Neurological Sciences
https://doi.org/10.1007/s10072-021-05568-6

ORIGINAL ARTICLE

Multi-auxiliary domain transfer learning for diagnosis of MCI conversion

Bo Cheng[1,2] · Bingli Zhu[2] · Shuchang Pu[3]

Received: 28 March 2021 / Accepted: 14 August 2021

Abstract
In the early stage of Alzheimer's disease (AD), mild cognitive impairment (MCI) has a higher risk of progression to AD, so the prediction of whether an MCI subject will progress to AD (known as progressive MCI, PMCI) or not (known as stable MCI, SMCI) within a certain period is particularly important in practice. It is known that such a task could benefit from jointly learning-related auxiliary tasks such as differentiating AD from PMCI or PMCI from normal control (NC) in order to take full advantage of their shared commonality. However, few existing methods along this line fully consider the correlations between the target and auxiliary tasks according to the clinical practice of AD pathology for diagnosis. To deal with this problem, in this paper, treating each task domain as a different one, we borrow the idea from transfer learning and propose a novel multi-auxiliary domain transfer learning (MaDTL) method, which explicitly utilizes the correlations between the target domain (task) and multi-auxiliary domains (tasks) according to the clinical practice. Specifically, the proposed MaDTL method incorporates two key modules. The first one is a multi-auxiliary domain transfer-based feature selection (MaDTFS) model, which can select a discriminative feature subset shared by the target domain and the multi-auxiliary domains. In the MaDTFS model, to combine more training data from multi-auxiliary domains and simultaneously suppress the negative effects resulting from the irrelevant parts of multi-auxiliary domains, we proposed a sparse group correlation Lasso that includes a proposed group correlation Lasso penalty (i.e., $\|\mathbf{WH}\|_{2,1}$) and a proposed correlation Lasso penalty (i.e., $\|\mathbf{WH}\|_{1,1}$). The second module in MaDTL is a multi-auxiliary domain transfer-based classification (MaDTC) model that improves the voting with linear weighting-based ensemble learning. This model extends the constraints of the linear weighting method so that it can simultaneously combine training data from multi-auxiliary domains and achieve a robust classifier by minimizing negative effects from the irrelevant part of multi-auxiliary domains. Experimental results on 409 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database with the baseline magnetic resonance imaging (MRI) and cerebrospinal fluid (CSF) data validate the effectiveness of the proposed method by significantly improving the classification accuracy to 80.37% for the identification of MCI-to-AD conversion, outperforming the state-of-the-art methods.

Keywords Transfer learning · Multi-auxiliary domain · Feature selection · Mild cognitive impairment · Alzheimer's disease

Introduction

Alzheimer's disease (AD) is a type of degenerative brain disease, which can be characterized by a decline in memory, language, problem-solving, and cognitive skills about everyday activities [1]. In clinical practice, it is of great importance to identify dementia at the stage of mild cognitive impairment (MCI) for timely diagnosis and intervention of AD. MCI can be further subdivided into progressive MCI (PMCI) and stable MCI (SMCI) based on its conversion to AD. A practical challenge for MCI identification is that not all cases of MCI will eventually progress to AD. Thus, accurate prediction of MCI progression (i.e., PMCI) is

✉ Bo Cheng
cb729@nuaa.edu.cn

[1] Key Laboratory of Intelligent Information Processing and Control of Chongqing Municipal Institutions of Higher Education, Chongqing Three Gorges University, Chongqing 404100, China

[2] College of Computer Science and Engineering, Chongqing Three Gorges University, Chongqing 404100, China

[3] Department of Logistics Management, Chongqing Three Gorges University, Chongqing 404100, China

Published online: 12 September 2021

fundamentally essential for timely therapy, disease-modifying drug development, and possible delay of the disease. In recent years, many machine learning methods based on neuroimaging analysis have been proposed to recognize the early stage of AD [2–14]. For example, magnetic resonance imaging (MRI) scans [15–19] can measure structural brain atrophy and have been widely applied to the diagnosis of MCI [20]. Also, before the appearance of atrophy, the biological cerebrospinal fluid (CSF) levels of Aβ_{42}, total tau (t-tau), and phosphorylated tau (p-tau) have also been considered effective biomarkers in tracking MCI progression [21–25]. Many studies have shown that the combination of imaging and biological biomarkers can achieve better diagnosis performances than the methods using only the single-modal biomarker [9, 14, 26–31]. Accordingly, in this paper, we combine MRI and CSF biomarkers to identify the MCI-to-AD conversion.

Recently, deep learning methods based on neuroimaging data analysis have been used for the early diagnosis of AD [4, 8, 10, 32–35]. Although these deep learning-based studies can achieve better performances, the requirements of using a large training dataset and more powerful computational devices will lead to certain limitations in some applications. Actually, available training samples are generally very small, while the dimensionality of sample features is often very high, which makes it very challenging to train an accurate classifier model. This so-called small-sample-size problem has been one of the main challenges in neuroimaging data analysis. To address this problem, some studies have proposed advanced feature learning methods to reduce feature dimensionality [2, 5, 6, 11–13, 27, 29, 30, 36]. Feature selection methods are widely used in neuroimaging data analysis [6, 11, 27, 29, 30, 36–39]. Some of these studies have employed multi-task learning strategies for feature selection [11, 27, 29, 30, 37, 39], i.e., learning a common feature subset that can be well-generalized from a set of multi-task training data with a strong relationship among training data, e.g., multi-tasks sharing the same input data. On the other hand, transfer learning, especially domain adaptation strategy, can also be used to design feature selection or classification models in several recent studies [2, 3, 6, 13, 30, 40]. The basic idea of transfer learning is to utilize the knowledge learned from one or more auxiliary domains to aid the learning task in the target domain, with the assumption that these auxiliary domains are related to the target domain. Different from multi-task learning, transfer learning methods can somewhat relax the relationship required between target and auxiliary domains by explicitly handling the domain gap. Therefore, the transfer learning strategy is increasingly used in the early diagnosis of AD based on neuroimaging data analysis [2, 3, 6, 13, 30, 40]. In these studies on early diagnosis of AD based on transfer learning, some approaches focus on the instance-transfer approach that can work on new related datasets through training learning models on original dataset [3, 6, 40]; other studies focus on feature representation learning that can select a common set of features from the target domain and one or more related auxiliary domain(s) [2, 13, 30].

In this paper, we propose a learning framework that combines the ideas from multi-task learning and transfer learning to produce an accurate and robust classifier to predict PMCI and SMCI subjects. In particular, we consider five auxiliary tasks (i.e., AD vs. normal controls (NC), AD vs. PMCI, AD vs. SMCI, PMCI vs. NC, and SMCI vs. NC) to aid the classification of PMCI and SMCI subjects. It is noted that each task has a different data domain with different data distribution. We therefore further utilize transfer learning to explicitly handle the domain gap via requiring the model learned in the target domain to be close to those learned in the auxiliary domains, given the implicit relationship between them. Compared with our previous works in [2, 13, 30], the work in this paper improves the transfer feature selection model and the transfer classification model based on the concept of ensemble learning. More importantly, we explicitly consider the different correlations between the target domain and each auxiliary domain in our current work. Also, we aim to incorporate more data from multi-auxiliary domains to improve limitations in our previous work [13, 30]. However, when more data from multi-auxiliary domains are used, the irrelevant parts among these auxiliary domains will impose a negative impact on the classification performance. To solve this problem, we developed a novel multi-auxiliary domain transfer learning (MaDTL) method, which can effectively incorporate more training data from multi-auxiliary domains while mitigating the negative effects from several irrelevant parts of the auxiliary domains during the course of feature selection and classification.

In the proposed MaDTL model, we not only model the correlation between the target domain and multi-auxiliary domains based on the AD pathology but also improve the performance of conventional ensemble learning. Specifically, the proposed MaDTL method consists of two key modules. The first module is the multi-auxiliary domain transfer feature selection (MaDTFS). In this module, we propose a group correlation Lasso penalty (i.e., $\|\mathbf{WH}\|_{2,1}$) and a correlation Lasso penalty (i.e.,$\|\mathbf{WH}\|_{1,1}$) as regularizers, which improve the conventional sparse group Lasso [41] for feature selection. By doing so, our MaDTFS model can effectively combine training data from multi-auxiliary domains as well as suppress the negative effects resulting from irrelevant parts of these domains. The second one is a multi-auxiliary domain transfer-based classification (MaDTC) model, which improves voting with the conventional linear weighting-based ensemble learning. We extended the constraints of conventional linear weighting method and employed the hierarchical optimization-based

[13, 30] grid search method to learn the optimized weight variable on training data. The MaDTC model can achieve a robust and accurate classifier by utilizing more auxiliary domains for training and minimizing the negative effects from the irrelevant parts of these domains. The proposed method is evaluated on the baseline Alzheimer's Disease Neuroimaging Initiative (ADNI) database of 409 subjects with baseline magnetic resonance imaging (MRI) and CSF data. The experimental results demonstrate that the proposed method can further improve the diagnosis performance of MCI-to-AD conversion, compared with several state-of-the-art methods.

Subjects and method

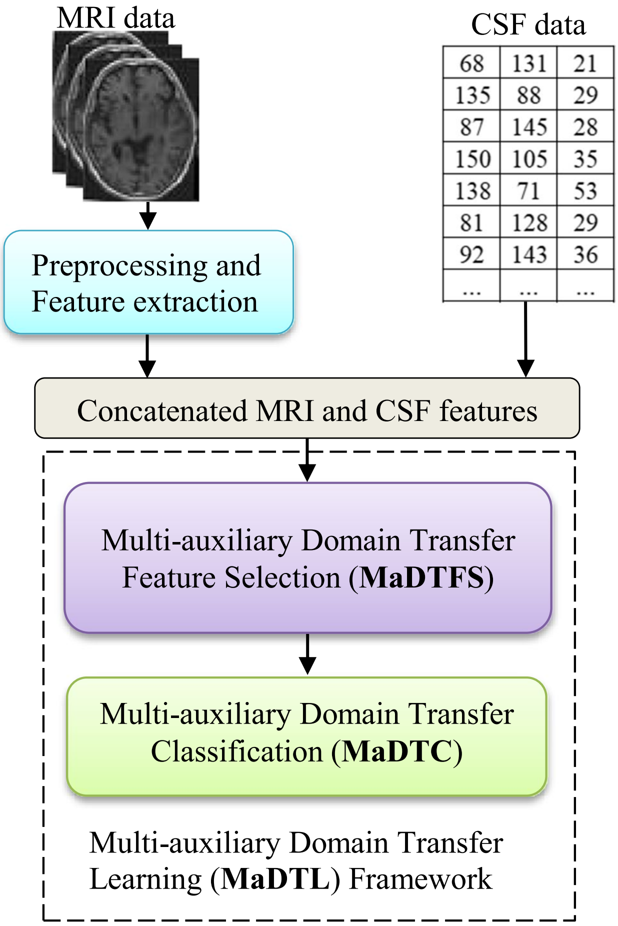
In this section, we first briefly introduce the experimental dataset from the ADNI database and then present our proposed MaDTL method framework and the mathematical theory for MaDTL.

Subjects

In this paper, we evaluate our method on the baseline structural MRI and CSF data of 409 subjects extracted from the ADNI database. A more detailed description of ADNI can be found in [13]. The ADNI study assesses participants in the following four stages: NC (i.e., normal aging/cognitively normal), SMC (i.e., significant memory concern), MCI (i.e., mild cognitive impairment), and AD (i.e., Alzheimer's disease). Since a new cohort is added to the ADNI database and denoted as ADNI2, the original ADNI database is denoted as ADNI1. In this work, we focus on using 409 subjects from the ADNI1 database with baseline MRI and CSF data. The 409 subjects include 102 AD subjects, 195 MCI subjects, and 112 normal control (NC) subjects. Among all 195 MCI subjects, during the 24-month follow-up period, 89 MCI subjects converted to AD, which are denoted as PMCI for short, and 106 MCI subjects remained stable, which are denoted as SMCI for short. We used the baseline CSF Aβ_{42}, t-tau, and p-tau data from the ADNI1 database. A more detailed description can be found in [14]. In this paper, CSF Aβ_{42}, CSF t-tau, and CSF p-tau are used as the features.

