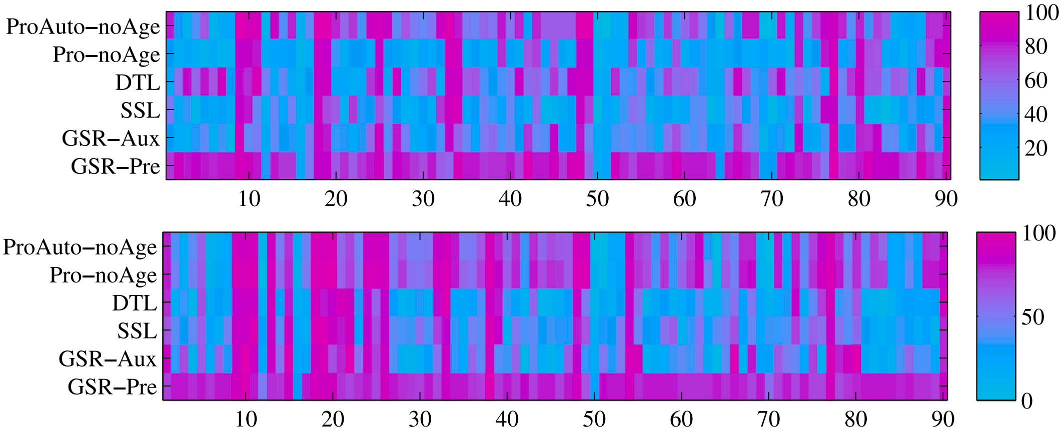


Fig. 4. The feature selection frequency map of MRI features in the pMCI/sMCI classification. In this map, MRI and PET features are used as the predictor data on Data1 (upper) and Data2 (bottom). The horizontal axis indicates the indices of ROIs..

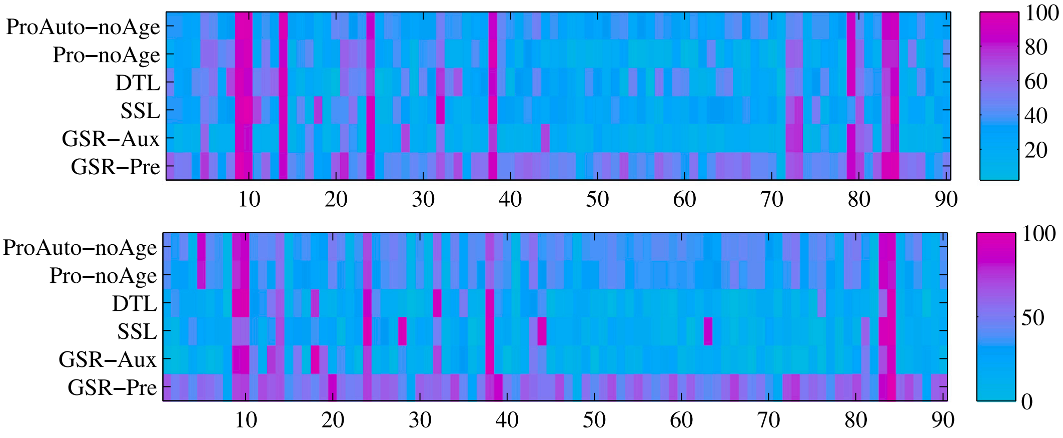
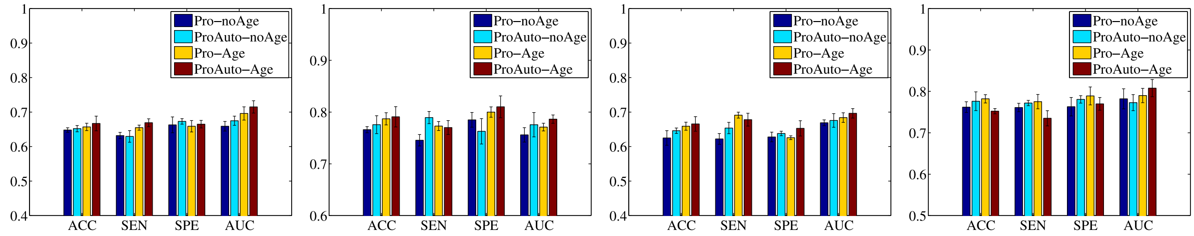


Fig. 5. The feature selection frequency map of PET features in the pMCI/sMCI classification. In this map, MRI and PET features are used as the predictor data on Data1 (upper) and Data2 (bottom). The horizontal axis indicates the indices of ROIs..

all 4 experiments, compared to GSR-Pre, since they select different features to conduct classification tasks. It is noteworthy that the features selected by GSR-Pre were more than the features selected by GSR-Aux. The reason may be that GSR-Pre could not capture subtle structure difference among ROIs with a limited number of high-dimensional samples.

