Computer Methods and Programs in Biomedicine 208 (2021) 106286

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Feature level-based group lasso method for amnestic mild cognitive impairment diagnosis

Leiming Jin[a], Wenying Du[b], Baoqiang Ma[a], Debin Zeng[a], Ying Han[b,c,d,e,**], Shuyu Li[a,f,*]

[a] School of Biological Science and Medical Engineering, Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Beijing 100083, China
[b] Department of Neurology, Xuan Wu Hospital of Capital Medical University, Beijing 100053, China
[c] School of Biomedical Engineering, Hainan University, Haikou 570228, China
[d] Center of Alzheimer's Disease, Beijing Institute for Brain Disorders, Beijing 100053, China
[e] National Clinical Research Center for Geriatric Disorders, Beijing 100053, China
[f] State Key Lab of Cognition Neuroscience and Learning, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 2 January 2021
Accepted 10 July 2021

Keywords:
MCI
Multitask learning
Feature level correlation
Feature selection
Ensemble classification

ABSTRACT

Background and Objective: Previous studies have indicated that brain morphological measures change in patients with amnestic mild cognitive impairment (aMCI). However, most existing classification methods cannot take full advantage of these measures. In this study, we improve traditional multitask learning framework by fully considering the relevance among related tasks and supplementary information from other unrelated tasks at the same time.

Methods: We propose a feature level-based group lasso (FL-GL) method in which a feature represents the average value of each ROI for each measure. First, we design a correlation matrix in which each row represents the relationship among different measures for each ROI. And this matrix is used to guide the feature selection based on a group lasso framework. Then, we train specific support vector machine (SVM) classifiers with the selected features for each measure. Finally, a weighted voting strategy is applied to combine these classifiers for a final prediction of aMCI from normal control (NC).

Results: We use the leave-one-out cross-validation strategy to verify our method on two datasets, the Xuan Wu Hospital dataset and the ADNI dataset. Compared with the traditional method, the results show that the classification accuracies can be improved by 6.12 and 4.92% with the FL-GL method on the two datasets.

Conclusions: The results of an ablation study indicated that feature level-based group sparsity term was the core of our method. So, considering correlation at the feature level could improve the traditional multitask learning framework and our FL-GL method obtained better classification performance of patients with MCI and NCs.

1. Introduction

Alzheimer's disease (AD) is one of the most common types of dementia [1]. Previous studies have found that there will be nearly 100 million individuals with AD in 2050 worldwide [2]. Thus, considering that earlier detection might delay the progression of AD, it is important to diagnose the early stage of AD, mild cognitive impairment (MCI). In addition, amnestic MCI (aMCI), a subtype of MCI, is more likely to progress to AD because of primary memory impairment [3,4]. Therefore, early detection of aMCI is especially important because timely treatment can effectively delay the development of AD.

Surface measures have been widely used for the early detection of AD such as aMCI detection [5–7]. These measures have unique neuropathological and genetic bases and could be represented by the volume and geometry of the cerebral cortex [8–11]. They can be divided into two categories: volumetric and geometric measures. Volumetric measures include cortical thickness, surface area and gray matter (GM) volume, while geometric measures include sulcal depth, metric distortion, and average curvature.

[*] Corresponding author at: School of Biological Science and Medical Engineering, Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Beijing 100083, China.
[**] Corresponding author at: Department of Neurology, Xuan Wu Hospital of Capital Medical University, Beijing 100053, China.
E-mail addresses: hanying@xwh.ccmu.edu.cn (Y. Han), shuyuli@buaa.edu.cn (S. Li).

https://doi.org/10.1016/j.cmpb.2021.106286
0169-2607/© 2021 Elsevier B.V. All rights reserved.

Compared with NCs, aMCI patients show abnormal changes in these measures [12], such as thinning of the global cortex, widening of the sulcus [13] and a decrease in the average curvature of the temporal lobe. Moreover, all these measures have played a crucial role in aMCI and NC classification [14].

Generally, the brain can be partitioned into multiple regions of interest (ROIs) and different kinds of volumetric and geometric measures are extracted for each ROI. For each measure, a $d \times N$ data matrix can be obtained where d is the number of brain ROIs and N is the number of subjects. Previous methods did not take advantage of these measures when training the classifier model between aMCI and NC group. They simply concatenate or average these data matrices, which could introduce redundant information and ignore potential relationships among different measures. Thus, our previous work embed our data into a multitask learning framework by treating each measure as a task [15], which aims at learning related tasks simultaneously. A group least absolute shrinkage and selection operator (lasso) method is a typical multitask learning method that selects a subset of features for all related tasks by making whole rows in model W zeros [16]. This method assumes that all tasks are related to each other, which is not always true in practice. Therefore, our previous work used the robust multitask feature learning (rMTFL) method, which divides model W into two parts: the related tasks feature selection model P and the outlier task detection structure Q [17]. However, there are two shortcomings. First, this approach considers relevance at the task level. This means that if two tasks are related, then all the features in these tasks are related, which may not always be true in practice. Additionally, the nonzero columns in structure Q represent the outlier task, and in the feature selection part, the nonzero elements indicate that the features need to be selected. Thus, the rMTFL method selects almost all the features in outlier tasks, and this strategy is obviously not reasonable. To solve the above problems, a feature level-based group lasso (FL-GL) method is proposed. First, for each feature, we compute a vector that represents the correlation between different tasks. To eliminate the impact of other tasks, the partial correlation is used instead of the Pearson correlation. Then, we connect these vectors to a correlation matrix. Finally, we use this matrix to guide us to select features at the feature level. For related tasks, we select the same features across these tasks with group sparsity. Moreover, we also sparse other tasks to capture supplementary information.