Overview of method

In Fig. 1, we provide an illustration of our proposed MaDTL framework for the diagnosis of MCI-to-AD conversion. At first, all structural MRI images are preprocessed for extracting features, and then these imaging-based features are concatenated with the CSF feature to form a new feature vector. The concatenated features are input into our proposed MaDTL framework for the classification of PMCI and SMCI subjects. Specifically, our transfer learning framework consists of two main modules, i.e., (1) MaDTFS module, which uses data from the target domain (i.e., consisting of PMCI vs. SMCI subjects) and five auxiliary domains (i.e., consisting of AD vs. NC, AD vs. PMCI, AD vs. SMCI, PMCI vs. NC, and SMCI vs. NC subjects, respectively) to form the training set and then learns the dimension-reduced feature vectors for the target and multi-auxiliary domains; and it should be noted that the PMCI and SMCI subjects in auxiliary domains are from the training set of the target domain; (2) MaDTC module, which uses the dimension-reduced feature vectors output by the MaDTFS module as


MRI data
CSF data
Preprocessing and Feature extraction
Concatenated MRI and CSF features
Multi-auxiliary Domain Transfer Feature Selection (MaDTFS)
Multi-auxiliary Domain Transfer Classification (MaDTC)
Multi-auxiliary Domain Transfer Learning (MaDTL) Framework

68	131	21
135	88	29
87	145	28
150	105	35
138	71	53
81	128	29
92	143	36
...	...	...

Fig. 1 Illustration of our proposed multi-auxiliary domain transfer learning (MaDTL) framework for the diagnosis of MCI-to-AD conversion (e.g., PMCI vs. SMCI). First, all MRI data of 409 subjects are preprocessed and features are extracted, and then these extracted features are concatenated with the original CSF feature to form a new feature vector. And then concatenated features are input into our proposed MaDTL framework for the classification of PMCI and SMCI subjects. The MaDTL framework includes two modules: (1) a MaDTFS model can combine data from the target domain and multi-auxiliary domains to select a discriminative feature subset; (2) a MaDTC model can achieve a robust classifier (PMCI vs. SMCI) by simultaneously combining data from the target domain and multi-auxiliary domains

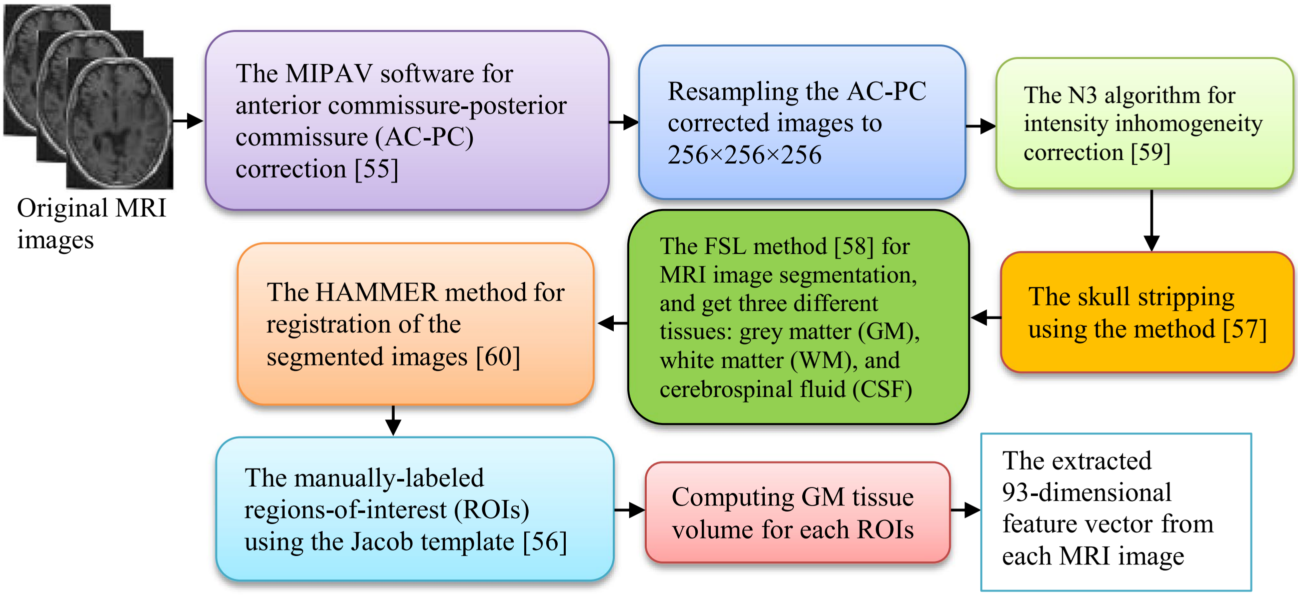

Original MRI images
The MIPAV software for anterior commissure-posterior commissure (AC-PC) correction [55]
Resampling the AC-PC corrected images to 256×256×256
The N3 algorithm for intensity inhomogeneity correction [59]
The skull stripping using the method [57]
The FSL method [58] for MRI image segmentation, and get three different tissues: grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF)
The HAMMER method for registration of the segmented images [60]
The manually-labeled regions-of-interest (ROIs) using the Jacob template [56]
Computing GM tissue volume for each ROIs
The extracted 93-dimensional feature vector from each MRI image

Fig. 2 The preprocessing flow of structural MRI images

the training data to learn support vector machine (SVM) classifier for each of the multi-auxiliary domains and the target domain, and output a final label vector. In order to properly utilize data from multi-auxiliary domains, we employ ensemble learning by voting with the linear weighting method to compute a final vector of the predicted value. Combining the MaDTFS module with the MaDTC module, we call it MaDTL. It is worth noting that hyperparameters of the MaDTFS and MaDTL models should be optimized by performing the MaDTL module with nested tenfold cross-validation on the current training data.

Image preprocessing and feature extraction

All structural MRI images are preprocessed by following the pipeline in the literature [14]. Specifically, the preprocessing flow is shown in Fig. 2. After registration, each subject's MRI image is labeled into 93 regions of interest (ROIs). Then, for each of the 93 ROIs, we compute its gray matter (GM) tissue volume as a feature. As a result, for each subject, we have a 93-dimensional feature vector to represent it. To fuse MRI and CSF features, we simply concatenated them into a long feature vector.

Multi-auxiliary domain transfer learning

To make use of the data from multi-auxiliary domains and simultaneously restrain negative effects from irrelevant parts of the multi-auxiliary domains, we propose a MaDTL model which simultaneously utilizes data from multi-auxiliary domains during the course of feature selection and classification. Specifically, as mentioned above, the proposed MaDTL model consists of two modules: (1) MaDTFS module; and (2) MaDTC module. Significantly, in the process of training the MaDTFS model, the PMCI and SMCI subjects in multi-auxiliary domains are from the training set of the target domain.

Multi-auxiliary domain transfer feature selection (MaDTFS)

Inspired by sparse group Lasso [41] and fused sparse group Lasso [36], we propose the MaDTFS module, which can capture an informative set of common features among data from both multi-auxiliary domains and the target domain. In the following subsections, we first introduce the formulation of MaDTFS and then employ the accelerated gradient descent (AGD) method [42, 43] to solve the optimization problem of the proposed MaDTFS model.

Assume that we have training data from the target domain $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$, with the i-th element $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ and its class label $y_i \in \{+1, -1\}$, where n denotes the number of training samples from the target domain and d denotes the dimensionality of sample feature vectors. Then, $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^{n\times 1}$ is the class label vector of the training set from the target domain $\mathbf{X}$. Also, we have a group of training data from multi-auxiliary domains $\left\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k\right\}$ with the i-th auxiliary domain data matrix $\mathbf{X}_i \in \mathbb{R}^{n_i\times d}$ and its class label vector $\mathbf{y}_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,n_i}]^T \in \mathbb{R}^{n_i\times 1}$, where n_i denotes the number of training samples from the i-th auxiliary domain and k is the number of auxiliary domains. It is noted that the dimensionality of feature vector d in the target domain is the same as that in each auxiliary domain. Therefore, some research employed the multi-task learning method by group Lasso penalty to capture a common set of features [27, 29, 37, 39]. Specifically, multi-task learning based on the group Lasso penalty (i.e., $L_{2,1}$-norm regularization term) adopted data from the target domain and multi-auxiliary domains $\{\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_k\}$ as well as a group of class label vectors $\{\mathbf{y}, \mathbf{y}_1, \dots, \mathbf{y}_k\}$ as training data and optimized the weight matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{k+1}] \in \mathbb{R}^{d\times(k+1)}$ ($\mathbf{w}_i$ is the i-th column vector of the weight matrix $\mathbf{W}$) to learn a common set of features. Formally, the multi-task learning model based on the group Lasso penalty [44] is formulated as:

$$\min_{\mathbf{W}} \sum_{i=1}^{k+1} \|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|_2^2 + \lambda\|\mathbf{W}\|_{2,1} \tag{1}$$

where $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{k+1} w_{i,j}^2}$ ($w_{i,j}$ is an element of the weight matrix $\mathbf{W}$) is a group Lasso penalty, the first term in Eq. (1) is an empirical loss function of the training data from all domains, and $\lambda > 0$ is a regularization parameter controlling the group sparseness of the weight matrix $\mathbf{W}$.

The group Lasso penalty of Eq. (1) tends to select features based on the strength and the commonality of the features over both the target domain and multi-auxiliary domains, and thus cannot select a specific set of features for each domain. In order to simultaneously select domain-specific features for multiple domains, several studies employed the multi-task learning-based sparse group Lasso penalty [41]. Formally, the multi-task learning-based sparse group Lasso penalty [41] is formulated as:

$$\min_{\mathbf{W}} \sum_{i=1}^{k+1} \|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|_2^2 + \lambda_1\|\mathbf{W}\|_{1,1} + \lambda_2\|\mathbf{W}\|_{2,1} \tag{2}$$

where $\|\mathbf{W}\|_{1,1} = \sum_{i=1}^{d} \sum_{j=1}^{k+1} \left|w_{i,j}\right|$ is well known as the Lasso penalty, and $\lambda_1, \lambda_2 > 0$ are regularization parameters.

In the model of Eq. (2), the combination of Lasso and group Lasso penalties is also known as a sparse group Lasso penalty, which allows simultaneous joint feature selection for all domains as well as the selection of a specific set of features for each domain. On the other hand, the models of Eq. (1) and (2) implicitly assume that these domains have sufficient correlation and similarity, regardless of whether the correlation of each pair of domains has a difference. However, in our task, the correlation between the target domain and each auxiliary domain could be different. Inspired by fused sparse group Lasso [36], we improve the model of Eq. (2) by introducing new regularizers by explicitly handling the correlation information between the target domain and each auxiliary domain and propose the MaDTFS model. Formally, the MaDTFS model is formulated as:

$$\min_{\mathbf{W}} \sum_{i=1}^{k+1} \|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|_2^2 + \lambda_1 \sum_{i=1}^{d} \sum_{j=2}^{k+1} \left|w_{i,1} - w_{i,j}\right| + \lambda_2 \sum_{i=1}^{d} \sqrt{\sum_{j=2}^{k+1} (w_{i,1} - w_{i,j})^2} \tag{3}$$

Here, the last term $\sum_{i=1}^{d} \sqrt{\sum_{j=2}^{k+1} (w_{i,1} - w_{i,j})^2}$ is a regularizer to represent the correlation between the target domain and all auxiliary domains, which is called a group correlation Lasso penalty in this paper. This regularizer can keep features with correlation on both the target domain and all auxiliary domains, and we called these features domain-common correlative features. Since we consider that the correlations between the target domain and each auxiliary domain are different, those selected domain-specific features are very important and thus we proposed a second term $\sum_{i=1}^{d} \sum_{j=2}^{k+1} \left|w_{i,1} - w_{i,j}\right|$. The regularizer is called a correlation Lasso penalty. The combination of group correlation Lasso and correlation Lasso penalties is called a sparse group correlation Lasso penalty, which can select those united features, i.e., both domain-common correlative features and domain-specific correlative features, and simultaneously restrain negative effects from several irrelevant auxiliary domains. For clarity, we illustrate the process of the MaDTFS model to learn an optimized weight matrix $\mathbf{W}$ in Fig. 3 and then acquire discriminative features via the optimized weight matrix $\mathbf{W}$.

