Computers in Biology and Medicine 152 (2023) 106367

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Exploiting task relationships for Alzheimer's disease cognitive score prediction via multi-task learning

Wei Liang [a], Kai Zhang [a], Peng Cao [a,b,*], Xiaoli Liu [c], Jinzhu Yang [a,b], Osmar R. Zaiane [d]

[a] *Computer Science and Engineering, Northeastern University, Shenyang, China*
[b] *Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China*
[c] *DAMO Academy, Alibaba Group, Hangzhou, China*
[d] *Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada*

ARTICLE INFO

Keywords:
Alzheimer's disease
Multi-task learning
Sparse learning
Feature selection
Biomarker identification

ABSTRACT

Alzheimer's disease (AD) is highly prevalent and a significant cause of dementia and death in elderly individuals. Motivated by breakthroughs of multi-task learning (MTL), efforts have been made to extend MTL to improve the Alzheimer's disease cognitive score prediction by exploiting structure correlation. Though important and well-studied, three key aspects are yet to be fully handled in an unified framework: (*i*) appropriately modeling the inherent task relationship; (*ii*) fully exploiting the task relatedness by considering the underlying feature structure. (*iii*) automatically determining the weight of each task. To this end, we present the Bi-Graph guided self-Paced Multi-Task Feature Learning (BGP-MTFL) framework for exploring the relationship among multiple tasks to improve overall learning performance of cognitive score prediction. The framework consists of the two correlation regularization for features and tasks, $\ell_{2,1}$ regularization and self-paced learning scheme. Moreover, we design an efficient optimization method to solve the non-smooth objective function of our approach based on the Alternating Direction Method of Multipliers (ADMM) combined with accelerated proximal gradient (APG). The proposed model is comprehensively evaluated on the Alzheimer's disease neuroimaging initiative (ADNI) datasets. Overall, the proposed algorithm achieves an nMSE (normalized Mean Squared Error) of 3.923 and an wR (weighted R-value) of 0.416 for predicting eighteen cognitive scores, respectively. The empirical study demonstrates that the proposed BGP-MTFL model outperforms the state-of-the-art AD prediction approaches and enables identifying more stable biomarkers.

1. Introduction

Alzheimer's disease (AD) is a major degenerative brain disease that is characterized by slow and irreversible progression [1,2]. The alleviation of symptoms is possible through early diagnosis and intervention [3–8]. Recently, tremendous research has been done towards diagnosing Alzheimer's disease states by using Neuroimaging [9–14]. Magnetic resonance imaging (MRI) is characterized with non-invasive and reveals changes in the cerebral such as parietal lobe atrophy [15]. These changes are regarded as sensitive and stable biomarkers for Alzheimer's disease diagnosis [16–21]. There is thus an urgent need to accurately predict the progression of the disease measured by cognitive scores. Many objective cognitive assessments are designed to evaluate the patients' cognitive scores, e.g., the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) and Mini-Mental State Examination (MMSE). Predicting subjects' cognitive performance from MRI and identifying relevant imaging biomarkers are considered to be significant research directions in Alzheimer's disease research [20,22,23].

The relationship between imaging biomarkers and cognitive scores is previously studied by regression models [9,22]. When there exist multiple cognitive score prediction tasks, we aim to learn the multiple tasks concurrently due to the inherent correlations among multiple cognitive states. Multi-task learning (MTL) [24] is a learning paradigm, which aims to leverage the shared information in related tasks. Furthermore, to alleviate the influence of high dimensionality in MRI data to MTL, Multi-Task Feature Learning (MTFL) [25] is proposed to jointly select features from the multiple tasks with the $\ell_{2,1}$ regularization [26]. The assumptions of MTFL are: (*i*) the brain imaging measures are uncorrelated; (*ii*) the correlation among all tasks is uniform; (*iii*) the weight of each task is fixed in the learning process. However, we argue that such an assumption is too restrictive, because the true correlation among the features and tasks, and the learning order among tasks for

* Corresponding author at: Computer Science and Engineering, Northeastern University, Shenyang, China.
E-mail address: caopeng@cse.neu.edu.cn (P. Cao).

https://doi.org/10.1016/j.compbiomed.2022.106367
Received 12 June 2022; Received in revised form 16 November 2022; Accepted 25 November 2022
Available online 7 December 2022

different clinical score predictions are actually unknown in practice. To more flexibly model the relationship between the imaging biomarkers and cognitive scores, several efforts are made by introducing specific regularization as prior knowledge. Although these multi-task regression approaches have been applied to AD research and achieve promising results, to the best of our knowledge, relatively few studies consider both the relationships among imaging markers and the relationships among prediction tasks at the same time. On the one hand, the previous studies have shown that brain regions in AD patients are inter-connected, leading to cognitive decline in AD patients [27,28]. The prior intrinsic correlation is crucial to guide the learning process and is required to be incorporated into the MTFL model. On the other hand, there is evidence that the correlation among cognitive outcomes is unequal [20,29]. Cognitive tests evaluating different cognitive functions have low correlation and prefer different brain regions, such as ANART and RAVLT. On the contrary, some cognitive tests belonging to the same category thus have higher correlations, such as TOTAL and TOT6. Besides, the existing approaches ignore that the difficulty of each task is different. It is inappropriate to jointly learn these tasks without considering the learning order. Taking these into consideration, two questions naturally rise:

(1) How to incorporate the feature or task correlation into the MTFL?

(2) How to take into account the task difficulty and learn the dynamic weight for each task to control the learning order of the tasks?

To overcome these challenges, we develop an unified multi-task feature learning framework, called BGP-MTFL (Bi-Graph guided self-Paced Multi-Task Feature Learning), to model the relationship between the MRI features and between the cognitive scores. Considering that graphs are effective in modeling the correlation, we choose the graph structure to construct the interdependencies between the tasks or the features in this study. Then, we propose a dual graph lasso regularization to model the underlying structures, involving a graph structure for the predictive tasks and a graph structure for the image features, respectively. These two regularization terms are based upon the assumption that stronger correlated tasks or features are required to share more similar weights. Next, the dual fused lasso regularization is incorporated into the MTFL formulation to guide the learning of multiple tasks and capture the underlying structures at the level of tasks and features. The two regularization terms penalize large differences in the learned weights for highly correlated tasks or imaging biomarkers. Furthermore, the weight of each task is not fixed during the model training procedure. It is desirable for the multi-task learning paradigm to dynamically consider the curriculum defined by the learner rather than a fixed and constant task learning order. Inspired by the self-paced learning [30,31], we capture the dynamic order of learned tasks to boost the multi-task learning performance. At last, optimizing the formulation with the multiple non-smooth regularization is challenging. We solve the complicated formulation based on the Alternating Direction Method of Multipliers (ADMM) [32] combined with the accelerated proximal gradient (APG) [33] method in our work. We conclude our major contributions as follows:

(1) **Two graph-guided regularization terms are proposed.** Considering that the feature correlation information and task correlation information help improve the regression performance and identify critical biomarkers significantly, we propose Bi-graph guided regularization, which takes task correlation and feature correlation into consideration at the same time and model these dependencies. The proposed regularization allows the MTFL model to identify the optimal imaging markers with high prediction power.

(2) **A self-paced multi-task feature learning scheme.** A key issue in multi-task learning is to understand the relationship and learning order between tasks since this understanding can be incorporated into the learning process to improve the generalization performance of all the tasks. Besides the exploration of the inherent correlation within the tasks or features, we also design a self-paced learning strategy to improve the training performance of multiple tasks. It gradually produces more reliable predictive tasks for learning better model weights, which in turn, boost the final multi-task feature learning process.

(3) **An effective optimization algorithm.** To estimate the weight parameters in the proposed BGP-MTFL method, an effective optimization algorithm based on ADMM combined with APG is proposed.

(4) **Comprehensive evaluation on the effectiveness of the proposed model.** The experiments on the ADNI dataset verified that the proposed framework performs better than the single-task regression models and the multi-task learning models. Furthermore, the proposed approach is able to find the more stable biomarkers associated with AD through the proposed multi-task learning scheme.

The rest of this work is organized as follows. In Section 2, a thorough formulation and the optimization approach are introduced in detail. In Section 3, the effectiveness of the proposed model is evaluated. In Section 4, comprehensive experiments are discussed to further demonstrate the effectiveness of our proposed model. Finally, we conclude this paper in Section 5.

2. Methods

2.1. Notation

In the process of AD clinical diagnosis, multiple cognitive tests are conducted for comprehensively evaluations. We view the prediction of each cognitive score as a single task and formulate the multiple prediction tasks as a multi-task learning paradigm. In this work, we concentrate on predicting k tasks (corresponding to k cognitive scores) with p features. The number of samples is n. The covariates matrix is represented as $X \in \mathbb{R}^{n\times p}$, ith column in X is $\boldsymbol{x}_i$. The responses matrix is represented as $Y \in \mathbb{R}^{n\times k}$, jth column in Y is $\boldsymbol{y}_j$ and each row in Y represents cognitive scores from one sample. $\Theta \in \mathbb{R}^{p\times k}$ is the parameter matrix to be learned, where mth column is $\boldsymbol{\theta}_{\cdot,m} \in \mathbb{R}^p$, corresponding to task m $(m = 1,\dots,k)$, jth row in Θ is $\boldsymbol{\theta}_{j,\cdot} \in \mathbb{R}^k$, corresponding to feature $j(j = 1,\dots,p)$. For simply, we name $\boldsymbol{\theta}_{\cdot,m}$ as $\boldsymbol{\theta}_m$ later. The multi-task regression is formulated as:

$$\min_{\Theta} \quad L(X,Y,\Theta) + \lambda R(\Theta) \;, \tag{1}$$

where $L(Y,X,\Theta) = \frac{1}{2}\|Y - X\Theta\|_F^2$ in our work, $R(\cdot)$ is a regularization term that aims to discover more stable biomarkers by introducing prior knowledge into the model, and $\lambda > 0$ is a regularization parameter to control the balance between the regularization term and the regression loss.

2.2. Constructing the feature graph and task graph

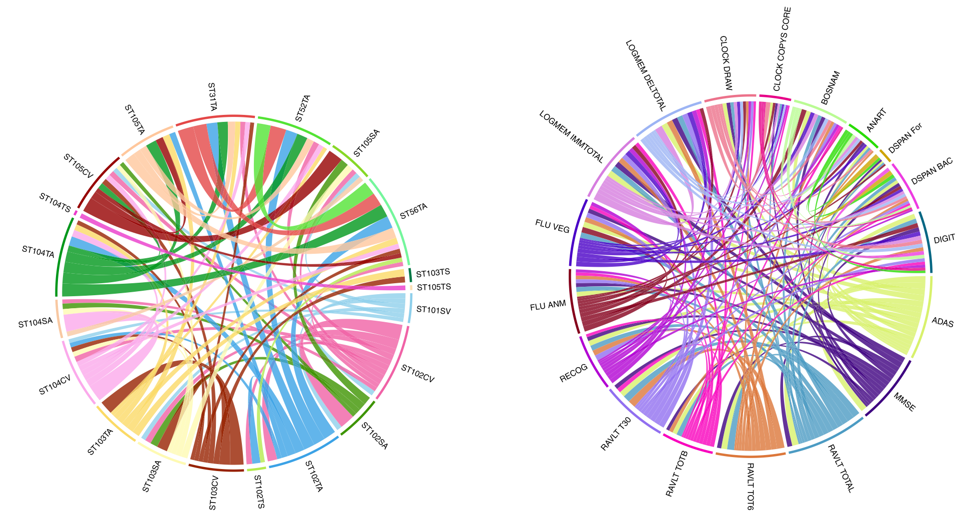
The traditional MTFL method has two limitations in AD study: (i). The MRI features are structurally interrelated, and the highly interrelated features have similar weight parameters. However, the traditional MTFL method ignores this interrelationship among features; (ii). MTFL forces all the tasks to share the same features with $\ell_{2,1}$ norm. However, the different cognitive test tasks should be related with the different imaging biomarkers.

To solve limitation (i), a feature correlation matrix $C^f \in \mathbb{R}^{p\times p}$ is constructed to model the interrelationship among features. Element $c^f_{i,j}$ in C^f indicates the Pearson Correlation Coefficient (PCC) between feature i and feature j, which can be computed as:

$$c^f_{i,j} = \frac{cov(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sigma_{\boldsymbol{x}_i}\sigma_{\boldsymbol{x}_j}} = \frac{E[(\boldsymbol{x}_i - \mu_{\boldsymbol{x}_i})(\boldsymbol{x}_j - \mu_{\boldsymbol{x}_j})]}{\sigma_{\boldsymbol{x}_i}\sigma_{\boldsymbol{x}_j}} \;, \tag{2}$$

where $i,j = 1,\dots,p$ and $c^f_{i,j} \in [-1,1]$.

A bigger value of $|c^f_{i,j}|$ indicates a stronger correlation between feature i and feature j. The visualization of the feature correlation matrix is shown in Fig. 1(a), where the arcs indicate the correlations among



(a) Visualization of the feature correlation. (b) Visualization of the task correlation.