In this study, the FL-GL method is used to identify MCI from NC based on multidimensional surface measures. Specifically, we first extract multidimensional surface measures for each subject with the FreeSurfer software and treat each measure as a task. Then, the FL-GL method is used to select features by comprehensively considering both measure relatedness and supplementary information provided by other unrelated measures at the feature level. Next, for each measure, we train a support vector machine (SVM) classifier [18] with the selected features and obtain the most suitable feature for the corresponding task. Finally, a weighted voting method is used to make a final prediction by fusing all the classifiers. We adopt this integration strategy since it is very popular and easy to implement [19,20]. To evaluate the validity of the above method, we conduct experiments on two datasets, and the results show the efficacy of our method in improving the diagnosis of AD. Moreover, we also perform an ablation study, and the results indicate that the feature level-based group sparsity term obviously improves the traditional multitask learning method.

In summary, the main contributions of our work are as follows: (1) we calculate task correlation at the feature level; (2) based on the correlation, we present a novel FL-GL method to classify MCI and AD and to obtain better results; and (3) we validate the proposed method on two datasets to obtain more comprehensive results.

Table 1
Demographic and clinical characteristics of the subjects.

	Xuan Wu Hospital dataset		ADNI dataset	
Subjects	aMCI ($n = 46$)	NC ($n = 52$)	MCI ($n = 69$)	NC ($n = 53$)
Gender (M/F)	24/22	21/31	32/37	20/33
Age	65.4 ± 9.5	63.0 ± 8.6	74.5 ± 6.3	75.3 ± 4.9
Education	10.1 ± 4.4	11.7 ± 4.4	15.6 ± 2.8	15.3 ± 3.5
MMSE	24.6 ± 4.0	28.5 ± 2.0	27.0 ± 1.6	29.1 ± 1.1

Age, education, and MMSE are shown as the mean ± SD. MMSE: mini-mental state examination. aMCI: amnestic mild cognitive impairment.

2. Materials and methods

To better verify the proposed feature selection method, we use the same flowchart proposed in our previous paper [15] it consists of three parts: feature extraction, feature selection and ensemble classification. An overview is provided in Fig. 1. In the following section, we describe each step in detail.

3. Datasets

3.1. Xuan Wu Hospital dataset

The first dataset in this research is the Xuan Wu Hospital dataset, which was approved by the Research Ethics Review Board of Xuan Wu Hospital. The inclusive criteria of aMCI were proposed by Petersen [21] and the diagnose of aMCI was affirmed by two experienced neurologists [22]. The measures used in this dataset were taken by the FreeSurfer software. Specifically, we first normalized the MRI data and corrected inhomogeneities. Afterwards, the skull was stripped by the watershed algorithm. Next, we segmented the images and performed deformation procedures. For example, the surface was inflated [23] and registered to a spherical atlas [24], and the cerebral cortex was partitioned into 148 regions [25]. Finally, the surface measures were computed in these regions.

This dataset contains 46 aMCI patients (24 males and 22 females) and 52 NCs (20 males and 32 females). The subjects were between the ages of 43 and 82 years and were right-handed. We carried out a statistical test, and the results showed that there were no significant differences ($p > 0.05$) between the aMCI patients and NCs in gender, age, or years of education, while the two groups showed significant differences in mini-mental state examination (MMSE) scores ($p < 0.01$). The statistical p values were analyzed using t tests for age, education and MMSE and chi-square tests for gender. The detailed demographic information and clinical characteristics involved in this research are summarized in Table 1.

3.2. ADNI dataset

The second dataset used in this research is the ADNI dataset (adni.loni.usc.edu). The ADNI dataset was created in 2003, and the primary goal of creating this dataset was to test and verify the possibility of the progression of MCI and early AD with various kinds of data, other biological markers, and clinical neuropsychological assessments. The measures used in this study are from the University of California San Francisco (UCSF) team, who used the FreeSurfer software for cortical reconstruction and volumetric segmentation on the imaging data according to the atlas in [26]. These measures are cortical volume, surface area, cortical thickness average and cortical thickness standard deviation, which are abbreviated as CV, SA, TA and TS, respectively.

We only selected the data without any missing values, and the selected dataset included 69 MCI patients (32 males and 37 females) and 53 NCs (20 males and 33 females). Then, a statistical test was carried out. The results show that there were no signifi-

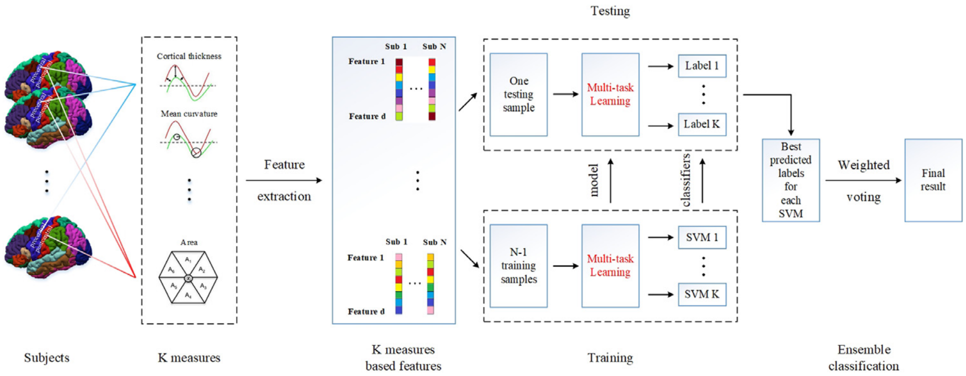


Fig. 1. The flowchart of our study. We first use FreeSurfer software to extract surface measures from each brain image. Then, we use the LOOCV strategy. These different feature selection methods are trained on the training set, and the feature selection models are applied on the testing set. Afterwards, for each measure, we train a specific SVM classifier. Finally, we use the weighted voting strategy to fuse the results of the above classifiers to make a final prediction.