3.5. Top selected ROIs

In this section, we list the top features (*i.e.,* the ROIs) selected by all methods in Figs. 3–5, which could help the clinicians to improve the efficiency and the effectiveness of the disease diagnosis. To do this, we first obtained the totally selected number for each feature across 1000



(a) MRI of Data1 (b) MRI+PET of Data1 (c) MRI of Data2 (d) MRI+PET of Data2

Fig. 6. Classification performance of our proposed methods on two data sets with different kinds of features.

experiments, *i.e.,* repeating the 10-fold cross validation scheme 100 times, and then reported the frequencies (*i.e.,* the selected times divided by 1000 for each feature) for all features. We list our observations as follows.

All methods select similar brain regions as the top regions. For example, for only using MRI features on the two data sets, all methods selected the following MRI features, including globus palladus right, subthalamic nucleus right, uncus right, occipital lobe WM right, nucleus accumbens left, occipital lobe WM left, and fornix right. For using both MRI and PET features, all methods outputted the MRI features including globus palladus right, globus palladus left, uncus right, occipital lobe WM right, nucleus accumbens left, and fornix right, and PET features including globus palladus right, globus palladus left, precuneus right, subthalamic nucleus left, precuneus left, middle occipital gyrus left, and angular gyrus left. Those selected ROIs were verified to be related to AD in previous studies [29,48]. Although all methods selected similar features as the top features, our methods are with the highest frequencies, compared to all comparison methods. This indicates the effectiveness of our proposed methods. Moreover, the ROIs, *e.g.,* globus palladus right, occipital lobe WM right, nucleus accumbens left, and fornix right, were selected by both the data sets with MRI features and the data set with MRI+PET features. The ROIs, *e.g.,* globus palladus right and globus palladus left, were selected by both MRI and PET features on the data sets with MRI+PET features. In particular, the ROI, *i.e.,* globus palladus right, was selected from the data sets with MRI features, as well as from both MRI and PET features on the data sets with MRI+PET features. The overlapped features across different data sets should be very important in the pMCI/sMCI classification.

All methods do not select some brains for the pMCI/sMCI classification, such as MRI features (*e.g.,* angular gyrus right and postcentral gyrus left) for the classification task with the MRI features, and MRI features (*e.g.,* angular gyrus right and postcentral gyrus left) and PET features (*e.g.,* nucleus accumbens left, lingual gyrus right, and thalamus right) for the classification task with the MRI + PET features. This indicates that these features have little contributions for the pMCI/sMCI classification.

It is noteworthy that our methods selected some ROIs more often than the comparison methods, such as hippocampal formation right, middle temporal gyrus left, and hippocampal formation left, which have been demonstrated to be related to AD in [29,48]. We believe that the selection of these ROIs could contribute to enhance the performance in our methods.

3.6. Discussion

In this section, we investigate the influence of the subject age as we assume that the ages of the subjects may influence the pMCI/sMCI classification. To do this, we report the classification performance of four methods of our proposed framework (*i.e.,* Pro-noAge, ProAuto-noAge, Pro-Age, and ProAuto-Age) in Fig. 6.

First, Pro-Age and ProAuto-Age outperform Pro-noAge and ProAuto-noAge, respectively. For example, Pro-Age improved on average by 1.96% and 4.60%, respectively, compared to Pro-NoAge and the best comparison method DTL, in terms of the classification accuracy across all 4 experiments. Moreover, based on the p-values from the paired t-tests at 95% significance level, Pro-Age statistically outperforms Pro-noAge and ProAuto-Age is statistically superior to ProAuto-noAge. Furthermore, each of them (*e.g.,* either Pro-Age or ProAuto-Age) is statistically superior to every comparison method. This indicates that age is one of the risk factor for AD, consistent with the conclusion in [17] that has validated the importance of removing aging effect. However, the small improvement between our methods with the age feature and our method without the age feature shows that the age can affect the classification result but not much. This could be due to the fact that age is only used as one of the features, and other features and auxiliary information may dominant over the aging effect.