Fig. 1. Visualization of the feature correlation and task correlation. The arcs indicate the features/tasks. The connections represent the correlations between features/tasks, and the width of each connection indicates the correlation strength.

features. According to Fig. 1(a), we can clearly observe that there exist strong correlations among several features, such as ST52TA(Cortical Thickness Average of LeftPrecuneus) and ST31TA(Cortical Thickness Average of LeftInferiorParietal), ST52TA(Cortical Thickness Average of LeftPrecuneus) and ST56TA(Cortical Thickness Average of LeftSuperiorFrontal). In general, more than 40% of correlations in the feature correlation matrix are greater than 0.2. Inspired by the above analysis, a feature correlation guided regularization is crucial to control the difference between two correlated features. To this end, an undirected graph structure $\mathbb{G}^f = (V^f, E^f)$ is constructed to represent the correlations among features. The vertexes V^f denote features, each edge in E^f denotes the correlation between two features. For example, edge $e^f(i,j) \in E^f$ corresponds to the correlation between feature i and feature j. Feature correlation matrix C^f is then normalized into S by dividing k^f, where k^f is the number of edges in Graph $\mathbb{G}^f$. The normalized feature correlation S is obtained as:

$$s_{i,j} = \begin{cases} -\frac{c^f_{i,j}}{k^f} & (i,j) \in E^f, i \neq j \\ \frac{\sum_{i=1, i\neq j}^{p} |c^f_{i,j}|}{k^f} & (i,j) \in E^f, i = j \\ 0 & otherwise \end{cases} \tag{3}$$

To solve the limitation (2), we assume that the intrinsic correlation exits in the multiple tasks. To discover such dependent structures among each pair of the target response variables, a task correlation matrix $C^t \in \mathbb{R}^{k\times k}$ is constructed. The correlation between task m and task n is calculated by Eq. (4):

$$c^t_{m,n} = \frac{cov(\boldsymbol{y}_m, \boldsymbol{y}_n)}{\sigma_{\boldsymbol{y}_m}\sigma_{\boldsymbol{y}_n}} = \frac{E[(\boldsymbol{y}_m - \mu_{\boldsymbol{y}_m})(\boldsymbol{y}_n - \mu_{\boldsymbol{y}_n})]}{\sigma_{\boldsymbol{y}_m}\sigma_{\boldsymbol{y}_n}} \tag{4}$$

where $m, n = 1, \dots, k$.

Similar to the feature graph, an undirected task graph $\mathbb{G}^t = (V^t, E^t)$ is developed to represent the task correlation in our study. Different from the feature correlation matrix, vertexes in undirected graph $\mathbb{G}^t$ denote tasks. An undirected edges $e^t(m,n) \in E^t$ connect task m and task n. The estimated graph is shown in Fig. 1(b). In Fig. 1(b), we can observe different correlation strengths exist between any two tasks. For instance, RAVLT TOTB and RAVLT TOT6 have strong correlations owing to the fact that they aim to evaluate the same cognitive function, whereas the prediction tasks of DSPAN and ANART have weaker correlations due to the evaluation for different cognitive functions. These observations motivate us to incorporate task correlation matrix into multi-task learning model and guide the multi-task feature learning. The task correlation matrix C^t is normalized as $Z = C^t/k^t$ ($Z \in \mathbb{R}^{k\times k}$), in which k^t is the number of edge in graph $\mathbb{G}^t$. The mth row and nth column of Z is formulated as:

$$z_{m,n} = \begin{cases} \frac{-c^t_{m,n}}{k^t} & (m,n) \in E^t, m \neq n \\ \frac{\sum_{m=1, m\neq n}^{k} |c^t_{m,n}|}{k^t} & (m,n) \in E^t, m = n \\ 0 & otherwise \end{cases} \tag{5}$$

2.3. Formulation

2.3.1. Feature/Task graph guided structure regularization

To take into account the correlations among features or tasks during the multi-task learning process, both the feature graph guided regularization and task graph guided regularization are proposed. More specifically, the feature graph guided regularization shrinks the difference of the weight parameters between correlated features towards zero, so that more features that are inherently beneficial to the cognitive predictions are considered. With the feature graph guided regularization, the features correlated with the selected features but ignored by $\ell_{2,1}$ norm can be preserved, which is important for the accurate cognitive score predictions and the biological interpretation. The feature graph guided regularization is formulated as:

$$\|S\Theta\|_1 = \sum_{e^f(i,j)\in E^f} |s_{i,j}| \|\theta_{i,\cdot} - sign(s_{i,j})\theta_{j,\cdot}\|_1 \tag{6}$$

The task graph guided regularization encourages the highly correlated tasks to share similar features. It allows for more consistent identification of the relevant imaging markers through a better understanding of the underlying associations among predictive tasks. If task m and task n are strongly correlated, the difference between the corresponding regression coefficients θ_m and θ_n should be towards to zero. The task graph guided regularization can be written as the following formula:

$$\|\Theta Z\|_1 = \sum_{e^t(m,n)\in E^t} |z_{m,n}| \|\theta_m - sign(z_{m,n})\theta_n\|_1 \tag{7}$$

2.3.2. Self-paced learning strategy

Bi-graph guided multi-task feature learning learns the multiple tasks simultaneously with the assumption that the multiple tasks have the same learning difficulty level. We thoroughly analyze the difficulties of all tasks, through which we demonstrate the potential gain of self-paced

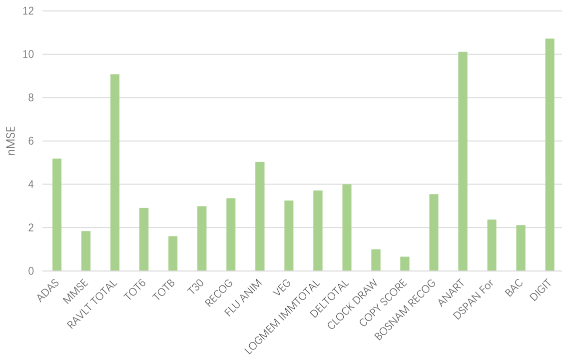


Fig. 2. Performance of single-task learning on eighteen cognitive scores in terms of nMSE. Tasks with higher nMSE are usually difficult to learn.

learning. Fig. 2 shows the nMSE performance of single-task learning on the eighteen cognitive scores. Tasks with higher nMSE are considered to be more difficult to be learned. As can be seen in Fig. 2, some tasks such as ANART and DIGIT are difficult to be learned, while some tasks such as CLOCK DRAW and COPYSCORE are relatively easier to be learned. Hence, the complex tasks need to be solved by leveraging the knowledge previously learned from the easier tasks. To solve it, we propose to utilize self-pace learning to monitor the learning procedure and automatically learn a policy to adjust the relative weights of the tasks with different difficulty levels. Learning easier tasks first can provide better initialization for subsequent learning with harder tasks, which helps improve the generalization performance of the multi-task learning.

2.3.3. Overall loss function

Here, we introduce the loss function in our BGP-MTFL algorithm involving $\ell_{2,1}$ regularization, feature graph guided regularization, task graph guided regularization and self-paced learning strategy. The overall loss function for our BGP-MTFL algorithm can be formulated as Eq. (8). The self-paced learning process and the Bi-graph guided regularization are illustrated in Fig. 3.

$$\begin{aligned}\min_{\Theta,B}\ &\frac{1}{2}\|(Y-X\Theta)B\|_F^2+\lambda_g\|\Theta\|_{2,1}+\lambda_f\|S\Theta\|_1\\&+\lambda_t\|\Theta Z\|_1-\lambda_{sp}\mathrm{Tr}(B)\end{aligned}\tag{8}$$

In the objective function, the first term $\frac{1}{2}\|(Y-X\Theta)B\|_F^2$ denotes the training losses over all the tasks. The second term $\|\Theta\|_{2,1}$ is an $\ell_{2,1}$ regularization, the advantage of which is to promote the sparsity and discover the relevant features. In the $\ell_{2,1}$ regularization, the inner ℓ_2 norm enforce column non-sparsed, while the outer ℓ_1 norm enforces row sparsity. Due to the existing correlations among different tasks, it is reasonable to encourage the parameters from multiple tasks to share similar sparsity patterns. In other words, if some features are redundant and negligible, $\ell_{2,1}$ norm tends to reduce the associated rows of output weight to zero. The third and fourth term act as regularization for preserving the feature correlation and task correlation with the constructed feature graph and task graph. λ_g, λ_f and λ_t are the regularization parameters to control the balance between the regression loss and the multiple regularization. λ_{sp} is used to identify the task difficulty. As λ_{sp} grows, the tasks with larger losses are gradually engaged to train a more sophisticated model. $B\in\mathbb{R}^{k\times k}$ is a diagonal matrix with elements on the diagonal (B_{mm}) representing the weight of each task. All tasks are categorized into two types: easy tasks and hard tasks. We assume that a task whose loss is smaller than a certain threshold λ_{sp} is regarded as an easy task, because it achieves a higher accuracy in the training dataset. Otherwise, a task whose loss is larger than a certain threshold λ_{sp} is regarded as a hard task. The easy task is selected in training and the weight for the easy tasks, named B_e is set to 1. On the contrary, the hard task is unselected in training and the weight for hard tasks, named B_h is set to 0. However, for the correctness of the algorithm, we set B_h to a number close to 0 (B_h =0.01) for the hard task. The effects of different values of B_e and B_h on regression performance are explored in Section 4.2.

2.4. Optimization of BGP-MTFL

The optimization of Eq. (8) is challenging due to the existence of the non-smooth regularization terms. In this work, we alternatively update the multiple parameters to simplify the process of optimizing BGP-MTFL. More specifically, we introduce slack variables $Q=\Theta$, $P=S\Theta$ and $R=\Theta Z$ into Eq. (8) and rewrite Eq. (8) as an augmented Lagrangian function:

$$\begin{aligned}&L_\rho(\Theta,B,Q,P,R,U_{(1)},U_{(2)},U_{(3)})=\frac{1}{2}\|(Y-X\Theta)B\|_F^2+\lambda_g\|Q\|_{2,1}+\lambda_f\|P\|_1\\&+\lambda_t\|R\|_1+<U_{(1)},\Theta-Q>+\frac{\rho}{2}\|\Theta-Q\|^2+<U_{(2)},S\Theta-P>\\&+\frac{\rho}{2}\|S\Theta-P\|^2+<U_{(3)},\Theta Z-R>+\frac{\rho}{2}\|\Theta Z-R\|^2-\lambda_{sp}Tr(B)\ ,\end{aligned}\tag{9}$$

where $U_{(1)}$, $U_{(2)}$ and $U_{(3)}$ are augmented Lagrangian multipliers.

2.4.1. Optimizing B,λ_{sp} when the other parameters are fixed

The value of parameter $B_{m,m}$ depends on the training error of task m. We update $B_{m,m}$ at the $(t+1)$-th iteration according to Eq. (10). The update of λ_{sp} at the $(t+1)$-th iteration is formulated as Eq. (11).

$$B_{m,m}^{(t+1)}=\begin{cases}\tau_h & t=0\\ \tau_e & t>0 \text{ and } (\boldsymbol{y}_m-X\boldsymbol{\theta}_m)<\lambda_{sp}^{(t+1)}\\ \tau_h & t>0 \text{ and } (\boldsymbol{y}_m-X\boldsymbol{\theta}_m)\geq\lambda_{sp}^{(t+1)}\end{cases}\tag{10}$$

$$\lambda_{sp}^{(t+1)}=\begin{cases}0.1 & t=0\\ c\lambda_{sp}^{(t)} & t>0\ ,\end{cases}\tag{11}$$

where c is used to control the learning pace. The value of c is usually greater than 1 such that λ_{sp} is relaxed to include more tasks at each iteration (In our study, c is set to 1.1).