cant differences ($p > 0.05$) between MCI patients and NCs in gender, age, or years of education, but the two groups showed significant differences in MMSE scores ($p < 0.01$). The detailed demographic information and clinical characteristics involved in this research are summarized in Table 1.

3.3. Feature extraction

As shown in the left part of Fig. 1, we partition the brain into multiple ROIs and extract different kinds of volumetric and geometric measures for each ROI. These measures include cortical thickness, surface area, gray matter volume, sulcal depth, metric distortion, and average curvature. For each measure, we compute a d-dimensional feature vector in which a feature represents the average value of each ROI, where d is the number of brain ROIs. Thus, for each subject, we have a feature matrix $X_n \in R^{K \times d}$, where K represents the number of measures. Obviously, we can get the data matrix $X_k \in R^{d \times N}$ of all subjects and all brain ROIs for each measure, where N is the number of subjects. In this way, each measure has a data matrix X_k.

3.4. Feature selection

The features obtained from the two datasets contain considerable redundant information, so we cannot directly use them for classification. Considering the relationship among these measures, in our study, we embed our data into a multitask learning framework by treating each measure as a task. In the following sections, we first briefly describe previous multitask feature learning methods. Then, the FL-GL method is presented.

3.4.1. Previous multitask feature learning method

In our study, there are K ($K = 6$ in the first dataset and $K = 4$ in the second dataset) measures, and we treat them as tasks. For the kth measure, $X_k \in R^{d \times N}$is denoted as input feature data that has N subjects, and each subject has a d-dimensional feature vector. Moreover, for these input data, we denote $Y_k \in R^N$ as a label vector. We also indicate the feature selection model $W \in R^{d \times K}$. The ith row and jth column of W are denoted as w^i and w_j, respectively. Then, by treating each measure as a task, we can embed our data into traditional multitask feature learning method that is proposed as follows [27–29]:

$$\min_{W} \frac{1}{2} \sum_{k=1}^{K} \left\| X_k^T w_k - Y_k \right\|_F^2 + \alpha \|W\|_{2,1}. \tag{1}$$

The first term of function (1) is the loss function, which computes the least square error between the predicted label and the true label. The second term is used to reduce the complexity of the model by penalizing the rows of the weight matrix W, where $\|W\|_{2,1} = \sum_{i=1}^{d} \|w^i\|_2$ is the $l_{2,1}$-norm of W. It can be computed by calculating the sum of the l_2-norms of w^i [30], which results in many rows with all zero elements. Thus, we only select the same features across all different tasks [30].

The above feature selection method assumes that there are no unrelated tasks, which may not be reasonable in practice. For this purpose, the rMTFL method, which can help us select a set of features across different related tasks and identify unrelated tasks simultaneously, is proposed. However, there are two shortcomings. First, this approach discusses relevance at the task level. This means that if two tasks are related, then all the features in these tasks are related. Additionally, the nonzero columns in structure Q represent the outlier tasks, and in feature selection, the nonzero elements indicate that the features need to be selected. Therefore, this method selects almost all the features in outlier tasks.

3.4.2. Feature level-based group lasso method

To solve the above problems, the FL-GL method is proposed. The key to this method is to find the feature level-based correlation matrix, and an overview of this part is provided in Fig. 2. First, for each feature, we compute the correlation between each pair of tasks. To eliminate the impact of other tasks, we use the partial correlation instead of the Pearson correlation. Then, we set the diagonal elements to 0 and add all rows together. In this way, we obtain a row vector, and the ith element represents the relevance between the ith task and all other tasks at the corresponding feature level. Next, we connect all these vectors into a matrix and set a threshold for it. Finally, we use this feature level-based correlation matrix to guide us to select features at the feature level with the help of the Hadamard product. The formulation is as follows:

$$\min_{W} \sum_{k=1}^{K} \left\| X_k^T w_k - Y_k \right\|_F^2 + \rho_1 \|A \odot W\|_{2,1} + \rho_2 \|(I - A) \odot W\|_1, \tag{2}$$

where $\odot$ is the Hadamard product, which denotes the product of the corresponding elements of two matrices. Matrix A is the feature-based correlation matrix, and all elements of matrix I are one.