Second, we need to tune 5 and 2 parameters, respectively, for the optimization of Eq. (5) (*i.e.,* Pro-Age and Pro-noAge) and Eq. (6) (*i.e.,* ProAuto-Age and ProAuto-noAge). The training process in the former methods need more running time than the latter methods. For example, we set the search rang of $\{10^{-5}, \ldots, 10^{5}\}$ for each parameter so that the former methods need to run the code at least 11^5 times while the latter methods only run the code 11^2 times, *i.e.,* the former methods is 1000 times as fast as the latter methods. However, the latter methods cannot guarantee to outperform the former methods. For example, ProAuto-Age improved on average by −0.34%, compared to Pro-Age in terms of classification accuracy, while ProAuto-noAge improved on average by 1.23%, compared to Pro-noAge in terms of classification accuracy, for all 4 experiments. Moreover, the p-values are 0.066 and 0.078, respectively, for the comparison between ProAuto-Age and Pro-Age and the comparison between ProAuto-noAge and Pro-noAge. This indicates that the comparison groups have no statistically significant difference. Hence, by regarding the efficiency, it is reasonable to replace Eq. (5) with Eq. (6).

3.7. Limitations

Although our proposed framework achieved the best classification performance compared to the comparison methods, there are still other issues that should be tackled for further performance improvement.

First, the proposed framework can be extended to handle samples with missing values. The samples in the ADNI cohort were collected in a longitudinal way at different time phases and for some subjects, there are no values available due to high measurement cost, poor data quality, and unwillingness of the patients to receive invasive tests [23,25,49]. Specifically, while all the subjects in ADNI-I have MRI data, PET data are available for only about half of them. In such a case, it is possible to use the available data (such as complete MRI data) to estimate the missing PET data by the proposed framework. In our future work, we plan to extend our proposed framework to conduct the pMCI/sMCI classification with missing data.

Second, our proposed methods are only devised to measure the linear relations between the neuroimaging-based features and the clinical labels. Hence, it can fail to capture their inherent complex relations, *e.g.,* nonlinear-based high-order relationship. To circumvent this limitation, it will be possible to use kernel methods by measuring pairwise subject relations among data to further incorporate the structural differences between pMCI subjects and sMCI subjects.

Third, the prediction accuracy for the pMCI/sMCI classification varies in the range of 56%–82% in recent studies. The variations in the reported results can be caused by two reasons, *i.e.,* the data set selection and the technique selection. The key factors of the data set selection include different biomarkers, different subsets from ADNI, different definitions of the pMCI and the sMCI. The factors of the technique selection include different methods to deal with the small-size samples, different cross validations, *etc.* In the future work, we plan to design new techniques to generate new samples to solve the issue of small-sized samples, such as over-sampling methods and generative adversarial networks [50].

4. Conclusion

In this paper, we proposed to use the auxiliary information from samples of auxiliary group subjects (*i.e.,* AD/NC), of other imaging modality (*i.e.,* PET), and of a subject's age, to improve the diagnostic accuracy for the pMCI/sMCI identification. The proposed methods used three ways to incorporate the auxiliary data and the predictor data into one formulation, *i.e.,* $\ell_{2,1}$-norm on the weight matrices for joint feature selection, $\ell_{2,1}$-norm loss function for outliers robustness, and including the age factor in the feature matrix for removing aging-effect. In our experiments, the proposed methods outperformed all comparison methods.

CRediT authorship contribution statement

Heng Tao Shen: Conceptualization, Code, Writing - original draft. **Xiaofeng Zhu:** Methodology, Experiments, Writing - review & editing. **Zheng Zhang:** Data curation, Writing - review & editing. **Shui-Hua Wang:** Data curation, Writing - review & editing. **Yi Chen:** Writnng reviewing. **Xing Xu:** Software, Writing - review & editing. **Jie Shao:** Data curation, Writing - review & editing.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] S. O'Sullivan, H. Heinsen, L.T. Grinberg, L. Chimelli, E. Amaro, P.H. do Nascimento Saldiva, F. Jeanquartier, C. Jean-Quartier, M.d.G.M. Martin, M.I. Sajid, et al., The role of artificial intelligence and machine learning in harmonization of high-resolution post-mortem MRI (virtopsy) with respect to brain microstructure, Brain Inf. 6 (1) (2019) 3.

[2] J.M. Gorriz, S. Group, et al., Statistical agnostic mapping: a framework in neuroimaging based on concentration inequalities, 2019, arXiv preprint arXiv:1912.12274.

[3] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 9 (4) (2019) e1312.

[4] S. Hao, Y. Zhou, Y. Guo, A brief survey on semantic segmentation with deep learning, Neurocomputing (2020) http://dx.doi.org/10.1016/j.neucom.2019.11.118.

[5] H. Shu, X. Wang, H. Zhu, D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets, J. Amer. Statist. Assoc. (2019) 1–29.

