Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Low-rank sparse feature selection with incomplete labels for Alzheimer's disease progression prediction

Zhi Chen [a], Yongguo Liu [a,*], Yun Zhang [a], Rongjiang Jin [b], Jing Tao [c], Lidian Chen [c]

[a] *Knowledge and Data Engineering Laboratory of Chinese Medicine, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, 610054, China*
[b] *College of Health Preservation and Rehabilitation, Chengdu University of Traditional Chinese Medicine, Chengdu, 610075, China*
[c] *College of Rehabilitation Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou, 350122, China*

ARTICLE INFO

Keywords:
Alzheimer's disease
Disease progression
Feature selection
Incomplete labeled data

ABSTRACT

Background: How to predict the cognitive performance of Alzheimer's disease (AD) and identify the informative neuroimaging markers is essential for timely treatment and possible delay of the disease. However, incomplete labeled samples and noises in neuroimaging data pose challenges to building reliable and robust prediction models. In this paper, we present a model named Low-rank Sparse Feature Selection with Incomplete Labels (LSFSIL) for predicting cognitive performance and identifying informative neuroimaging markers with MRI data and incomplete cognitive scores.
Method: We propose a sparse matrix decomposition method to decompose the incomplete cognitive score matrix into two parts for recovering missing scores and utilizing incomplete labeled data. The former is the recovered cognitive score matrix without missing values. To make the recovered scores close to the real ones, a manifold regularizer is devised to fit the label distribution for capturing the label correlations locally. The latter is a ℓ_1-norm regularized matrix which represents the associated errors. Next, a low-rank regression model that regards the recovered matrix as the target is developed to increase the robustness to noises and outliers. Besides, $\ell_{2,1}$-norm is introduced into the objective function as a sparse regularization to identify the important features.
Results: Experimental results demonstrate that LSFSIL achieves higher performance and outperforms several state-of-the-art feature selection approaches. Moreover, the neuroimaging markers selected by LSFSIL are consistent with the previous AD studies.
Conclusions: LSFSIL is effective in informative neuroimaging marker identification for cognitive performance prediction with incomplete labeled data.

1. Introduction

Alzheimer's disease (AD), the most prevalent cause of dementia, is an irreversible and progressive neurodegenerative disease [1]. Recent studies have shown that approximately 46.8 million people were living with AD in 2016 in the world [2] and the prevalence of AD may reach over 100 million in the world by 2050 [3]. In advanced stages, due to the massive death of cells in brain, patients are unable to perform even basic tasks required for daily living, resulting in a need for constant monitoring and care [4]. Unfortunately, there is no effective cure for this debilitating and ultimately fatal disease until now [5]. However, the timely AD diagnosis and treatment in its early stages can defer or stop the disease progression [6–9]. Therefore, various cognitive tests have been designed to help the diagnosis of AD and the evaluation of treatment effect, including mini-mental state examination (MMSE) [10], AD assessment scale-cognitive subscale (ADAS-Cog) [11], clinical dementia rating-sum of the boxes scale (CDR-SB) [12], and so on. These tests focus on quantitively measuring the cognitive status of patients based on the performance in a series of tasks or questions, such as identifying a picture of an animal and counting backward [13,14]. They evaluate particular aspects in one or more cognitive domains, such as memory, language, and the ability to recognize objects. For example, MMSE [10], which requires about 10 min to administer, assesses cognitive function in the areas of orientation, memory, attention, calculation, language, and visual construction. ADAS-Cog [11] is a detailed cognitive test covering all cognitive areas in dementia. It takes about 40 min to

[*] Corresponding author.
E-mail address: liuyg@uestc.edu.cn (Y. Liu).

https://doi.org/10.1016/j.compbiomed.2022.105705
Received 10 March 2022; Received in revised form 16 May 2022; Accepted 4 June 2022
Available online 9 June 2022

administer. CDR-SB [12] is used to assess both cognition and basic-instrumental activities of daily living. It is relatively easy to administer and requires only a few minutes to complete. Considering that different cognitive tests have different sensitivities and specificities and each test evaluates particular aspects in cognitive domains, clinicians usually combine multiple tests for cognitive evaluation [14]. However, conducting cognitive assessments by clinicians is a highly time-consuming task and can only reveal the current cognitive status but is not able to predict the cognitive trajectories in the future. Therefore, there has been a growing interest in building machine learning models for estimating current and future cognitive scores based on brain magnetic resonance imaging (MRI) data [5,9,15,16].

However, the high feature dimensionality of MRI data poses a big challenge for existing models. The number of features extracted from MRI usually reaches hundreds to thousands but only a few features are related to AD [17]. As a result, feature selection, which aims to select the most discriminative and informative elements from the original feature set to represent the data and reduce feature dimensionality, is essential for cognitive score prediction [18]. Recently, an abundance of feature selection approaches for cognitive score prediction have been proposed [6,19–24]. Most methods build linear regression models based on the sparsity-inducing norm on the projection matrix to select relevant and discriminative features. For example, Zhou et al. [19] formulated cognitive score prediction as a multi-task learning problem and selected features with $\ell_{2,1}$-norm and ℓ_1-norm. Considering that different cognitive scores may prefer different brain regions, Cao et al. [20] developed a generalized fused group lasso to model the relations among cognitive scores and MRI features based on prior knowledge. Zhang et al. [24] presented a $\ell_{2,1}$-norm regularized regression model to select features so as to perform AD classification and cognitive score prediction jointly. Despite the promising results achieved by the existing studies, there are still two main limitations, regarding the output labels (cognitive scores) and input features (MRI features) of the model, respectively:

1) For output labels, the aforementioned methods assume the cognitive scores of all subjects are complete. However, some subjects may not be assessed as scheduled due to various reasons and miss the ground-truth cognitive scores at some time-points. For example, among the 814 subjects in AD Neuroimaging Initiative (ADNI) dataset, only 534 subjects own full MMSE scores for all time-points during the two-year follow-up period. Existing models usually discard the subjects with incomplete cognitive scores, which makes the overfitting problem serious and degrades the accuracy and robustness of the prediction models [25]. Hence, dealing with the data with incomplete cognitive scores is of great significance in cognitive score prediction.
2) For input features, MRI data may be affected by a wide variety of noises in the procedure of acquisition and preprocessing [26]. For example, we usually perform image segmentation, *i.e.*, partition an image into distinct regions, and then extract the features of these regions for the following tasks. However, the segmentation may fail due to the noises in MRI, leading to the inaccurate partition of some regions. As a result, there exist features and samples which are contaminated and unsuitable for the following learning tasks. Most previous studies adopt ℓ_2 or F-norm to characterize prediction errors, which are sensitive to noises and usually fail to build reliable models [27,28].

To deal with the aforementioned limitations, we present a model named Low-rank Sparse Feature Selection with Incomplete Labels (LSFSIL) to select informative MRI features for predicting cognitive scores at multiple time-points. For employing the incomplete labeled samples, we decompose the incomplete cognitive score matrix into two parts: a recovered cognitive score matrix and a sparse error matrix. The former is assumed to be the recovered cognitive score matrix without missing values and is regarded as the regression target while the latter is a sparse matrix that corresponds to the recovery errors. In this way, all available samples can be utilized for training, which leads to a substantial number of samples and yields the proper modeling of the intrinsic relations between MRI features and cognitive scores. Here, a manifold regularization term is designed to capture the cognitive score correlations locally and guide the decomposition of the incomplete cognitive score matrix by ensuring that similar subjects have similar recovered cognitive scores. Moreover, to improve the robustness to noises and outliers in MRI data, we design a low-rank sparse regression model which adopts the nuclear norm as the basic metric of the loss function. An effective optimization algorithm is developed to solve the optimization problem. The main contributions of this paper are summarized as:

1) We develop a low-rank sparse feature selection model with incomplete labeled MRI data which can select informative MRI features for predicting multiple cognitive scores at multiple time-points.
2) To employ incomplete labeled subjects, a sparse matrix decomposition method is designed to recover missing scores. To preserve the local neighborhood of the cognitive scores and capture the correlations among cognitive scores after recovering the missing values, a manifold regularization term is integrated into the framework.
3) To improve the robustness to noises and outliers, we design a low-rank sparse regression model that adopts the nuclear norm as the basic metric to measure the regression loss. Moreover, an efficient iterative algorithm is developed to solve the optimization of the proposed formulation.
4) We conduct experiments on the ADNI dataset to verify the effectiveness of the proposed method. Experimental results show that the features selected by the proposed LSFSIL perform better than previous methods in most cases for cognitive prediction.

The rest of this paper is organized as follows: In Section 2, the related works are reviewed. In Section 3, we first propose the LSFSIL model and then provide the optimization algorithm as well as its computational complexity analysis. Experimental results and comparisons with other approaches are presented in Section 4. Section 5 presents the discussion and Section 6 finally concludes the paper.