The first term is the loss function without any changes. The second term is a feature level-based group sparsity term, and the third term penalizes the uncorrelated features. According to the previous introduction of matrix A, we can find that for each feature, an element equal to one or zero in A indicates a high or low correlation, respectively. Therefore, the second term is used to

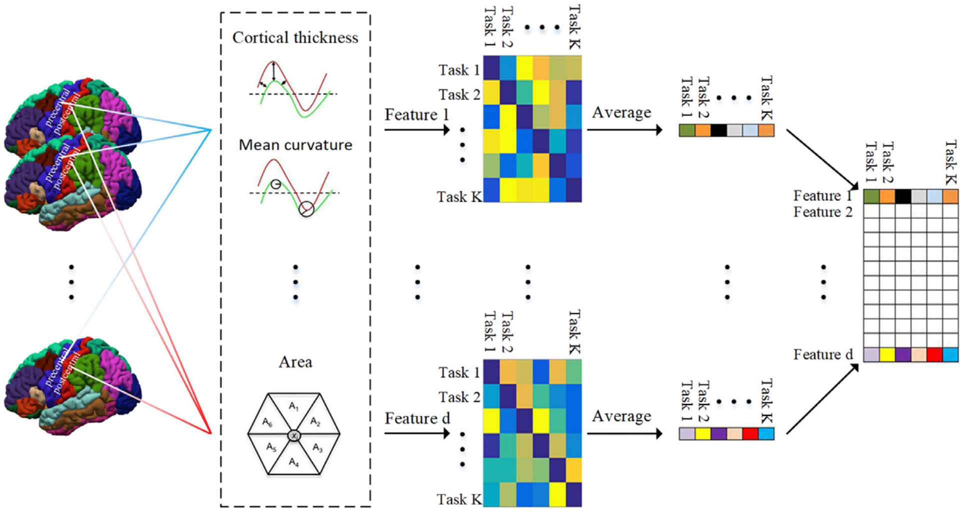


Fig. 2. The flowchart of the feature-based correlation matrix. First, for each feature (the value of each ROI for each measure), we use the partial correlation to compute the correlation between each pair of tasks and obtain a correlation matrix. Then, we set the diagonal element to zero and add all rows together to obtain a row vector. The *i*th element of this vector represents the relevance of the *i*th task and all other tasks. In this way, we can obtain several vectors, each corresponding to a specific feature. Finally, we connect all the vectors into a matrix and set a threshold for it.

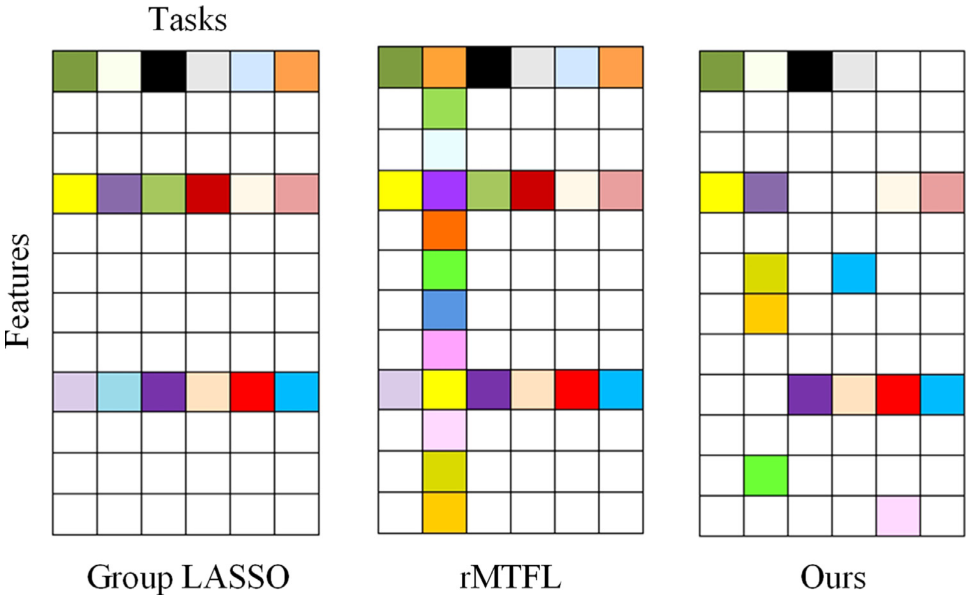


Fig. 3. The difference among the three predictive models. The group lasso method selects features across all different tasks. The rMTFL can select a set of features among related tasks and identify unrelated tasks simultaneously. However, this approach discusses relevance at the task level and selects almost all the features in the outlier tasks. To overcome these two shortcomings, our method can select shared features from related tasks and capture supplementary information from other tasks at the feature level.

select features across tasks with high correlations. Moreover, *I–A* denotes the low-correlation tasks, and these tasks are also useful for classification. Therefore, we use the l_1-norm to select features. Above all, our method selects shared features from related tasks and captures supplementary information from other tasks at the feature level. The differences among these three predictive models are shown in Fig. 3.

3.5. Optimization algorithms and convergence analysis

Although function (2) is convex, it is hard to solve due to the last two nonsmooth terms. Therefore, we use an efficient algorithm to solve it and prove convergence.

Taking the derivative of $w_i (1 \le i \le k)$ and setting it equal to 0, we obtain:

$$X_i X_i^T w_i - X_i y_i + \rho_1 D_i w_i + \rho_2 \bar{D}_i w_i = 0$$

where $D_i (1 \le i \le k)$ and $\bar{D}_i (1 \le i \le k)$ are diagonal matrices and their *k*th elements are $\frac{a_{ki}^2}{2\|(A \odot W)^k\|}$ and $\frac{|1-a_{ki}|}{2|w_{ki}|}$, respectively. Therefore,

$$w_i = \left(X_i X_i^T + \rho_1 D_i + \rho_2 \bar{D}_i\right)^{-1} X_i y_i. \quad (4)$$

Considering that D_i and $\bar{D}_i$ depend on *W*, we cannot obtain them directly. Hence, we adopt an iterative algorithm, listed in Algorithm 1, to solve the above problem.