To conveniently solve the optimization problem of the MaDTFS model, the model of Eq. (3) can be expressed as:

$$\min_{\mathbf{W}} \sum_{i=1}^{k+1} \|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|_2^2 + \lambda_1\|\mathbf{W}\mathbf{H}\|_{1,1} + \lambda_2\|\mathbf{W}\mathbf{H}\|_{2,1} \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{(k+1)\times k}$ is a $(k+1) \times k$ is a sparse matrix and defined as follows: $h_{i,j} = 1$ if $i = 1$ ($h_{i,j}$ is an element of the sparse matrix $\mathbf{H}$), $h_{i,j} = -1$ if $i = j + 1$, and $h_{i,j} = 0$ otherwise. By minimizing Eq. (4), we can learn a converged $\mathbf{W}$ from the target domain and multi-auxiliary domains, and then the elements of the optimized weight matrix $\mathbf{W}$ will be zero. For feature selection, we just keep those features with nonzero weights.

To solve the optimization problem of Eq. (4), we employ the accelerated gradient descent algorithm [42, 43]. To be specific, we decompose the objective function $F(\mathbf{W})$ in Eq. (4) into two parts, i.e., a differential term $L(\mathbf{W})$ and a non-differential term $R(\mathbf{W})$, as follows:

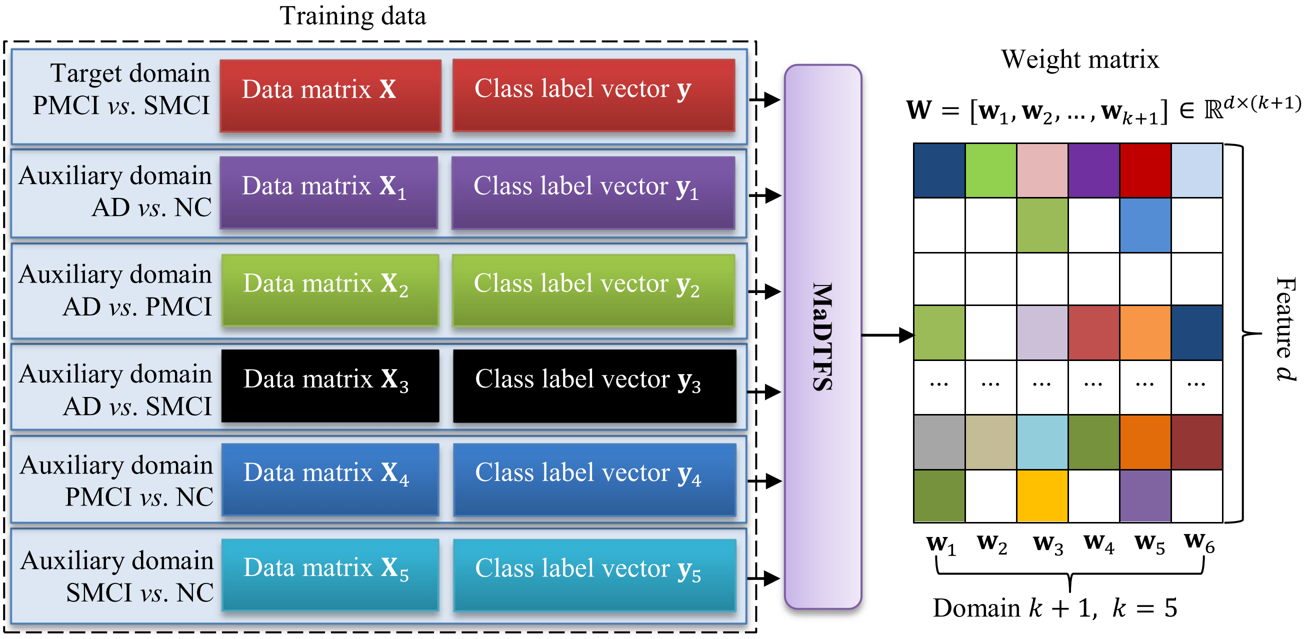


Fig. 3 Illustration of the MaDTFS model for selecting the discriminative features on training data from the target domain and multi-auxiliary domains. It is worth noting that the PMCI and SMCI subjects in multi-auxiliary domains are from the training set of the target domain

$$L(\mathbf{W}) = \sum_{i=1}^{k+1} \|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|_2^2, R(\mathbf{W}) = \lambda_1\|\mathbf{WH}\|_{1,1} + \lambda_2\|\mathbf{WH}\|_{2,1}$$
$$F(\mathbf{W}) = L(\mathbf{W}) + R(\mathbf{W}) \tag{5}$$

Then, we define the generalized gradient update rule as follows:

$$Q(\mathbf{W},\mathbf{W}_t) = L(\mathbf{W}_t) + tr\left((\mathbf{W}-\mathbf{W}_t)^T \nabla L(\mathbf{W}_t)\right) + \frac{1}{2}\|\mathbf{W}-\mathbf{W}_t\|_F^2 + R(\mathbf{W}) \tag{6}$$

In addition, we set $q(\mathbf{W}_t)$ as the following:

$$q(\mathbf{W}_t) = \mathrm{arg}min_{\mathbf{W}} Q(\mathbf{W},\mathbf{W}_t) \tag{7}$$

Specifically, we summarize the details of the AGD method for the optimization problem of the MaDTFS model in Algorithm 1.

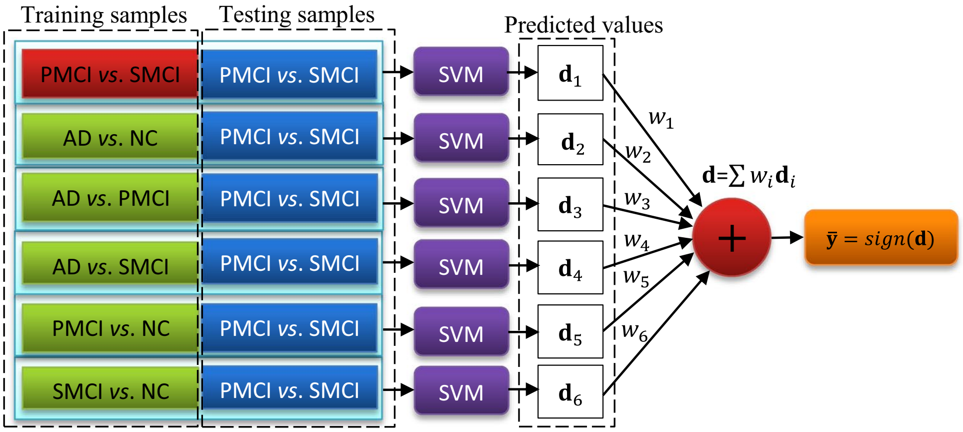


Fig. 4 Illustration of the multi-auxiliary domain transfer classification (MaDTC) model for classification on training data from the target domain and multi-auxiliary domains and testing data only from the target domain. In the process of MaDTFS, we have acquired a group of the most discriminative features to reduce the dimension of training and testing samples. To make better use of data from multi-auxiliary domains, we train the six SVM models on training data from the target domain and five auxiliary domains of feature reduction, then we apply the six learned SVM to test testing samples from the target domain and acquire six vectors of predicted value $(d_1, \ldots, d_6)$. Finally, we employ voting with the linear weighting method to compute a final vector of predicted value d and acquire a final classification label vector of testing samples $\bar{y}$

Algorithm 1. AGD algorithm for MaDTFS in Eq. (4)

```
Initialization: $\mathbf{W}_0 \in \mathbb{R}^{d\times(k+1)}$, $\mathbf{V}_0 = \mathbf{W}_0$, $\eta > 0$, and $\alpha_0 = 1$.
for $t = 0,1,2,\ldots$ until convergence of $\mathbf{W}_t$ do:
    while $F(q(\mathbf{V}_t)) > Q(q(\mathbf{V}_t), \mathbf{V}_t)$ and Compute:
            $\mathbf{W}_{t+1} = q(\mathbf{V}_t)$
            $\alpha_{t+1} = \frac{2}{t+3}$, $\beta_{t+1} = \mathbf{W}_{t+1} - \mathbf{W}_t$
            $\mathbf{V}_{t+1} = \mathbf{W}_{t+1} + \frac{1-\alpha_t}{\alpha_t}\alpha_{t+1}\beta_{t+1}$
    end-while
end-for
```

To solve the generalized gradient update efficiently, according to [42], $q(\mathbf{V}_t)$ in Algorithm 1 can be rewritten as

$$q(\mathbf{V}_t) = \mathrm{argmin}_{\mathbf{W}}\left(\frac{1}{2}\|\mathbf{W} - (\mathbf{V}_t - \eta\nabla L(\mathbf{V}_t))\|_F^2 + \lambda_1\|\mathbf{WH}\|_{1,1} + \lambda_2\|\mathbf{WH}\|_{2,1}\right) \quad (8)$$

Setting $\mathbf{Z} = \mathbf{V}_t - \eta\nabla L(\mathbf{V}_t)$, Eq. (8) can be expressed as

$$q(\mathbf{V}_t) = \arg\min_{\mathbf{W}}\left(\frac{1}{2}\|\mathbf{W} - \mathbf{Z}\|_F^2 + \lambda_1\|\mathbf{WH}\|_{1,1} + \lambda_2\|\mathbf{WH}\|_{2,1}\right)$$
$$= \arg\min_{\mathbf{w}^1,\ldots,\mathbf{w}^d} \sum_{i=1}^{d}\left(\frac{1}{2}\|\mathbf{w}^i - \mathbf{z}^i\|_2^2 + \lambda_1\|\mathbf{w}^i\mathbf{H}\|_1 + \lambda_2\|\mathbf{w}^i\mathbf{H}\|_2\right) \quad (9)$$

where $\mathbf{w}^i$ and$\mathbf{z}^i$ denote the i-th row vector of the matrix $\mathbf{W}, \mathbf{Z}$, respectively. Therefore, Eq. (8) can be decomposed into d separate subproblems of dimension $k + 1$.

For each subproblem of Eq. (9):

$$\min_{\mathbf{w}}\frac{1}{2}\|\mathbf{w} - \mathbf{z}\|_2^2 + \lambda_1\|\mathbf{wH}\|_1 + \lambda_2\|\mathbf{wH}\|_2 \quad (10)$$

Inspired by solving the proximal operator associated with the fused sparse group Lasso [36], solving the optimization problem of Eq. (10) is presented in Algorithm 2. It is worth reminding that the vectors w, u, z are row vectors in Algorithm 2 and Eq. (9).

Algorithm 2. Proximal operator associated with the MaDTFS

```
Input: $\mathbf{Z}, \mathbf{H}, \lambda_1, \lambda_2$
Output: $\mathbf{W}$
for $t = 0,1,2,\ldots,t$ do:
    $\mathbf{u}^i = \arg min_{\mathbf{w}}\frac{1}{2}\|\mathbf{w} - \mathbf{z}^i\|_2^2 + \lambda_1\|\mathbf{wH}\|_1$
    $\mathbf{w}^i = \arg min_{\mathbf{w}}\frac{1}{2}\|\mathbf{w} - \mathbf{u}^i\|_2^2 + \lambda_2\|\mathbf{wH}\|_2$
end-for
```

According to Algorithm 1 and Algorithm 2, incorporating the research work of [36], the optimal solution $\mathbf{W}$ can be easily obtained. And then, the optimized weight matrix $\mathbf{W}$ can be used to select the most discriminative features. Specifically, we select those features corresponding to the nonzero rows in the first column of the optimized weight matrix $\mathbf{W}$.