[6] D. Kong, B. An, J. Zhang, H. Zhu, L2RM: Low-rank linear regression models for high-dimensional matrix responses, J. Amer. Statist. Assoc. (2019) 1–47.

[7] L. Khedher, J. Ramírez, J.M. Górriz, A. Brahim, F. Segovia, A.D.N. Initiative, et al., Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented mri images, Neurocomputing 151 (2015) 139–150.

[8] A. Ortiz, J. Munilla, J.M. Gorriz, J. Ramirez, Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease, Int. J. Neural Syst. 26 (07) (2016) 1650025.

[9] X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, D. Shen, A.D.N. Initiative, et al., A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, Med. Image Anal. 38 (2017) 205–214.

[10] L. Khedher, I.A. Illán, J.M. Górriz, J. Ramírez, A. Brahim, A. Meyer-Baese, Independent component analysis-support vector machine-based computer-aided diagnosis system for Alzheimer's with visual support, Int. J. Neural Syst. 27 (03) (2017) 1650050.

[11] F.J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, D. Castillo-Barnes, Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders, IEEE J. Biomed. Health Inf. 24 (1) (2019) 17–26.

[12] G. Lombardi, G. Crescioli, E. Cavedo, E. Lucenteforte, G. Casazza, A.-G. Bellatorre, C. Lista, G. Costantino, G. Frisoni, G. Virgili, et al., Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment, Cochrane Database Syst. Rev. (3) (2020).

[13] Y. Guo, Y. Gao, D. Shen, Deformable MR prostate segmentation via deep feature learning and sparse patch matching, IEEE Trans. Med. Imaging 35 (4) (2016) 1077–1089.

[14] M. Lorenzi, M. Filippone, G.B. Frisoni, D.C. Alexander, S. Ourselin, A.D.N. Initiative, et al., Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease, NeuroImage 190 (2019) 56–68.

[15] Y. Guo, Z. Wu, D. Shen, Learning longitudinal classification-regression model for infant hippocampus segmentation, Neurocomputing (2019) http://dx.doi.org/10.1016/j.neucom.2019.01.108.

[16] S. Janelidze, N. Mattsson, S. Palmqvist, R. Smith, T.G. Beach, G.E. Serrano, X. Chai, N.K. Proctor, U. Eichenlaub, H. Zetterberg, et al., Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia, Nat. Med. 26 (3) (2020) 379–386.

[17] E. Moradi, A. Pepe, C. Gaser, et al., Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, NeuroImage 104 (2015) 398–412.

[18] L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q.Z. Sheng, M. Sharaf, Mining personal health index from annual geriatric medical examinations, in: ICDM, 2014, pp. 761–766.

[19] X. Hao, Y. Bao, Y. Guo, M. Yu, D. Zhang, S.L. Risacher, A.J. Saykin, X. Yao, L. Shen, A.D.N. Initiative, et al., Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease, Med. Image Anal. 60 (2020) 101625.

[20] X. Zhu, Prediction of mild cognitive impairment conversion using auxiliary information, in: IJCAI, 2019, pp. 4475–4481.

[21] X. Zhu, Y. Zhu, W. Zheng, Spectral rotation for deep one-step clustering, Pattern Recognit. 105 (2020) 107175.

[22] J. Gui, Z. Sun, S. Ji, D. Tao, T. Tan, Feature selection based on structured sparsity: A comprehensive study, IEEE Trans. Neural Netw. Learn. Syst. 28 (7) (2016) 1490–1507.

[23] X. Zhu, J. Yang, C. Zhang, S. Zhang, Efficient utilization of missing data in cost-sensitive learning, IEEE Trans. Knowl. Data Eng. (2019) http://dx.doi.org/10.1109/TKDE.2019.2956530.

[24] S. Wang, F. Nie, X. Chang, L. Yao, X. Li, Q.Z. Sheng, Unsupervised feature analysis with class margin optimization, in: ECML/PKDD, 2015, pp. 383–398.

[25] H.T. Shen, Y. Zhu, W. Zheng, X. Zhu, Half-quadratic minimization for unsupervised feature selection on incomplete data, IEEE Trans. Neural Netw. Learn. Syst. (2020) http://dx.doi.org/10.1109/TNNLS.2020.3009632.

[26] M.W. Weiner, D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, D. Harvey, C.R. Jack, W. Jagust, J.C. Morris, R.C. Petersen, A.J. Saykin, L.M. Shaw, A.W. Toga, J.Q. Trojanowski, Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials, Alzheimer's Dement. (ISSN: 1552-5260) 13 (4) (2017) e1 – e85.