In the following, we prove convergence.

Theorem 1 Algorithm 1 decreases the objective value of function (4) in each iteration.

Prove: According to Step 2 in the algorithm and function (4), we know that:

$$W^{(t+1)} = \min_W \mathrm{Tr}\left(X^T W - Y\right)^T \left(X^T W - Y\right) + \rho_1 \sum_{i=1}^{K} w_i^T D_i^{(t)} w_i + \rho_2 \sum_{i=1}^{K} w_i^T \bar{D}_i^{(t)} w_i. \quad (5)$$

Algorithm 1

```
Input: $X, Y, A$
Initial $\mathrm{W}^1 \in \mathrm{R}^{d \times K}$, $t = 1$;
While not converge do
Calculate the diagonal matrices $\mathrm{D}_i^{(t)}$ and $\bar{\mathrm{D}}_i^{(t)}$, whose $k$th diagonal elements are $\frac{a_{ki}^2}{2\|(A \odot W)^k\|}$ and $\frac{|1-a_{ki}|}{2|w_{ki}|}$, respectively;
For each $i(1 \le i \le k)$,
$w_i = (X_i X_i^T + \rho_1 D_i + \rho_2 \bar{D}_i)^{-1} X_i y_{i\cdot}$;
$t = t + 1$
Output: $\mathrm{W}^t \in \mathrm{R}^{d \times K}$.
```

Therefore, we have

$$
\begin{aligned}
&\mathrm{Tr}\left(X^T W^{(t+1)} - Y\right)^T \left(X^T W^{(t+1)} - Y\right) + \rho_1 \sum_{i=1}^{K} \left(w_i^{(t+1)}\right)^T D_i^{(t)} w_i^{(t+1)} + \rho_2 \sum_{i=1}^{K} \left(w_i^{(t+1)}\right)^T \bar{D}_i^{(t)} w_i^{(t+1)} \\
&\le \mathrm{Tr}\left(X^T W^{(t)} - Y\right)^T \left(X^T W^{(t)} - Y\right) + \rho_1 \sum_{i=1}^{K} \left(w_i^{(t)}\right)^T D_i^{(t)} w_i^{(t)} + \rho_2 \sum_{i=1}^{K} \left(w_i^{(t)}\right)^T \bar{D}_i^{(t)} w_i^{(t)}.
\end{aligned} \tag{6}
$$

According to the definition of $D_i^{(t)}$ and $\bar{D}_i^{(t)}$, we can rewrite (6) as

$$
\begin{aligned}
&\mathrm{Tr}\left(X^T W^{(t+1)} - Y\right)^T \left(X^T W^{(t+1)} - Y\right) + \rho_1 \sum_{k=1}^{d} \left(\left\| \left((A \odot W)^{(t+1)}\right)^k \right\|_2 + \frac{\left\| \left((A \odot W)^{(t+1)}\right)^k \right\|_2^2}{2\left\| \left((A \odot W)^{(t)}\right)^k \right\|_2} - \left\| \left((A \odot W)^{(t+1)}\right)^k \right\|_2 \right) \\
&+ \rho_2 \sum_{i=1}^{d} \sum_{j=1}^{K} |1 - a_{ij}| \left(\|w_{ij}^{(t+1)}\| + \frac{\left(w_{ij}^{(t+1)}\right)^2}{2\|w_{ij}^{(t)}\|} - \|w_{ij}^{(t+1)}\| \right) \\
&\mathrm{Tr}\left(X^T W^{(t)} - Y\right)^T \left(X^T W^{(t)} - Y\right) + \rho_1 \sum_{k=1}^{d} \left(\left\| \left((A \odot W)^{(t)}\right)^k \right\|_2 + \frac{\left\| \left((A \odot W)^{(t)}\right)^k \right\|_2^2}{2\left\| \left((A \odot W)^{(t)}\right)^k \right\|_2} - \left\| \left((A \odot W)^{(t)}\right)^k \right\|_2 \right) \\
&+ \rho_2 \sum_{i=1}^{d} \sum_{j=1}^{K} |1 - a_{ij}| \left(\|w_{ij}^{(t)}\| + \frac{\left(w_{ij}^{(t)}\right)^2}{2\|w_{ij}^{(t)}\|} - \|w_{ij}^{(t)}\| \right)
\end{aligned} \tag{7}
$$

Following [31], for any vector w and w_0, we have $\frac{\|w\|_2^2}{2\|w_0\|_2} - \|w\|_2 \ge \frac{\|w_0\|_2^2}{2\|w_0\|_2} - \|w_0\|_2$.
Therefore, we can rewrite (7) as

$$
\begin{aligned}
&\mathrm{Tr}\left(X^T W^{(t+1)} - Y\right)^T \left(X^T W^{(t+1)} - Y\right) + \rho_1 \sum_{k=1}^{d} \left\| \left((A \odot W)^{(t+1)}\right)^k \right\|_2 + \rho_2 \sum_{i=1}^{d} \sum_{j=1}^{K} \left\| ((I - A) \odot W)_{ij}^{(t+1)} \right\| \le \\
&\mathrm{Tr}\left(X^T W^{(t)} - Y\right)^T \left(X^T W^{(t)} - Y\right) + \rho_1 \sum_{k=1}^{d} \left\| \left((A \odot W)^{(t)}\right)^k \right\|_2 + \rho_2 \sum_{i=1}^{d} \sum_{j=1}^{K} \left\| ((I - A) \odot W)_{ij}^{(t)} \right\|.
\end{aligned} \tag{8}
$$

Function (2) is a convex problem and satisfies function (4). This means that *W* is a globally optimum solution. Hence, the proposed algorithm can decrease the training error and converges to the global optimum of function (2).