2.4.2. Optimizing Θ when the other parameters are fixed

When the other variables except Θ are fixed, the update of Θ at the $(t+1)$-th iteration is carried out by:

$$\begin{aligned}\Theta^{t+1}=\arg\min_{\Theta}\ &\frac{1}{2}\|(Y-X\Theta)B^{(t)}\|_F^2+<U_{(1)}^{(t)},\Theta-Q^{(t)}>\\&+\frac{\rho}{2}\|\Theta-Q^{(t)}\|^2+<U_{(2)}^{(t)},S\Theta-P^{(t)}>+\frac{\rho}{2}\|S\Theta-P^{(t)}\|^2\\&+<U_{(3)}^{(t)},\Theta Z-R^{(t)}>+\frac{\rho}{2}\|\Theta Z-R^{(t)}\|^2\end{aligned}\tag{12}$$

Eq. (12) is regarded as a convex function and its solution can be obtained by setting its derivation to zero.

$$\begin{aligned}0=&-X^{\mathrm{T}}(Y-X\Theta)B^{(t)}B^{(t)T}+U_{(1)}^{(t)}+\rho(\Theta-Q^{(t)})\\&+SU_{(2)}^{(t)}+\rho S(S\Theta-P^{(t)})+U_{(3)}^{(t)}Z+\rho(\Theta Z-R^{(t)})Z\end{aligned}\tag{13}$$

Furthermore, we define a new function: $\Phi=ZZ$. The mth row and mth column of Φ is $\phi_{m,m}$. Both Z and Φ are symmetric matrix. Θ can be update concurrently by the θ_m. Therefore, Eq. (13) is rewrite as Eq. (14).

$$\begin{aligned}0=&-X^{\mathrm{T}}(\boldsymbol{y}_m-X\boldsymbol{\theta}_m)B_{m,m}^{(t)2}+\boldsymbol{u}_{(1)m}^{(t)}+\rho(\boldsymbol{\theta}_m-\boldsymbol{q}_m^{(t)})\\&+S\boldsymbol{u}_{(2)m}^{(t)}+\rho SS\boldsymbol{\theta}_m-\rho S\boldsymbol{p}_m^{(t)}+(\boldsymbol{u}_{(3)m}^{(t)}-\sum_{m=1,m\neq l}^{k}z_{m,l}\boldsymbol{u}_{(3)m}^{(t)})\\&+\rho(\phi_{m,m}\boldsymbol{\theta}_m-\sum_{m=1,m\neq l}^{k}\phi_{m,l}\boldsymbol{\theta}_l)-\rho(\boldsymbol{r}_m^{(t)}-\sum_{m=1,m\neq l}^{k}z_{m,l}\boldsymbol{r}_l^{(t)})\end{aligned}\tag{14}$$

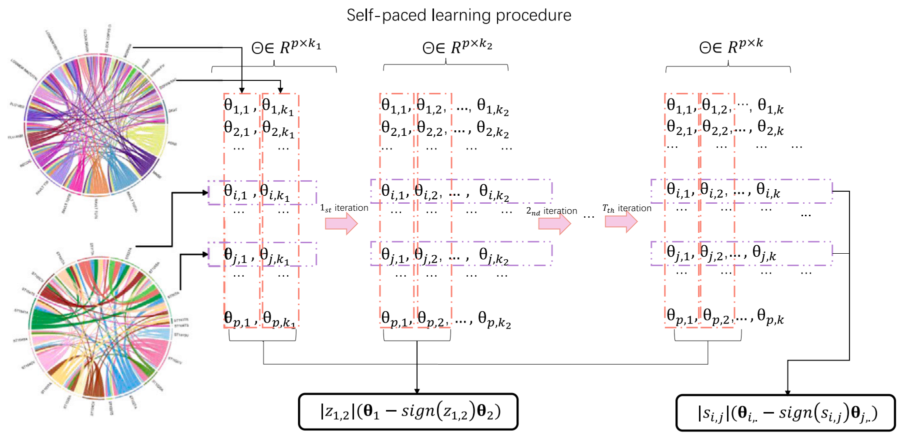


Fig. 3. Illustration of the self-paced learning process and Bi-graph guided regularization. Each column of Θ corresponds to a single task and each row of Θ represents a feature dimension. Tasks are learned dynamically in T iterations. k_1, k_2 and k are the number of tasks in the 1st, 2nd, Tth iterations, respectively. In each iteration, feature graph and task graph guided regularization are integrated to the multi-task learning model to guide the learning.

The update of $\boldsymbol{\theta}_m$ at the $(t+1)$-th iteration is given by Eqs. (15) to (17).

$$\boldsymbol{\theta}_m^{(t+1)} = F_m^{(t)-1}\boldsymbol{b}_m^{(t)} \tag{15}$$

$$F_m^{(t)} = X^{\mathrm{T}}XB_{m,m}^{(t)2} + \rho SS + \rho(1+\phi_{m,m})I \tag{16}$$

$$\begin{aligned}\boldsymbol{b}_m^{(t)} =& X^{\mathrm{T}}\boldsymbol{y}_m B_{m,m}^{(t)2} - \boldsymbol{u}_{(1)m}^{(t)} + \rho\boldsymbol{q}_m^{(t)} - S\boldsymbol{u}_{(2)m} + \rho S\boldsymbol{p}_m \\ &- (\boldsymbol{u}_{(3)m}^{(t)} - \sum_{m=1,m\neq l}^{k} Z_{m,l}\boldsymbol{u}_{(3)l}^{(t)}) + \rho(\boldsymbol{r}_m^{(t)} - \sum_{m=1,m\neq l}^{k} Z_{m,n}\boldsymbol{r}_l^{(t)})\end{aligned} \tag{17}$$

2.4.3. Update $\boldsymbol{Q}$ when the other parameters are fixed

When the other parameters are given and fixed, Q can be solved by the following equation.

$$\begin{aligned}Q^{(t+1)} &= \arg\min_{Q} \lambda_g\|Q\|_{2,1} + < U_{(1)}^{(t)}, \Theta^{(t+1)} - Q > + \frac{\rho}{2}\|\Theta^{(t+1)} - Q\|^2 \\ &= \arg\min_{Q} \frac{\lambda_g}{\rho}\|Q\|_{2,1} - \frac{1}{\rho} < U_{(1)}^{(t)}, Q - \Theta^{(t+1)} > \\ &\quad + \frac{1}{2}\|Q - \Theta^{(t+1)}\|^2 + \frac{1}{2\rho^2}U_{(1)}^{(t)2} \\ &= \arg\min_{Q} \frac{1}{2}\|Q - \Theta^{(t+1)} - \frac{U_{(1)}^{(t)}}{\rho}\|^2 + \frac{\lambda_g}{\rho}\|Q\|_{2,1}\end{aligned} \tag{18}$$

By replacing $(\Theta^{(t+1)} + \frac{U_{(1)}^{(t)}}{\rho})$ with $\Lambda_{(1)}^{(t+1)}$, the Eq. (18) is rewritten as Eq. (19).

$$Q^{(t+1)} = \arg\min_{Q} \frac{1}{2}\|Q - \Lambda_{(1)}^{(t+1)}\|^2 + \frac{\lambda_g}{\rho}\|Q\|_{2,1} \tag{19}$$

Each row of Q is optimized by:

$$\boldsymbol{q}_i^{(t+1)} = \arg\min_{\boldsymbol{q}_i} \frac{1}{2}\|\boldsymbol{q}_i - \boldsymbol{\alpha}_{(1)i}^{(t+1)}\|^2 + \frac{\lambda_g}{\rho}\|\boldsymbol{q}_i\|_1 \tag{20}$$

where $\boldsymbol{\alpha}_{(1)i}^{(t+1)}$ and $\boldsymbol{q}_i^{(t+1)}$ are the ith row of $\Lambda_{(1)}^{(t+1)}$ and $Q^{(t+1)}$, respectively. $\boldsymbol{q}_i^{(t+1)}$ can be updated by Lemma 1 considering that it is strictly convex [34].

Lemma 1. *For any $\lambda_g \geq 0$, Eq. (20) can be calculated by the following:*

$$\boldsymbol{q}_i^{(t+1)} = \frac{\max(\|\boldsymbol{\alpha}_{(1)i}^{(t+1)}\|_2 - \frac{\lambda_g}{\rho}, 0)}{\|\boldsymbol{\alpha}_{(1)i}^{(t+1)}\|_2}\boldsymbol{\alpha}_{(1)i}^{(t+1)} \tag{21}$$

2.4.4. Optimizing P/R when the other parameters are fixed

Owning to the introduction of similarity weights S and Z, the optimization of P and R is in the same way. Here, we introduce the optimization of P in detail. P is updated as follows:

$$\begin{aligned}P^{(t+1)} &= \arg\min_{P} \lambda_f\|P\|_1 + < U_{(2)}^{(t)}, S\Theta^{(t+1)} - P > + \frac{\rho}{2}\|S\Theta^{(t+1)} - P\|^2 \\ &= \arg\min_{P} \frac{\lambda_f}{\rho}\|P\|_1 - \frac{1}{\rho} < U_{(2)}^{(t)}, P - S\Theta^{(t+1)} > + \frac{1}{2}\|P - S\Theta^{(t+1)}\|^2 \\ &\quad + \frac{1}{2\rho_2}U_{(2)}^{(t)2} \\ &= \arg\min_{P} \frac{1}{2}\|P - S\Theta^{(t+1)} - \frac{U_{(2)}^{(t)}}{\rho}\|^2 + \frac{\lambda_f}{\rho}\|P\|_1\end{aligned} \tag{22}$$

Eq. (22) is rewritten as Eq. (23) by replacing $(S\Theta^{(t+1)} + \frac{U_{(2)}^{(t)}}{\rho})$ with $\Lambda_{(2)}^{(t+1)}$.

$$P^{(t+1)} = \arg\min_{P} \frac{1}{2}\|P - \Lambda_{(2)}^{(t+1)}\|^2 + \frac{\lambda_f}{\rho}\|P\|_1 \tag{23}$$

Each row of P is computed by Eq. (24).

$$\boldsymbol{p}_i^{(t+1)} = \arg\min_{\boldsymbol{p}_i} \frac{1}{2}\|\boldsymbol{p}_i - \boldsymbol{\alpha}_{(2)i}^{(t+1)}\|^2 + \frac{\lambda_f}{\rho}\|\boldsymbol{p}_i\|_1 \tag{24}$$

where $\boldsymbol{\alpha}_{(2)i}^{(t+1)}$ and $\boldsymbol{p}_i^{(t+1)}$ are the ith row of $\Lambda_{(2)}^{(t+1)}$ and $P^{(t+1)}$ respectively. For any $\lambda_f \geq 0$, we can calculate Eq. (24) by the following:

$$p_{i,j}^{(t+1)} = sign(\alpha_{(2)i,j}^{(t+1)})\max(|\alpha_{(2)i,j}^{(t+1)}| - \frac{\lambda_f}{\rho}, 0) \tag{25}$$

2.4.5. Update $U_{(1)}^{(t)}$, $U_{(2)}^{(t)}$ and $U_{(3)}^{(t)}$

The updates of augmented Lagrangian multipliers are performed by Eq. (26) according to the standard ADMM.

$$\begin{aligned}U_{(1)}^{(t+1)} &= U_{(1)}^{(t)} + \rho(\Theta^{(t+1)} - Q^{(t+1)}) \\ U_{(2)}^{(t+1)} &= U_{(2)}^{(t)} + \rho(S\Theta^{(t+1)} - P^{(t+1)}) \\ U_{(3)}^{(t+1)} &= U_{(3)}^{(t)} + \rho(\Theta^{(t+1)}Z - R^{(t+1)})\end{aligned} \tag{26}$$

Generally, the optimization process of BGP-MTFL can be summarized as Algorithm 1.

3. Experiment

3.1. Data

We conduct experiments on Alzheimer's Disease Neuroimaging Initiative(ADNI) database, which is publicly available and comprises 1.5 T structural MRI. The MRI data is processed with FreeSurfer Software Suite developed by the University of California at San Francisco team (http://surfer.nmr.mgh.harvard.edu/). In total, 319 features

Algorithm 1 ADMM optimization of BGP-MTFL

Input: Feature matrix X, Target matrix Y, λ_g, λ_f, λ_t, ρ, task number k.
Output: Θ.
Initialization: $\Theta^{(0)} \leftarrow 0$, $Q^{(0)} \leftarrow 0$, $P^{(0)} \leftarrow 0, R^{(0)} \leftarrow 0$, $U^{(0)}_{(1)} \leftarrow 0$, $U^{(0)}_{(2)} \leftarrow 0$, $U^{(0)}_{(3)} \leftarrow 0, \lambda_{sp} \leftarrow 0.1$, $c \leftarrow 1.1$.
repeat
Update $B^{(t+1)}, \lambda_{sp}^{(t+1)}$ by Eqs. (10) and (11).
Update $\Theta^{(t+1)}$ by Eqs. (15) to (17).
Update $Q^{(t+1)}$ by Eqs. (18) to (21).
Update $P^{(t+1)}$ and $R^{(t+1)}$ by Eqs. (22) to (25).
Update $U^{(t+1)}_{(1)}$, $U^{(t+1)}_{(2)}$, $U^{(t+1)}_{(3)}$ by Eq. (26).
until Convergence.

(68 × 4+2 × 1+1+44) are generated from 70 cortical regions and 44 subcortical regions. In the 70 cortical regions, 68 cortical regions contain four features: cortical thickness average (TA), the standard deviation of a thickness (TS), surface area (SA) and cortical volume (CV). In addition, left and right hemisphere regions contain a single SA feature, respectively. Besides, Total Intracranial Volume (ICV) denotes the volume of the cranial cavity as taken from a 3D T1 MRI. For the 44 subcortical regions, only feature of subcortical volume (SV) is calculated in each subcortical region. The feature description is listed in Table 1. The target variables explored in this study are 18 clinic cognitive scores: ADAS, MMSE, RAVLT TOTAL, RAVLT TOT6, RAVAL TOTB, RAVLT T30, RAVLT T30, RAVLT RECOG, FLU ANIM, FLU VEG, LOGMEM IMMTOTAL, LOGMEM DELTOTAL, BOSNAM, ANART, DSPAN For, DSPAN BAC, DIGIT. All of the 18 cognitive scores are given by the experts based on the gold standard.

