



Published in final edited form as:

Pattern Recognit. 2019 April ; 88: 370–382. doi:10.1016/j.patcog.2018.11.027.

Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis

Jialin Peng^{a,b,c,*}, Xiaofeng Zhu^c, Ye Wang^a, Le An^c, and Dinggang Shen^{c,d,*}

^aCollege of Computer Science and Technology, Huaqiao University, Xiamen, China

^bXiamen Key Laboratory of CVPR, Huaqiao University, Xiamen, China

^cDepartment of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

^dDepartment of Brain and Cognitive Engineering, Korea University, Seoul, Korea

Abstract

Multimodal data fusion has shown great advantages in uncovering information that could be overlooked by using single modality. In this paper, we consider the integration of high-dimensional *multi-modality imaging* and *genetic* data for Alzheimer's disease (AD) diagnosis. With a focus on taking advantage of both phenotype and genotype information, a novel structured sparsity, defined by $\ell_{1,p}$ -norm ($p > 1$), regularized multiple kernel learning method is designed. Specifically, to facilitate structured feature selection and fusion from heterogeneous modalities and also capture feature-wise importance, we represent each feature with a distinct kernel as a basis, followed by grouping the kernels according to modalities. Then, an optimally combined kernel presentation of multimodal features is learned in a data-driven approach. Contrary to the Group Lasso (i.e., $\ell_{2,1}$ -norm penalty) which performs sparse group selection, the proposed regularizer enforced on kernel weights is to *sparsely* select concise feature set within each homogenous group and fuse the heterogeneous feature groups by taking advantage of dense norms. We have evaluated our method using data of subjects from Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The effectiveness of the method is demonstrated by the clearly improved prediction diagnosis and also the discovered brain regions and SNPs relevant to AD.

Keywords

Structured sparsity; Multimodal features; Multiple kernel learning; Feature selection; Alzheimer's disease diagnosis

1. Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disorder, resulting in a gradual loss of memory and cognitive function [1]. Recognized as an early stage of AD, mild cognitive impairment (MCI) has a high risk of progressing to AD [2]. Advances in acquiring

*Corresponding authors. 2004pj1@163.com (J. Peng), dgshen@med.unc.edu (D. Shen).

multimodal imaging, such as magnetic resonance imaging (MRI) and positron emission tomography (PET) for the brain, provide unprecedented opportunities for early prediction of the disease. In this context, many machine learning methods [3–9] have been introduced to identify diverse biomarkers for AD and MCI using different image modalities, yielding important insights into the progression patterns of AD [1]. Actually, when used together for diagnosis of AD or MCI, different data modalities provide different *yet* complementary information [6,7,10–12]. On the other hand, genotype information has also played an increasingly important role in AD research over the past few decades [1,13,14]. Among the many factors that increase the risk of getting AD, the genetic variation has been identified as an important one [1,13] as AD is heritable. In particular, single nucleotide polymorphisms (SNPs) are the most common type of genetic variation [13]. Currently, identifying SNPs associated with AD have attracted a lot of attentions [13,14].

Therefore, it is important and beneficial to build prediction models by leveraging both phenotype and genotype data, e.g., MRI, PET, and SNPs, for improving diagnosis performance. Using biomarkers of multiple modalities may reveal hidden information that may be overlooked by using single modality. However, the integration of multimodal data is burdened by a number of challenges, such as limited observations, highly-redundant high-dimensional data, and the heterogeneous nature of the multimodal data. For instance, the genetic data usually contain thousands of SNP features with many irrelevant ones, while subjects with all modality data are relatively scarce. In fact, high-dimensional problem has represented a critical challenge in many fields [17]. While regularization by sparsity-inducing norms such as ℓ_1 norm [15] is a fruitful way of avoiding overfitting through variable or feature selection, taking into account structure information among features is another important topic [18–21]. Meanwhile, the phenotype and genotype data encode different level of knowledge for the disease, and different imaging modalities further capture different phenotype information. While directly pooling features together will treat multimodal features equally, the strategy that independently selects features from each modality and then combines them together is also suboptimal, as most discriminative features in one modal may not the best candidate ones for combination with features from other modalities. Therefore, feature selection and fusion by taking advantage of the multimodal nature of the data become essential to prepare clean, interpretable data and build simpler and more powerful models.