3.6. Ensemble classification

After the feature selection part, we use an ensemble classification method to obtain a better classification performance. Specifically, the dataset is divided into a training set and a testing set by the leave-one-out cross-validation (LOOCV) method. Different feature selection methods are trained on the training set, and these selected features are used to train the individual SVM classifier for each measure. And then the trained feature selection models and SVM classifiers are applied on the testing set. Finally, we use a weighted voting method to fuse the outputs of the classifiers and obtain the final prediction. Here, the weight is the normalized accuracy of each trained SVM classifier on the training set. In this study, we choose a linear SVM as the classifier because it is suitable for various training datasets from different fields [32–35].

3.7. Experiments and results

3.7.1. Experimental settings

We evaluate our proposed method based on the classification results of patients with MCI and NCs. To better evaluate the performance results of different methods, we use the LOOCV method in this study because it has two obvious advantages. (1) In each cross-validation process, almost all subjects take part in the training, so most of the data information is retained, and the results are convincing. (2) There are no random factors. Regardless of the number of experiments, the results do not change. Therefore, the experiment can be repeated by any researcher. Specifically, for each experiment, only one sample is selected as the testing set, and the others are the training set. We use only the training set to calculate matrix *A*, select features and train the SVM classifiers. The feature selection model is used in the testing set, and the selected features are input into the trained classifiers to obtain a label. We independently repeat this process *N* times to eliminate any biases caused by randomly segmenting the datasets in the LOOCV process. Then, we can obtain labels and the corresponding accuracy for each SVM. Finally, the weighted voting ensemble method is used to fuse the predicted labels of different classifiers based on their accuracies to make a final prediction. To obtain a better result, we use a grid search strategy to choose hyperparameters. There are three kinds of parameters in our method: the penalty term of the SVM classifier, feature selection parameters and the threshold of correlation matrix *A*. The range of the first two parameters is {2^{-10},........,2^{10}}, while the range of the last parameter is {0, 0.1,........, 1}.

We compare our approach with some existing multitask learning methods, including group lasso and rMTFL, on both datasets. Moreover, we perform an ablation study to better prove the effectiveness of the proposed method by setting the two parameters of feature selection (ρ_1,ρ_2) to zero. Specifically, we use a multitask learning framework to select features. For the above five methods, the overall process is the same. The only differences are in feature selection. Then, we train the SVM classifiers with the selected features, and each SVM corresponds to one measure. Finally, we use a weighted voting strategy to fuse all classifiers and give a final prediction. All the above feature selection methods are implemented by the MALSAR toolbox [36].

To compare the different methods, we adopted four criteria: classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and the area under the curve (AUC). Specifically, accuracy is the ratio of correctly classified samples to all samples and reflects the performance of the classifier. Sensitivity is defined as the ratio of correctly classified patients to all patients, and it can reflect the classifier's ability to distinguish patients. Similarly, specificity is the ratio of correctly classified NCs to all NCs. The receiver operating characteristic (ROC) curve can avoid the bias caused by the classification threshold in the SVM, and each point on this curve represents a specific decision threshold. Therefore, we use the AUC to comprehensively show the classification performance.

3.7.2. *Classification results*

The classification results of aMCI vs. NC on the Xuan Wu Hospital dataset and MCI vs. NC on the ADNI dataset can be found in Table 2.

It is clearly shown that the performance of our FL-GL method is better on both datasets. Specifically, on the first dataset, the accuracy, sensitivity, specificity, and AUC are 83.67%, 78.26%, 88.46% and 0.8311, respectively. The accuracies are 6.12 and 3.06% higher than those of the other two methods. Moreover, the AUC is apparently higher than those of the other methods. On the ADNI dataset, all methods perform better than they do on the first dataset, and the proposed method is still superior to the other methods. The accuracy, sensitivity, specificity, and AUC are 85.25%, 86.96%, 83.02% and 0.8690, respectively. The accuracies are 4.92 and 4.10% higher than those of the other two methods. To comprehensively show the classification performance, we also plot the ROC curves. Fig. 4 shows the curves of the different methods. From these figures, especially Fig. 4 (a), we can see that the curves of our method are closer to the (0,1) point (upper left corner), which means that the accuracy of the experiments is better. In summary, all criteria are better with our method than with the other methods on both datasets.

Table 2
Classification performance on the MRI datasets. (ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE) and Area Under Curve (AUC)).