These data is further processed by the following six steps: (1) samples without baseline MRI records are removed. (2) ROIs whose name is "unknown" are deleted. (3) Features with a missing rate greater than 10% are deleted. (4) Samples with missing labels are deleted. (5) The missing values are replaced with average values. (6) All data are normalized by standard score. Finally, 788 subjects are involved in our study. All the subjects include three groups: Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI) and Normal Control (NC). The summary of subject information is shown in Table 2.

3.2. Settings

A nested cross-validation procedure is employed to evaluate the performance of different models and select the optimal hyperparameter value. The nested cross-validation has an inner loop cross-validation nested in each outer cross-validation. The inner loop is responsible for model hyperparameter tuning, while the outer loop is responsible for performance evaluation. Specifically, the outer loop in our study is a 10-fold cross-validation procedure, in each fold of which involves a training set (90%) and test set (10%). It must be noted that the dataset is split by class for training and testing. The model is trained in the training set, and the trained model is evaluated in the unseen test set. In the training set of each fold, the inner loop is conducted. The inner loop is a 5-fold cross-validation procedure for the hyperparameter optimization, such as grid search. The scope of the regularization parameters of our model $\lambda_g, \lambda_f, \lambda_t$ and λ_{sp} is [0.001,0.01,0.1,1,10,100,1000]. Grid search is considered a very traditional hyperparameter optimization method since we are basically "brute-forcing" all possible combinations. The evaluation of each hyperparameter is performed using a 5-fold cross-validation to guarantee the stability of the hyperparameter selection. With the best hyperparameter chosen on the training set, the performance on the test set is evaluated and the mean score over all 10 folds is reported as the overall performance.

Note that both the feature correlation graph and task correlation graph are calculated based on the trainval set. For quantitative evaluation, correlation coefficient (CC) and root mean squared error (rMSE) are chosen to evaluate the performance on each task. Moreover, the nMSE and wR are applied to evaluate the comprehensive performance of the proposed model on all the tasks. The measures are formulated as Eqs. (27) to (30).

$$\mathrm{CC}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{cov(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sigma(\boldsymbol{y})\sigma(\hat{\boldsymbol{y}})} \tag{27}$$

$$\mathrm{rMSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2}{n} \tag{28}$$

$$\mathrm{nMSE}\left(Y, \hat{Y}\right) = \frac{\sum_{h=1}^{k} \frac{\|\boldsymbol{y}_h - \hat{\boldsymbol{y}}_h\|_2^2}{\sigma(\boldsymbol{y}_h)}}{\sum_{h=1}^{k} n_h} \tag{29}$$

$$\mathrm{wR}\left(Y, \hat{Y}\right) = \frac{\sum_{h=1}^{k} \mathrm{CC}(\boldsymbol{y}_h, \hat{\boldsymbol{y}}_h) n_h}{\sum_{h=1}^{k} n_h} \tag{30}$$

In Eqs. (27)–(30), $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are the ground truth value and predicted value of a single task. Y and $\hat{Y}$ are the ground truth and predicted value of all tasks. $\boldsymbol{y}_h$ and $\hat{\boldsymbol{y}}_h$ are the hth row of Y and $\hat{Y}$, respectively. n_h is the number of subjects in task h. k is the number of tasks.

3.3. Result

To verify the effectiveness of the proposed BGP-MTFL model, we compare BGP-MTFL with two single-task learning methods (Ridge, Lasso, SVR and Random Forest) and seven multi-task learning methods.

1. Single-task learning methods

- Ridge: Ridge regularization, also called an ℓ_2 penalty, presents only one solution and converges the problem to a certain extent.

$$\min_{\theta_i} \frac{1}{2}\|\boldsymbol{y}_i - X\boldsymbol{\theta}_i\|_F^2 + \lambda\|\boldsymbol{\theta}_i\|_2 \tag{31}$$

- Lasso: Lasso regularization, also called an ℓ_1 penalty, is going to take the absolute value of coefficients, it helps in feature selection by reducing certain rows when not important. It may have multiple solutions.

$$\min_{\theta_i} \frac{1}{2}\|\boldsymbol{y}_i - X\boldsymbol{\theta}_i\|_F^2 + \lambda\|\boldsymbol{\theta}_i\|_1 \tag{32}$$

2. Multi-task learning methods

- Multi-task feature learning (MTFL): MTFL utilizes $\ell_{2,1}$ regularization to exploit the inherent relationship among all tasks.

$$\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda\|\Theta\|_{2,1} \tag{33}$$

- Robust Multi-task Learning (RMTL) [35]: RMTL applies low-rank regularization and group-sparse regularization to capture the task relationships and identifies the anomalous tasks respectively.

$$\min_{\Theta} \frac{1}{2}\|Y - X(L+S)\|_F^2 + \lambda_1\|L\|_* + \lambda_2\|S\|_{1,2} \tag{34}$$

- Robust Multi-task Feature Learning (rMTFL) [36]: For the same problem as RMTL, rMTFL decomposes the weight matrix Θ into two components L and S. regularization on L capture the shared features among tasks and regularization on S discovers the outlier tasks.

$$\min_{\Theta} \frac{1}{2}\|Y - X(L+S)\|_F^2 + \lambda_1\|L\|_{2,1} + \lambda_2\|S^T\|_{2,1} \tag{35}$$

- Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) [37]: G-SMuRFS group relevant features together with a group-level $\ell_{2,1}$ norm strategy to guide the learning process.

$$\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{G_{2,1}} + \lambda_2\|\Theta\|_{2,1} \tag{36}$$

Table 1
The description of features generated from 70 cortical regions and 44 subcortical regions. L or R denotes the left or right hemisphere.

ID	Brain region	Feature	ID	ROI name	Feature
1	Banks superior temporal sulcus (L, R)	CV, SA, TA, TS	1	Accumbens area (L, R)	SV
2	Caudal anterior cingulate cortex (L, R)	CV, SA, TA, TS	2	Amygdala (L, R)	SV
3	Caudal middle frontal gyrus (L, R)	CV, SA, TA, TS	3	Caudate (L, R)	SV
4	Cuneus cortex (L, R)	CV, SA, TA, TS	4	Cerebellum cortex (L, R)	SV
5	Entorhinal cortex (L, R)	CV, SA, TA, TS	5	Cerebellum white matter(L, R)	SV
6	Frontal pole (L, R)	CV, SA, TA, TS	6	Cerebral cortex (L, R)	SV
7	Fusiform gyrus (L, R)	CV, SA, TA, TS	7	Cerebral white matter(L, R)	SV
8	Inferior parietal cortex (L, R)	CV, SA, TA, TS	8	Choroid plexus (L, R)	SV
9	Inferior temporal gyrus (L, R)	CV, SA, TA, TS	9	Hippocampus(L, R)	SV
10	Insula (L, R)	CV, SA, TA, TS	10	Inferior lateral ventricle (L, R)	SV
11	IsthmusCingulate (L, R)	CV, SA, TA, TS	11	Lateral ventricle(L, R)	SV
12	Lateral occipital cortex (L, R)	CV, SA, TA, TS	12	Pallidum (L, R)	SV
13	Lateral orbital frontal cortex (L, R)	CV, SA, TA, TS	13	Putamen(L, R)	SV
14	Lingual gyrus (L, R)	CV, SA, TA, TS	14	Thalamus (L, R)	SV
15	Medial orbital frontal cortex (L, R)	CV, SA, TA, TS	15	Ventricle diencephalon(L, R)	SV
16	Middle temporal gyrus (L, R)	CV, SA, TA, TS	16	Vessel (L, R)	SV
17	Paracentral lobule (L, R)	CV, SA, TA, TS	17	Brain stem (Bilateral)	SV
18	Parahippocampal gyrus (L, R)	CV, SA, TA, TS	18	Corpus callosum anterior (Bilateral)	SV
19	Pars opercularis(L, R)	CV, SA, TA, TS	19	Corpus callosum central (Bilateral)	SV
20	Pars orbitalis (L, R)	CV, SA, TA, TS	20	Corpus callosum middle anterior (Bilateral)	SV
21	Pars triangularis (L, R)	CV, SA, TA, TS	21	Corpus callosum middle posterior (Bilateral)	SV
22	Pericalcarine cortex (L, R)	CV, SA, TA, TS	22	Corpus callosum posterior (Bilateral)	SV
23	Postcentral gyrus (L, R)	CV, SA, TA, TS	23	Cerebrospinal fluid (Bilateral)	SV
24	Posterior cingulate cortex (L, R)	CV, SA, TA, TS	24	Fourth ventricle (Bilateral)	SV
25	Precentral gyrus (L, R)	CV, SA, TA, TS	25	Non white matter hypointensities (Bilateral)	SV
26	Precuneus cortex (L, R)	CV, SA, TA, TS	26	Optic chiasm (Bilateral)	SV
27	Rostral anterior cingulate cortex (L, R)	CV, SA, TA, TS	27	Third ventricle (Bilateral)	SV
28	Rostral middle frontal gyrus (L, R)	CV, SA, TA, TS	28	White matter hypointensities (Bilateral)	SV
29	Superior frontal gyrus (L, R)	CV, SA, TA, TS			
30	Superior parietal cortex (L, R)	CV, SA, TA, TS			
31	Superior temporal gyrus (L, R)	CV, SA, TA, TS			
32	Supramarginal gyrus (L, R)	CV, SA, TA, TS			
33	Temporal pole (L, R)	CV, SA, TA, TS			
34	Transverse temporal cortex(L, R)	CV, SA, TA, TS			
35	Hemisphere (L, R)	SA			
36	Total intracranial volume (Bilateral)	ICV			

Table 2
Summary of subject information.

Category	CN	MCI	AD
Number	225	390	173
Gender(M/F)	116/109	252/138	88/85
Age(ave ± std)	75.87 ± 5.04	74.75 ± 7.39	75.42 ± 7.25
Educatio(ave ± std)	16.03 ± 2.85	15.67 ± 2.95	14.65 ± 3.17

F: female, M: male, ave: average, std: standard deviation.

- Trace-norm Multi-task Learning (Trace) [38]: Trace norm regularization is proposed in [38] to constrain multiple tasks share certain information.

$$\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda\|\Theta\|_* \tag{37}$$

- Sparse feature decomPosition for muLti-vIew multi-Task learning (SPLIT) [39]: SPLIT assume tasks can be reconstructed to several latent topics. It decomposes Θ into three multiplicative components A, B and H to model task correlation.

$$\begin{aligned}&\min_{\Theta} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|A\|_{1,1} + \lambda_2\|B\|_F^2 + \eta\|H\|_F^2 \\ &\quad \text{s.t.} \quad \Theta = (A \circ B)H, B \geq 0\end{aligned} \tag{38}$$

- fAst and robust method with Group sparsIty for multi-view multi-task LEarning (AGILE) [40]: AGILE decomposing feature parameters into two components, one for saving relevant features and the other detecting outlier task.

$$\min_{\Theta=W+H} \frac{1}{2}\|Y - X\Theta\|_2^2 + \lambda_1\|W\|_{2,1} + \lambda_2\|UW\|_* + \eta\|H\|_{G_1} \tag{39}$$

The results are shown in Tables 3 and 4. We can clearly observe that our proposed method significantly outperforms the other methods except MTFL, G-sMuRFS and AGILE on all of the 18 tasks. Our method significantly outperform MTFL, G-sMuRFS and AGILE on 13 tasks (except on TOTB, RECOG, CLOCK DRAW, COPYSCORE and BAC), 16 tasks (except on RECOG,RAVLT TOTAL) and 16 tasks (except on RECOG and DIGIT) respectively. Specifically, we observe the following: (1) For the four single-task learning methods, Lasso performs better than Ridge due to the sparse solution obtained by Lasso, so that more irrelevant features can be removed. (2) Multi-task learning is not generally better than single-task learning. This is not surprising as the task correlation is not improperly captured. For example, as can be seen in Tables 3 and 4, RMTL, rMTFL and Trace perform even worse than Lasso. The reason is that RMTL, rMTFL, SPLIT and Trace restricted Θ low rank to make Θ has many rows (columns) that are linearly correlated. In other words, they assume that some tasks are linear related, which is not always appropriate in practice. (3) Our proposed BGP-MTFL framework achieves the best performance, which is benefit from the incorporation of both task and feature correlations. Similar to our framework, G-SMuRFS incorporates the group lasso regularization into MTFL and achieves better results than MTFL, which also implies that introducing the appropriate regularization consistent with prior knowledge can significantly improve multi-task learning performance.

4. Discussion

In this section, We have conducted a comprehensive experiments to answer four questions:

- How does feature graph guided regularization, task graph guided regularization and self-paced scheme affect the model performance (Section 4.1)?

Table 3
Performance comparison in terms of rMSE and nMSE. The best results are bold.