In view of the complementary information contained in different modalities in our case, all modalities are expected to contribute to AD prediction, although with the possible different levels of importance. In fact, for some modalities, their features as a whole are relatively weaker than those in other modalities. Therefore, when utilizing the sparsity inducing Lasso/Group Lasso for feature/group selection [14,22,23], features from weak modalities may have less chance to be selected, as illustrated in Fig. 1. To address this issue, we propose to learn a better multimodal feature combination by jointly selecting subsets of discriminative features from each modality with a novel structured sparsity regularizer.

In this paper, a novel route¹ is introduced for multimodal data based AD diagnosis (see Fig. 2). Specifically, we propose a (weighted) $\ell_{1,p}$ norm ($p > 1$) regularized multiple kernel learning (MKL) method for multimodal feature selection and fusion (note that a special case

(ℓ_2 norm) was already considered for regression [25] and multitask learning [26] and also analyzed by Kong et al. [27]). Instead of representing each modality with a kernel as previous works [10,22,28], we assign each feature with a distinct kernel through its own feature mapping and capture feature-wise importance by learning the weight for each kernel. Further, the kernels are grouped according to task specific criteria, e.g., feature modalities in this multimodal diagnosis task. Then, the proposed structured sparsity regularizer is utilized for feature selection through enforcing both feature-level and group-level constraints on kernel weights. More specifically, 1) by enforcing ℓ_1 sparsity on kernel weights within each modality, we can select the informative features from each modality in which the features are relatively homogenous, and 2) by performing dense ℓ_2 regularization across different modalities, which has the advantage of better combining complementary features than ℓ_1 norm [29], we can better fuse all modalities. In this way, the sparse feature selections from different modalities are performed simultaneously and also constrained by each other. Accordingly, features that are *not only* discriminative *but also* complementary will be more likely selected.

1.1. Background and related works

For AD diagnosis, many methods towards heterogeneous multimodal feature selection and fusion have been developed [6,9,11,19,30–32]. A straightforward way to fusion multimodality data is to pool features from multiple modalities together [33,34], and then feed them to train a classifier. However, this simple concatenation strategy has shown to be suboptimal to integrate heterogeneous modalities [10], as it tends to ignore relationships within and/or across modalities [9,28,35].

An alternative way to leverage multi-modal data is the kernel or graph based methods. In Zhang et al. [10] and Hinrichs et al. [36], each modality was represented with a kernel, and modality fusion is achieved through linear kernel combination. Many works [23,28,37–40] followed this strategy for multimodality fusion. In Gray et al. [41], each modality was represented by a graph encoding subject similarity; then, the modality fusion is achieved through linear combination of the graphs. Recently, Tong et al. [42] utilized cross diffusion among graphs to achieve nonlinear modality fusion.

Feature selection methods (e.g., Lasso [15], t -test, and Fisher Score [43]) are usually applied on individual modality before concatenation or on the concatenated feature vector. While selecting features individually will ignore the presence of other modalities, selecting features from concatenated features treats features from all modalities equally. For improved feature selection by incorporation of modality relationship, the concept of multitask learning has been introduced for AD diagnosis [6]. In Zu and Co-authors [37–39], each modality was regarded a task, and multimodal features are jointly selected. Specifically, they assumed different modalities have the same number of features characterizing group of brain subregions. With a sparsity regularization, different types of features for the same subregion are jointly selected or discarded. However, this assumption prohibits its application to the joint selection of genotype and phenotype features, where no direct correspondences

¹A preliminary version of this paper was appeared in Peng et al. [24].

between them exist. In [14], Wang et al. combined quantitative traits regression and multi-class classification as multi-task. Multimodal data integration is achieved through two Group Lasso regularizers ($\ell_{2,1}$ norm) [16] which performed sparse selection at the group level and dense combination features within each group. Specifically, one Group Lasso regularizer selected the most discriminative modality; the other one selected the most discriminative features from the kept modalities and also enforces the same feature to be kept across different modalities.