2. Related work

In this section, we briefly review the related works of machine learning methods for computer-aided AD diagnosis. In recent years, computer-aided diagnosis techniques based on machine learning approaches have been widely applied to detect AD at early stages and predict the disease progression using neuroimaging data. For example, Zhang et al. [24] proposed a multi-modal multi-task method for simultaneous AD classification and cognitive score prediction. Duchesne et al. [29] employed a robust linear regression model to estimate one-year changes in MMSE from MRI. Wang et al. [30] designed a high-dimensional kernel regression method to estimate the scores of ADAS-Cog and MMSE. As the dimension of neuroimaging data is normally far larger than the sample size, many dimensionality reduction methods that aim to reduce the dimension of data features have been proposed [6,31]. Among these methods, with the ability to select the discriminative features subset and provide interpretable results, feature selection has become a better alternative method in AD diagnosis. For example, Zhou et al. [19] built a multi-task learning model with $\ell_{2,1}$-norm and ℓ_1-norm to select feature subsets that are important for cognitive score prediction. Cao et al. [20] proposed to model the relationships among cognitive scores and MRI features based on prior knowledge to select important features. Zhu et al. [22] designed a feature selection method that considers relational information inherent in the observations for joint regression and classification in AD diagnosis.

However, the methods mentioned above usually assume each sample has all cognitive scores at all time-points and cannot directly utilize the samples with incomplete cognitive scores for training. Some patients do not receive cognitive assessments at the agreed time for various reasons and these patients have incomplete cognitive scores. For employing the samples with incomplete cognitive scores, Liu et al. [25] proposed to discard the missing scores at certain time-points and make use of the remaining time-points. However, there are often correlations between cognitive tests and time-points. For example, it is reported that different cognitive tests evaluate some overlapping cognitive abilities [32]. The correlations are helpful for improving prediction performance [33,34], which will be lost if the missing cognitive scores are discarded. In this paper, we propose a sparse matrix decomposition method to decompose the incomplete cognitive score matrix into two parts for recovering missing scores. Thus, both incomplete labeled samples and the correlations between cognitive tests and time-points can be utilized.

Meanwhile, the neuroimaging data are usually contaminated by various noises and many samples are affected. Adeli et al. [35,36] proposed to take advantage of testing samples as unlabeled data during the training phase to deal with noises and outliers simultaneously. Zhu et al. [22] exploited relational information inherent in the observations to develop a feature selection method robust to noises and outliers. However, most existing regression methods minimize the regression term using the ℓ_2 or F-norm, which is sensitive to noises and outliers in the data. To enhance the robustness and discrimination of selected features, we encode the regression errors using nuclear norm and propose a low-rank sparse feature selection method for cognitive score prediction. Although several low-rank sparse regression methods have been proposed in machine learning community [37,38], recent progress on handling noisy data in AD diagnosis has gone largely unnoticed. Besides, these low-rank sparse regression methods can not take advantage of the incomplete labeled samples.

3. Methods

In this section, we first introduce some notations and model formulation. Next, we describe the LSFSIL model. Then, an iterative optimization algorithm is presented. Finally, we provide the complexity analysis for the proposed method.

3.1. Notations and model formulation

For matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, the (i,j)-th element, i-th row and j-th column are denoted by a_{ij}, $\mathbf{a}^i$ and $\mathbf{a}_j$, respectively. The trace and transpose of $\mathbf{A}$ are $\mathrm{tr}(\mathbf{A})$ and $\mathbf{A}^T$, respectively. The F-norm of matrix $\mathbf{A}$ is

$$\left\|\mathbf{A}\right\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}. \tag{1}$$

The ℓ_1-norm of matrix $\mathbf{A}$ is

$$\left\|\mathbf{A}\right\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left|a_{ij}\right|. \tag{2}$$

The $\ell_{2,1}$-norm of matrix $\mathbf{A}$ is

$$\left\|\mathbf{A}\right\|_{2,1} = \sum_{i=1}^{n}\left\|\mathbf{a}^i\right\|_2, \tag{3}$$

where $\left\|\mathbf{a}^i\right\|_2 = \mathrm{sqrt}(\sum_{j=1}^{m} a_{ij}^2)$. The nuclear norm of matrix $\mathbf{A}$ is

$$\left\|\mathbf{A}\right\|_* = \sum_i \sigma_i(\mathbf{A}), \tag{4}$$

where $\sigma_i(\mathbf{A})$ denotes the i-th singular value of $\mathbf{A}$. The ∞-norm of matrix $\mathbf{A}$ is.

$$\left\|\mathbf{A}\right\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{m}\left|a_{ij}\right|. \tag{5}$$

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ describes the MRI feature matrix and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]^T \in \mathbb{R}^{n\times c}$ records the corresponding cognitive scores, where n is the number of samples, d is the number of features, and c is the number of cognitive scores. It is worth noting that several elements are missing in $\mathbf{Y}$ and the missing elements are recorded as -1. In this paper, we aim to develop a feature selection method that is robust to noises and can make full use of incomplete labeled samples. The illustration of LSFSIL is shown in Fig. 1. As shown in Fig. 1, the original target matrix $\mathbf{Y}$, which contains the existing and missing cognitive scores, is decomposed into two parts: recovered target matrix $\mathbf{Z}$ and associated error matrix $\mathbf{E}$. Matrix $\mathbf{Z}$ is assumed to contain the recovered cognitive scores without missing values. That is, the missing elements in $\mathbf{Y}$ are recovered in $\mathbf{Z}$ by the decomposition. The term $\mathrm{tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z})$ guides the decomposition of the incomplete cognitive score matrix by ensuring that similar subjects have similar recovered cognitive scores. Error matrix $\mathbf{E}$ records the missing scores and the term $\left\|\mathbf{E}\right\|_1$ is utilized to constrain its sparsity. Then, the recovered matrix $\mathbf{Z}$ is regarded as the prediction target and the input feature matrix $\mathbf{X}$ is projected into the target space by matrix $\mathbf{W}$. The regularizer $\left\|\mathbf{W}\right\|_{2,1}$ is used to select informative features across all samples with joint sparsity. Finally, the nuclear norm is adopted to characterize the prediction loss. In the following subsection, we describe the formulation of LSFSIL in detail.

3.2. Objective function

The objective function of feature selection for cognitive score prediction task can be defined as a linear regression model:

$$\min_{\mathbf{W}}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1}, \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the projection matrix, loss function $\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2$ is the element-wise difference between the target values and the predicted ones, and α is a scalar regularization hyperparameter. The $\ell_{2,1}$-norm regularization term $\|\mathbf{W}\|_{2,1}$ penalizes the root sum square of the rows in $\mathbf{W}$ by making some rows easily shrink to zero so as to select informative and important features.

It is noted that (6) directly utilizes a half-baked cognitive score matrix as the regression target. However, in clinical practice, some subjects may not have all ground-truth cognitive scores/labels due to various reasons. For example, due to time conflict, some subjects could not be evaluated as scheduled and the cognitive scores at certain time-points are missing. Equation (6) only considers the subjects with complete cognitive scores during training. In this way, the number of available training samples is decreased, resulting in the overfitting problem [39]. In order to avoid the disturbance of missing cognitive scores and make full use of all subjects, we assume that the target matrix $\mathbf{Y}$ contains two parts, one includes the ground-truth cognitive scores without any missing values and the other includes the missing values. Accordingly, we decompose target matrix $\mathbf{Y}$ into $\mathbf{Z} + \mathbf{E}$, where $\mathbf{Z} \in \mathbb{R}^{n\times c}$ represents the recovered cognitive score matrix and $\mathbf{E} \in \mathbb{R}^{n\times c}$ is the associated error matrix. We add the ℓ_1-norm regularizer on $\mathbf{E}$ so as to encourage the sparsity of the error matrix. Therefore, we have the following objective function

$$\min_{\mathbf{Z},\mathbf{E},\mathbf{W}}\left\|\mathbf{Z} - \mathbf{X}\mathbf{W}\right\|_F^2 + \alpha\left\|\mathbf{W}\right\|_{2,1} + \beta\left\|\mathbf{E}\right\|_1 \; s.t.\; \mathbf{Y} = \mathbf{Z} + \mathbf{E}, \tag{7}$$

where β is a regularization parameter. As can be seen, the recovered cognitive score matrix $\mathbf{Z}$ is used as the regression target in (7). In this way, both the subjects with complete cognitive scores and the ones with

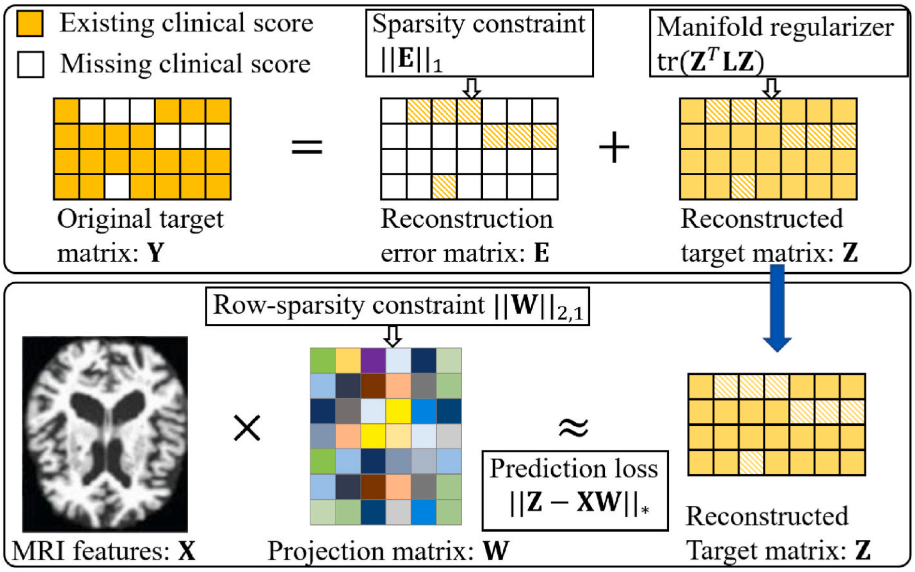

Existing clinical score
Missing clinical score
Sparsity constraint $||\mathbf{E}||_1$
Manifold regularizer $\mathrm{tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z})$
Original target matrix: **Y**
=
Reconstruction error matrix: **E**
+
Reconstructed target matrix: **Z**
Row-sparsity constraint $||\mathbf{W}||_{2,1}$
MRI features: **X**
×
Projection matrix: **W**
≈
Prediction loss $||\mathbf{Z}-\mathbf{XW}||_*$
Reconstructed Target matrix: **Z**

Fig. 1. Illustration of the proposed LSFSIL model. The original target matrix **Y** is decomposed into two parts: recovered target matrix **Z** and associated error matrix **E**. Then, the MRI data **X** are projected into the recovered target matrix **Z** by $\ell_{2,1}$-norm regularized projection matrix **W** to make use of all available samples and select informative features.

incomplete cognitive scores can be employed to train the model.