Multi-auxiliary domain transfer classification

After performing the process of MaDTFS, we have acquired a group of the most discriminative features for reducing the dimension of training and testing samples, upon which we will build a MaDTC for the final classification. In Fig. 4, we provide an illustration of the MaDTC model for the final classification. Inspired by the technique of transductive transfer learning [45, 46], we developed the MaDTC model in order to properly utilize the dimension-reduced training data from multi-auxiliary domains. Specifically, there are three steps to compute the final class label of testing samples. In the first step, we treat data from each auxiliary domain and target domain as a training set in turn, while the testing set is the same and from the target domain. In the second step, we train a SVM classifier with the training set, then get $k + 1$ SVM classifier models, and output $k + 1$ groups of predicted value vectors. We denote the $k + 1$ groups of predicted value vectors as $\{\mathbf{d}_i\}_{i=1}^{k+1}$. Finally, in order to achieve better classification performance, we employ ensemble learning by voting with the linear weighting method to combine these $k + 1$ vectors of the predicted value. Formally, the voting with linear weighting method to fuse the $k + 1$ groups of predicted value vectors is formulated as:

$$\mathbf{d} = \sum_{i=1}^{k+1} w_i\mathbf{d}_i \quad (11)$$

where the learnable variable w_i is the weight corresponding to the i-th vector of predicted value $\mathbf{d}_i$, $\mathbf{d}$ is a final vector of predicted value, and a hierarchical optimization method-based [13, 30] grid search is employed to learn the optimized weight variables w_i in training the MaDTL model. Then, we directly call the sign function to compute the final classification label vector of testing samples $\overline{\mathbf{y}}$ (i.e.,$\overline{\mathbf{y}} = \mathrm{sign}(\mathbf{d})$).

To achieve better classification performance, we constrain the weights to be $\sum_{i=1}^{k+1} w_i = 1$ and $-1 \le w_i \le 1$. There is a reason that we constrain the weight w_i. Specifically, in data from multi-auxiliary domains, usually, there are one or more auxiliary domains that can be weakly correlative or irrelevant to the target domain, so if we directly use the voting with linear weighting method in conventional ensemble learning, we may have the negative transfer effect on classification performance of target domain. Different from conventional ensemble learning, those weak correlative or irrelevant auxiliary domain data can be used for learning performance improvement of the target domain in transfer learning. According to the pathology of AD, we consider that correlations between target domain and each auxiliary domain have different, and a few irrelevant auxiliary domains often exist in all auxiliary domains but these irrelevant auxiliary domains may result in negative effects on classification performance. To avoid negative effects from irrelevant auxiliary domains, we extend the restricted condition of conventional voting with linear weighting method and set the range of the weight w_i as between − 1 and 1.

Hierarchical optimization method for hyperparameter learning

In the subsections "Multi-auxiliary domain transfer feature selection (MaDTFS)" and "Multi-auxiliary domain transfer classification," we have described in detail our proposed MaDTL method and its components (i.e., MaDTFS and MaDTC). However, we do not describe how to use our proposed MaDTL method for the diagnosis of MCI-to-AD conversion in real-world applications. In order to more visually understand the process of training and testing our proposed MaDTL method for diagnosis of MCI-to-AD conversion, we provide an illustration for showing this process in Fig. 5.

In Fig. 5, there are three steps to training and testing the MaDTL model. Firstly, we use a tenfold cross-validation strategy to partition the target domain (i.e., 195 subjects of PMCI and SMCI) samples into training and testing subsets; these training samples from the target domain are added to 214 AD and NC samples and then form five auxiliary learning tasks. Secondly, to search reasonable values of hyperparameters for the MaDTFS and MaDTC models, we employ a hierarchical optimization method-based [30] grid search to search the optimized regularization parameters (λ_1, λ_2) and weight variables w_i. Specifically, for each fold of training samples from the target domain, we again employ the tenfold cross-validation (i.e., nested tenfold cross-validation) strategy to partition the current training samples from the target domain and combine auxiliary domain samples into a new set of training samples, then we perform the hierarchical optimization method-based grid search for hyperparameter learning. Thirdly, through hyperparameter learning, these optimized hyperparameters ($\lambda_1, \lambda_2, w_1, \ldots, w_6$) have been acquired then input into the MaDTL (MaDTFS + MaDTC) model and obtain a group of performance measures (ACC, SEN, SPE, and AUC) on each fold. Through performing tenfold cross-validation, we obtain ten groups of performance measures and then compute the mean value of all groups of performance measures. It is worth noting that we perform this process of tenfold cross-validation 10 times in random order and report the average of all performance measures (ACC, SEN, SPE, and AUC).

The hierarchical optimization method-based grid search is employed to search for more optimal hyperparameters of the MaDTL model. Specifically, there are two regularization parameters (i.e., λ_1, λ_2) in the MaDTFS model and six weight parameters (i.e., $w_1, w_2, w_3, w_4, w_5, w_6$) in the MaDTC model. Firstly, we use the grid search method to optimize the first regularization parameter λ_1, while we fix default values for other rest parameters and acquire the optimized parameter λ_1. Then, we use the grid search method to optimize the parameter λ_2 with the optimized parameter λ_1 and the fixed default values for other rest parameters. All hyperparameters are optimized one time, which is an iteration for the hierarchical optimization method-based grid search. In general, through multiple iterations, the performance measures would converge, and we set the number of iterations as a constant.

Experimental settings

The classification of PMCI and SMCI subjects is the target domain in our task, i.e., PMCI (+ 1) vs. SMCI (− 1) classification. In addition, we further consider five binary classification tasks as auxiliary domains, i.e., AD (+ 1) vs. NC (− 1) classification, PMCI (+ 1) vs. NC (− 1) classification, AD (+ 1) vs. PMCI (− 1) classification, SMCI (+ 1) vs. NC (− 1) classification, and AD (+ 1) vs. SMCI (− 1) classification, to improve our ultimate task of PMCI and SMCI classification. In order to avoid the possible bias that occurred during sample partitioning, a tenfold cross-validation strategy is used to partition the target domain data into the training and testing subsets in all experiments. In the tenfold cross-validation, we train and test our model 10 times in random order and report the average performances in terms of area under the receiver operating characteristic curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE).

We compare the proposed MaDTL method with the standard SVM (denoted as SVM) and other state-of-the-art methods, including those using multi-task learning and transfer learning. These methods are as follows: (1) MKSVM [14], (2) MTFS [37], (3) cFSGL [36], (4) Lasso [47], (5) sgLasso [41], (6) rMLTFL [2], and (7) MDTL [13]. Experiment settings of these methods are listed as follows.

SVM: The training data are only from the target domain, without any feature selection stage. The linear SVM with $C = 1$ is used as the classifier.

MKSVM: The training data are only from the target domain. According to the study of [14], we do not simply concatenate the features of MRI and CSF into a long feature vector; instead, we adopt the multi-kernel learning to

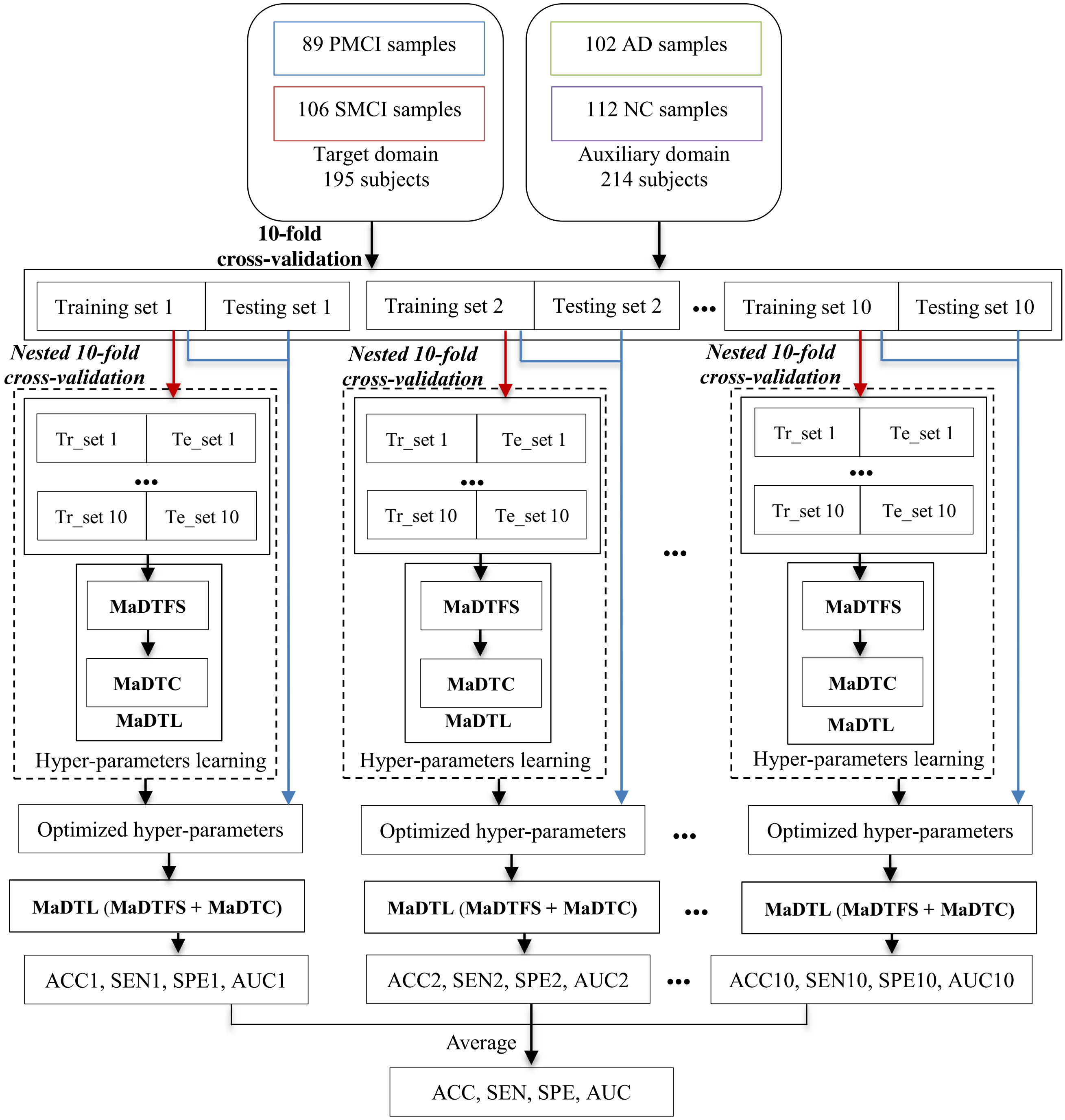


Fig. 5 Illustration that shows the process of training and testing of our proposed MaDTL method for the classification of PMCI and SMCI subjects. The abbreviations of "training set" and "testing set" are "Tr_set" and "Te_set," respectively. Hyperparameter learning is that we use the hierarchical optimization method-based grid search to search the optimized hyperparameters of MaDTFS and MaDTC on training data. The area under the receiver operating characteristic curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE)

combine the features of MRI and CSF. All experiment settings proceeded in strict accordance with the study of [14]. More details can be found in [14].