The most crucial element for a kernel based method is the kernel construction. Multiple kernel learning (MKL) [44,45] provides an elegant framework for learning a data-dependent kernel representation [29,46]. Specifically, MKL learns from data an optimal combination of a set of basis kernels. The selection of certain regularization methods yields different kernel selection and/or combination approaches [22,23,47]. In particular, ℓ_1 -MKL [46] with the sparsity inducing ℓ_1 norm [15] constraint on kernel weights, is able to sparsely select a few most relevant and discriminative kernels. Usually by encoding information of each whole modality with a base kernel, the kernel selection yields discriminative modality selection. In Yeh et al. [48] and Liu et al. [23], each modality was affiliated with several types of kernels, forming kernel groups; then the Group Lasso regularized MKL [47] was utilized to simultaneously select a few most relevant modalities and densely combine different kernel types within each group. It should be noted that, although some modalities (e.g., SNPs) contain large number of irrelevant features, they also encode critical aspects of pathological changes associated with AD.

In this paper, we address the problem of the multimodal feature selection and combination by kernel representation learning. To facilitate feature selection in kernel space, we firstly represent each feature with a basis kernel. Secondly, we explore groups (e.g., modality groups) in the multimodal features and use a novel $\ell_{1,p}$ norm, $1 < p$, to achieve structured feature selection and modality fusion simultaneously. The advantage of our structured sparsity method over the popular Lasso and Group Lasso is illustrated in Fig. 1. Specifically, although Lasso supports interpretability and scalability, it can only select the most discriminative features without being aware of any group information. Moreover, it is less effective to combine complementary features, which has been noticed in many studies [16,17] including MKL related methods [29]. Group Lasso can only select some most discriminative groups of features. In contrast, the proposed one will select discriminative features while being aware of heterogeneous groups and the effect of group combination.

Structured sparsity by $\ell_{q,p}(1-p, q)$ mixed norm [25], which can explore inherent group structure of data, has considered in many tasks [18,49,50]. In addition to the most notable Group Lasso ($\ell_{2,1}$ norm), more general cases $\ell_{q,1}$ norm with $q > 1$ and $p = 1$ has been considered in Zhao et al. [49]. In Rakotomamonjy et al. [51], the $\ell_{q,1}$ norm with $1 < q \leq 2$ was combined with MKL for multitask learning, and more generally the nonconvex case with $0 < p < 1$ was also considered. However, with $1 < q$ and $p = 1$, the regularizer only introduces group-level sparsity and all variables within the selected group will be selected. In some applications, it is desirable to also enforce sparsity within groups while keeping all groups. Accordingly, Kowalski et al. [25] investigated the effect of general $\ell_{q,p}(1-p, q)$ including the cases $q = 1$ for regression task with least square fidelity loss. For the case $\ell_{1,2}$,

in each group, at least one variable will be selected. The same $\ell_{1,2}$ norm penalty is also proposed in Zhou et al. [26] for multitask learning to achieve the competition among tasks. In this paper, we consider the combination of the more general case $\ell_{1,p}$ with $1 < p$ and MKL for multimodal data based diagnosis. A main challenge for this penalty, even for the apparently easier case $p = 2$, is that there is no analytical solution for its proximal operator [25], which is usually an essential tool to optimize with sparsity-inducing penalties [18]. By employing model structure, a block coordinate descent algorithm applicable to any $1 < p$ is introduced.

The remainder of the paper is organized as follows. Section 2 presents the proposed method and optimization algorithm. A number of experimental and comparative results are presented in Section 3, followed by discussion and conclusion in Sections 4 and 5.

2. Method

An overview of our framework is illustrated in Fig. 2. Let $\{x^{(i)}, y^{(i)}\}_{i=1}^N$ be the training data, where $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_M^{(i)})^T \in R^M$ is the data sample, M is the number of all features from all modalities, and $y^{(i)} \in \{1, -1\}$ is a class label. The aim is to simultaneously learn an optimal feature representation and a max-margin classifier in kernel space because of its efficient and elegant way of modeling complicated patterns. Specifically, the optimal kernel for feature representation is a linear combination of a set of basis kernels, each of which is a kernel representation of one raw feature. In this way, the selection of basis kernel amounts to feature selection. As both classification and feature representation can benefit from effective feature selection, we introduce structured sparse penalty on the weights for kernel combination. The learned optimal kernel is employed for classifier learning.