It is obvious that there is no instruction for the decomposition of the target matrix **Y** in (7). That is, the recovered matrix **Z** may be arbitrary and far away from the real cognitive score matrix. Thus, training the model with such an arbitrary **Z** is not beneficial for performance promotion. To encourage that the recovered cognitive scores are close to the real cognitive scores, we preserve the local neighborhood of the cognitive scores after the decomposition. We expect that if samples are close to each other in feature space, then their respective recovered cognitive scores should be also similar to each other and add a regularization term in (7) to preserve the local structure:

$$\sum_{i,j} s_{ij}\left\|\mathbf{z}^i-\mathbf{z}^j\right\|_F^2, \tag{8}$$

where s_{ij} represents the similarity between the i-th subject and the j-th subject, which is calculated as

$$s_{ij}=e^{-\frac{\left\|x^i-x^j\right\|_2^2}{\sigma}}. \tag{9}$$

Equation (9) can be reformulated as $\mathrm{tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z})$, where **L** is the Laplacian matrix given as $\mathbf{D}_v-\mathbf{S}_v$, $\mathbf{D}_v$ is a diagonal matrix in which the i-th diagonal element is the sum of the i-th row of $\mathbf{S}_v$, and $\mathbf{S}_v$ is the similarity matrix in which the (i,j)-th element is s_{ij}. Then, the objective function can be reformulated as

$$\min_{\mathbf{Z},\mathbf{E},\mathbf{W}}\|\mathbf{Z}-\mathbf{XW}\|_F^2+\alpha\|\mathbf{W}\|_{2,1}+\beta\|\mathbf{E}\|_1+\gamma\mathrm{tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z})\, s.t.\ \mathbf{Y}=\mathbf{Z}+\mathbf{E}, \tag{10}$$

where γ is a regularization hyperparameter.

It is found that (10) adopts the F-norm that is sensitive to noises to measure the regression loss [27,28]. However, MRI data are affected by a wide variety of noises in the procedure of acquisition and preprocessing, which blurs images and disturbs MRI segmentation. Consequently, the partition of some brain regions may be inaccurate, which contaminates the features related to these brain regions. The F-norm usually fails to build reliable models due to the influence of noises [27]. It is known that the nuclear norm, the sum of all singular values of the matrix, is more robust to noises than F-norm [28]. This motivates us to adopt the nuclear norm to characterize the regression loss. Finally, the objective function is formulated as

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{E},\mathbf{W}}\left\|\mathbf{Z}-\mathbf{XW}\right\|_*+\alpha\left\|\mathbf{W}\right\|_{2,1}+\beta\left\|\mathbf{E}\right\|_1+\gamma\mathrm{tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z})\\&s.t.\ \mathbf{Y}=\mathbf{Z}+\mathbf{E}.\end{aligned} \tag{11}$$

3.3. Optimization

To solve the nonconvex optimization of (11), we adopt an alternating direction method of multipliers (ADMM) algorithm [40]. Relaxation variable **K** is introduced and (1) can be rewritten as

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{E},\mathbf{W},\mathbf{K}}||\mathbf{K}||_*+\alpha\|\mathbf{W}\|_{2,1}+\beta\|\mathbf{E}\|_1+\gamma\mathrm{tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z})\\&s.t.\ \mathbf{Y}=\mathbf{Z}+\mathbf{E},\ \mathbf{K}=\mathbf{Z}-\mathbf{XW}.\end{aligned} \tag{12}$$

Since that $||\mathbf{K}||_*$ is equivalent to $\min_{\mathbf{P},\mathbf{Q}}\frac{1}{2}(\|\mathbf{P}\|_F^2+\|\mathbf{Q}\|_F^2)$, $\mathbf{P}\in\mathbb{R}^{n\times r}$, $\mathbf{Q}\in\mathbb{R}^{r\times c}$, and $r\le\min\{n,c\}$ [28], (12) can be converted to the following equivalent form

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{E},\mathbf{W},\mathbf{K},\mathbf{P},\mathbf{Q}}\frac{1}{2}\left(\|\mathbf{P}\|_F^2+\|\mathbf{Q}\|_F^2\right)+\alpha\|\mathbf{W}\|_{2,1}+\beta\|\mathbf{E}\|_1+\gamma\mathrm{tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z}).\\&s.t.\ \mathbf{Y}=\mathbf{Z}+\mathbf{E},\mathbf{K}=\mathbf{Z}-\mathbf{XW},\mathbf{K}=\mathbf{PQ}.\end{aligned} \tag{13}$$

Solving (13) is equivalent to minimizing the following augmented Lagrange function

$$\mathscr{L} = \frac{1}{2}\left(\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2\right) + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{E}\|_1 + \gamma \mathrm{tr}\left(\mathbf{ZLZ}^T\right)$$
$$+tr\left(\mathbf{U}_1{}^T(\mathbf{Y}-\mathbf{Z}-\mathbf{E})\right) + tr\left(\mathbf{U}_2{}^T(\mathbf{K}-\mathbf{Z}+\mathbf{XW})\right) + tr\left(\mathbf{U}_3{}^T(\mathbf{K}-\mathbf{PQ})\right)$$
$$+\frac{\mu}{2}\left(\left|\left|\mathbf{Y}-\mathbf{Z}-\mathbf{E}\right|\right|_F^2 + \left|\left|\mathbf{K}-\mathbf{Z}+\mathbf{XW}\right|\right|_F^2 + \left|\left|\mathbf{K}-\mathbf{PQ}\right|\right|_F^2\right), \tag{14}$$

where $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$ are the Lagrange multipliers, and penalty parameter $\mu > 0$. In ADMM, the augmented Lagrange function is minimized by solving the subproblems w.r.t. each unknown variable iteratively and each subproblem can be solved efficiently. It contains the following steps to update all variables in each iteration:

Step 1. Update **K**: we fix other variables, and update **K** by solving the following problem

$$\min_{\mathbf{K}} tr\left(\mathbf{U}_2{}^T(\mathbf{K}-\mathbf{Z}+\mathbf{XW})\right) + tr\left(\mathbf{U}_3{}^T(\mathbf{K}-\mathbf{PQ})\right)$$
$$+\frac{\mu}{2}\left(\left|\left|\mathbf{K}-\mathbf{Z}+\mathbf{XW}\right|\right|_F^2 + \left|\left|\mathbf{K}-\mathbf{PQ}\right|\right|_F^2\right) \tag{15}$$

By setting the derivative of (15) w.r.t. **K** to zero, we can obtain its optimal solution

$$\mathbf{K} = \frac{\mathbf{M}_1 + \mathbf{M}_2}{2}, \tag{16}$$

where $\mathbf{M}_1 = \mathbf{Z} - \mathbf{XW} - \mathbf{U}_2/\mu$ and $\mathbf{M}_2 = \mathbf{PQ} - \mathbf{U}_3/\mu$.

Step 2. Update **Z**: updating **Z** by optimizing (14) is equivalent to minimizing the following problem

$$\min_{\mathbf{Z}} \gamma \mathrm{tr}\left(\mathbf{Z}^T\mathbf{LZ}\right) + tr\left(\mathbf{U}_1{}^T(\mathbf{Y}-\mathbf{Z}-\mathbf{E})\right)$$
$$+tr\left(\mathbf{U}_2{}^T(\mathbf{K}-\mathbf{Z}+\mathbf{XW})\right) \tag{17}$$
$$+\frac{\mu}{2}\left(\left|\left|\mathbf{Y}-\mathbf{Z}-\mathbf{E}\right|\right|_F^2 + \left|\left|\mathbf{K}-\mathbf{Z}+\mathbf{XW}\right|\right|_F^2\right).$$

We take the derivative of (17) w.r.t. **Z** and set it to zero. Then, we obtain

$$\mathbf{Z} = (2\mu\mathbf{I} + 2\gamma\mathbf{L})^{-1}(\mu\mathbf{R}_1 + \mu\mathbf{R}_2), \tag{18}$$

where $\mathbf{R}_1 = \mathbf{Y} - \mathbf{E} + \mathbf{U}_1/\mu$ and $\mathbf{R}_2 = \mathbf{K} + \mathbf{XW} + \mathbf{U}_2/\mu$.