MTFS: Multi-task Lasso feature selection (MTFS) is employed in the study of [37]. The training data are from both the target domain and multi-auxiliary domains, and the MTFS algorithm is conducted for feature selection before using linear SVM for classification.

cFSGL: Convex fused sparse group Lasso (cFSGL) is proposed in the study of [36]. The training data are from

Table 1 Comparison of our proposed MaDTL method and eight state-of-the-art methods (SVM, MKSVM, Lasso, MTFS, sgLasso, cFSGL, MDTFS, rMLTFL) for diagnosis of MCI to AD conversion (mean ± std)

Method	ACC %	SEN %	SPE %	AUC %	p-value
SVM	66.60 ± 2.12	62.63 ± 2.40	69.80 ± 1.91	71.33 ± 0.94	< 0.00001
MKSVM	69.11 ± 1.53	65.45 ± 1.76	72.06 ± 1.36	73.51 ± 0.71	< 0.00001
MTFS	70.67 ± 2.11	67.18 ± 2.38	73.48 ± 1.90	76.75 ± 1.27	< 0.00001
cFSGL	70.71 ± 2.37	67.23 ± 2.67	73.51 ± 2.13	76.57 ± 1.23	< 0.00001
Lasso	73.24 ± 2.19	70.18 ± 2.51	75.71 ± 1.95	80.32 ± 1.48	< 0.0001
sgLasso	74.51 ± 1.70	71.48 ± 1.96	76.95 ± 1.50	82.14 ± 1.06	< 0.0001
rMLTFL	75.07 ± 1.98	72.11 ± 2.26	77.45 ± 1.77	79.37 ± 0.82	< 0.001
MDTL	76.02 ± 1.63	73.16 ± 1.85	78.33 ± 1.47	80.79 ± 1.09	< 0.001
MaDTL	**80.37 ± 1.28**	**76.61 ± 2.74**	**83.39 ± 2.59**	**88.16 ± 0.61**	-

ACC, accuracy; *SEN*, sensitivity; *SPE*, specificity; *AUC*, area under the receiver operating characteristic curve.

both the target domain and multi-auxiliary domains. The cFSGL method is used for feature selection and then linear SVM is used for classification.

Lasso: The training data are only from the target domain, and the L_1-norm-based feature selection is performed before classification. Finally, a linear SVM is used for classification.

sgLasso: Sparse group Lasso (sgLasso) is proposed in the study of [41]. The training data are from both the target domain and multi-auxiliary domains, and the sgLasso method is used for feature selection and then a linear SVM is used for classification.

rMLTFL: Robust multi-label transfer feature learning (rMLTFL) is proposed in the study of [2]. The training data are from both the target domain and multi-auxiliary domains. The rMLTFL method is used for feature selection and then linear SVM is used for classification.

MDTL: Multi-domain transfer learning (MDTL) is proposed in the study of [13], including components of multi-domain transfer feature selection (MDTFS) and multi-domain transfer classification (MDTC). The training data are from both the target domain and multi-auxiliary domains. MDTFS is used for feature selection and then MDTC is used for classification.

The SVM is implemented using the LIBSVM[1] toolbox with a linear kernel and a default value for the parameter $C = 1$. For Lasso, MTFS, sgLasso, and cFSGL methods, we adopt the MALSAR toolbox to solve the optimization problem. There are multiple regularization parameters of the above methods (apart from SVM) to be optimized. All regularization parameters of these methods are chosen from the range of Ω[2] by a nested tenfold cross-validation on the training data. Before training models, we normalized features by following the study of [14].

For hyperparameter learning of our proposed MaDTL method, we employ a hierarchical optimization method-based grid search to search the optimal parameters, as used in our previous work [30]. Specifically, we first optimize the regularization parameter λ_1, while we fix default values $(\lambda_2 = 1, w_1 = 0.5, w_2 = w_3 = w_4 = w_5 = w_6 = 0.1)$ for other seven parameters, and then we optimize the parameter λ_2 with the optimized parameter λ_1 and the fixed default values for other six parameters $(w_1, \ldots, w_6)$. Then, we successively optimize those remaining parameters by the aforementioned way and iterate this process t (with the default setting of t as 10 in this paper) times until the values of classification accuracy stop changing. In addition, the range of regularization parameters is Ω $(\lambda_1, \lambda_2 \in \Omega)$, and the range of six weight parameters $(w_1, \ldots, w_6)$ is $\mathrm{P} = \{-1 : 0.01 : 1\}$.

Results

In this section, we first provide experimental results of comparison between MaDTL and related state-of-the-art methods, then provide results of comparison between MaDTL and its variant methods, and for evaluating the effect of multi-auxiliary domains we provide classification results in which we use different numbers of auxiliary domains in the process of training the MaDTL model, the last is the performance of discriminative feature detection, and list experimental results of selected features.

Comparison between MaDTL and other methods

To evaluate the classification performance on the diagnosis of MCI conversion with the MaDTL model, we employ several state-of-the-art related methods and run these methods on our used dataset in this paper, and the classification results are listed in Table 1. Specifically, these related methods include SVM, MKSVM [14], MTFS [37], cFSGL [36], Lasso [47], sgLasso [41], rMLTFL [2], and MDTL [13]. Note that each value in Table 1 is the averaged result of the tenfold cross-validation 10 times. Also, to further evaluate

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2] $\Omega = \{0.0001, 0.0005, 0.0009, 0.001 : 0.001 : 0.009, 0.01 : 0.01 : 0.09, 0.1 : 0.1 : 2\}$

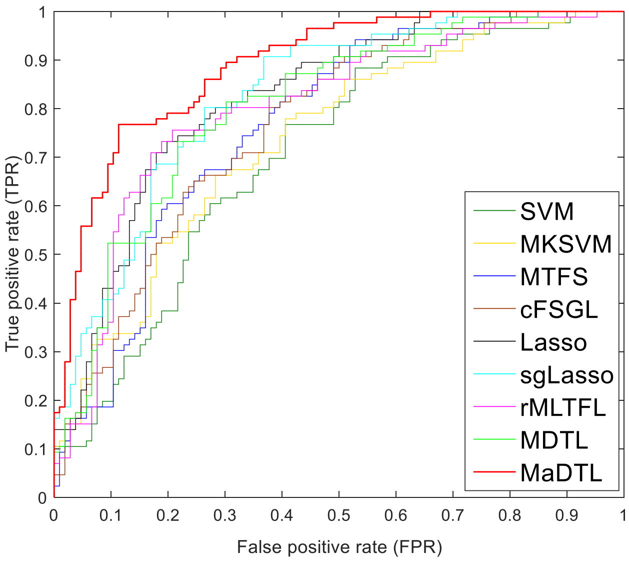


Fig. 6 ROC curves of different methods for the classification of PMCI and SMCI subjects

Table 2 The comparison of our proposed method (MaDTL), its two variant methods (MaDTC and MaDTFS + SVM), and SVM (as a baseline method) (mean ± std)

Method	ACC %	SEN %	SPE %	AUC %	*p*-value
SVM	66.60 ± 2.12	62.63 ± 2.40	69.80 ± 1.91	71.33 ± 0.94	< 0.000001
MaDTC	72.73 ± 2.10	69.64 ± 2.07	75.21 ± 2.54	78.86 ± 0.41	< 0.0001
MaDTFS + SVM	76.62 ± 2.28	73.86 ± 2.59	78.85 ± 2.03	81.64 ± 1.09	< 0.001
MaDTL(i.e., MaDTFS + MaDTC)	80.37 ± 1.28	76.61 ± 2.74	83.39 ± 2.59	88.16 ± 0.61	-

the availability of the MaDTL model, we use DeLong's method [48] on the AUC between the proposed method and each of the other compared methods and list the corresponding *p*-values in Table 1. In addition, to understand more visually the effectiveness of the MaDTL model, we plot the ROC curves of all the methods in Fig. 6.

From Table 1 and Fig. 6, we have the following observations. First, the proposed MaDTL method consistently outperforms those eight competing methods regarding all measures, demonstrating the effectiveness of the MaDTL method for the diagnosis of MCI-to-AD conversion. Second, our proposed MaDTL method and our previous works (i.e., MDTL and rMLTFL methods) are superior to other multi-task learning methods (i.e., MTFS, sgLasso, and cFSGL), which shows that modeling correlations between the target and the multi-auxiliary domains has the advantage over those multi-task learning methods that only implicitly utilize the relationships between tasks. Third, MaDTL consistently achieves better classification performance than our previous works (i.e., MDTL and rMLTFL methods), suggesting that the sparse group correlation Lasso penalties (i.e., regularization terms of $\|\mathbf{WH}\|_{1,1}$ and $\|\mathbf{WH}\|_{2,1}$) proposed in MaDTL are more useful in promoting classification performance. Finally, the Lasso method without using any auxiliary domain data has significantly better classification performance than the two multi-task learning methods (i.e., MTFS and cFSGL) using auxiliary domain data, implying that the existence of several irrelevant auxiliary domains may cause negative effects and restrict the improvement of classification performance.

Comparison with MaDTL and its variants

In our proposed MaDTL model, it consists of two parts: a MaDTFS module for feature selection and a MaDTC module for classification. In order to evaluate the contributions of each component, we proposed two variant MaDTL methods (i.e., MaDTC and MaDTFS + SVM) and performed these variant methods on the ADNI database of 409 subjects. In Table 2, we list the classification results of MaDTC and MaDTFS + SVM methods and compare them with MaDTL and SVM (as a baseline method). It should be noted that the MaDTL method first performs MaDTFS for feature selection and then performs MaDTC for classification (i.e., MaDTFS + MaDTC), the MaDTC method only performs the MaDTC module without feature selection, and the MaDTFS + SVM method first performs the MaDTFS module and then performs SVM for classification. For intuitive comparison, we also plot the ROC curves achieved by these methods in Fig. 7 and perform the DeLong's test method [48] on AUC between the MaDTC method and its two variant methods as well as the baseline method. As we can see from Table 2 and Fig. 7, each component can boost the classification performance compared with the baseline SVM method, using our feature selection module (i.e., MaDTFS) can achieve better improvement than the MaDTC method for classification, and the MaDTL model integrates MaDTFS and MaDTC modules together to achieve a performance better than that of each variant method.

In addition, there are interesting observations from Table 1 and Table 2. First, the MaDTC method in Table 2 is significantly better than the MKSVM [14] method in Table 1, which suggests that the MaDTC method provides a better way than MKSVM to utilize multi-modality data for classification. Second, the MaDTFS + SVM method in

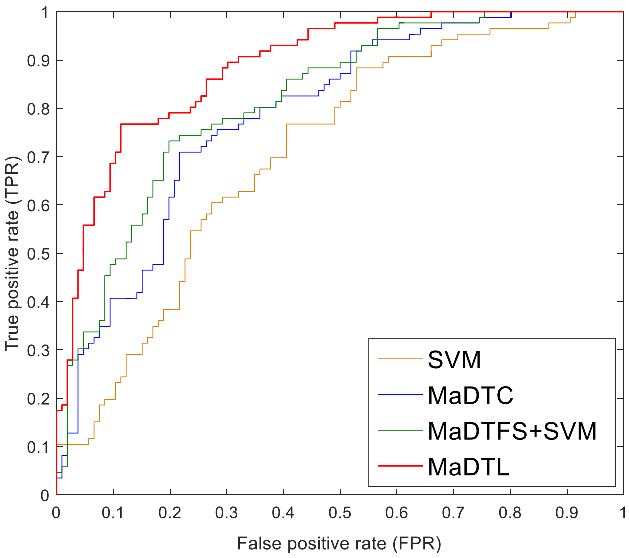


Fig. 7 ROC curves of the MaDTL method, its two variant methods, and SVM

Table 2 slightly outperforms the MDTL (MDTFS + MDTC) method in Table 1, but the *p*-value computed by DeLong's method [48] on the AUC values between these two methods is 0.036, showing the advantage of MaDTFS + SVM is statistically significant. This indicates that our proposed regularization terms on $\|\mathbf{WH}\|_{1,1}$ and $\|\mathbf{WH}\|_{2,1}$ are more effective than the regularization terms (i.e., a combination of $\|\mathbf{WH}\|_{2,2}$, $\|\mathbf{W}\|_{1,1}$, and $\|\mathbf{W}\|_{2,1}$) of the MDTL method for feature selection in diagnosing MCI-to-AD conversion. Third, the MaDTFS + SVM method is superior to the rMLTFL method in Table 1, and the *p*-value on the AUC values between these two methods is less than 0.01, indicating the significance of our proposed MaDTFS in effectively using multi-auxiliary domain data to improve the performance of the target learning domain than the rMLTFL method.