In the following, we denote vectors as boldface letters (e.g., θ) and vector elements as non-bold letters with subscripts (e.g., θ_j), the transpose of vector by the superscript T and a vector with all entries equal to a constant C as \mathbf{C} (e.g., $\mathbf{1}$). We also denote $\|\cdot\|_p$ with $p = 1$ as the ℓ_p norm of vector and $|\cdot|$ as the absolute value of scalar. Symbol \triangleq means definition.

2.1. Structured sparsity feature selection and kernel learning

In this section, we introduce the group structured sparsity penalized multiple kernel learning. In detail, as shown in Fig. 2 we separately transform the M -dimensional features into new feature spaces via M different feature mappings $\{\phi_m\}_{m=1}^M$, such that the originally complicated task is transformed into an easy linear one. By employing kernel trick [45], each ϕ_m gives rise to a basis kernel $K_m = 0$ defined by inner products in the new feature space. In the model computation, no feature mappings $\{\phi_m\}_{m=1}^M$ but kernels $\{K_m\}_{m=1}^M$ are needed to explicitly specify, which will be clear in Algorithm 1. Exploiting inherent group structures [16,18,52] has shown to improve the performance and interpretability of the learned models. Let $\mathcal{S} = \{1, 2, \dots, M\}$ be the feature index set which is partitioned into L non-overlapping groups $\{\mathcal{S}_l\}_{l=1}^L$ according to task-specific knowledge. For AD diagnosis, multimodal heterogeneous features are naturally partitioned into groups according to the number of

modalities. Although each modality contains large number of irrelevant features, all modalities contain crucial factors complementary to each other about AD.

Given the transformed feature space defined by the joint feature mapping $\Phi(\mathbf{x}) = (\phi_1(x_1), \phi_2(x_2), \dots, \phi_M(x_M))^T$ the objective is to learn a linear discriminant function $f(\mathbf{x})$ that can generalize well on unseen data,

$$f(\mathbf{x}) = \sum_{l=1}^L \sum_{m \in \mathcal{G}_l} \sqrt{\theta_m} \tilde{\mathbf{w}}_m^T \phi_m(x_m) + b. \quad (1)$$

Here, we have explicitly written out the group structure $\{\mathcal{G}_l\}_{l=1}^L$. Further,

$\tilde{\mathbf{w}} = (\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_M)^T$ is the normal vector of the decision hyperplane $f(\mathbf{x})$, b encodes the bias, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_M)^T$ contains the feature mapping weights. Thus, features with zero feature mapping weights would not be active in the discriminant function $f(\mathbf{x})$.

In order to obtain a filtered set of features, we propose to enforce an $\ell_{1,p}$ mixed norm on the weights (i.e., $\boldsymbol{\theta}$) of the feature mappings. More generally, we can further introduce 1) M positive weights $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_M)^T$ on the elements in $\boldsymbol{\theta}$, and 2) L positive weights $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_L)^T$ on groups in $\boldsymbol{\theta}$ to encode prior information. For example, for ROIs or ROI groups in brain known to be less relevant to AD, we can specify larger weights. If we have no knowledge about feature and/or group importance, we can set $\boldsymbol{\beta} = \mathbf{1}$ and/or $\boldsymbol{\gamma} = \mathbf{1}$. Accordingly, our generalized MKL model with a structured sparsity-inducing constraint can be formulated as below:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \tilde{\mathbf{w}}, b} C \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}^{(i)}), y^{(i)}) + \frac{1}{2} \sum_{l=1}^L \sum_{m \in \mathcal{G}_l} \left\| \tilde{\mathbf{w}}_m \right\|_2^2 \quad (2) \\ \text{s.t. } \left\| \boldsymbol{\theta} \right\|_{1,p; \boldsymbol{\beta}, \boldsymbol{\gamma}} \triangleq \left(\sum_{l=1}^L \gamma_l \left(\sum_{m \in \mathcal{G}_l} \beta_m |\theta_m| \right)^p \right)^{\frac{1}{p}} \leq \tau, \\ 0 \leq \boldsymbol{\theta}, \end{aligned}$$