Step 3. Update **E**: after other variables are fixed, **E** can be calculated by solving the following minimization problem

$$\min_{\mathbf{E}} \beta\|\mathbf{E}\|_1 + \frac{\mu}{2}\left|\left|\mathbf{Y}-\mathbf{Z}-\mathbf{E}+\frac{\mathbf{U}_1}{\mu}\right|\right|_F^2. \tag{19}$$

Since that $\mathbf{R}_1 = \mathbf{Y} - \mathbf{Z} + \mathbf{U}_1/\mu$, (19) can be rewritten as

$$\min_{\mathbf{E}} \beta\|\mathbf{E}\|_1 + \frac{\mu}{2}\left|\left|\mathbf{E}-\mathbf{R}_1\right|\right|_F^2. \tag{20}$$

Equation (20) has a closed-form solution according to the shrinkage operator:

$$\mathbf{E} = S_{\frac{\beta}{\mu}}(\mathbf{R}_1), \tag{21}$$

where $S_\eta(x) = \mathrm{sign}(x)\max(|x| - \eta, 0)$.

Step 4. Update **W**: after other variables are fixed, we can calculate **W** by solving the following problem

$$\min_{\mathbf{W}} \alpha\|\mathbf{W}\|_{2,1} + \frac{\mu}{2}\left|\left|\mathbf{K}-\mathbf{Z}+\mathbf{XW}+\frac{\mathbf{U}_2}{\mu}\right|\right|_F^2. \tag{22}$$

Let $\mathbf{D} = \mathbf{Z} - \mathbf{K} - \mathbf{U}_2/\mu$, we have

$$\alpha\left|\left|\mathbf{W}\right|\right|_{2,1} + \frac{\mu}{2}\left|\left|\mathbf{K}-\mathbf{Z}+\mathbf{XW}\right.\right.$$
$$\left.\left.+\frac{\mathbf{U}_2}{\mu}\right|\right|_F^2 = \alpha\left|\left|\mathbf{W}\right|\right|_{2,1} + \frac{\mu}{2}\left|\left|\mathbf{XW}-\mathbf{D}\right|\right|_F^2 = \alpha\|\mathbf{W}\|_{2,1} + \frac{\mu}{2}tr\left((\mathbf{XW}-\mathbf{D})^T\right.$$
$$\left.(\mathbf{XW}-\mathbf{D})\right) = \alpha tr\left(\mathbf{W}^T\mathbf{GW}\right) + \frac{\mu}{2}tr\left(\mathbf{W}^T\mathbf{X}^T\mathbf{XW} + \mathbf{D}^T\mathbf{D} - 2\mathbf{D}^T\mathbf{XW}\right), \tag{23}$$

where **G** is the diagonal matrix with the i-th diagonal element $g_{ii} = 1/2\left|\left|\mathbf{w}^i\right|\right|_2$. Then, (22) can be rewritten as

$$\min_{\mathbf{W}} \alpha tr\left(\mathbf{W}^T\mathbf{GW}\right) + \frac{\mu}{2}tr\left(\mathbf{X}^T\mathbf{XWW}^T - 2\mathbf{D}^T\mathbf{XW}\right). \tag{24}$$

By setting the derivative of (24) w.r.t. **W** to zero, we can obtain its optimal solution

$$\mathbf{W} = \mu\left(2\alpha\mathbf{G} + \mu\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}. \tag{25}$$

Step 5. Update **P**: with other variables fixed, **P** can be solved by solving the following problem

$$\min_{\mathbf{P}} \frac{1}{2}\left|\left|\mathbf{P}\right|\right|_F^2 + \frac{\mu}{2}\left|\left|\mathbf{K}-\mathbf{PQ}+\frac{\mathbf{U}_3}{\mu}\right|\right|_F^2. \tag{26}$$

By setting the derivative of (26) w.r.t. **P** equal to zero, it is obvious that

$$\mathbf{P} = \mu\left(\mathbf{K}+\frac{\mathbf{U}_3}{\mu}\right)\mathbf{Q}^T\left(\mathbf{I}+\mu\mathbf{QQ}^T\right)^{-1}. \tag{27}$$

Step 6. Update **Q**: after other variables are fixed, we can obtain **Q** by solving the following problem

$$\min_{\mathbf{P}} \frac{1}{2}\left|\left|\mathbf{Q}\right|\right|_F^2 + \frac{\mu}{2}\left|\left|\mathbf{K}-\mathbf{PQ}+\frac{\mathbf{U}_3}{\mu}\right|\right|_F^2. \tag{28}$$

Similar to the optimization strategy of **P**, we can get the closed-form solution of problem (28)

$$\mathbf{Q} = \mu\left(\mathbf{I}+\mu\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\left(\mathbf{K}+\frac{\mathbf{U}_3}{\mu}\right). \tag{29}$$

Step 7. the Lagrange multipliers and the penalty factor are updated as

$$\begin{aligned}
\mathbf{U}_1 &= \mathbf{U}_1 + \mu(\mathbf{Y}-\mathbf{Z}-\mathbf{E}),\\
\mathbf{U}_2 &= \mathbf{U}_2 + \mu(\mathbf{K}-\mathbf{Z}+\mathbf{XW}),\\
\mathbf{U}_3 &= \mathbf{U}_3 + \mu(\mathbf{K}-\mathbf{PQ}),\\
\mu &= \min(\rho\mu, \mu_{\max}),
\end{aligned} \tag{30}$$

where $\rho > 1$ and $\mu_{\max}$ are manually set constants.

Convergence criteria. The ADMM algorithm solves the original objective function by solving a sequence of subproblems w.r.t. each unknown variable iteratively. It is important to adopt proper stopping criteria for achieving the optimal solution. Following the suggestions in Ref. [40], the stopping criteria are defined as

$$\begin{aligned}
&\left|\left|\mathbf{Y}-\mathbf{Z}-\mathbf{E}\right|\right|_\infty < \varepsilon,\\
&\left|\left|\mathbf{K}-\mathbf{Z}+\mathbf{XW}\right|\right|_\infty < \varepsilon,\\
&\left|\left|\mathbf{K}-\mathbf{PQ}\right|\right|_\infty < \varepsilon,
\end{aligned} \tag{31}$$

where $\varepsilon > 0$ is a small tolerance error.

Algorithm 1. Solving (11) by ADMM

```
Input: MRI data $\mathbf{X}$, target cognitive scores $\mathbf{Y}$, model parameters $\alpha, \beta, \gamma, \rho = 1.1, \mu_{\max} = 10^7, \mu = 10^{-7}, \varepsilon = 10^{-8}$, convergence criteria parameter $\varepsilon$, and the number of selected features $k$.
1:  Initialization: $\mathbf{K} = \mathbf{0}, \mathbf{Z} = \mathbf{Y}, \mathbf{E} = \mathbf{0}, \mathbf{W} = \mathbf{0}, \mathbf{P} = \mathbf{0}, \mathbf{Q} = \mathbf{0}, \mathbf{A} = \mathbf{0}, \mathbf{U}_1 = \mathbf{0}, \mathbf{U}_2 = \mathbf{0}, \mathbf{U}_3 = \mathbf{0}$.
2:  While not converged do
3:      Update $\mathbf{K}$ by (16);
4:      Update $\mathbf{Z}$ by (18);
5:      Update $\mathbf{E}$ by (20);
6:      Update $\mathbf{W}$ by (25);
7:      Update $\mathbf{P}$ by (27);
8:      Update $\mathbf{Q}$ by (29);
9:      Update Lagrange multipliers $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ and penalty factor $\mu$ by (30);
10:     Check the convergence criteria (31).
11: end while
Output: Calculate and sort $\|\mathbf{w}^i\|_2$ $(i = 1,2,\ldots,d)$ in the descending order, and select the top $k$ ranked features.
```

3.4. Computational analysis

Here, we analyze the computational cost of Algorithm 1. We assume that the number of iterations for ADMM is τ. The major computational complexity lies on Step 1, Step 2, and Step 4. The major computational complexity for Step 1, Step 2, and Step 4 are $\mathscr{O}(ndc + nrc)$, $\mathscr{O}(n^3 + ndc + n^2c)$, and $\mathscr{O}(nd^2 + d^3 + ndc)$, respectively. Then, the computational complexity of Algorithm 1 is $\mathscr{O}(\tau(n^3 + n^2c + ndc + d^3))$. As $c \ll d$ and $c \ll n$, the total complexity is $\mathscr{O}(\tau(n^3 + d^3))$.

4. Results

In this section, the experimental dataset, comparison methods, and experimental settings are firstly described. Then, the LSFSIL algorithm is compared with its variants and state-of-the-art methods.

4.1. Dataset

The ADNI dataset (http://adni.loni.usc.edu) [41] is used to validate the performance of LSFSIL. The ADNI project was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration. Its primary goal is to test whether clinical, imaging, genetic, and biochemical biomarkers can be used to detect AD at the earliest possible stage. The ADNI project is a longitudinal study, in which biomarkers and cognitive scores are collected every 6 or 12 months. We aim to infer three clinical assessments: ADAS-Cog, MMSE, and CDR-SB, at four time-points given the baseline MRI scans. The four ime-points include baseline (BL), 6 months (M06), 12 months (M12), and 24 months (M24) after BL. To provide a high degree of reproducibility, we adopt the MRI features extracted by a team from University of California at San Francisco (data acquired from http://adni.loni.usc.edu/). They perform cortical reconstruction and volumetric segmentation with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). Pre-processed ADNI1 1.5T T1 weighted image data in NiFTI format (warping, scaling, B1 correction and N3 inhomogeneity correction) are run with FreeSurfer version 4.3. Each scan is processed by the following steps:

Table 1
The detailed demographic information and clinical characteristics of subjects.