Effect of multi-auxiliary domains

To investigate the influence of multi-auxiliary domain data on the performances of the MaDTL model, we further performed a set of experiments by using different numbers of auxiliary domains in the steps of feature selection and classification. Specifically, we evenly select a number of auxiliary domain data from all auxiliary domains (i.e., five auxiliary domains in our work) to train the MaDTL model, test this process multiple times to select a certain number of auxiliary domains, and report the average and the standard deviation of accuracy, sensitivity, specificity, and AUC in Fig. 8. For a clear observation of the change of classification measures, we also added the classification measures of the MaDTL method using all five auxiliary domains in Fig. 8 for reference. Since only up to five auxiliary domains are used in our work, there is no standard deviation of classification measures at the abscissa value of 5 in Fig. 8. To properly explore the effect of multi-auxiliary domain data, we repeated multiple times to select a given number of auxiliary domains for training, specifically, 5 times for selecting one auxiliary domain, 10 times for selecting two auxiliary domains, 10 times for selecting three auxiliary domains, and 5 times for selecting four auxiliary domains. The classification results using different numbers of auxiliary domains are reported in Fig. 8 and Table 3.

As can be seen from Fig. 8, with the increase of the number of auxiliary domains, the four classification measures (i.e., accuracy, sensitivity, specificity, and AUC) of the MaDTL method rise monotonically. While we select two and three auxiliary domains, the values of standard deviation are greater than those using one and four auxiliary domains. That is because we tested 10 times to select two and three auxiliary domains, greater than the testing times of selecting one and four auxiliary domains. In addition, this result also reveals that the correlations between each group of the multi-auxiliary and the target domains have a significant difference. This reinforces the necessity for us to develop the regularization terms on $\|\mathbf{WH}\|_{1,1}$ and $\|\mathbf{WH}\|_{2,1}$, which can restrain irrelevant auxiliary domains that may cause negative effects for classification. In general, using more data from multi-auxiliary domains can effectively improve the performance of the target learning domain.

Discriminative feature detection

The proposed MaDTL method can identify the discriminative features (corresponding to ROIs or features of CSF levels) that are helpful for the diagnosis of MCI-to-AD conversion in clinical practice. Since we adopt a tenfold cross-validation strategy with 10 times repetition to evaluate the effectiveness of the MaDTL model and the feature selection in each fold is performed only based on the current training set, the selected features could vary across different folds and runs. We counted the frequency of the selected features across all folds and runs (i.e., a total of 100 times for tenfold cross-validation with 10 independent runs) by the MaDTL method using the concatenated MRI and CSF biomarkers and listed all the most discriminative features with the highest frequency of occurrence (i.e., each feature is selected across all folds and runs) in Table 4.

From Table 4, we can observe that our proposed MaDTL method successfully selects discriminative features, since the corresponding ROIs and the features of CSF biomarker are known to be related to the early diagnosis of AD [14, 25, 29, 37–39, 49]. Specifically, there are 26 features that are consistently selected across all folds and all runs, and an average of 32 features are selected via the tenfold cross-validation 10 times. The features

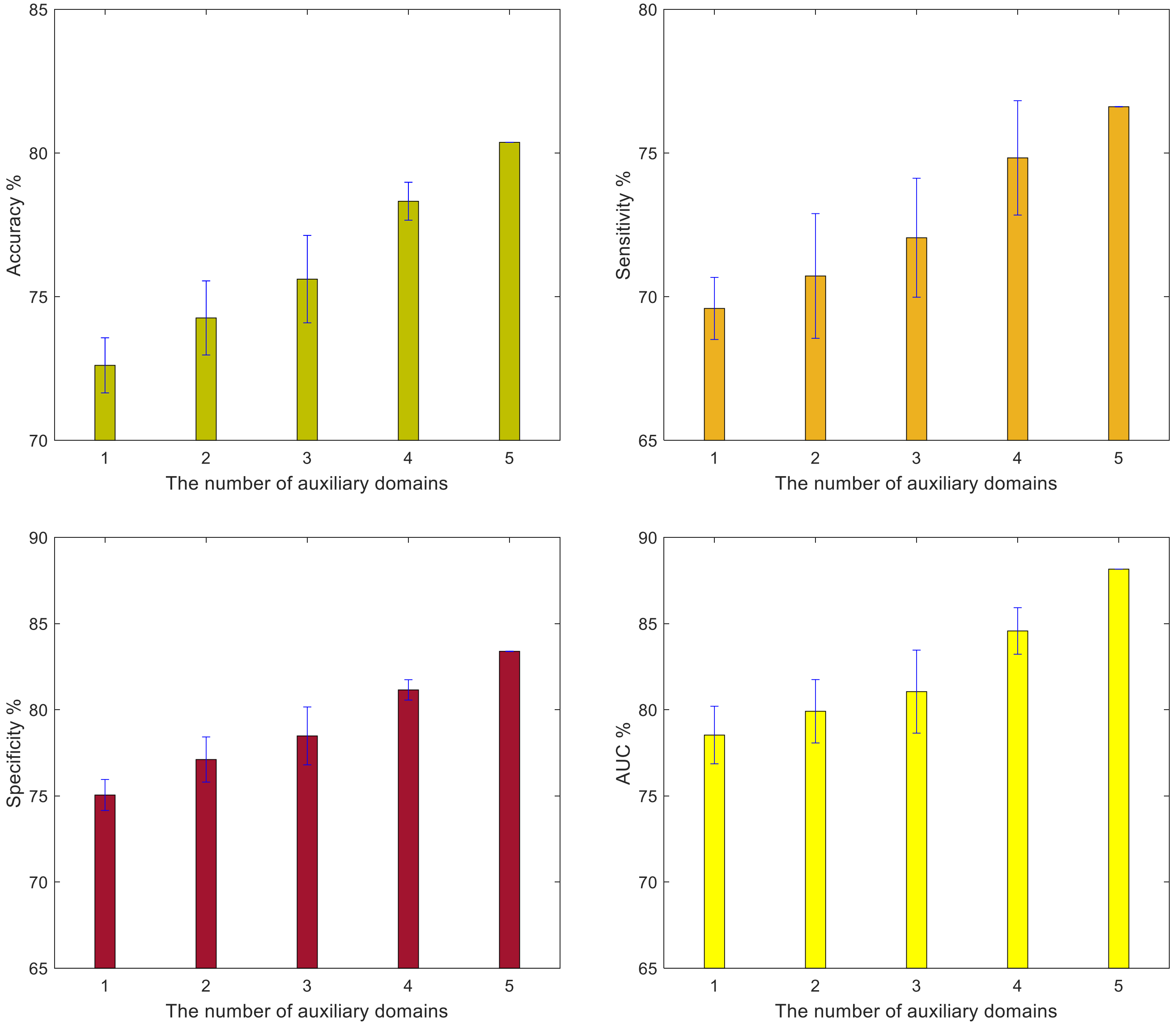


Fig. 8 The changes of four classification measures (accuracy, sensitivity, specificity, and AUC) of the MaDTL method with respect to the used number of auxiliary domains

(i.e., Aβ_{42}, t-tau, and p-tau) from the CSF biomarker are selected across all folds and runs, which imply that the biological cerebrospinal fluid levels of Aβ_{42}, t-tau, and p-tau have also been changed before the appearance of brain atrophy; thus, the CSF biomarker is well suited to the diagnosis of MCI-to-AD conversion. Also, a number of features of brain regions (e.g., hippocampal formation, amygdala, uncus, and cuneus) from MRI biomarkers are selected across all folds and runs, which shows that a number of brain regions have shown atrophy in the stage of MCI. Therefore, combining MRI and CSF biomarkers is able to provide complementary and discriminative information in the diagnosis of MCI-to-AD conversion. In a word, these observations suggest that brain structure and function have gradually changed with the progression of MCI.

Discussion

In this paper, we propose a MaDTL model to identify MCI-to-AD conversion patients, which can select discriminative feature subsets from the target domain and multi-auxiliary domains and combine these data to achieve a robust classifier. We have evaluated the performance of our method on 409 baseline subjects from the ADNI database, and the experimental results show that our proposed method can consistently and substantially improve

Table 3 Classification performances that we select the different number of auxiliary domains in the process of training the MaDTL model

Auxiliary domain	ACC %	SEN %	SPE %	AUC %
AD vs. NC	71.39	68.14	73.99	77.15
AD vs. PMCI	72.56	69.27	75.20	77.04
AD vs. SMCI	73.39	70.43	75.78	80.90
PMCI vs. NC	73.71	70.88	76.00	79.57
SMCI vs. NC	72.01	69.23	74.26	78.00
AD vs. NC + AD vs. PMCI	73.33	69.46	76.46	79.19
AD vs. NC + AD vs. SMCI	73.63	69.93	76.61	80.30
AD vs. NC + PMCI vs. NC	75.44	69.80	80.00	80.72
AD vs. NC + SMCI vs. NC	73.63	71.66	75.23	77.46
AD vs. PMCI + AD vs. SMCI	76.27	74.32	77.83	83.41
AD vs. PMCI + PMCI vs. NC	74.10	69.71	77.63	80.12
AD vs. PMCI + SMCI vs. NC	71.78	66.36	76.15	76.92
AD vs. SMCI + PMCI vs. NC	74.06	71.34	76.24	79.33
AD vs. SMCI + SMCI vs. NC	75.41	72.46	77.79	80.69
PMCI vs. NC + SMCI vs. NC	74.91	72.11	77.18	80.97
AD vs. NC + AD vs. PMCI + AD vs. SMCI	77.99	76.29	79.34	85.60
AD vs. NC + AD vs. PMCI + PMCI vs. NC	75.54	71.73	78.64	80.60
AD vs. NC + AD vs. PMCI + SMCI vs. NC	72.79	68.68	76.10	77.14
AD vs. NC + AD vs. SMCI + PMCI vs. NC	75.70	73.55	77.45	80.66
AD vs. NC + AD vs. SMCI + SMCI vs. NC	74.68	71.18	77.50	79.14
AD vs. NC + PMCI vs. NC + SMCI vs. NC	77.58	72.95	81.33	83.70
AD vs. PMCI + AD vs. SMCI + PMCI vs. NC	75.56	69.86	80.14	80.83
AD vs. PMCI + AD vs. SMCI + SMCI vs. NC	74.22	71.82	76.16	79.61
AD vs. PMCI + PMCI vs. NC + SMCI vs. NC	75.78	71.75	79.04	80.52
AD vs. SMCI + PMCI vs. NC + SMCI vs. NC	76.26	72.70	79.14	82.71
AD vs. NC + AD vs. PMCI + AD vs. SMCI + PMCI vs. NC	79.17	77.73	80.34	85.16
AD vs. NC + AD vs. PMCI + AD vs. SMCI + SMCI vs. NC	78.38	74.23	81.71	83.83
AD vs. NC + AD vs. PMCI + PMCI vs. NC + SMCI vs. NC	77.49	73.14	81.01	82.62
AD vs. NC + AD vs. SMCI + PMCI vs. NC + SMCI vs. NC	77.88	73.11	81.74	85.21
AD vs. PMCI + AD vs. SMCI + PMCI vs. NC + SMCI vs. NC	78.70	75.93	80.94	86.05

Table 4 The most discriminative features identified by the proposed MaDTL method

Features (brain regions)	
Parahippocampal gyrus left	Perirhinal cortex left
Angular gyrus right	Entorhinal cortex left
Uncus right	Hippocampal formation left
Fornix left	Middle temporal gyrus right
Precuneus right	Corpus callosum
Hippocampal formation right	Amygdala right
Inferior occipital gyrus left	Inferior temporal gyrus right
Cuneus left	Lateral occipitotemporal gyrus left
Supramarginal gyrus right	Thalamus right
Uncus left	Occipital pole left
Middle temporal gyrus left	Aβ_{42}
Precentral gyrus left	t-tau
Perirhinal cortex right	p-tau

the classification performance, with an overall classification accuracy of 80.37% to differentiate PMCI and SMCI subjects.