where the first term of the objective function measures the classification error with the hinge loss $\mathcal{L}(t, y) = \max(0, 1 - ty)$, the second term ensures max-margin classification, the nonnegative parameter C is a trade-off weight of the two terms. The weighted $\ell_{1,p}$ mixed norm, i.e., $\|\cdot\|_{1,p; \boldsymbol{\beta}, \boldsymbol{\gamma}}$ in the inequality constraint simultaneously promotes sparse selection inside groups with the inner weighted ℓ_1 norm, and pursues dense combination of groups with the outer weighted ℓ_p norm ($p > 1$) which is not a sparsity-inducing norm. As has been discussed by Kloft et al. [29], although the non-sparse ℓ_p norm cannot promote feature

selection, it has the advantage of better combining complementary features than ℓ_1 norm. The nonnegative parameter τ in Eq. (2) is used to control the sparsity level of θ . As will be shown in Section 2.2, the parameters C and τ can be fold into one parameter and set $\tau = 1$.

Similar to the classical MKL [46], the subproblem about θ in (Eq. 2) is equivalent to learning an optimally combined kernel $K = \sum_{m=1}^M \theta_m K_m$ (see Eq. (12) and the Appendix for a proof). Therefore, θ also acts as weights for kernels. With the one-to-one correspondence between the M features and the M feature mappings (or M kernels), through optimizing Eq. (2), we can obtain the optimal coefficients θ^* which is implicitly related to all the features. The coefficients in θ^* indicate the feature contributions.

Algorithm 1

Block Coordinate Descent Algorithm for the Proposed Model.

-
- 1: **initial input:** $C', \{K_m\}, \beta, \gamma$, feasible θ , such as $\theta_m = \left(\sum_{l=1}^L \gamma_l \left(\sum_{m' \in \mathcal{G}_l} \beta_{m'} \right)^p \right)^{-\frac{1}{p}}$
 - 2: **while** optimality condition is not satisfied **do**
 - 3: Compute α in Eq. (12) and b using SVM solver [54]
 - 4: Compute $\|w_m\|_2$ for all $m = 1, \dots, M$ according to Eq. (13)
 - 5: Update θ_m for all $m = 1, \dots, M$ according to Eq. (4), i.e.,

$$\theta_m = \frac{\|w_m\|_2}{\beta_m^2 \gamma_m^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1; \beta}^{\frac{p-1}{p+1}}} \cdot \frac{1}{\left(\sum_{l=1}^L \gamma_l^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1; \beta}^{\frac{2p}{p+1}} \right)^{\frac{1}{p}}}$$
 - 6: **end while**
-

The rationale of using the proposed structured sparsity constraint is that, each individual modality contains redundant high-dimensional features and meanwhile offers complementary information to other modalities. As for the AD, the genetic and anatomical variations encoded by SNPs and imaging modalities are different but crucial measurements for the brain structure and function. Consequently, each modality is crucial. Accordingly, the proposed structured sparsity constraint 1) promotes sparse feature selection within each modality, which is desirable to interpret the results and obtain a simplified decision rule, and 2) encourages dense ℓ_p combination across groups to leverage the synergy between different modalities. As will be shown later in Eq. (4), θ attains optimal value at the contour of constant τ , i.e., $\|\theta^*\|_{1,p;\beta,\gamma} = \tau$, moreover, the optimal value tends to attain at regions with high curvature [17,53]. Specifically, for $p = 2$, an ℓ_2 norm regularizer encourages all feature groups to have similar degrees of importance; for $1 < p < 2$, ℓ_p norm regularizer encourages different degrees of importance for different groups, as the high curvature regions are nearby the axis. for $p = 1$, it is well known that the ℓ_1 norm promotes some groups to have zero weights/contributions. In view of different modality's strength for AD classification, we take a compromise of Lasso (ℓ_1 norm) and Ridge (ℓ_2 norm) regularization and intuitively set $p = 1.5$ for inter-group regularization, thus allowing the assignment of larger weights for leading groups/modalities.