	Normal	MCI	AD
Gender	118/109	256/141	100/90
Age(mean ± std)	76.0 ± 5.0	74.8 ± 7.4	75.3 ± 7.5
Edu(mean ± std)	16.0 ± 2.9	15.6 ± 3.0	14.7 ± 3.1
ADAS-Cog(mean ± std)	9.5 ± 4.2	18.5 ± 6.5	28.3 ± 8.8
MMSE(mean ± std)	29.1 ± 1.0	27.0 ± 1.8	23.3 ± 2.0
CDRSB(mean ± std)	0.0 ± 0.1	1.6 ± 0.9	4.3 ± 1.6

> Step 1: In this step, 1) motion correction and registration, 2) non-uniform intensity normalization, 3) talairach transform computation, and 4) intensity normalization and skull strip are initiated. Step 2: This step creates the white-matter and pial surfaces and then segments the gray and white matter, and the sub-cortical structures. Step 3: cortical parcellation is created in this step.

The cortical volume (CV), surface area (SA), cortical thickness average (TA), and standard deviation of thickness (TS) of cortical regions and subcortical regions are extracted as features. Meanwhile, total intracranial volume (ICV) and left and right hemisphere SA are also extracted. We remove the features with missing entries and a total of 327 MRI features are used in our experiments.

We use 814 subjects from the ADNI dataset, including 227 normal controls (NC), 397 MCI subjects, and 190 AD subjects. The detailed information of these subjects is summarized in Table 1. Among them, some subjects may miss ground-truth cognitive scores at certain time-points. We list the number of subjects with three types of cognitive assessments at four time-points in Table 2.

4.2. Algorithms for comparison

The LSFSIL method is compared with the following existing approaches:

- SVR: In this method, the support vector regression (SVR) is trained using the original MRI features without feature selection to infer the cognitive scores.
- MSL [21]: In this method, a matrix similarity-based regularizer is designed to take into account the relations between labels and between samples.
- L2PSC [42]: It uses $\ell_{2,p}$-norm to measure loss and select features. Moreover, a regularizer is introduced to preserve local structure information between samples and labels.
- RRFS [22]: In RRFS, relational regularizers, which incorporate feature-feature relation, sample-sample relation, and response-response relation, are combined with a $\ell_{2,1}$- norm regularizer for feature selection.
- RRDSL [43]: RRDSL is a discriminative learning method that incorporates relational information to explore the relations among features and training subjects.
- CSL [6]: In CSL, correlation-aware ℓ_1-norm is developed to explore the relations between imaging markers and cognitive scores for selecting informative features.

We also compare LSFSIL with three state-of-the-art semi-supervised feature selection methods:

- SFSGL [44]: SFSGL exploits graph Laplacian-based scatter matrix to make use of both labeled and unlabeled samples for feature selection in regression problems.

Table 2
The number of subjects with three types of cognitive scores (*i.e.*, ADAS-Cog, MMSE, and CDR-SB) at four time-points (*i.e.*, BL, M06, M12, and M24).

	ADAS-Cog				MMSE				CDR-SB			
	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24
Normal	227	219	206	199	227	219	209	201	227	215	205	196
MCI	394	374	354	299	397	378	356	302	397	378	356	300
AD	186	172	156	123	190	179	161	135	190	178	159	134

Table 3
Comparison of LSFSIL with existing approaches. SFSGL is a linear discriminant analysis method and utilized a trace operation based objective function.

	Does the method utilize incomplete labeled samples?	Does the method utilize the remaining scores in incomplete labeled samples?	The norm used to encode the regression errors
MSL			F-norm
L2PSC			$\ell_{2,p}$-norm
RRFS			F-norm
RRDSL			F-norm
CSL			F-norm
SFSGL	✓		–
GSFS	✓		$\ell_{2,p}$-norm
FSLCLC	✓		*F*-norm
LSFSIL	✓	✓	Nuclear norm

- GSFS [45]: GSFS applies $\ell_{2,p}$-norm on both loss function and regularization for utilizing labeled and unlabeled data based on manifold regularization.
- FSLCLC [46]: FSLCLC adopts the low-rank matrix factorization on the label matrix to compress labels and recover the missing labels.

Table 3 summarizes the advantages of LSFSIL over existing approaches. As illustrated in Table 3, only LSFSIL can take advantage of incomplete labeled samples and the remaining scores in incomplete labeled samples simultaneously. Besides, LSFSIL adopts the nuclear norm to encode the regression errors, which is robust to noises and outliers.

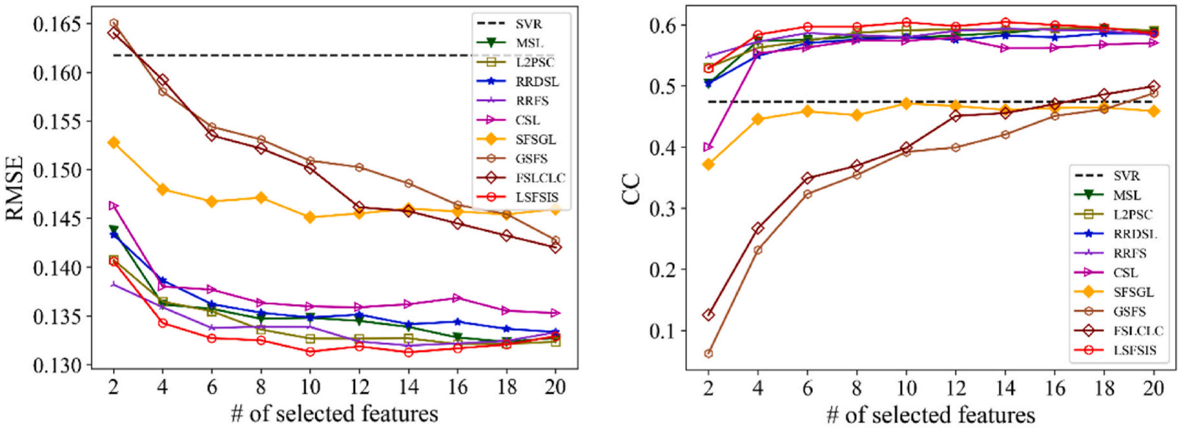
Besides, we compare LSFSIL with its variants to further validate the effectiveness of each component in LSFSIL.

- LSFSIL-S: It does not utilize subjects with incomplete cognitive scores.
- LSFSIL-R, LSFSIL-E, and LSFSIL-E: The three variants discard the second term, third term and fourth term in LSFSIL, respectively.

4.3. Experimental settings

We normalize all features by subtracting the minimum value and dividing the result by the difference between the maximum value and the minimum value so that all feature values are between 0 and 1. All cognitive scores are normalized (subtracted the maximum value) to avoid different response scales. Two metrics, *i.e.*, root mean square error (RMSE) and correlation coefficient (CC), are employed for performance evaluation.

As a convention [21,22], we use the SVR model with a linear kernel to evaluate the features selected by each algorithm. Default values for the SVR are used for both training and testing phases. The number of features is {2, 4, 6, …, 16, 18, 20} since there is no further performance improvement for larger values. We finetune parameters α, β, and γ by a grid-search strategy from $\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$. We use the five-fold cross-validation to evaluate all approaches. It means that all samples are equally divided into five portions. The samples in one portion are successively chosen as the testing data, and the rest are utilized as the training data. For semi-supervised methods (SFSGL, GSFS, and FSLCLC) and our proposed LSFSIL, all samples in the training set are used for training. For other methods, only the samples with complete cognitive scores can be used for training. When the training is finished, we check the prediction performance of the selected features in cognitive score prediction using the support vector regression (SVR) model. Specifically, a linear SVR model is trained with the selected features for each cognitive score and each method. For a fair comparison, we train the SVR models using the complete labeled samples for all methods. Then, the trained SVR is adopted to predict the scores of each sample in the testing set (only samples that have ground-truth scores can be predicted). The results of each fold are averaged. We repeat the whole process 10 times to avoid possible bias during dataset partitioning for cross-validation. All experiments are implemented with Python and conducted on an Intel Core (TM) i3-8100 CPU with 3.6 GHz processing speed and 8 GB main memory. Our code has been released at https://github.com/chenz96/LSFSIL.



(a) RMSE

(b) CC

Fig. 2. SVR prediction results of the comparing feature selection methods with varying numbers of selected features on the ADNI dataset. SVR in the legend means the SVR model on all features. The mean (a) RMSE and (b) CC are reported for each number of selected features.