Method	ADAS	MMSE	RAVLT TOTAL	RAVLT TOT6	RAVLT TOTB	RAVLT T30	RAVLT RECOG
Ridge	7.433±0.477*	2.783±0.179*	11.18±0.788*	3.859±0.380*	1.984±0.117*	4.018±0.298*	4.283±0.427*
Lasso	6.936±0.670*	2.258±0.169*	10.43±0.767*	3.422±0.303*	1.731±0.199*	3.517±0.210*	3.776±0.281*
RF	9.643 ±0.692*	3.056 ±0.164*	13.54 ±1.424*	4.678 ±0.318*	2.377 ±0.143*	4.678 ±0.462*	5.062 ±0.375*
SVR	7.835 ±0.438*	2.928 ±0.258*	12.03 ±0.956*	4.115 ±0.343*	2.309 ±0.082*	4.347 ±0.282*	4.847 ±0.435*
MTFL	6.881±0.489*	2.248±0.105*	9.715±0.776*	3.339±0.255*	**1.651 ± 0.162**	3.471±0.270*	**3.608 ± 0.181**
RMTL	7.048±0.473*	2.813±0.390*	10.93±0.751*	3.594±0.372*	1.782±0.140*	3.727±0.293*	3.929±0.420*
rMTFL	6.991±0.443*	2.375±0.235*	10.79±0.686*	3.468±0.330*	1.695±0.155*	3.602±0.253*	3.836±0.401*
G-SMuRFS	6.899±0.533*	2.258±0.102*	9.673 ± 0.794	3.324±0.256*	1.654±0.158*	3.442±0.296*	**3.608 ± 0.202**
Trace	6.885±0.551*	2.932±0.132*	10.55±0.777*	3.481±0.298*	1.729±0.136*	3.619±0.242*	3.748±0.278*
SPLIT	7.052±0.547*	2.932±0.152*	10.571±0.724*	3.513±0.236*	1.736±0.125*	3.651±0.202*	3.795±0.307*
AGILE	6.709±0.608*	2.491±0.308*	9.763±0.844*	3.446±0.376*	1.761±0.182*	3.553±0.438*	3.715 ± 0.155
BGP-MTFL(ours)	**6.652 ± 0.486**	**2.194 ± 0.092**	**9.669 ± 0.677**	**3.316 ± 0.274**	1.684 ± 0.171	**3.432 ± 0.280**	3.619 ± 0.208

Method	FLU ANIM	FLU VEG	LOGMEM IMMTOTAL	LOGMEM DELTOTAL	CLOCK DRAW	CLOCK COPYSCORE	BOSNAM RECOG
Ridge	6.312±0.603*	4.284±0.391*	4.673±0.399*	5.211±0.542*	1.155±0.104*	0.779±0.041*	4.675±0.423*
Lasso	5.554±0.434*	3.755±0.181*	4.382±0.424*	4.778±0.514*	1.022±0.093*	0.665±0.079*	4.113±0.553*
RF	7.264 ±0.530*	5.252 ±0.392*	6.153 ±0.640*	6.691 ±0.424*	1.374 ±0.126*	0.875 ±0.130*	5.477 ±0.494*
SVR	7.047 ±0.463*	4.864 ±0.316*	5.047 ±0.411*	5.626 ±0.542*	1.314 ±0.149*	0.776 ±0.076*	4.868 ± 0.583
MTFL	5.251±0.492*	3.729±0.237*	4.142±0.377*	4.560±0.509*	**0.971 ± 0.110**	**0.648 ± 0.882**	4.044±0.501*
RMTL	5.861±0.605*	3.993±0.283*	4.442±0.366*	4.897±0.507*	1.054±0.091*	0.775±0.120*	4.484±0.386*
rMTFL	5.599±0.493*	3.846±0.281*	4.299±0.307*	4.768±0.491*	1.004±0.154*	0.688±0.156*	4.211±0.511*
G-SMuRFS	5.245±0.482*	3.717±0.232*	4.162±0.374*	4.565±0.523*	0.973±0.108*	0.649±0.086*	4.044±0.526*
Trace	5.532±0.531*	3.873±0.281*	4.334±0.376*	4.747±0.422*	1.013±0.118*	0.716±0.072*	4.382±0.364*
SPLIT	5.513±0.439*	3.861±0.234*	4.375±0.340*	4.805±0.366*	1.006±0.114*	0.710±0.077*	4.399±0.362*
AGILE	5.364±0.494*	3.838±0.284*	4.308±0.399*	4.715±0.577*	1.143±0.069*	0.888±0.062*	4.104±0.513*
BGP-MTFL(ours)	**5.236 ± 0.485**	**3.668 ± 0.223**	**4.134 ± 0.364**	**4.517 ± 0.517**	0.999 ± 0.094	0.675 ± 0.071	**3.954 ± 0.480**

Method	ANART	DSPAN For	DSPAN BAC	DIGIT	nMSE
Ridge	11.21±0.731*	2.405±0.207*	2.571±0.188*	12.76±1.305*	5.354±0.325*
Lasso	10.39±1.233*	2.072±0.235*	2.192±0.186*	12.26±1.524*	4.419±0.530*
RF	13.32 ±1.488*	2.761 ±0.187*	2.999 ±0.324*	16.41 ±1.060*	7.954 ±0.469*
SVR	12.12 ±0.864*	2.848 ±0.222*	3.062 ±0.138*	13.64 ±1.268*	6.370 ±0.355*
MTFL	9.434±0.698*	2.004±0.151*	**2.117 ± 0.183**	11.58±1.275*	3.991±0.229*
RMTL	10.51±0.696*	2.174±0.150*	2.266±0.199*	12.59±1.219*	4.815±0.318*
rMTFL	10.39±0.730*	2.036±0.730*	2.167±0.208*	12.44±1.169*	4.512±0.278*
G-SMuRFS	9.425±0.694*	2.010±0.154*	2.123±0.190*	11.57±1.297*	3.984±0.216*
Trace	10.01±0.666*	2.124±0.126*	2.214±0.206*	12.00±1.306*	4.485±0.250*
SPLIT	9.908±0.682*	2.123±0.105*	2.226 ± 0.185	12.049±1.239*	4.508±0.241*
AGILE	9.483±0.756*	2.113±0.180*	2.245±0.213*	11.466 ± 1.291	4.177±0.251*
BGP-MTFL(ours)	**9.450 ± 0.679**	**2.000 ± 0.154**	2.131 ± 0.195	**11.324 ± 1.263**	**3.923 ± 0.213**

Superscript $*$ indicate that the proposed approach significantly outperforms that method. (Note that Student's t-test at a level of 0.05 is used.).

- How parameters of B_e and B_h affect model performance (Section 4.2)?
- Which biomarkers and brain regions can be identified (Section 4.3)?
- What task-feature correlations can be explored (Section 4.4)?
- How the model performs on the longitudinal multiple tasks (Section 4.5)?

4.1. Ablation study

To verify the effect of individual component in our framework and show the contribution of individual components, we evaluate the three components of our approach: MTFL ($\lambda_t = \lambda_f = \lambda_{sp} = 0$ and $B = I_n$) with only $\ell_{2,1}$ regularization, fG-MTFL ($\lambda_t = \lambda_{sp} = 0$ and $B = I_n$) with $\ell_{2,1}$ regularization and feature graph guided regularization, tG-MTFL ($\lambda_f = \lambda_{sp} = 0$ and $B = I_n$) with $\ell_{2,1}$ regularization and task graph guided regularization and BG-MTFL ($\lambda_{sp} = 0$ and $B = I_n$) with $\ell_{2,1}$ regularization and Bi-graph (feature graph and task graph) guided regularization.

Table 5 shows the performance of BGP-MTFL and its four variants in terms of rMSE and nMSE. From Table 5, we draw the following conclusions:

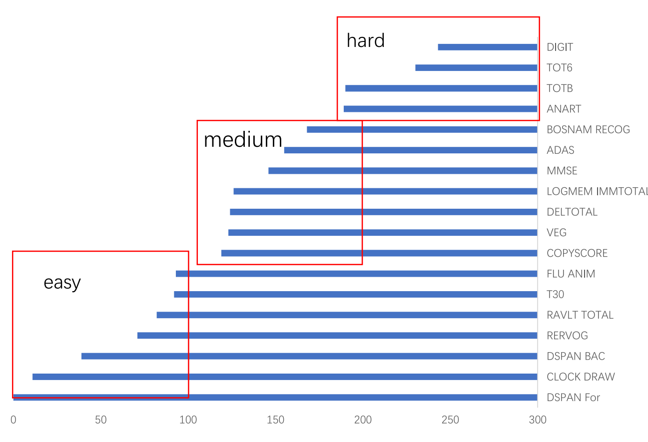


Fig. 4. The task sequence obtained by our model. The start position of each bar represents that the corresponding task starts to be learned. Tasks to be learned early are taken as easy tasks. Otherwise, tasks to be learned in the late iterations are taken as hard tasks.

Table 4
Performance comparison in terms of CC and wR. The best results are bold.

Method	ADAS	MMSE	RAVLT				
			TOTAL	TOT6	TOTB	T30	RECOG
Ridge	0.601±0.053*	0.421±0.067*	0.407±0.124*	0.362±0.133*	0.141±0.090*	0.375±0.135*	0.268±0.112*
Lasso	0.638±0.071*	0.510±0.057*	0.455±0.104*	0.466±0.111*	0.271±0.124*	0.489±0.100*	0.375±0.131*
RF	0.455 ±0.069*	0.331 ±0.045*	0.304 ±0.137*	0.253 ±0.080*	0.110 ±0.098*	0.317 ±0.098*	0.221 ±0.053*
SVR	0.571 ±0.054*	0.345 ±0.086*	0.353 ±0.131*	0.354 ±0.125*	0.063 ±0.107*	0.342 ±0.125*	0.203 ±0.103*
MTFL	0.638±0.077*	0.541±0.066*	0.512±0.107*	0.488±0.123*	**0.331 ± 0.087**	0.495±0.109*	**0.419 ± 0.124**
RMTL	0.629±0.052*	0.408±0.067*	0.421±0.131*	0.414±0.145*	0.209±0.104*	0.434±0.124*	0.330±0.114*
rMTFL	0.636±0.051*	0.506±0.056*	0.429±0.122*	0.443±0.122*	0.275±0.082*	0.455±0.117*	0.343±0.115*
G-SMuRFS	0.638±0.077*	0.542±0.065*	**0.522 ± 0.097**	0.497±0.118*	0.327±0.080*	0.511±0.104*	**0.419 ± 0.127**
Trace	0.645±0.054*	0.366±0.086*	0.443±0.128*	0.444±0.135*	0.244±0.119*	0.454±0.116*	0.373±0.122*
SPLIT	0.626±0.056*	0.337±0.081*	0.439±0.122*	0.436±0.125*	0.237±0.106*	0.448±0.114*	0.354±0.121*
AGILE	0.600±0.063*	0.527±0.047*	0.512±0.105*	0.473±0.124*	0.293±0.096*	0.487±0.114*	0.416 ± 0.111
BGP-MTFL(ours)	**0.667 ± 0.070**†	**0.547 ± 0.068**†	0.518±0.097†	**0.503 ± 0.122**†	0.311±0.072†	**0.516 ± 0.116**†	0.417±0.126†

Method	FLU		LOGMEM		CLOCK		BOSNAM
	ANIM	VEG	IMMTOTAL	DELTOTAL	DRAW	COPYSCORE	RECOG
Ridge	0.201±0.130*	0.389±0.128*	0.418±0.111*	0.433±0.121*	0.227±0.107*	0.133±0.095*	0.363±0.145*
Lasso	0.315±0.097*	0.495±0.076*	0.473±0.115*	0.507±0.123*	0.334±0.055*	0.068±0.115*	0.444±0.101*
RF	0.181 ±0.118*	0.272 ±0.073*	0.215 ±0.100*	0.276 ±0.109*	0.174 ±0.148*	0.114 ±0.119*	0.306 ±0.087*
SVR I	0.134 ±0.118*	0.314 ±0.122*	0.385 ±0.096*	0.388 ±0.130*	0.174 ±0.117*	0.135 ±0.106*	0.333 ±0.135*
MTFL	0.395±0.084*	0.490±0.091*	0.511±0.084*	0.531±0.094*	0.389 ± 0.085	0.223 ± 0.097	0.465±0.103*
RMTL	0.265±0.129*	0.441±0.107*	0.455±0.100*	0.475±0.112*	0.340±0.104*	0.166±0.088*	0.386±0.129*
rMTFL	0.299±0.126*	0.473±0.104*	0.474±0.103*	0.488±0.120*	0.373±0.090*	0.230±0.097*	0.424±0.138*
G-SMuRFS	**0.396 ± 0.073**	0.498±0.086*	0.508±0.087*	0.534±0.092*	0.379±0.081*	0.224±0.113*	0.458±0.082*
Trace	0.318±0.112*	0.456±0.089*	0.467±0.094*	0.496±0.092*	0.333±0.111*	0.165±0.094*	0.378±0.124*
SPLIT	0.322±0.092*	0.461±0.067*	0.457±0.086*	0.484±0.080*	0.338±0.102*	0.165±0.096*	0.361±0.093*
AGILE	0.382±0.112*	0.492±0.101*	0.501±0.096*	0.527±0.109*	0.291±0.098*	−0.038±0.076*	0.456±0.123*
BGP-MTFL(ours)	0.389±0.094†	**0.514 ± 0.082**†	**0.514 ± 0.091**†	**0.543 ± 0.099**†	**0.392 ± 0.077**†	**0.241 ± 0.098**†	**0.481 ± 0.088**†