Note that the proposed $\ell_{1,p}$ -norm based regularization is completely different from Group Lasso ($\ell_{2,1}$ norm) which gives a sparse set of groups but performs no feature selection within each group [16,47]. When all features form a single group, the proposed model degenerates to ℓ_1 -MKL [46]. A comparison of different sparsity patterns selected by Lasso, Group Lasso and the proposed structured sparsity is shown in Fig. 1. Specifically, the Lasso and Group Lasso tend to sparsely select few most discriminative features and groups, respectively. Thus features in some relatively-weak modalities may be mostly or totally discarded. In contrast, the proposed model can *not only* keep information from each modality with the outer

nonsparse regularization *but also* support variable interpretability and scalability with the inner sparse feature selection. Moreover, feature selections in different modalities interact with each other in our method to finally obtain a better combined feature set.

2.2. Model analysis and computation

For further understanding of the proposed model, we introduce variable changing. Specifically, let $w_m = \sqrt{\theta_m} \bar{w}_m$ for each m and $w = (w_1, w_2, \dots, w_M)^T$, the proposed model (2) can be reformulate into the following convex optimization problem,

$$\begin{aligned} \min_{\theta \geq 0, w, b} \quad & C \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}^{(i)}), y^{(i)}) + \frac{1}{2} \sum_{l=1}^L \sum_{m \in \mathcal{G}_l} \frac{\|w_m\|_2^2}{\theta_m} \quad (3) \\ \text{s.t.} \quad & \left\| \boldsymbol{\theta} \right\|_{1,p;\beta,\gamma} \leq \tau, \end{aligned}$$

where, here and in what follows, we use the convention that $0/0 = 0$. We have the following lemma, which also gives light on the computation of the sub-problem about $\boldsymbol{\theta}$ in Eq. (2).

Lemma 1. (Solution for the subproblem of $\boldsymbol{\theta}$) Given $p \geq 1$, positive weights $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Let $\mathbf{W} = (\|w_1\|_2, \|w_2\|_2, \dots, \|w_M\|_2)^T$. For fixed $\mathbf{W} \geq 0$ and b , the minimal $\boldsymbol{\theta}$ in Eq. (3) is attained at

$$\theta_m^* = \frac{\|w_m\|_2}{\beta_m^{\frac{1}{2}} \gamma_l^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1;\beta}^{\frac{p-1}{p+1}}} \cdot \frac{\tau}{\left(\sum_{l=1}^L \gamma_l^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1;\beta}^{\frac{2p}{p+1}} \right)^{\frac{1}{p}}}, \quad \forall m = 1, 2, \dots, M \quad (4)$$

where $\left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1;\beta} = \sum_{m' \in \mathcal{G}_l} \beta_{m'}^{\frac{1}{2}} \|w_{m'}\|_2$, and \mathcal{G}_l is the index set that m belongs to. Moreover, $\|\boldsymbol{\theta}^*\|_{1,p;\beta,\gamma} = \tau$.

Proof. The partial Lagrangian function associated to Eq. (3) is

$$L = C \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}^{(i)}), y^{(i)}) + \frac{1}{2} \sum_{l=1}^L \sum_{m \in \mathcal{G}_l} \frac{\|w_m\|_2^2}{\theta_m} + \frac{\mu}{2} (\|\boldsymbol{\theta}\|_{1,p;\beta,\gamma} - \tau), \quad (5)$$

where $\mu \geq 0$ is the Lagrange multiplier. As the objective function with respect to $\boldsymbol{\theta}$ is monotone, the convex constraint is active and $\|\boldsymbol{\theta}\|_{1,p;\beta,\gamma} = \tau$. So, the Lagrange multiplier $\mu > 0$. According to the first order optimality conditions, i.e., the Karush-Kuhn-Tucker (KKT) conditions [54], we have

$$\begin{aligned}
 0 &= \frac{\partial L}{\partial \theta_m} = -\frac{\|\mathbf{w}_m\|_2^2}{\theta_m^2} + \mu \left\| \boldsymbol{\theta} \right\|_{1,p;\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{\partial \|\boldsymbol{\theta}\|_{1,p;\boldsymbol{\beta},\boldsymbol{\gamma}}}{\partial \theta_m} \quad (6) \\
 &= -\frac{\|\mathbf{w}_m\|_2^2}{\theta_m^2} + \mu \left\| \boldsymbol{\theta} \right\|_{1,p;\boldsymbol{\beta},\boldsymbol{\gamma}}^{2-p} \left\| \boldsymbol{\theta}_{\mathcal{G}_{lm}} \right\|_1^{p-1} \gamma_{l_m} \boldsymbol{\beta}_m.
 \end{aligned}$$