4.4. *Experimental results*

Fig. 2 shows the variation of results of different methods with different numbers of selected features and we have the following observation. First of all, it is clear that, with the limited features, LSFSIL is superior to other approaches in terms of RMSE and CC, which demonstrates its effectiveness in cognitive score prediction. Moreover, the performance of these feature selection methods is not always improved as the number of selected features increases. As seen from Fig. 2, the performance of feature selection methods decreases slightly after reaching their peaks, showing that more features contain redundancy and noises in the selected feature subset. If we further increase the number of features, redundant and noisy features are selected, which may yield inaccurate modeling between MRI features and cognitive scores and then degrade prediction performance. However, the performance of LSFSIL decreases more significantly than other methods. The reason is that, for LSFSIL, almost all of the discriminative and key features that play important roles in the prediction task have been selected when the performance reaches its peak. If we further increase the number of selected features, redundant and noisy features are inevitably selected, which decreases the prediction performance. As a result, our method has an obviously local optimal performance. For other methods, several important features have not been selected at the peak. As we continue to increase the number of selected features, these important features that are previously omitted may be selected. Therefore, after reaching the peak, the performance degradation of other methods is smaller than that of our method. Besides, almost all feature selection methods outperform the baseline method, *i.e.*, SVR. This can be attributed to the fact that the redundant and noisy features in MRI data cause interference to the prediction model. The superior performance of feature selection methods demonstrates that they remove these features and select discriminative features. Consequently, it is necessary to select informative features before performing the cognitive score prediction. Compared with other semi-supervised methods (GSFS, SFSGL, and FSLCLC), LSFSIL yields better RMSE and CC. The reason may be that other semi-supervised methods are devised for the single-task learning problem. That is, these methods do not consider the correlations among cognitive scores in their formulations. In contrast, LSFSIL explores the correlations among cognitive scores with the Laplacian regularization term.

Tables 4 and 5 report the RMSE and CC results of different methods with top-10 selected features, respectively. As can be seen, LSFSIL outperforms all the other feature selection methods in terms of RMSE and CC. For example, the average RMSE and CC values of LSFSIL are 0.1323 and 0.5940, respectively. The results of the best comparison method (*i. e.*, L2PSC) are 0.1337 and 0.5900, respectively, and those of the worst one (*i.e.*, SVR) are 0.1617 and 0.4742, respectively. Besides, the performance of all methods decreases overtime. For example, LSFSIL obtains an RMSE value of 0.1454, which is higher than that at BL (*i.e.*, 0.1329) in predicting ADAS-Cog scores at M24. The reason may be that we use the MRI data at BL to predict the scores at four time-points but the brain structure may slightly change over time after BL and the individual differences in patients, such as age and education, may influence the progression of cognitive function. Therefore, it is reasonable that the performance slightly decreases overtime. Moreover, a paired *t*-test [47] at a significance level of 0.05 is performed to determine whether the performance differences between LSFSIL and the other feature selection methods are significant. We mark statistically significant differences with the superscript symbol *. As can be seen, LSFSIL is statistically better than SVR, MSL, RRDSL, RRFS, CSL, SFSGL, GSFS, and FSLCLC, and is at least comparable to L2PSC.

We list the RMSE and CC values of LSFSIL and its variants, *i.e.*, LSFSIL-S, LSFSIL-R, LSFSIL-E, and LSFSIL-M in Table 6. As can be seen, LSFSIL achieves better performance than its variants. In LSFSIL-S, the subjects with incomplete cognitive scores are discarded. In contrast, LSFSIL employs these subjects by clinical score matrix decomposition and therefore the relations between MRI features and clinical scores can be better described with more samples. Compared with LSFSIL-R, LSFSIL employs the $\ell_{2,1}$-norm to make $\mathbf{W}$ is sparse in rows. In this way, the most relevant features can be selected according to the norms of the rows in $\mathbf{W}$ and the performance can be improved. In LSFSIL-E, since the error matrix is unconstrained, its most elements are nonzero. However, only part of the clinical scores are missing and the existing clinical scores may be affected by the corresponding elements in the error matrix. As a result, the performance of LSFSIL-E is lower than that of LSFSIL. Compared with LSFSIL-M, LSFSIL adopts the manifold regularization term to guide similar subjects to own similar denoised cognitive scores. In this way, the local neighborhood of the cognitive scores is preserved in LSFSIL and therefore LSFSIL provides superior performance in feature selection.

5. Discussion

In this section, we first analyze the parameter sensitivity and convergence of LSFSIL. Then we show the discriminative brain regions identified by LSFSIL. Finally, we analyze the limitations and possible future directions of our work.

5.1. *Parameter analysis and convergence analysis*

Fig. 3 shows the performance variety under different α and the number of selected features when β and γ are fixed to 10^{-4} and 10^{-3}, respectively. As can be seen, as parameter α increases, the performance first increases and then decreases. When α is small, the $\ell_{2,1}$-norm regularization has little effect on $\mathbf{W}$ and $\mathbf{W}$ is not sparse in rows. Therefore, it is difficult to select important features based on $\mathbf{W}$.

$\mathbf{W}$. On the other hand, $\mathbf{W}$ is excessively sparse and all elements are mostly close to zero if α is too large, which makes it is no sense to select the features according to $\mathbf{W}$. According to Fig. 3, parameter α is set to 10^{-3}. Under the condition of parameters $\alpha = 10^{-3}$ and $\gamma = 10^{-3}$, the RMSE and CC results w.r.t. the number of selected features under the varying β are shown in Fig. 3. We observe that the prediction

Table 4

Prediction performance measured by RMSE. Superscript symbol * indicates that LSLFSIL significantly outperformed that method. Paired t -test at a level of 0.05 is used.

Method	ADAS-Cog				MMSE				CDR-SB				Avg.
	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24	
SVR	0.1710	0.1614	0.1620	0.1867	0.0797	0.1102	0.1437	0.1712	0.2120	0.1676	0.1820	0.1933	0.1617*
MSL	0.1376	0.1333	0.1341	0.1502	0.0781	0.1021	0.1162	0.1479	0.1743	0.1386	0.1459	0.1713	0.1358*
L2PSC	0.1359	0.1302	0.1305	0.1474	0.0779	0.1013	0.1130	**0.1417**	0.1748	0.1380	0.1460	0.1675	0.1337
RRDSL	0.1372	0.1342	0.1336	0.1516	0.0780	0.1021	0.1165	0.1482	0.1737	0.1388	0.1455	0.1706	0.1358*
RRFS	0.1351	0.1316	0.1326	0.1494	0.0775	0.1019	0.1148	0.1431	0.1759	0.1383	0.1479	0.1704	0.1349*
CSL	0.1408	0.1365	0.1357	0.1499	0.0785	0.1021	0.1163	0.1442	0.1802	0.1404	0.1479	0.1713	0.1370*
SFSGL	0.1491	0.1453	0.1446	0.1642	0.0805	0.1061	0.1252	0.1631	0.1845	0.1445	0.1526	0.1807	0.1450*
GSFS	0.1557	0.1502	0.1499	0.1731	0.0830	0.1085	0.1282	0.1668	0.1917	0.1513	0.1599	0.1927	0.1509*
FSLCLC	0.1559	0.1504	0.1506	0.1663	0.0832	0.1084	0.1294	0.1652	0.1914	0.1507	0.1610	0.1891	0.1501*
LSFSIL	**0.1329**	**0.1297**	**0.1291**	**0.1454**	**0.0762**	**0.1003**	**0.1128**	0.1426	**0.1730**	**0.1363**	**0.1442**	**0.1655**	**0.1323***

Table 5
Prediction performance measured by CC. Superscript symbol * indicates that LSLFSIL significantly outperformed that method. Paired t -test at a level of 0.05 is used.

Method	ADAS-Cog				MMSE				CDR-SB				Avg.
	BL	M06	M12	M24	BL	M06	M12	M24	BL	M06	M12	M24	
SVR	0.4517	0.5012	0.5055	0.5521	0.454	0.4534	0.3974	0.5327	0.4076	0.4379	0.4318	0.565	0.4732*
MSL	0.5972	0.6015	0.5991	0.6248	0.4923	0.5287	0.55	0.586	0.5436	0.5579	0.5625	0.5799	0.5776*
L2PSC	0.6051	0.6246	0.6223	0.6417	0.4942	0.5349	0.5799	0.6219	0.5354	0.5528	0.5552	0.6035	0.5900
RRDSL	0.5991	0.599	0.6048	0.6207	0.494	0.5332	0.5457	0.5854	0.5488	0.5591	0.5608	0.5904	0.5791*
RRFS	0.6105	0.6099	0.6056	0.6285	0.4958	0.5265	0.5581	0.6083	0.527	0.544	0.5349	0.5841	0.5784*
CSL	0.5736	0.5812	0.591	0.6321	0.4845	0.534	0.553	0.6141	0.5058	0.5451	0.5537	0.5974	0.5728*
SFSGL	0.5072	0.5105	0.5052	0.5286	0.4246	0.4352	0.4041	0.3775	0.4711	0.4867	0.4957	0.506	0.4700*
GSFS	0.4131	0.4263	0.4342	0.4385	0.3294	0.3709	0.3394	0.318	0.3923	0.4062	0.4196	0.4175	0.3911*
FSLCLC	0.4115	0.4354	0.4312	0.5025	0.3402	0.3827	0.3059	0.3353	0.3899	0.4129	0.3983	0.4392	0.3978*
LSFSIL	**0.6272**	**0.6273**	**0.6312**	**0.649**	**0.529**	**0.5519**	**0.5821**	**0.6236**	**0.5496**	**0.574**	**0.5682**	**0.6142**	**0.5940**

Table 6
Prediction performance of LSFSIL and its variants measured by RMSE and CC.