[27] H. Zhu, Z. Khondker, Z. Lu, J.G. Ibrahim, Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers, J. Amer. Statist. Assoc. 109 (507) (2014) 977–990.

[28] R. Wang, F. Nie, R. Hong, X. Chang, X. Yang, W. Yu, Fast and orthogonal locality preserving projections for dimensionality reduction, IEEE Trans. Image Process. 26 (10) (2017) 5019–5030.

[29] B. Cheng, M. Liu, D. Zhang, B.C. Munsell, D. Shen, Domain transfer learning for MCI conversion prediction, IEEE Trans. Biomed. Eng. 62 (7) (2015) 1805–1817.

[30] J. Gui, P. Li, Multi-view feature selection for heterogeneous face recognition, in: ICDM, 2018, pp. 983–988.

[31] Z. Kang, X. Zhao, Shi, c. Peng, H. Zhu, J.T. Zhou, X. Peng, W. Chen, Z. Xu, Partition level multiview subspace clustering, Neural Netw. 122 (2020) 279–288.

[32] S.F. Eskildsen, P. Coupé, D. García-Lorenzo, et al., Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning, NeuroImage 65 (2013) 511–521.

[33] D.H. Ye, K.M. Pohl, C. Davatzikos, Semi-supervised pattern classification: application to structural MRI of Alzheimer's disease, in: PRNI, 2011, pp. 1–4.

[34] R. Hu, X. Zhu, Y. Zhu, J. Gan, Robust SVM with adaptive graph learning, World Wide Web 23 (2020) 1945–1968.

[35] Y. Zhou, L. Tian, C. Zhu, X. Jin, Y. Sun, Video coding optimization for virtual reality 360-degree source, J. Sel. Top. Signal Process. 14 (1) (2020) 118–129.

[36] S. Zhang, X. Li, M. Zong, X. Zhu, D. Cheng, Learning k for knn classification, ACM Trans. Intell. Syst. Technol. (TIST) 8 (3) (2017) 1–19.

[37] X. Zhu, X. Li, S. Zhang, C. Ju, X. Wu, Robust joint graph sparse coding for unsupervised spectral feature selection, IEEE Trans. Neural Netw. Learn. Syst. 28 (6) (2017) 1263–1275.

[38] H.T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, R. Hong, Exploiting subspace relation in semantic labels for cross-modal hashing, IEEE Trans. Knowl. Data Eng. (2020) http://dx.doi.org/10.1109/TKDE.2020.2970050.

[39] X. Zhu, J. Gan, G. Lu, J. Li, S. Zhang, Spectral clustering via half-quadratic optimization, World Wide Web 23 (2020) 1969–1988.

[40] Z. Kang, H. Pan, S.C. Hoi, Z. Xu, Robust graph learning from noisy data, IEEE Trans. Cybern. 50 (5) (2020) 1833–1843.

[41] P. Coupé, S.F. Eskildsen, J.V. Manjón, et al., Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease, NeuroImage: Clin. 1 (1) (2012) 141–152.

[42] J. Young, M. Modat, M.J. Cardoso, et al., Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment, NeuroImage: Clin. 2 (2013) 735–745.

[43] Z. Kang, X. Lu, Y. Lu, c. Peng, W. Chen, Z. Xu, Structure learning with similarity preserving, Neural Netw. (2020) http://dx.doi.org/10.1016/j.neunet.2020.05.030.

[44] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, H.T. Shen, Unsupervised deep hashing with similarity-adaptive and discrete optimization, IEEE Trans. Pattern Anal. Mach. Intell. 40 (12) (2018) 3034–3044.

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

[46] I. Daubechies, R. Devore, M. Fornasier, Iteratively reweighted least squares minimization for sparse recovery, Commun. Pure Appl. Math. 63 (1) (2010) 1–38.

[47] H. Peng, Y. Fan, A general framework for sparsity regularized feature selection via iteratively reweighted least square minimization, in: AAAI, 2017, pp. 2471–2477.

[48] C. Misra, Y. Fan, C. Davatzikos, Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI, Neuroimage 44 (4) (2009) 1415–1422.

[49] K. Thung, C. Wee, P. Yap, D. Shen, Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion, NeuroImage 91 (2014) 386–400.

[50] X. Xu, H. Lu, J. Song, Y. Yang, H.T. Shen, X. Li, Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval, IEEE Trans. Cybern. (2019) http://dx.doi.org/10.1109/TCYB.2019.2928180.