Method	ANART	DSPAN		DIGIT	wR
		For	BAC		
Ridge	0.049±0.083*	0.011±0.060*	0.031±0.118*	0.390±0.045*	0.290±0.055*
Lasso	0.100±0.087*	0.026±0.073*	0.129±0.094*	0.402±0.077*	0.361±0.051*
RF	0.057 ±0.168*	0.025 ±0.099*	0.01 ±0.114*	0.175 ±0.118*	0.210 ±0.029*
SVR	0.038 ±0.106*	0.03 ±0.056*	0.04 ±0.103*	0.325 ±0.080*	0.244 ±0.055*
MTFL	0.160 ± 0.121	0.027±0.075*	**0.210 ± 0.129**	0.429±0.114*	0.403±0.063*
RMTL	0.075±0.073*	0.021±0.066*	0.121±0.106*	0.401±0.066*	0.333±0.064*
rMTFL	0.085±0.071*	0.092±0.092*	0.157±0.122*	0.402±0.043*	0.366±0.056*
G-SMuRFS	0.161±0.116*	0.001±0.062*	0.189±0.140*	0.432±0.107*	0.402±0.061*
Trace	0.098±0.071*	−0.03±0.110*	0.138±0.086*	0.418±0.049*	0.345±0.065*
SPLIT	0.115±0.083*	0.072±0.068*	0.121±0.083*	0.414±0.049*	0.338±0.061*
AGILE	0.149±0.077*	−0.047±0.068*	0.144±0.108*	**0.472 ± 0.050**	0.372±0.058*
BGP-MTFL(ours)	**0.167 ± 0.108**†	**0.106 ± 0.121**†	0.207±0.122†	0.463±0.067†	**0.416 ± 0.058**†

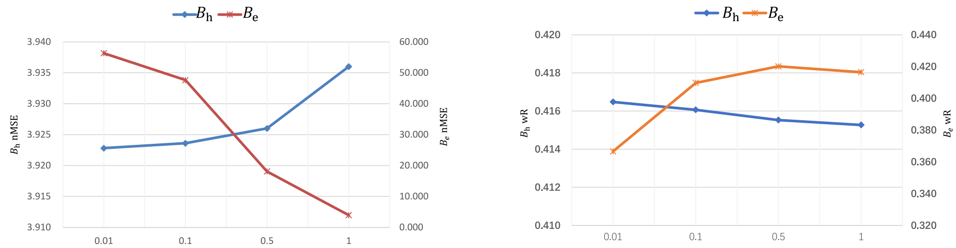
Superscript $*$ indicates that the proposed approach significantly outperforms the competitive method (Student's t-test at a level of 0.05 is used.). Superscript † indicates that there is significantly relationship between the predict results of our proposed approach and the true cognitive score values (a *p*-value less than 0.05 is statistically significant).

Table 5
Performance comparison of ablation studies in terms of rMSE and nMSE.

Method	ADAS	MMSE	RAVLT				
			TOTAL	TOT6	TOTB	T30	RECOG
MTFL	6.881 ± 0.489	2.248 ± 0.105	9.715 ± 0.776	3.339 ± 0.255	**1.651 ± 0.162**	3.471 ± 0.270	**3.608 ± 0.181**
fG-MTFL	6.681 ± 0.438	2.168 ± 0.087	9.857 ± 0.598	3.320 ± 0.266	1.664 ± 0.153	3.431 ± 0.216	3.631 ± 0.243
tG-MTFL	6.665 ± 0.475	2.168 ± 0.092	9.810 ± 0.640	**3.311 ± 0.276**	1.664 ± 0.165	**3.424 ± 0.238**	3.619 ± 0.241
BG-MTFL	6.667 ± 0.475	2.195 ± 0.093	9.703 ± 0.643	3.314 ± 0.283	1.681 ± 0.173	3.434 ± 0.281	3.617 ± 0.209
BGP-MTFL(ours)	**6.652 ± 0.486**	**2.194 ± 0.092**	**9.669 ± 0.677**	3.316 ± 0.274	1.684 ± 0.171	3.432 ± 0.280	3.619 ± 0.208

Method	FLU		LOGMEM		CLOCK		BOSNAM
	ANIM	VEG	IMMTOTAL	DELTOTAL	DRAW	COPYSCORE	RECOG
MTFL	5.251 ± 0.492	3.729 ± 0.237	4.142 ± 0.377	4.560 ± 0.509	0.971 ± 0.110	0.648 ± 0.088	4.044 ± 0.501
fG-MTFL	5.288 ± 0.456	3.672 ± 0.196	4.171 ± 0.307	4.571 ± 0.437	**0.975 ± 0.102**	**0.649 ± 0.081**	3.937 ± 0.427
tG-MTFL	5.263 ± 0.484	**3.663 ± 0.210**	4.158 ± 0.334	4.554 ± 0.468	0.977 ± 0.105	0.652 ± 0.082	**3.943 ± 0.461**
BG-MTFL(ours)	5.244 ± 0.483	3.673 ± 0.221	4.146 ± 0.359	4.526 ± 0.508	0.997 ± 0.094	0.674 ± 0.069	3.954 ± 0.480
BGP-MTFL(ours)	**5.236 ± 0.485**	3.668 ± 0.223	**4.134 ± 0.364**	**4.517 ± 0.517**	0.999 ± 0.094	0.675 ± 0.071	3.954 ± 0.480

Method	ANART	DSPAN		DIGIT	nMSE
		For	BAC		
MTFL	**9.434 ± 0.698**	2.004 ± 0.151	**2.117 ± 0.183**	11.58 ± 1.275	3.991 ± 0.229
fG-MTFL	9.599 ± 0.644	1.990 ± 0.138	2.131 ± 0.180	11.442 ± 1.245	3.977 ± 0.216
tG-MTFL	9.538 ± 0.682	**1.996 ± 0.148**	2.124 ± 0.189	11.405 ± 1.309	3.952 ± 0.217
BG-MTFL(ours)	9.469 ± 0.677	2.003 ± 0.154	2.130 ± 0.195	11.360 ± 1.281	3.936 ± 0.201
BGP-MTFL(ours)	9.450 ± 0.679	**2.000 ± 0.154**	2.131 ± 0.195	**11.324 ± 1.263**	**3.923 ± 0.213**



(a) The performance of nMSE. (b) The performance of wR.

Fig. 5. Results of parameter analysis in the self-paced learning process. Curve in orange means that B_e varies within the range of [0.01, 0.1, 0.5, 1] when B_h is fixed as 0.01. Similarly, curve in blue means that B_h varies within the range of [0.01, 0.1, 0.5, 1] when B_e is fixed as 1.

(1) The method with any single graph regularization, fG-MTFL or tG-MTFL performs better than MTFL without considering the feature/task correlation. Besides, the multi-task learning methods with Bi-graph guided sparse regularization (e.g., BG-MTFL, BGP-MTFL) are obviously better than fG-MTFL and tG-MTFL. The results demonstrate that these prior knowledge benefits the prediction performance of multi-task feature learning and the two regularizations are complementary.

(2) When comparing the two methods with only single graph regularization, we find that tG-MTFL outperforms fG-MTFL. The reasons are that only 5% of features correlation is bigger than 0.5 for feature correlation, whereas more than 18% of task correlation is bigger than 0.5 in the task correlation. It is obvious that the task correlation is more significant than the feature correlation. Another reason is that the incorporation of feature correlation may inevitably introduce redundant features and reduces the sparsity level.

(3) BGP-MTFL is generally better than BG-MTFL. The difference between BG-MTFL and BGP-MTFL is that BGP-MTFL begins to learn from the easier set of tasks and gradually incorporates more difficult ones to build the shared knowledge. Through the comparison, we observe that the tasks with the most significant improvements are: RAVLT TOTB, DIGIT, ANART and DSPAN For. Among them, DSPAN For is an easier task and has weak correlations with the other tasks, thus it is learned early (Fig. 4) without being negative influenced by the complicated tasks. TOTB, DIGIT and ANART belong to hard-learning tasks and are subsequently learned, so that the shared knowledge from easier tasks can be leveraged by the harder tasks. This demonstrates that incorporating the easy-to-hard strategy into the multi-task learning process can improve the prediction performance. Fig. 4 shows the task sequence obtained by our model. We can divide the 18 tasks into three categories (easy, medium and hard) based on their learning difficulty.

4.2. Parameter analysis

In the experiments above, the weight of easy task B_e is set as 1 and the weight of hard task B_h is set as 0.01. In this section, we evaluate how changes to the parameters B_e and B_h affect model performance and show the results in Fig. 5.

In Fig. 5, we can observe that: (1) When B_e is fixed as 1, the model performs better with the parameter of B_h decreasing. On the contrary, when B_h is fixed as 0.01, the model performs better with the parameter B_e increasing. These observations suggest that B_e and B_h significantly influence the model performance, and the order of task learning affects the performance of the model. (2) The model achieves the best performance when B_e=1 and B_h=0.01, whereas it obtains the worst performance when B_e=0.01 and B_h=1. These observations are consistence with our assumption that the easy task should be selected in training (B_e=1). On the contrary, the hard task is unselected in training (B_h=0.01). (3) When B_e=1 and B_h=1, the model degenerates into BG-MTFL. In this case, all tasks are learned at the same time, causing that BG-MTFL is inferior to BGP-MTFL. This result demonstrates that self-paced learning process, which learns the multiple tasks in sequence (from easy to hard) instead of learning multiple tasks simultaneously, helps in improving the generalization performance of the multi-task learning.

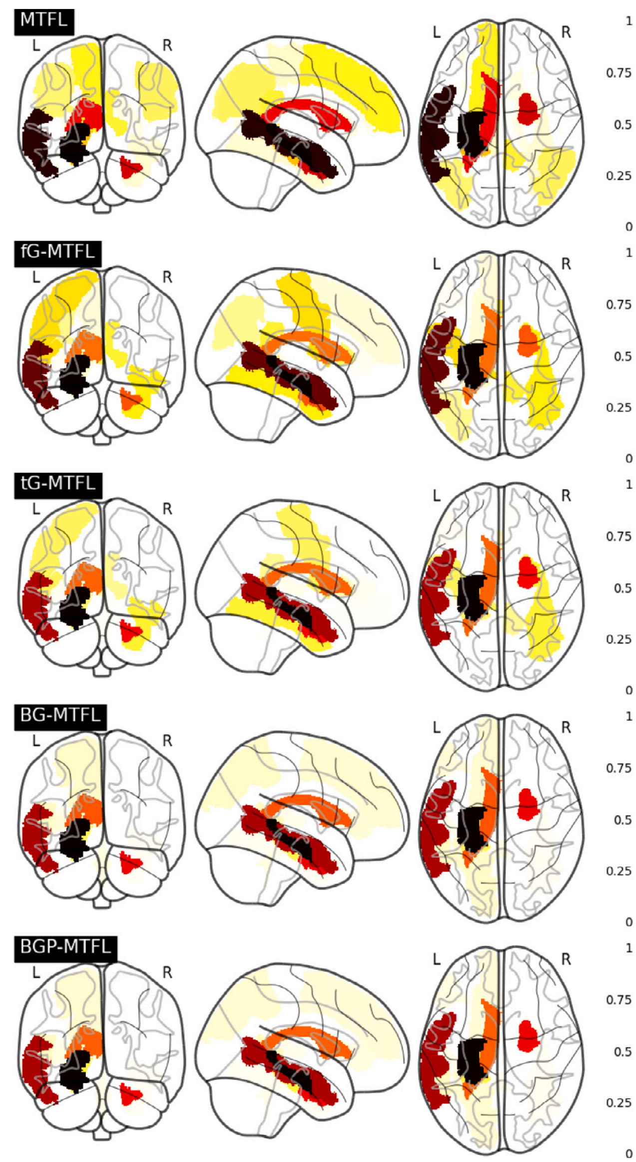


Fig. 6. Location of the selected important ROIs. These are the most relevant areas for predicting all cognitive scores jointly. The color encodes the weight at each ROI. The brain regions are segmented based on Desikan-Killiany atlas.