Method	RMSE	CC
LSFSIL-S	0.1356	0.5647
LSFSIL-R	0.1607	0.2700
LSFSIL-E	0.1340	0.5840
LSFSIL-M	0.1330	0.5909
LSFSIL	**0.1323**	**0.5940**

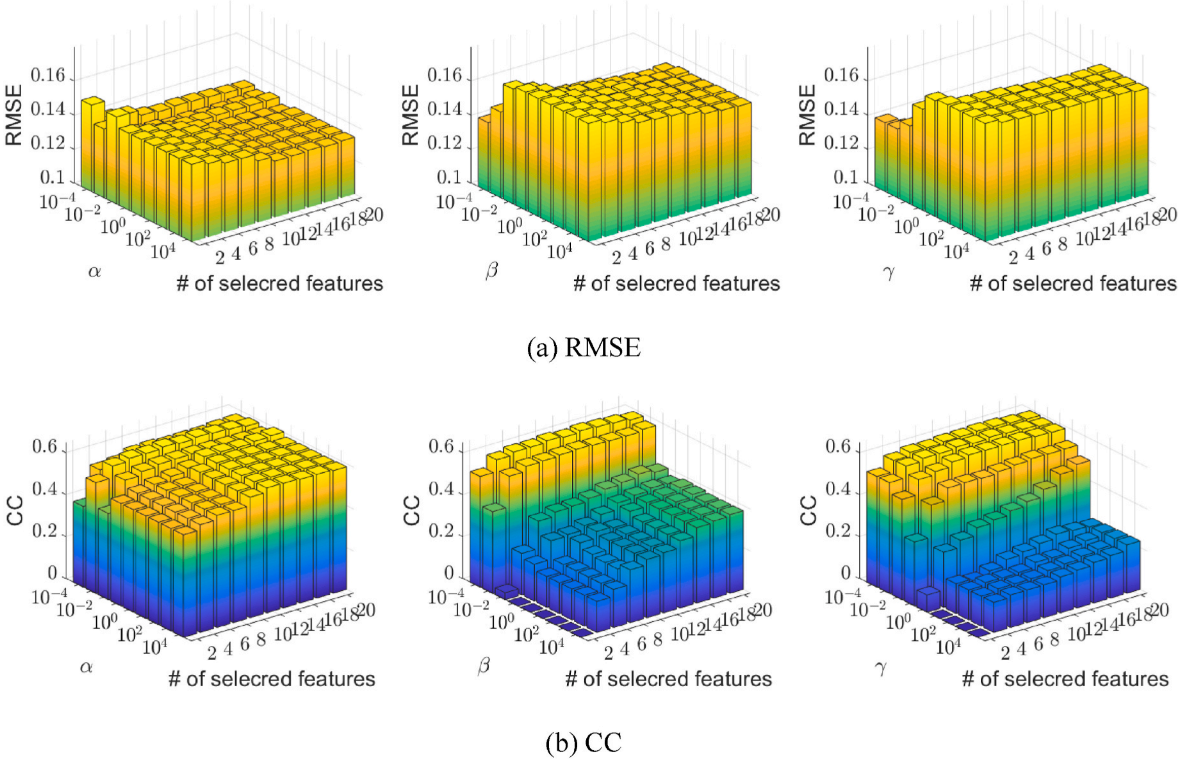
performance decreases with respect to parameter β. The reason may be that when β is too larger, $\mathbf{E}$ is excessively sparse and the recovered cognitive score matrix $\mathbf{Z}$ is very close to the original matrix $\mathbf{Y}$. Thus, it is no sense to optimize the loss function with $\mathbf{Z}$ as the target. Here, we set parameter β to 10^{-4}. We also show the prediction performance of γ and the number of selected features when α and β are fixed to 10^{-3} and 10^{-4} in Fig. 3. LSFSIL is robust to parameter γ in a wide range, *i.e.*, $\gamma \in [10^{-4}, 10^{-2}]$. When γ is too large, the performance of LSFSIL is significantly degraded. The reason may be that a large γ reduces the effects of other terms on the objective function and the selected features are not relevant to the task.

Here, we set γ to 10^{-3}.

Then, we experimentally explore the convergence of Algorithm 1. Fig. 4 shows the variation of the residual value with the increase of iterations. From Fig. 4, we can observe that the residual values reach a convergence within 200 iterations. The experimental results in Fig. 4 demonstrate the convergence of the proposed optimization algorithm.

5.2. Top selected brain regions

We analyze the top selected brain regions by LSFSIL. The brain regions with top occurrence frequency in all cross-validation are shown in Fig. 5. Some important brain regions are selected, such as hippocampus [48], middle temporal [49], entorhinal [50], and corpus callosum [51]. These ROIs are known to be highly related to AD and cognitive impairment in many previous studies. For example, the size of the hippocampus can be used to predict whether MCI will progress into AD [48]. Patients with temporal lobe epilepsy usually suffer significant memory deficits that appear similar to those seen in amnestic MCI [49]. Desikan et al. found that hippocampus volume and entorhinal cortex



(a) RMSE

(b) CC

Fig. 3. Average prediction performance of LSFSIL versus hyperparameters on the ADNI dataset. The y-axis represents α, β, and γ (from left to right), while x-axis represents the number of the selected features, and z-axis represents (a) RMSE and (b) CC.

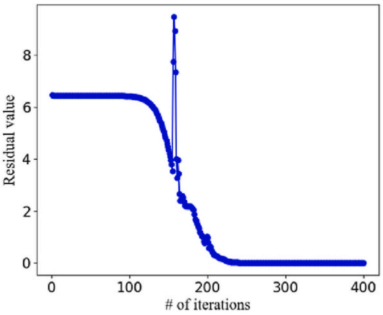
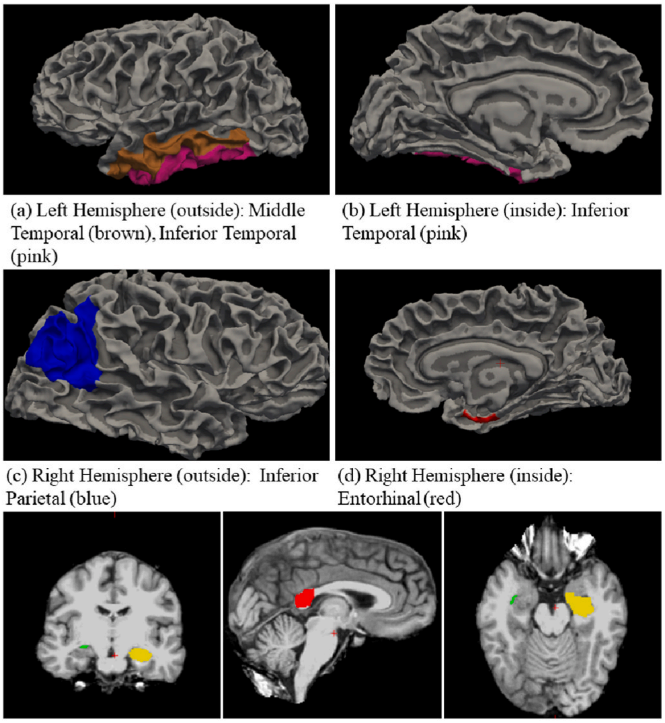


Fig. 4. Convergence curve of Algorithm 1 on the ADNI dataset.


(a) Left Hemisphere (outside): Middle Temporal (brown), Inferior Temporal (pink)

(b) Left Hemisphere (inside): Inferior Temporal (pink)

(c) Right Hemisphere (outside): Inferior Parietal (blue)

(d) Right Hemisphere (inside): Entorhinal (red)

(e) Subcortical: Right Inferior Lateral Ventricle (green), Left Hippocampus (yellow), Corpus Callosum Posterior (red)

Fig. 5. Most discriminative regions detected by LSFSIL. (a)–(d) are cortical ROIs and (e) is sub-cortical ROIs.

thickness can be used to identify MCI and AD individuals with high accuracy [50]. These results show that our findings are consistent with the results reported in previous studies [48–51], demonstrating the effectiveness of LSFSIL in identifying related features relevant to cognitive impairment.

5.3. Limitations and future directions

There are still several limitations in LSFSIL. First, we build the prediction model based on a single modality. Multi-modality data provide complementary information to each other and may help to promote prediction performance [53,54]. One important future direction is to develop regularization terms that can utilize the complementary and consensus properties of multi-modality data for the cognitive score prediction task. Besides, we assume that the features related to different cognitive assessments are the same. However, different assessments may prefer different features. For example, ADAS-Cog includes several additional assessment components targeting memory, praxis, and language compared with MMSE [11]. In future, we will focus on learning an assessment-specific projection matrix for each assessment so as to select an assessment-specific feature subset.

6. Conclusions

In this paper, we propose a new feature selection method named LSFSIL for AD progression prediction with incomplete cognitive scores, which is different from most existing methods that only employ the subjects with complete cognitive scores. To make full use of all available samples, we recover the real cognitive scores by decomposing the input original target matrix into two parts. The former is assumed to be the real cognitive score matrix without missing values and is regarded as the regression target. The latter is a matrix regularized by ℓ_1-norm for characterizing recovery errors. A manifold regularization term is developed to guide the decomposition by ensuring that similar subjects own similar recovered cognitive scores. Experimental results for the ADNI dataset suggest the superiority of our method in cognitive score prediction. Besides, we identify some important brain regions consistent with the previous studies.