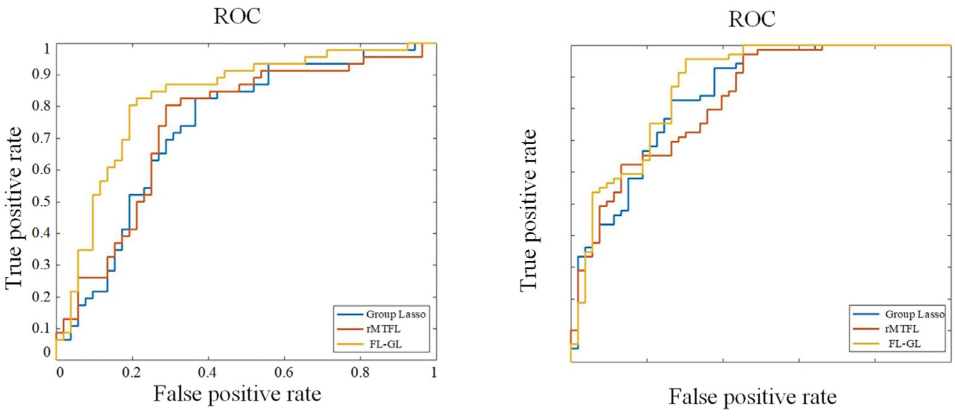
Method		ACC (%)	SEN (%)	SPE (%)	AUC
Xuan Wu	Group lasso	77.55	73.91	80.77	0.7366
Hospital	rMTFL	80.61	**78.26**	82.69	0.7446
dataset	Ours	**83.67**	**78.26**	**88.46**	**0.8311**
ADNI	Group lasso	80.33	82.61	77.36	0.8389
dataset	rMTFL	81.15	**86.96**	73.58	0.8253
	Ours	**85.25**	**86.96**	**83.02**	**0.8690**

Table 3
Classification performance of ablation experiments on MRI datasets. (ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE) and Area Under Curve (AUC)).

Method		ACC (%)	SEN (%)	SPE (%)	AUC
Xuan Wu	rho1=0	73.47	58.70	86.54	0.7646
Hospital	rho2=0	80.61	76.09	84.62	0.8303
dataset	Ours	**83.67**	**78.26**	**88.46**	**0.8311**
ADNI	rho1=0	80.33	82.61	77.36	0.8515
Dataset	rho2=0	84.43	**86.96**	81.13	0.8682
	Ours	**85.25**	**86.96**	**83.02**	**0.8690**

Moreover, to further evaluate our method, we performed an ablation study by setting the two parameters of feature selection (ρ_1,ρ_2) to zero. The classification results can be found in Table 3.

For both datasets, it can be clearly shown that when we set ρ_1 to zero, the accuracy drops sharply. This means that the second term of function (2) is much more important than the third term. It is the feature level-based group sparsity term used to extract the relationship among tasks. The third term is used to extract additional information from unrelated tasks. Therefore, the second term is much more important than the third term. Moreover, when we set ρ_1 to zero, the accuracy drops by 10.2% and 4.92% on the



(a) (b)

Fig. 4. (a) and (b) represent the ROC curves of the different methods on the Xuan Wu Hospital dataset and ADNI dataset, respectively. The blue line represents the group lasso method. The red line represents the robust multitask feature learning method. The last line represents our proposed method.

datasets, respectively, which means that when the task number is larger, the second term is particularly crucial. Finally, when we set ρ_2 to zero, our method is similar to the group lasso method, but the accuracy increases by 3.06% and 4.1%, respectively. Therefore, the feature level-based correlation matrix is a good guide for selecting features at the feature level.

4. Discussion

In the present study, we use the FL-GL method to classify patients with aMCI and NCs, and the results show that our method improves classification performance. Here, we interpret the reasons in two ways: (1) Traditional methods discuss relevance at the task level. This means that if two tasks are related, then the method assumes all features in these tasks are related, which may not always be true in practice. However, in our proposed method, we use a feature level-based relationship matrix to guide us to select a group of features across tasks at the feature level. (2) With the help of the l_1 norm, we also capture supplementary information from other unrelated tasks at the feature level.

Moreover, by setting the two feature selection parameters to zero, we find that the feature level-based group sparsity term is very important, and this is especially true for a large number of tasks. Moreover, the feature level-based relationship matrix plays an important role in the feature selection model.

However, we acknowledge that our study has two limitations. (1) The number of samples is relatively small in both datasets and cannot represent the pathological characteristics of a large number of patients. (2) For each feature, we average the features from all voxels, which may ignore some important information. Therefore, in our future work, we will consider applying our method to large datasets or other voxel-based datasets to reveal more significant results.

5. Conclusion

In this paper, the FL-GL method is proposed to classify patients with MCI and NCs; the FL-GL method makes full use of the task relationship and supplementary information from other unrelated tasks at the feature level. Specifically, we first extract features from several measures. Next, we compute the feature level-based correlation matrix, which is used for feature selection. Then, we train a specific SVM classifier for each measure. Finally, we use a weighted voting strategy to fuse the results of the above classifiers for a final prediction. The results show that the performance of our method is superior to previous methods on both datasets and that the feature level-based group sparsity term, which can significantly improve the results, is the core of the method. The last term is also important and can supply supplementary information from other unrelated tasks.

Declaration of Competing Interest

The authors have no relevant conflicts of interest to disclose.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 81972160, 61633018, 82020108013 and No. 81622025).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (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, the Alzheimer's Association; the Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; the Bristol-Myers Squibb Company; CereSpir, Inc.; 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; the Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; the 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

References

[1] D. Baskar, V.S. Jayanthi, A.N. Jayanthi, An efficient classification approach for detection of Alzheimer's disease from biomedical imaging modalities, Multimed. Tools Appl. 78 (10) (2018) 12883–12915.

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

[3] R.C. Petersen, et al., Current concepts in mild cognitive impairment, Arch. Neurol. 58 (12) (2001) 1985–1992.

[4] K.A. Jellinger, Mild cognitive impairment. Aging to Alzheimer's disease, Am. J. Psychiatry 10 (4) (2003) 466.