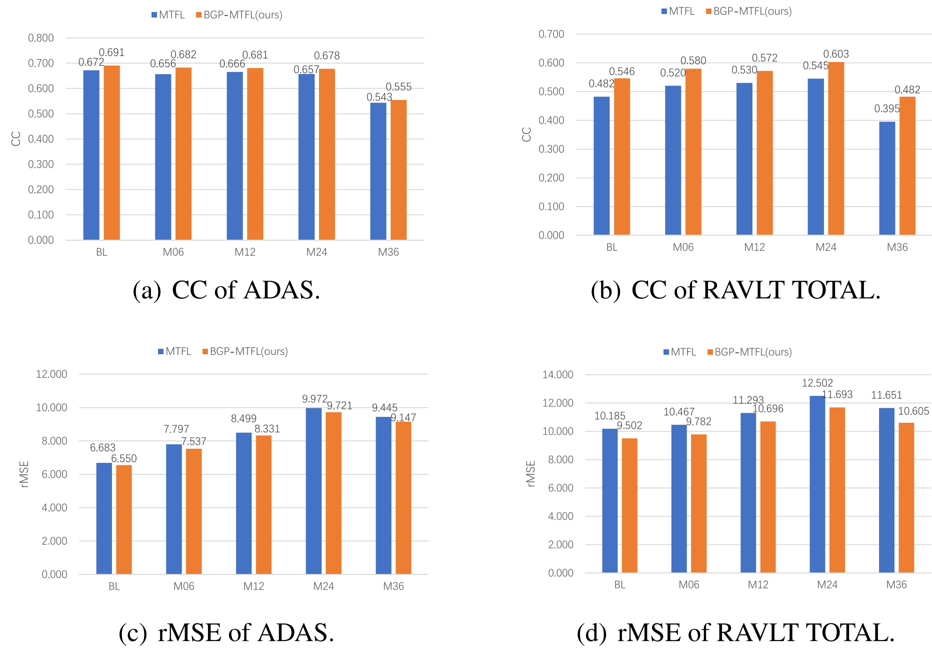


Fig. 7. Comparison of our proposed methods and MTFL on longitudinal prediction in terms of rMSE and CC. It is importance to notice that higher CC and lower rMSE mean better result.

4.3. Biomarker analysis

To capture the biomarkers and ROIs that are related to AD identified by the proposed BGP-MTFL model and the traditional MTFL, in this section, we take the ℓ_2-norm of Θ as the weight of each feature. Larger weights mean more important features. The identified top 10 important features are shown in Table 6. On the other hand, to identify the import ROIs that are related to AD, we consider the average of feature weights in the same ROI as the weight of this ROI. The top 10 important ROIs are shown in Table 7.

In Tables 6 and 7, features/ROIs colored in red indicate that they are discovered by BGP-MTFL while missed by MTFL. Features/ROIs colored in blue are selected by MTFL while ignored by BGP-MTFL. From Table 6, we can derive several interesting observations: (1) Some common features are identified by MTFL and BG-MTFL, such as SV of L.Hippocampus, TA of L.MidTemporal, CV of R.Entorhinal, etc, which are presented to be pathological associated with AD in the previous works [41–43]. (2) BGP-MTFL identified three specific features: SA of L.SuperiorParietal, SV of CorpusCallosumAnterior and CV of R.Fusiform. These three features are also presented to be effective in the process of tracking AD pathology by several studies [44,45]. The discovery of these three important features benefit from the introduce of feature graph guided regularization. Due to SA of L.SuperiorParietal, SV of CorpusCallosumAnterior and CV of R.Fusiform are highly related with some important features (e.g., high correlation between SuperiorParietal and SuperiorFrontal) and feature graph guided regularization enforce them share similar weights. (3) Compared with MTFL, the features selected by BGP-MTFL are assigned larger weights, which means that the BGP-MTFL has better discriminative feature extraction ability, which is benefit from the incorporation of the proposed bi-graph guided regularization and self-paced learning scheme.

Table 7 shows the top 10 important ROIs identified by the proposed BGP-MTFL and MTFL. Similar to the above observations, BGP-MTFL is able to identified subtle change in the progression of AD, such as CorpusCallosumAnterior, LateralVentricle and CorpusCallosumMidPos. Although MTFL selects three ROIs that are ignored by BGP-MTFL, these ROIs have not been identified to be effective for tracking the progress of AD. The identified top 10 important ROIs are visualized in Fig. 6.

Table 6
The identified top fifteen important features when predicting eighteen cognitive scores. Features/ROIs colored in magenta (blue) indicate that they are discovered by BGP-MTFL (MTFL) specifically.

MTFL		BGP-MTFL	
Features name	Weight	Features name	Weight
TA of L.MidTemporal	3.568	SV of L.Hippocampus	3.640
SV of L.Hippocampus	3.398	TA of L.MiddleTemporal	2.931
CV of R.Entorhinal	2.444	CV of R.Entorhinal	2.577
SV of L.InferiorLateralVen	2.251	SV of L.InferiorLateralVen	2.251
TA of R.Entorhinal	1.578	TA of R.Entorhinal	1.620
TA of L.Parahippocampal	1.302	TA of L.Parahippocampal	1.370
TS of L.SuperiorFrontal	0.752	SA of L.SuperiorParietal	1.033
TA of R.IsthmusCingulate	0.741	TS of L.SuperiorFrontal	1.019
TA of R.InferiorParietal	0.650	SV of CorpusCallosumAnterior	0.968
TA of L.InferiorParietal	0.598	CV of R.Fusiform	0.948

Table 7
The identified important ROIs. ROIs colored in magenta (blue) indicate that they are discovered by BGP-MTFL (MTFL) specifically.

MTFL		BGP-MTFL	
ROIs name	Weight	Features name	Weight
L.Hippocampus	3.398	L.Hippocampus	3.640
L.InferiorLateralVen	2.251	L.InferiorLateralVen	2.251
R.EntorhinalCortex	1.005	R.EntorhinalCortex	1.219
L.MiddleTemporalGyrus	0.892	L.MiddleTemporalGyrus	0.998
L.ParahippocampalGyrus	0.325	CorpusCallosumAnterior	0.968
L.SuperiorFrontalGyrus	0.188	R.LateralVentricle	0.735
R.IsthmusCingulate	0.185	L.Cerebellum	0.722
R.InferiorParietalCortex	0.162	CorpusCallosumMidPos	0.710
L.InferiorParietalCortex	0.150	R.IsthmusCingulate	0.519
L.Cerebellum	0.130	L.ParahippocampalGyrus	0.515

4.4. Task-feature relationship analysis

Besides the identified common biomarkers for all the tasks, we also investigate the identified features for each task to explore the task-feature relationship. The task-feature relationship is shown in Table 8. In Table 8, each column corresponds to one task and the top 15 important features that are related to this task. From Table 8, we can draw the following conclusions: (1) some features are shared among almost

Table 8
The identified features for the eighteen tasks. Note that the features marked with dark blue denotes the features shared by all the eighteen tasks, and the features marked with other colors indicate the ones locally shared by several tasks rather than all the tasks.

ADAS	RAVLT				
	TOTAL	TOT6	T30	RECOG	TOTB
CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal
SV of L.Hippocampus	SV of L.Hippocampus	SV of L.Hippocampus	SV of L.Hippocampus	SV of L.Hippocampus	SV of L.Hippocampus
SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen
CV of L.TemporalPole	TA of L.Parahippocampal	TA of L.Parahippocampal	TA of L.Parahippocampal	TA of L.Parahippocampal	TA of L.Parahippocampal
TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal
TA of R.Entorhinal	TA of R.Entorhinal	TA of R.Entorhinal	TA of R.Entorhinal	TA of R.Entorhinal	TA of R.Entorhinal
TA of L.Entorhinal	TA of L.InferiorTemporal	TA of L.Entorhinal	CV of L.Entorhinal	TA of L.Entorhinal	TA of L.InferiorParietal
CV of R.TransverseTemporal	CV of L.LateralOrbitofrontal	CV of L.ParsOpercularis	CV of L.ParsOpercularis	CV of L.ParsOpercularis	TA of L.InferiorTemporal
TS of L.ParsTriangularis	TA of L.Precuneus	TA of L.Precuneus	TA of L.Precuneus	TA of L.Precuneus	TA of L.Precuneus
TA of L.InferiorParietal	CV of L.Precentral	CV of L.Precentral	CV of L.Precentral	CV of L.InferiorParietal	SV of R.Hippocampus
TS of L.Paracentral	SA of R.CaudalMiddleFrontal	CV of L.RostralAnteriorCingulate	SA of L.Fusiform	SV of L.ChoroidPlexus	TA of R.IsthmusCingulate
CV of L.InferiorParietal	SA 0f L.Bankssts	SV of L.CerebellumCortex	SV of L.CerebellumCorte	SV of L.CerebellumCorte	CV of R.Fusiform
CV of L.Precentral	SV of CorpusCallosumMidPos	TS of R.Fusiform	SV of CorpusCallosumMidPos	SV of R.Amygdala	TA of R.SuperiorFrontal
TA of R.InferiorParietal	CV of R.LateralOccipital	TA of R.InferiorParietal	CV of L.LateralOrbitofrontal	SV of R.Thalamus	TA of R.InferiorTemporal
TS of R.MedialOrbitofrontal	TS of R.MedialOrbitofrontal	TS of R.MedialOrbitofrontal	SV of R.Hippocampus	TA of R.Precentral	TA of R.InferiorParietal
FLU		CLOCK		DSPAN	
ANIM	VEG	DRAW	COPYSCORE	FOR	BAC
CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal
SV of L.Hippocampus	SV of L.Hippocampus	SV of L.Hippocampus	SV of L.Hippocampus	SV of L.Hippocampus	TA of L.Entorhinal
SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen
TA of L.Parahippocampal	TA of L.Parahippocampal	TA of L.Parahippocampal	TA of L.Parahippocampal	TA of L.Parahippocampal	TA of L.Parahippocampal
TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal
TA of R.Entorhinal	TA of R.Entorhinal	TA of R.Entorhinal	TA of R.Entorhinal	TA of R.tEntorhinal	TA of R.tEntorhinal
CV of L.MiddleTemporal	TS of L.Paracentral	TA of R.IsthmusCingulate	TA of R.IsthmusCingulate	TA of L.Entorhinal	TA of L.Entorhinal
TS of L.LateralOrbitofrontal	SA of R.RostralAnteriorCingulate	TA of L.Supramarginal	TA of L.Supramarginal	SV of R.Amygdala	TA of L.Fusiform
TS of L.Fusiform	TS of L.Fusiform	TA of R.InferiorParietal	TA of R.InferiorParietal	SA of R.Bankssts	TA of L.InferiorParietal
SV of R.Thalamus	SV of R.Thalamus	TS of L.SuperiorFrontal	TS of L.SuperiorFrontal	TS of L.TransverseTemporal	TA of LeftInferiorTemporal
TA of R.IsthmusCingulate	CV of R.LateralOccipital	TA of L.InferiorParietal	TA of L.InferiorParietal	TA of L.Lingual	TA of L.Bankssts
CV of R.Fusiform	CV of R.Fusiform	CV of R.Fusiform	CV of R.Fusiform	CV of R.Fusiform	CV of R.Fusiform
CV of L.InferiorParietal	TA of L.InferiorParietal	TA of L.Precuneus	SV of R.LateralVentricle	SA of R.FrontalPole	TA of R.IsthmusCingulate
TS of R.LateralOccipital	TS of L.SuperiorFrontal	TA of L.InferiorTemporal	CV of L.InferiorParietal	CV of R.FrontalPole	SV of R.LateralVentricle
CV of R.TransverseTemporal	SV of L.ChoroidPlexus	TA of R.SuperiorTemporal	TA of R.SuperiorFrontal	CV of R.IsthmusCingulate	TA of R.MedialOrbitofrontal
ANART	LOGMEM		DIGIT	MMSE	BOSNAM
	IMMTOTAL	DELTOTAL			RECOG
CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Entorhinal	CV of R.Fusiform	CV of R.Entorhinal	CV of R.Entorhinal
SV of L.ChoroidPlexus	TA of L.Entorhinal	TA of L.Entorhinal	TA of L.InferiorParietal	SV of L.Hippocampus	SV of L.ChoroidPlexus
CV of LeftCuneus	SV of L.InferiorLateralVen	CV of LeftCuneus	SV of L.InferiorLateralVen	SV of L.InferiorLateralVen	TA of L.Entorhinal
SV of CorpusCallosumMidPos	TA of L.Parahippocampal	TA of L.Parahippocampal	TS of LeftSuperiorFrontal	TA of L.Parahippocampal	TA of L.Parahippocampal
TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal	TA of L.MiddleTemporal
SA of R.CaudalMiddleFrontal	TA of RightEntorhinal	TA of RightEntorhinal	TA of RightEntorhinal	TA of R.Entorhinal	TA of R.Entorhinal
ST of LeftCuneus	CV of R.TransverseTemporal	CV of R.TransverseTemporal	TS of LeftIsthmusCingulate	TA of L.Entorhinal	TA of L.Fusiform
SA of LeftFrontalPole	CV of LeftParsOpercularis	CV of LeftParsOpercularis	SA of LeftSuperiorParietal	CV of R.TransverseTemporal	CV of L.InferiorTemporal
SA of LeftTransverseTemporal	CV of L.Precentral	CV of L.Precentral	TA of LeftSupramarginal	CV of L.InferiorParietal	CV of L.MiddleTemporal
SA of LeftTransverseTemporal	CV of LeftPrecentral	CV of LeftPrecentral	TA of LeftSupramarginal	TA of L.Fusiform	TA of L.InferiorTemporal
TS of RightEntorhinal	CV of LeftMiddleTemporal	SV of CorpusCallosumAnterior	SA of LeftTransverseTemporal	TA of L.InferiorTemporal	TA of R.CaudalAnteriorCingulate
TA of LeftBankssts	TA of LeftInferiorParietal	TA of LeftEntorhinal	ST of WMHypoIntensities	CV of R.Fusiform	CV of R.Fusiform
TA of RightInferiorParietal	TA of RightCaudalAnteriorCingulate	ST of CorpusCallosumMidPos	TA of RightIsthmusCingulate	TS of L.Paracentral	TA of .SuperiorTemporal
TS of RightIsthmusCingulate	SA of LeftRostralMiddleFrontal	TA of LeftPrecuneus	SV of RightLateralVentricle	CV of L.Entorhinal	SA of L.Bankssts
TS of RightMedialOrbitofrontal	TA of LeftInferiorTemporal	TA of LeftBankssts	SV of FourthVentricle	SV of R.Hippocampus	TS of R.MedialOrbitofrontal

all the tasks such as CV of R.Entorhinal and SV of L.Hippocampus. These findings are in accordance with the previous researches and existing knowledge in the pathological pathway of AD [46]. (2) The tasks of RAVLT TOTAL,TOT6,TOTB, T30 and RECOG involve more common features compared with the other tasks. The reason is that these tasks belong to the same category and have higher correlations with each other. Similarly, CLOSK DRAW and CLOCK COPYSCORE, as well as LOGMEM IMMTOTAL and LOGMEM DELTOTAL also have more shared features, respectively. The results confirm the effectiveness of our method again. (3) We observe that the tasks of ADAS, MMSE, ANART and DIGIT involve fewer common features with the other tasks. The evaluation of the MMSE, ANART and DIGIT tests aim to measure reading ability and short-term memory. They have lower correlations with the other tasks. In addition, a higher ADAS score indicates a higher probability of suffering from AD. It is opposite to the other measure tests. Hence, the ADAS prediction task is negatively correlated with the other tasks, thereby has more specific features.