Let $\xi = 1/(\mu \|\boldsymbol{\theta}\|_{1,p;\boldsymbol{\beta},\boldsymbol{\gamma}}^{2-p})$, we have

$$\theta_m \left\| \boldsymbol{\theta}_{\mathcal{G}_{lm}} \right\|_1^{\frac{p-1}{2}} = \xi \gamma_{l_m}^{-1/2} \boldsymbol{\beta}_m^{-1/2} \|\mathbf{w}_m\|_2. \quad (7)$$

Taking into account the definition of $\|\boldsymbol{\theta}\|_{1,p;\boldsymbol{\beta},\boldsymbol{\gamma}}$ we further have

$$\left\| \boldsymbol{\theta}_{\mathcal{G}_{lm}} \right\|_1 = \xi^{\frac{2}{p+1}} \gamma_{l_m}^{-\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_{lm}} \right\|_{1;\boldsymbol{\beta}}^{\frac{2}{p+1}}. \quad (8)$$

By using $\tau = \|\boldsymbol{\theta}\|_{1,p;\boldsymbol{\beta},\boldsymbol{\gamma}}$ and $\mathbf{W} = 0$, we can obtain

$$\tau = \left\| \boldsymbol{\theta} \right\|_{1,p;\boldsymbol{\beta},\boldsymbol{\gamma}} = \left(\sum_l \gamma_l \left\| \boldsymbol{\theta}_{\mathcal{G}_l} \right\|_1^p \right)^{1/p} = \xi^{\frac{2}{p+1}} \left(\sum_l \gamma_l^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1;\boldsymbol{\beta}}^{\frac{2p}{p-1}} \right)^{1/p}. \quad (9)$$

$$\xi^{\frac{2p}{p+1}} = \frac{\tau^p}{\sum_{l=1}^L \gamma_l^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1;\boldsymbol{\beta}}^{\frac{2p}{p+1}}} \quad (10)$$

Resubstitution of $\xi^{\frac{2p}{p+1}}$ and $\left\| \boldsymbol{\theta}_{\mathcal{G}_{lm}} \right\|_1$ into Eq. (7) yields the claimed results. By using the

definition of the weighted $\mathcal{L}_{1,p}$ mixed norm, we can obtain $\|\boldsymbol{\theta}^*\|_{1,p;\boldsymbol{\beta}} = \tau$. \square

For fixed \mathbf{W} , Eq. (2) gives an explicit solution for $\boldsymbol{\theta}$. Plugging Eq. (4) into Eq. (2) yields the following equivalent form for the proposed model.

Theorem 1. Let $p' = \frac{2p}{p+1}$. For $p \geq 1$, the model in Eq. (2) (also Eq. (3)) is equivalent to

$$\min_{\mathbf{w}, b} C \sum_{i=1}^N \mathcal{L} \left(\sum_{l=1}^L \sum_{m \in \mathcal{E}_l} \mathbf{w}_m^T \boldsymbol{\phi}_m(x_m^{(i)}) + b, y^{(i)} \right) + \frac{1}{2\tau} \left(\sum_{l=1}^L \gamma_l \frac{2-p'}{p'} \left(\sum_{m \in \mathcal{E}_l} \beta_m^2 \left\| \mathbf{w}_m \right\|_2 \right)^{p'} \right)^{\frac{2}{p'}} \quad (11)$$

The second term in Eq. (11) is again a weighted $q_{p'}$ norm penalty on \mathbf{W} with $p' \in [1, 2)$. By choosing $p = 1.5$ and thus $p' = 1.2$, it shares similar group-level regularization property with that in Eq. (2) on $\boldsymbol{\theta}$. As a result, in each group, only a small number of \mathbf{w}_m can contribute to the decision function $f(\mathbf{x})$ with nonzero values. Accordingly, only a few features in each group can be selected. Meanwhile, the sparsely filtered groups are densely combined, while allowing the presence of leading groups.