[5] J.P. Lerch, A.C. Evans, Cortical thickness analysis examined through power analysis and a population simulation, Neuroimage 24 (1) (2005) 163–173 Jan 1.

[6] L.G. Apostolova, et al., Three-dimensional gray matter atrophy mapping in mild cognitive impairment and mild Alzheimer disease, Arch. Neurol. 64 (10) (2007) 1489–1495.

[7] G.B. Frisoni, et al., The topography of grey matter involvement in early and late onset Alzheimer's disease, Brain 130 (Pt 3) (2007) 720–730 Mar.

[8] P. Rakic, Defects of neuronal migration and the pathogenesis of cortical malformations, Prog. Brain Res. 73 (73) (1988) 15–37.

[9] M.S. Panizzon, et al., Distinct genetic influences on cortical surface area and cortical thickness, Cereb. Cortex 19 (11) (2009) 2728–2735 Nov.

[10] P.R. Huttenlocher, Morphometric study of human cerebral cortex development, Neuropsychologia 28 (6) (1990) 517–527.

[11] D.C. Van Essen, A tension-based theory of morphogenesis and compact wiring in the central nervous system, Nature 385 (6614) (1997) 313–318.

[12] K. Im, J.M. Lee, S.W. Seo, S. Hyung Kim, S.I. Kim, D.L. Na, Sulcal morphology changes and their relationship with cortical thickness and gyral white matter volume in mild cognitive impairment and Alzheimer's disease, Neuroimage 43 (1) (2008) 103–113 Oct 15.

[13] T. Liu, et al., Longitudinal changes in sulcal morphology associated with late-life aging and MCI, Neuroimage 74 (2013) 337–342 Jul 1.

[14] S. Li, et al., Abnormal changes of multidimensional surface features using multivariate pattern classification in amnestic mild cognitive impairment patients, J. Neurosci. 34 (32) (2014) 10541–10553 Aug 6.

[15] L. Jin, X. Wang, P. Jiang, Q. Li, D. Zen, S. Li, Robust multitask feature learning for amnestic mild cognitive impairment diagnosis based on multidimensional surface measures, Med. Novel Technol. Devices 6 (100035) (2020) 1–8.

[16] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient L2,1-norm minimization, in: Proceedings of the Conference on Uncertainty in Artificial Intelligence AUAI Press, 2012, pp. 339–348.

[17] P. Gong, J. Ye, C. Zhang, Robust multi-task feature learning, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 895–903.

[18] C.C. Chang, C.J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–27.

[19] S.Modak Kumar, Singh, Jha, Vijay Kumar, Multibiometric fusion strategy and its applications: a review, Inf. Fus. 49 (2019) 174–204.

[20] M. Imran, A. Rao, G.Hemantha Kumar, Multibiometric systems: a comparative study of multi-algorithmic and multimodal approaches, Proc. Compt. Sci. 2 (2010) 207–212.

[21] R. Petersen, Mild cognitive impairment as a diagnostic entity, J. Intern. Med. 256 (3) (2010) 183–194.

[22] R.C. Petersen, G.E. Smith, S.C. Waring, R.J. Ivnik, E.G. Tangalos, E. Kokmen, Mild cognitive impairment: clinical characterization and outcome, Arch. Neurol. 56 (3) (1999) 303–308.

[23] A. Dale, Cortical surface-based analysis I. segmentation and surface reconstruction, Neuroimage 9 (2) (1999) 179–194.

[24] B. Fischl, Cortical Surface-based analysis II: inflation, flattening, and a surface-based coordinate system, Neuroimage 9 (2) (1999) 195–207.

[25] B. Fischl, et al., Automatically parcellating the human cerebral cortex, Cereb. Cortex 14 (1) (2004) 11–22.

[26] R.S. Desikan, et al., An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, Neuroimage 31 (3) (Jul 1 2006) 968–980.

[27] J. Baxter, A bayesian/information theoretic model of learning to learn via multiple task sampling, Mach. Learn. 28 (1) (1997) 7–39.

[28] X. Zhu, H.I. Suk, S.W. Lee, D. Shen, Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis, Brain Imaging Behav. 10 (3) (Sep 2016) 818–828.

[29] D. Zhang, D. Shen, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, Neuroimage 59 (2) (Jan 16 2012) 895–907.

[30] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Statist. Soc. Ser. B 68 (1) (2006) 49–67.

[31] F. Nie, H. Huang, X. Cai, C.H.Q. Ding, Efficient and Robust Feature selection via joint ℓ2, 1-norms minimization,", Proceeding of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems, 2010 December 2010.

[32] S. Kloppel, et al., Automatic classification of MR scans in Alzheimer's disease, Brain 131 (Pt 3) (2008) 681–689 Mar.

[33] M. Liu, D. Zhang, D. Shen, Ensemble sparse classification of Alzheimer's disease, Neuroimage 60 (2) (2012) 1106–1116 Apr 2.

[34] M. Liu, D. Zhang, D. Shen, Relationship Induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment, IEEE Trans. Med. Imaging 35 (6) (2016) 1463–1474 Jun.

[35] B. Jie, D. Zhang, C.Y. Wee, D. Shen, Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification, Hum. Brain Mapp. 35 (7) (2014) 2876–2897 Jul.

[36] J. Zhou, J. Chen, and J. Ye. (2012). *MALSAR: multi-task learning via structural regularization*. Available: http://www.MALSAR.org.