Multi-auxiliary domain transfer learning

Recently, multi-task learning, deep learning, and transfer learning are used to diagnose the MCI-to-AD conversion [2–4, 6, 8, 10, 11, 13, 27, 29, 30, 32–35, 37, 39, 40]. Multi-task learning assumes that all these learning tasks should have a strong correlation and only implicitly uses this relationship to learn common features for all tasks. That would be limited in clinical practice because the links between these tasks may be weak correlative or irrelevant, but most of the existing multi-task learning studies have not sufficiently considered this case. In contrast, in our previous works [2, 13, 30], a transfer learning strategy is adopted for the early diagnosis of AD, which can effectively combine multiple auxiliary domains to further promote the performance of the target domain.

The research of this paper is to improve and extend our previous works [13, 30], and the feature learning and classification models are different from our previous works [13, 30]. Specifically, the significant differences between this work and our previous works are summarized as follows. First, multi-auxiliary domains are used in this paper, but only a single auxiliary domain was used in our previous work [13, 30]. Second, our current study considers possible weak correlative or irrelevance between auxiliary multi-domains and the target domain, but our previous work [13, 30] assumed that auxiliary domains have consistent relevance to the target domain. In addition, our previous works [13, 30] employed a conventional sparse group Lasso penalty (i.e., $\|\mathbf{W}\|_{1,1}$ and $\|\mathbf{W}\|_{2,1}$) for feature selection, and, according to prior knowledge of AD pathology, the current study proposed a group correlation Lasso penalty ($\|\mathbf{WH}\|_{1,1}$ and $\|\mathbf{WH}\|_{2,1}$) for feature selection. Results in Tables 1 and 2 show that our proposed group correlation Lasso penalties on $\|\mathbf{WH}\|_{1,1}$ and $\|\mathbf{WH}\|_{2,1}$ are more effective than the conventional sparse group Lasso penalties ($\|\mathbf{W}\|_{1,1}$ and $\|\mathbf{W}\|_{2,1}$) for selecting features for the diagnosis of MCI-to-AD conversion.

In the step of classification, our previous works [13, 30] employed the technique of adaptive SVMs to build the classifier, which is sensitive to the negative effects from a few irrelevant auxiliary domains. However, according to the ensemble learning strategy, our current study proposes a MaDTC module for classification, and we employ the conventional SVM to train base classifiers on each auxiliary domain and the target domain. To mitigate the negative effects from a few irrelevant auxiliary domains, we extend the restricted condition of conventional voting with the linear weighting method. Different from our previous works [13, 30], theoretically, the MaDTC module can employ any classifier to build the base classifiers while using SVMs would achieve better performance. For validation, we test the method of the previous work [13] to classify PMCI and SMCI subjects without any feature selection and achieve a classification accuracy of 70.03% and AUC of 74.64%, both of which are inferior to our proposed MaDTC module with a classification accuracy of 72.73% and AUC of 78.86%. A statistic test verified such a difference is significant according to the p-values that are less than 0.01. In general, our proposed MaDTC module can significantly improve our previous works [13, 30] to recognize PMCI and SMCI subjects.

In our MaDTL model, there are eight hyperparameters $(\lambda_1, \lambda_2, w_1, \dots, w_6)$ to be optimized. For simplicity, we employed a hierarchical optimization method-based grid search for hyperparameter optimization. To demonstrate the convergence of the algorithm, we plot a broken line graph of classification accuracy and AUC values with respect to a different number of iterations using the iterative optimization algorithm in Fig. 9. In Fig. 9, classification accuracy and AUC values first rise with the increasing number of iterations and then kept stable when the number of iterations is larger than 6, which can be proven by the convergence of hierarchical optimization algorithm-based grid search. In addition, we need to further evaluate the contribution of each operator term for the MaDTL model. In Table 5, we list the classification results of the MaDTL model by setting the respective parameters to 0. For instance, the regularization parameter λ_1 is set as 0 ($\lambda_1 = 0$), which can be used to evaluate the contribution of the first regularization term in the MaDTFS module, and a certain weight parameter w_i is set as 0 ($w_i = 0$), which can be used to evaluate the contribution of a certain domain data in the MaDTC module. These experimental results showed that each operator term can boost classification performance for the diagnosis of MCI-to-AD conversion and combining all operator terms can achieve better diagnosis performance.

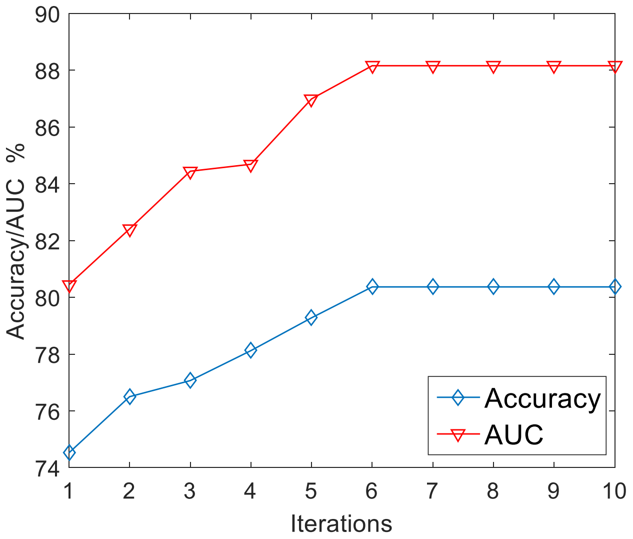


Fig. 9 Classification accuracy/AUC of our proposed MaDTC method with respect to a different number of iterations, achieved by the iterative optimization algorithm

Table 5 Comparison of our proposed MaDTL method using different settings of parameters (mean ± std)

Parameter	ACC %	SEN %	SPE %	AUC %
$\lambda_1 = 0$	78.77 ± 1.87	74.86 ± 3.41	81.94 ± 2.17	86.61 ± 0.47
$\lambda_2 = 0$	77.64 ± 1.92	75.43 ± 3.21	79.43 ± 2.43	85.51 ± 0.46
$w_1 = 0$	76.28 ± 2.05	73.41 ± 3.92	78.59 ± 4.49	84.14 ± 1.03
$w_2 = 0$	78.80 ± 1.54	75.63 ± 3.36	81.35 ± 1.91	87.55 ± 0.32
$w_3 = 0$	78.84 ± 1.44	71.43 ± 2.87	84.86 ± 3.05	86.89 ± 0.56
$w_4 = 0$	78.27 ± 1.83	71.55 ± 5.03	83.69 ± 2.55	85.25 ± 0.62
$w_5 = 0$	77.29 ± 1.16	72.27 ± 2.34	81.33 ± 1.68	85.81 ± 0.57
$w_6 = 0$	78.12 ± 1.82	70.14 ± 2.71	84.61 ± 3.53	86.91 ± 1.97
All	80.37 ± 1.28	76.61 ± 2.74	83.39 ± 2.59	88.16 ± 0.61

Table 6 Comparison of our proposed method (MaDTL) and eight state-of-the-art methods in classifying AD vs. NC, AD vs. PMCI, and SMCI vs. NC. (mean ± std)

Method	AD vs. NC			SMCI vs. NC			AD vs. PMCI		
	ACC %	AUC %	*p*-value	ACC %	AUC %	*p*-value	ACC %	AUC %	*p*-value
SVM	90.52 ± 1.23	96.53 ± 0.38	< 0.0001	67.84 ± 2.11	71.33 ± 1.51	< 0.0001	58.44 ± 2.05	61.12 ± 1.64	< 0.0001
MKSVM	90.01 ± 1.28	96.20 ± 0.34	< 0.0001	70.02 ± 1.83	72.98 ± 1.18	< 0.0001	61.04 ± 1.82	62.17 ± 1.81	< 0.0001
MTFS	92.69 ± 0.72	97.57 ± 0.31	< 0.001	70.85 ± 1.12	74.96 ± 0.72	< 0.0001	63.38 ± 1.85	66.38 ± 1.26	< 0.0001
cFSGL	92.03 ± 1.32	97.45 ± 0.42	< 0.001	73.27 ± 1.12	78.75 ± 0.88	< 0.001	66.59 ± 1.69	70.10 ± 1.05	< 0.001
Lasso	90.90 ± 1.11	96.51 ± 0.38	< 0.0001	68.25 ± 2.01	71.34 ± 1.76	< 0.0001	60.31 ± 1.65	62.40 ± 1.91	< 0.0001
sgLasso	94.12 ± 1.15	98.34 ± 0.32	< 0.01	73.20 ± 1.20	77.37 ± 0.64	< 0.001	67.57 ± 1.66	73.03 ± 0.98	< 0.001
rMLTFL	92.52 ± 0.62	97.47 ± 0.30	< 0.001	74.02 ± 1.50	80.84 ± 0.91	< 0.001	67.97 ± 2.06	73.14 ± 1.18	< 0.001
MDTL	93.93 ± 1.37	98.49 ± 0.30	< 0.01	75.06 ± 0.91	79.41 ± 0.59	< 0.01	69.11 ± 2.96	76.86 ± 2.42	< 0.001
MaDTL	96.45 ± 0.64	99.11 ± 0.36	-	77.85 ± 1.23	82.08 ± 0.48	-	73.34 ± 2.26	81.53 ± 1.33	-

In the MaDTL model, the weight w_i represents how one domain can be important. We compute the mean values of the weight w_i across all folds and runs and list all mean values of weight w_i (i.e., $w_1(0.16)$, $w_2(-0.23)$, $w_3(-0.02)$, $w_4(0.51)$, $w_5(0.56)$, and $w_6(0.02)$). These reported values of weight are optimized by the hierarchical optimization method-based grid search method on training data. According to the mean values of weightw_i, weight w_4 and weight w_5 are better than the other four mean values of weight, which shows the importance of auxiliary domains (AD vs. SMCI and PMCI vs. NC); weight w_2 and w_3 are negative, which suggests negative transfer appeared at the use of auxiliary domains (AD vs. NC and AD vs. PMCI); weight w_4 and w_5 are better than weightw_1, which implies auxiliary domains (AD vs. SMCI and PMCI vs. NC) can provide more useful decision information than the target domain (PMCI vs. SMCI). However, in Table 3, without use of auxiliary domains (AD vs. NC and AD vs. PMCI) cannot achieve better performance than the use of all auxiliary domains, which shows the combination of more training data from multi-auxiliary domains can provide more complementary inter-domain information. These experimental results confirm that the use of more training data from multi-auxiliary domains is useful to the diagnosis of MCI-to-AD conversion.

Extension for classifying AD/NC, AD/PMCI, and SMCI/NC

In the early diagnosis of AD, the recognition of MCI-to-AD conversion is becoming more and more important. Therefore, we only report classification results of classifying SMCI and PMCI subjects in Table 1. To further investigate the effectiveness of our proposed MaDTL model, we also apply our model to classify AD vs. NC, AD vs. PMCI, and SMCI vs. NC. Specifically, each target domain corresponds to the learning task of classifying AD vs. NC, AD vs. PMCI, or SMCI vs. NC, and the rest of the domains are used as multi-auxiliary domains. In Table 6, we provide two significant performance measurements (i.e., ACC and AUC) and the *p*-value that is computed by DeLong's method [48] on the AUC between the proposed model and competing methods.