4.5. Validation on the longitudinal multiple tasks

To estimate the effectiveness of the proposed BGP-MTFL model in the longitudinal multiple tasks, we evaluate our proposed BGP-MTFL model on the longitudinal multiple tasks and compare it with the multi-task learning method (MTFL) on two of the most common cognitive scores (ADAS and RAVLT TOTAL). The experiment setting is the same as Section 3.2. Experimental results of rMSE and CC are shown in Fig. 7. From Fig. 7, we can draw the following conclusions: (1) The result of our proposed BGP-MTFL is consistently better than the traditional MTFL method even in the longitudinal multiple tasks. This observation further demonstrates that the incorporation of feature correlation, task correlation, and self-paced process benefits the prediction performance of multi-task feature learning. (2) CC and rMSE at baseline (BL) in the longitudinal multiple tasks experiment perform better than in the cross-sectional multiple tasks experiment (experiment in Section 3.3). This implies that longitudinal multiple tasks (same tasks at multiple time points) have stronger correlations than cross-sectional tasks (multiple different tasks at the same time point) and our proposed method can achieve a higher CC value of 0.691 at baseline time.

5. Conclusion

In this work, we propose a dual-graph guided self-paced multi-task feature learning framework BGP-MTFL for accurately modeling the relationship between MRI and cognitive scores. The incorporation of feature graph and task graph enables the interdependencies among the tasks or the features to be considered. On the other hand, the introduction of self-paced learning enable the hard tasks to benefit from the easy tasks. Consequently, BGP-MTFL performs better regression performance and can identify more stable biomarkers. Furthermore, the proposed BGP-MTFL approach is capable of identifying both the features shared among all the tasks and features that are specific to each task.

Declaration of competing interest

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

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 62076059) and the Science Project of Liaoning Province, China (2021-MS-105).

References

[1] Alzheimer's disease, Lancet 388 (10043) (2016) 505–517, http://dx.doi.org/10.1016/S0140-6736(15)01124-1.
[2] World Health Organization and others, Risk reduction of cognitive decline and dementia: WHO guidelines, in: Risk Reduction of Cognitive Decline and Dementia: WHO Guidelines, 2019.
[3] Y. Yang, X. Li, P. Wang, Y. Xia, Q. Ye, Multi-Source Transfer Learning via Ensemble Approach for Initial Diagnosis of Alzheimer's Disease, IEEE J. Transl. Eng. Health Med. 8 (2020) 1–10.
[4] J. Weller, A. Budson, Current understanding of Alzheimer's disease diagnosis and treatment, F1000Research 7 (2018).
[5] M. Bilal, M. Barani, F. Sabir, A. Rahdar, G.Z. Kyzas, Nanomaterials for the treatment and diagnosis of Alzheimer's disease: An overview, NanoImpact (2020) 100251.
[6] M. Tanveer, B. Richhariya, R. Khan, A. Rashid, P. Khanna, M. Prasad, C. Lin, Machine learning techniques for the diagnosis of Alzheimer's disease: A review, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 16 (1s) (2020) 1–35.
[7] A. Li, L. Yue, S. Xiao, M. Liu, Cognitive Function Assessment and Prediction for Subjective Cognitive Decline and Mild Cognitive Impairment, Brain Imag. Behav. (2021) 1–14.
[8] P. Cao, X. Shan, D. Zhao, M. Huang, O. Zaiane, Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease, Pattern Recognit. 72 (2017) 219–235.
[9] G.B. Frisoni, N.C. Fox, C.R. Jack, P. Scheltens, P.M. Thompson, The clinical use of structural MRI in Alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 67–77.
[10] C.R. Jack Jr., D.S. Knopman, W.J. Jagust, R.C. Petersen, M.W. Weiner, P.S. Aisen, L.M. Shaw, P. Vemuri, H.J. Wiste, S.D. Weigand, et al., Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers, Lancet Neurol. 12 (2) (2013) 207–216.
[11] L. Harper, F. Barkhof, P. Scheltens, J.M. Schott, N.C. Fox, An algorithmic approach to structural imaging in dementia, J. Neurol. Neurosurg. Psychiatry 85 (6) (2014) 692–698.
[12] N. Mattsson, P.S. Insel, M. Donohue, J. Jögi, R. Ossenkoppele, T. Olsson, M. Schöll, R. Smith, O. Hansson, Predicting diagnosis and cognition with 18F-AV-1451 tau PET and structural MRI in Alzheimer's disease, Alzheimer's Dementia 15 (4) (2019) 570–580.
[13] R. Ossenkoppele, R. Smith, T. Ohlsson, O. Strandberg, N. Mattsson, P.S. Insel, S. Palmqvist, O. Hansson, Associations between tau, Aβ, and cortical thickness with cognition in Alzheimer disease, Neurology 92 (6) (2019) e601–e612.
[14] N.I. Bohnen, D.S. Djang, K. Herholz, Y. Anzai, S. Minoshima, Effectiveness and safety of 18F-FDG PET in the evaluation of dementia: a review of the recent literature, J. Nucl. Med. 53 (1) (2012) 59–71.
[15] J.L. Whitwell, D.W. Dickson, M.E. Murray, S.D. Weigand, N. Tosakulwong, M.L. Senjem, D.S. Knopman, B.F. Boeve, J.E. Parisi, R.C. Petersen, et al., Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study, Lancet Neurol. 11 (10) (2012) 868–877.
[16] K. Popuri, D. Ma, L. Wang, M.F. Beg, Using machine learning to quantify structural MRI neurodegeneration patterns of Alzheimer's disease into dementia score: Independent validation on 8,834 images from ADNI, AIBL, OASIS, and MIRIAD databases, Hum. Brain Mapping 41 (14) (2020) 4127–4147.
[17] K. Blennow, H. Zetterberg, Biomarkers for Alzheimer's disease: current status and prospects for the future, J. Internal Med. 284 (6) (2018) 643–663.
[18] J. Peng, X. Zhu, Y. Wang, L. An, D. Shen, Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis, Pattern Recognit. 88 (2019) 370–382.
[19] P. Cao, X. Liu, H. Liu, J. Yang, D. Zhao, M. Huang, O. Zaiane, Generalized fused group lasso regularized multi-task feature learning for predicting cognitive outcomes in Alzheimers disease, Comput. Methods Programs Biomed. 162 (2018) 19–45.
[20] X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, D. Shen, A.D.N. Initiative, et al., A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, Med. Image Anal. 38 (2017) 205–214.
[21] X. Liu, P. Cao, J. Wang, J. Kong, D. Zhao, Fused group lasso regularized multi-task feature learning and its application to the cognitive performance prediction of Alzheimer's disease, Neuroinformatics 17 (2) (2019) 271–294.
[22] C.M. Stonnington, C. Chu, S. Klöppel, C.R. Jack Jr., J. Ashburner, R.S. Frackowiak, A.D.N. Initiative, et al., Predicting clinical scores from magnetic resonance scans in Alzheimer's disease, Neuroimage 51 (4) (2010) 1405–1413.
[23] M.R. Sabuncu, E. Konukoglu, A.D.N. Initiative, et al., Clinical prediction from structural brain MRI scans: a large-scale empirical study, Neuroinformatics 13 (1) (2015) 31–46.
[24] R. Caruana, Multitask learning, Mach. Learn. 28 (1) (1997) 41–75.
[25] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient l2, 1-norm minimization, 2012, arXiv preprint arXiv:1205.2631.
[26] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, Adv. Neural Inf. Process. Syst. 19 (2006).
[27] P. Cao, S. Tang, M. Huang, J. Yang, D. Zhao, A. Trabelsi, O. Zaiane, Feature-aware Multi-task feature learning for Predicting Cognitive Outcomes in Alzheimer's disease, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, 2019, pp. 1–5, http://dx.doi.org/10.1109/BIBM47256.2019.8983258.
[28] X. Liu, P. Cao, A.R. Gonçalves, D. Zhao, A. Banerjee, Modeling Alzheimer's disease progression with fused laplacian sparse group lasso, ACM Trans. Knowl. Discov. Data (TKDD) 12 (6) (2018) 1–35.
[29] M. Wang, D. Zhang, D. Shen, M. Liu, Multi-task exclusive relationship learning for Alzheimer's disease progression prediction with longitudinal data, Med. Image Anal. 53 (2019) 111–122.
[30] H. Li, M. Gong, Self-paced Convolutional Neural Networks, in: IJCAI, 2017, pp. 2110–2116.
[31] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, A. Hauptmann, Self-paced learning with diversity, Adv. Neural Inf. Process. Syst. 27 (2014).
[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.
[33] P.L. Combettes, J.-C. Pesquet, Proximal splitting methods in signal processing, in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 185–212.
[34] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient l2, 1-norm minimization, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 339–348.
[35] J. Chen, J. Zhou, J. Ye, Integrating low-rank and group-sparse structures for robust multi-task learning, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 42–50.
[36] 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.
[37] J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, S. Kim, S.L. Risacher, A.J. Saykin, L. Shen, et al., Cortical surface biomarkers for predicting cognitive outcomes using group l2, 1 norm, Neurobiol. Aging 36 (2015) S185–S193.
[38] S. Ji, J. Ye, An accelerated gradient method for trace norm minimization, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 457–464.
[39] L. Sun, C.H. Nguyen, H. Mamitsuka, Multiplicative Sparse Feature Decomposition for Efficient Multi-View Multi-Task Learning., in: IJCAI, 2019, pp. 3506–3512.
[40] L. Sun, C.H. Nguyen, H. Mamitsuka, Fast and robust multi-view multi-task learning via group sparsity, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 3499–3505.
[41] I.S. Piras, J. Krate, E. Delvaux, J. Nolz, D.F. Mastroeni, A.M. Persico, W.M. Jepsen, T.G. Beach, M.J. Huentelman, P.D. Coleman, Transcriptome changes in the Alzheimer's disease middle temporal gyrus: importance of RNA metabolism and mitochondria-associated membrane genes, J. Alzheimer's Dis. 70 (3) (2019) 691–713.
[42] C. Pennanen, M. Kivipelto, S. Tuomainen, P. Hartikainen, T. Hänninen, M.P. Laakso, M. Hallikainen, M. Vanhanen, A. Nissinen, E.-L. Helkala, et al., Hippocampus and entorhinal cortex in mild cognitive impairment and early AD, Neurobiol. Aging 25 (3) (2004) 303–310.
[43] M. Laakso, M. Lehtovirta, K. Partanen, P. Riekkinen, H. Soininen, Hippocampus in Alzheimer's disease: A 3-year follow-up MRI study, Biol. Psychiat. 47 (2000) 557–561.
[44] K.S. Frederiksen, E. Garde, A. Skimminge, C. Ryberg, E. Rostrup, W.F. Baaré, H.R. Siebner, A.-M. Hejl, A.-M. Leffers, G. Waldemar, Corpus callosum atrophy in patients with mild Alzheimer's disease, Neurodegenerative Dis. 8 (6) (2011) 476–482.
[45] Disconnection of the right superior parietal lobule from the precuneus is associated with memory impairment in oldest-old Alzheimer's disease patients, Heliyon 6 (7) (2020) e04516.
[46] X. Liu, P. Cao, D. Zhao, O. Zaiane, et al., Group guided sparse group lasso multi-task learning for cognitive performance prediction of Alzheimer's disease, in: International Conference on Brain Informatics, Springer, 2017, pp. 202–212.