Acknowledgments

This research was supported in part by the National Key R&D Program of China under grant 2019YFC1710300, and the Sichuan Science and Technology Program under grants 2019YFS0019, 2020YFS0283, 2021YFS0152, and 2021YJ0184. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI), USA (National Institutes of Health Grant U01 AG024904) and DOD ADNI, USA (Department of Defense award number W81XWH-12-2-0012). 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; BristolMyers 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.

Author contribution

Zhi Chen: Conceptualization, Methodology, Software, Writing - original draft. **Yongguo Liu**: Conceptualization, Funding acquisition, Writing - review & editing. **Yun Zhang**: Investigation, Methodology, Software. **Rongjiang Jin**: Visualization, Methodology, Supervision. **Jing Tao**: Visualization, Validation, Supervision. **Lidian Chen**: Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationship that could have appeared to influence the work reported in this paper.

References

[1] Z.S. Khachaturian, Diagnosis of Alzheimer's disease, BMJ 302 (6773) (1985) 1097–1105.

[2] Alzheimer's Association, et al., 2016 Alzheimer's disease facts and figures, Alzheimer's Dementia 12 (4) (2016) 459–509.

[3] R. Brookmeyer, et al., Forecasting the global burden of Alzheimer's disease, Alzheimer's Dementia 3 (3) (2007) 186–191.

[4] C.R. Jack, et al., Introduction to the recommendations from the national institute on Aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's Dementia 7 (3) (2011) 257–262.

[5] S. Tang, P. Cao, M. Huang, X. Liu, O. Zaiane, Dual feature correlation guided multi-task learning for Alzheimer's disease prediction, Comput. Biol. Med. 40 (105090) (2022).

[6] P. Jiang, X. Wang, Q. Li, Correlation-aware sparse and low-rank constrained multi-task learning for longitudinal analysis of Alzheimer's disease, IEEE J. Biomed. Health Inform. 23 (4) (2019) 1450–1456.

[7] Y. Zhao, et al., Prediction of Alzheimer's disease progression with multi-information generative adversarial network, IEEE J. Biomed. Health Inform. 25 (3) (2021) 711–719.

[8] L. Brand, et al., Joint multimodal longitudinal regression and classification for Alzheimer's disease prediction, IEEE Trans. Med. Imag. 39 (6) (2020) 1845–1855.

[9] W. Liang, K. Zhang, P. Cao, X. Liu, J. Yang, O. Zaiane, Rethinking modeling Alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network, Comput. Biol. Med. 138 (104935) (2021).

[10] T.N. Tombaugh, N.J. McIntyre, The mini-mental state examination: a comprehensive review, J. Am. Geriatr. Soc. 40 (9) (1992) 922–935.

[11] W.G. Rosen, R.C. Mohs, K.L. Davis, A new rating scale for Alzheimer's disease, Am. J. Psychiatr. 141 (1984) 1356–1364.

[12] J.C. Morris, The Clinical Dementia Rating (CDR): current version and scoring rules, Neurology 43 (11) (1993) 2412–2414.

[13] A.J. Larner, Assessment with cognitive screening instruments, in: Dementia in Clinical Practice: A Neurological Perspective, Springer, Cham, 2018, pp. 73–136.

[14] L. Zhuang, Y. Yang, J. Gao, Cognitive assessment tools for mild cognitive impairment screening, J. Neurol. 268 (5) (2021) 1615–1622.

[15] P. Cao, et al., Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in Alzheimers disease, Comput. Methods Progr. Biomed. 162 (2018) 19–45.

[16] L. Huang, et al., Longitudinal cognitive score prediction in Alzheimer's disease with soft-split sparse regression based random forest, Neurobiol. Aging 46 (2016) 180–191.

[17] H. Fukunishi, M. Nishiyama, Y. Luo, M. Kubo, Y. Kobayashi, Alzheimer-type dementia prediction by sparse logistic regression using claim data, Comput. Methods Progr. Biomed. 196 (105582) (2020).

[18] Z. Li, et al., Clustering-guided sparse structural learning for unsupervised feature selection, IEEE Trans. Knowl. Data Eng. 26 (9) (2014) 2138–2150.

[19] J. Zhou, et al., Modeling disease progression via multi-task learning, Neuroimage 78 (2013) 233–248.

[20] P. Cao, et al., Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease, Pattern Recogn. 72 (2017) 219–235.

[21] X. Zhu, H. Suk, D. Shen, A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis, Neuroimage 100 (2014) 91–105.

[22] X. Zhu, et al., A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, Med. Image Anal. 38 (2017) 205–214.

[23] L. Brand, et al., Joint multi-modal longitudinal regression and classification for Alzheimer's disease prediction, IEEE Trans. Med. Imag. 39 (6) (2020) 1845–1855.

[24] D. Zhang, D. Shen, Multi-modal multi-task learning for joint prediction of multiple regression and classification, Neuroimage 59 (2) (2012) 895–907.

[25] M. Liu, et al., Weakly supervised deep learning for brain disease prognosis using mri and incomplete cognitive scores, IEEE Trans. Cybern. 50 (27) (2020) 3381–3392.

[26] S. Aja-Fernandez, C. Alberola-Lopez, C. Westin, Noise and signal estimation in magnitude MRI and Rician distributed images: a LMMSE approach, IEEE Trans. Image Process. 17 (8) (2008) 1383–1398.

[27] F. Zhang, et al., Nuclear norm-based 2-DPCA for extracting features from images, IEEE Transact. Neural Networks Learn. Syst. 26 (10) (2015) 2247–2260.

[28] Y. Lu, et al., Low-rank discriminative regression learning for image classification, Neural Network. 125 (2020) 245–257.

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

[30] Y. Wang, Y. Fan, P. Bhatt, C. Davatzikos, High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables, Neuroimage 50 (4) (2010) 1519–1535.

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

[32] J.H. Bobholz, J. Brand, Assessment of cognitive impairment: relationship of the dementia rating scale to the mini-mental state examination, J. Geriatr. Psychiatr. Neurol. 6 (4) (1993) 210–213.

[33] X. Wang, X. Zhen, Q. Li, D. Shen, H. Huang, Cognitive assessment prediction in alzheimer's disease by multi-layer multi-target regression, Neuroinformatics 16 (3–4) (2018) 285–294.

[34] X. Zhen, M. Yu, X. He, S. Li, Multi-target regression via robust low-rank learning, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 497–504.

[35] E. Adeli, et al., Robust feature-sample linear discriminant analysis for brain disorders diagnosis, in: Proc. NIPS, 2015.

[36] E. Adeli, et al., Semi-supervised discriminative classification robust to sample-outliers and feature-noises, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 515–522.

[37] M. Iliadis, H. Wang, R. Molina, A.K. Katsaggelos, Robust and low-rank representation for fast face identification with occlusions, IEEE Trans. Image Process. 26 (5) (2017) 2203–2218.

[38] Z. Zhang, F. Li, M. Zhao, L. Zhang, S. Yan, Robust neighborhood preserving projection by nuclear/L2,1-norm regularization for image feature extraction, IEEE Trans. Image Process. 26 (4) (2017) 1607–1622.

[39] S. Mutasa, S. Sun, R. Ha, Understanding artificial intelligence based radiology studies: what is overfitting? Clin. Imag. 65 (2020) 96–99.

[40] S. Boyd, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations Trends Machine Learn. 3 (1) (2011) 1–122.

[41] M.W. Weiner, et al., The Alzheimer's disease neuroimaging initiative: progress report and future plans, Alzheimer's Dementia 6 (2010) 202–211.

[42] M. Zhang, et al., $\ell_{2,p}$-norm and sample constraint based feature selection and classification for AD diagnosis, Neurocomputing 195 (2016) 104–111.

[43] B. Lei, et al., Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis, IEEE Trans. Cybern. 47 (4) (2017) 1102–1113.

[44] R. Sheikhpour, M.A. Sarram, E. Sheikhpour, Semi-supervised sparse feature selection via graph Laplacian based scatter matrix for regression problems, Inf. Sci. 468 (2018) 14–28.

[45] R. Sheikhpour, et al., A robust graph-based semi-supervised sparse feature selection method, Inf. Sci. 531 (2020) 13–30.

[46] L. Jiang, et al., Feature selection with missing labels based on label compression and local feature correlation, Neurocomputing 395 (2020) 95–106.

[47] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998) 1895–1923.

[48] H. Eichenbaum, Hippocampus: cognitive processes and neural representations that underlie declarative memory, Neuron 44 (1) (2004) 109–120.

[49] E. Kaestner, "Atrophy and cognitive profiles in older adults with temporal lobe epilepsy are similar to mild cognitive impairment," Brain, vol. awaa397, pp. 1-15.

[50] R.S. Desikan, et al., Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease, Brain 132 (8) (2009) 2048–2057.

[51] B.J. Hallam, et al., Regional atrophy of the corpus callosum in dementia, J. Int. Neuropsychol. Soc. 14 (3) (2008) 414–423.

[53] X. Bi, et al., Multimodal data analysis of alzheimer's disease based on clustering evolutionary random forest, IEEE J. Biomed. Health Inform. 24 (10) (2020) 2973–2983.

[54] W. Lin, Q. Gao, M. Du, W. Chen, T. Tong, Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data, Comput. Biol. Med. 134 (104478) (2021).