As can be seen from Table 6, our proposed MaDTL model consistently outperforms those competing methods regarding all measurements to classify AD vs. NC, AD vs. PMCI, and SMCI vs. NC. Compared with the tasks of classifying AD vs. NC and SMCI vs. NC, performance improvements in classifying AD vs. PMCI with the MaDTL model are more protruded. Specifically, compared with the SVM method, the use of the MaDTL model can achieve almost 15% and 10% improvements of accuracy for classifying AD vs. PMCI and SMCI vs. NC, and almost 6% improvement of accuracy for classifying AD vs. NC. These experimental results further verify the effectiveness of our proposed MaDTL model in diagnosing MCI conversion.

Limitations

The current study is limited by several factors. First, there are six weight parameters in the MaDTC model, which also takes more time to tune parameters and restricts more auxiliary domain data to be added. In our future work, we will improve the MaDTC model in order to introduce more data from auxiliary domains conveniently. Second, many data from status-unlabeled subjects can be available from the ADNI database, and we can extend our current method to use unlabeled subjects. We will also investigate whether adding status-unlabeled data can further improve the performance. Third, for the preprocessing of MR images, our current study only uses ROI features, while previous studies have shown the effectiveness of cortical thickness in the early diagnosis of AD [49–54]. In fact, considering the small number of training samples, as well as the sensitivity of those very local features (i.e., thickness and tissue density) to noises, as well as errors in processing pipeline (including skull stripping, tissue segmentation, image registration, and region-of-interest (ROI) labeling), our current study considers only using the mid-level features, such as regional features (or ROI features), and no surface-based cortical thickness features are extracted. It is interesting to take advantage of both cortical thickness features extracted from MR images and the ROI-based features for MCI-to-AD conversion prediction in the future.

Conclusion

In this paper, we propose a novel multi-auxiliary domain transfer learning (MaDTL) model for the diagnosis of MCI-to-AD conversion, which can select the discriminative feature subset from the target domain and multi-auxiliary domains and combine data from multi-auxiliary domains and target domain to achieve a robust classifier. The main idea of our method is to exploit data from multi-auxiliary domains to improve classification performance in the target domain. Specifically, we first develop a MaDTFS module to select the discriminative feature subset. Then, we propose a MaDTC module to train a robust classifier that can restrain negative effects from a few irrelevant auxiliary domains. We evaluate our model on the baseline ADNI database with MRI and CSF data, and the experimental results demonstrate the effectiveness of our MaDTL model.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10072-021-05568-6.

Acknowledgements ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffmann-La Roche, Schering-Plough, Synarc, Inc., as well as nonprofit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the US Food and Drug Administration. Private sector contributions to ADNI 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 of Neuroimaging at the University of California, Los Angeles.

Funding Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). This work was supported by the National Natural Science Foundation of China (No. 61602072), the Chongqing Cutting-edge and Applied Foundation Research Program (No. cstc2018jcyjAX0502), the Scientific and Technological Research Program of Chongqing Municipal Education Commission (No. KJQN202001222), and the Chongqing Three Gorges University Research Program (No.19QN08).

Data availability The datasets used in this paper are from the Alzheimer's Disease Neuroimaging Initiative (ADNI) which are available at http://adni.loni.usc.edu/. Source code and binary programs developed in this paper are available via email, cb729@nuaa.edu.cn.

Declarations

Ethics approval None.

Conflict of interest The authors declare no competing interests.

References

1. Association A s, (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dement* 15, 321–387.
2. Cheng B, Liu M, Shen D, Zhang D (2019) Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease. Brain Imaging Behav 13:138–153
3. Wee CY, Liu C, Lee A, Joann SP, Ji H, Qiu A (2019) Cortical graph neural network for AD and MCI diagnosis and transfer learning across populations. NeuroImage Clin 23:101929
4. Choia H, Jinb KH (2018) Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behav Brain Res 344:103–109
5. Liu X, Goncalves AR, Cao P, Zhao D, Banerjee A (2018) Modeling Alzheimer's disease cognitive scores using multi-task sparse group lasso. Comput Med Imaging Graph 66:100–114
6. Zhou K, He W, Xu Y, Xiong G, Cai J (2018) Feature selection and transfer learning for Alzheimer's disease clinical diagnosis. Appl Sci 8:1372
7. Hojjati SH, Ebrahimzadeh A, Khazaee A, Feremi AB (2017) Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. J Neurosci Methods 282:69–80
8. Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R, den Heeten A, Karssemeijer N (2017) Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 35:303–312
9. Li Q, Wu X, Xu L, Chen K, Yao L, Li R (2017) Multi-modal discriminative dictionary learning for Alzheimer's disease and mild cognitive impairment. Comput Methods Programs Biomed 150:1–8
10. Suk HI, Lee SW, Shen D (2017) Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal 37:101–113
11. Zhu X, Suk H, Wang L, Lee SW, Shen D (2017) A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Med Image Anal 38:205–214
12. Shi B, Chen Y, Zhang P, Smith CD, Liu J (2017) Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis. Pattern Recogn 63:487–498
13. Cheng B, Liu M, Shen D, Li Z, Zhang D, Alzheimer's Disease Neuroimaging I (2017) Multi-domain transfer learning for early diagnosis of Alzheimer's disease. Neuroinformatics 15:115–132
14. Zhang D, Wang Y, Zhou L, Yuan H, Shen D, ADNI, (2011) Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55:856–867
15. Chetelat G, Landeau B, Eustache F, Mezenge F, Viader F, de la Sayette V, Desgranges B, Baron JC (2005) Using voxel-based

morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. Neuroimage 27:934–946

16. Chao LL, Buckley ST, Kornak J, Schuff N, Madison C, Yaffe K, Miller BL, Kramer JH, Weiner MW (2010) ASL perfusion MRI predicts cognitive decline and conversion from MCI to dementia. Alzheimer Dis Assoc Disord 24:19–27
17. deToledo-Morrell L, Stoub TR, Bulgakova M, Wilson RS, Bennett DA, Leurgans S, Wuu J, Turner DA (2004) MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. Neurobiol Aging 25:1197–1203
18. Risacher SL, Saykin AJ, West JD, Shen L, Firpi HA, McDonald BC (2009) Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr Alzheimer Res 6:347–361
19. Misra C, Fan Y, Davatzikos C (2009) Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. Neuroimage 44:1415–1422
20. Liu M, Zhang D, Shen D (2016) Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans Med Imaging 35:1463–1474
21. Bouwman FH, Schoonenboom SNM, van der Flier WM, van Elk EJ, Kok A, Barkhof F, Blankenstein MA, Scheltens P (2007) CSF biomarkers and medial temporal lobe atrophy predict dementia in mild cognitive impairment. Neurobiol Aging 28:1070–1074
22. Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, Weiner MW, Knopman DS, Petersen RC, Jack CR, ADNI, (2009) MRI and CSF biomarkers in normal, MCI, and AD subjects predicting future clinical change. Neurology 73:294–301
23. Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, Weiner MW, Knopman DS, Petersen RC, Jack CR, ADNI, (2009) MRI and CSF biomarkers in normal, MCI, and AD subjects diagnostic discrimination and cognitive correlations. Neurology 73:287–293
24. Lehmann M, Koedam E L, Barnes J, Bartlett J W, Barkhof F, Wattjes M P, Schott J M, Scheltens P, Fox N C, (2012). Visual ratings of atrophy in MCI: prediction of conversion and relationship with CSF biomarkers. *Neurobiology of Aging*.
25. Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ (2011) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging 32:2322.e19-2322.e27
26. Hinrichs C, Singh V, Xu GF, Johnson SC, Neuroimaging AD (2011) Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55:574–589
27. Liu F, Wee CY, Chen HF, Shen DG, ADNI, (2014) Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. Neuroimage 84:466–475
28. Liu M, Zhang J, Yap PT, Shen D (2017) View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. Med Image Anal 36:123–134
29. Jie B, Zhang D, Cheng B, Shen D (2015) Manifold regularized multitask feature learning for multimodality disease classification. Hum Brain Mapp 36:489–507
30. Cheng B, Liu M, Zhang D, Munsell BC, Shen D (2015) Domain transfer learning for MCI conversion prediction. IEEE Trans Biomed Eng 62:1805–1817
31. Shi J, Zheng X, Li Y, Zhang Q, Ying S (2018) Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform 22:173–183
32. Liu M, Zhang J, Adeli E, Shen D (2017) Deep multi-task multi-channel learning for joint classification and regression of brain status. MICCAI 2017(3):3–11
33. Suk H, Lee SW, D. S, ADNI, (2014) Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 101:569–582
34. van der Burgh HK, Schmidt R, Westeneng HJ, de Reus MA, van den Berg LH, van den Heuvel MP (2017) Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. Neuroimage Clin 13:361–369
35. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A (2016) A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 191:214–223
36. Zhou J, Liu J, Narayan VA, Ye J, ADNI, (2013) Modeling disease progression via multi-task learning. Neuroimage 78:233–248
37. Zhang D, Shen D, ADNI, (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage* 59:895-907
38. Ye J, Farnum M, Yang E, Verbeeck R, Lobanov V, Raghavan N, Novak G, DiBernardo A, Narayan V A, ADNI, (2012). Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *Bmc Neurology* 12, 1471–2377–12–46.
39. Zhu X, Suk H, Shen D (2014) A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. Neuroimage 100:91–105

40 Wachinger C, Reuter M, ADNI, (2016) Domain adaptation for Alzheimer's disease diagnostics. NeuroImage 139:470–479

41. Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. J Comput Graph Stat 22:231–245
42. Chen X, Pan W, Kwok J T, Carbonell J G, (2009). Accelerated gradient method for multi-task sparse learning problem. *Proceeding of Ninth IEEE International Conference on Data Mining and Knowledge Discovery*, 746–751.
43. Nemirovski A, (2005). Efficient methods in convex programming.
44. Argyriou A, Evgeniou T, Pontil M (2008) Convex multi-task feature learning. Mach Learn 73:243–272
45. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359
46. Tan B, Song Y, Zhong E, Yang Q, 2015. Transitive transfer learning. the 21th ACM SIGKDD International Conference. *ACM*.
47. Tibshirani RJ (1996) Regression shrinkage and selection via the LASSO. J Roy Stat Soc B 58:267–288
48. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
49. Eskildsen SF, Coupé P, García-Lorenzo D, Fonov V, Pruessner JC, Collins DL (2013) Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. Neuroimage 65:511–521
50. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehericy S, Habert MO, Chupin M, Benali H, Colliot O (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56:766–781
51. Cho Y, Seong JK, Jeong Y, Shin SY, ADNI, (2012) Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. Neuroimage 59:2217–2230
52. Wolz R, Julkunen V, Koikkalainen J, Niskanen E, Zhang DP, Rueckert D, Soininen H, Lotjonen J (2011) Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. Plos One 6:e25446
53. Querbes O, Aubry F, Pariente J, Lotterie J-A, Demonet J-F, Duret V, Puel M, Berry I, Fort J-C, Celsis P, ADNI, (2009) Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain : a journal of neurology 132:2036–2047

54. Wee CY, Yap PT, Shen DG, ADNI, (2013) Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum Brain Mapp 34:3411–3425
55. National Institutes of Health Center for Information Technology (CIT) (2020) Medical image processing, analysis and visualization (MIPAV Version 10.0.0). https://mipav.cit.nih.gov/clickwrap.php
56. Kabani N, MacDonald D, Holmes CJ, Evans A (1998) A 3D atlas of the human brain. Neuroimage 7:S717
57. Wang Y, Nie J, Yap P T, Shi F, Guo L, Shen D, (2011). Deformable surface based skull-stripping for large-scale studies. in *Medical Image Computing and Computer-Assisted Intervention* 3, 635–642.
58. Zhang YY, Brady M, Smith S (2001) Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20:45–57
59. Sled JG, Zijdenbos AP, Evans AC (1998) A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17:87–97
60. Shen DG, Davatzikos C (2002) HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans Med Imaging 21:1421–1439

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.