After the variable changing, the model in Eq. (2) is convex w.r.t. to $\boldsymbol{\theta}$, \mathbf{w} and b , respectively. We can optimize it via block coordinate descent, which updates just one block of variables at a time. Moreover, from Eq. (11), it is obvious that we can fold τ and C into a single trade-off weight C' and set $\tau = 1$. In this way, we have single model parameter C' which *not only* acts as the soft margin parameter *but also* controls the sparsity of $\boldsymbol{\theta}$ and \mathbf{W} .

The detailed optimization procedure is shown in Algorithm 1. To run this algorithm, we need not to specify the feature mappings, and instead we just need to specify the base kernels by using the following dual form of Problem (3) (see Appendix for a proof),

$$\begin{aligned} \min_{\boldsymbol{\theta} \geq 0} \max_{\boldsymbol{\alpha}} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\sum_{m=1}^M \theta_m K_m(x_m^{(i)}, x_m^{(j)}) \right) \quad (12) \\ \text{s.t. } \|\boldsymbol{\theta}\|_{1,p}, \beta, \gamma \leq 1, \sum_{i=1}^N \alpha_i y^{(i)} = 0, \\ 0 \leq \alpha_i \leq C', i = 1, \dots, N. \end{aligned}$$

The vector $\boldsymbol{\alpha}$ is the dual variable of \mathbf{w} , and $\mathbf{w}_m = \theta_m \sum_{i=1}^N \alpha_i y^{(i)} \boldsymbol{\phi}_m(x_m^{(i)})$ in the optimal point for the fixed $\boldsymbol{\alpha}$, hence we have

$$\left\| \mathbf{w}_m \right\|_2^2 = \theta_m^2 \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y^{(i)} y^{(j)} K_m(x_m^{(i)}, x_m^{(j)}) \quad (13)$$

Accordingly, the decision function with kernel is as follows,

$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i y^{(i)} \left(\sum_{m=1}^M \theta_m K_m(x_m, x_m^{(i)}) \right) + b. \quad (14)$$

As each base kernel is based on one dimensional feature, the computational time for the summed kernel is $\mathcal{O}(M)$. Apparently, the running complexity needed to classify a sample is $\mathcal{O}(NM)$. However, as a max-margin classifier similar to SVM, only $N_{sv} (< N)$ support vectors involve the prediction. Moreover, due to the sparse selection of base kernels, the overall complexity is $\mathcal{O}(N_{sv}M_s)$, where $M_s (< M)$ is the number of selected base kernels.

3. Experiments

In this section, we study the proposed $\ell_{1,p}$ -norm ($p > 1$) regularized MKL method in terms of efficiency of feature selection and accuracy of diagnosis. To this end, a synthetic data experiment was firstly introduced to shed light on the efficiency of structured feature selection. For the evaluation of AD diagnosis, we employed data obtained from ADNI dataset².

3.1. Simulation study

The synthetic data with feature groups was generated for classification and feature selection tasks as follows. Suppose a dataset with $N = 100$ observations and $M = 100$ variables is collected in data matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, in which the M features have L groups. Specifically, we used $L = 5$ groups of equal size in the simulation, i.e., indexes in the l th group are from $(l - 1) * M/L + 1$ to $l * M/L$. The data matrix \mathbf{X} samples from a multivariate normal distribution $\mathcal{N}(\mathbf{1}, \Sigma)$ with a Toeplitz covariance matrix Σ that encourages different level of correlation between groups and within groups. For variables within the l th group, the entry of Σ is $\Sigma_{i,j} = c_l^{|i-j|}$, $l = 1, 2, \dots, L$; for variables between groups, the entry of Σ is $\Sigma_{i,j} = d^{|i-j|}$. With larger c_l , the correlation within the l th group will be larger. We set $\mathbf{c} = (0.1, 0.3, 0.5, 0.6, 0.7)^T$ to simulate varied within-group correlations, and the parameter d was set 0.1.

We consider the classification model $\mathbf{y} = \text{sign}(\mathbf{X}\boldsymbol{\xi} + \mathbf{b} + \boldsymbol{\epsilon})$, where \mathbf{y} contains the true class labels, $\boldsymbol{\xi}$ is true parameter with sparsity, \mathbf{b} is the bias and $\boldsymbol{\epsilon}$ is white Gaussian noise with a standard deviation of 0.3. In each group, $\boldsymbol{\xi}$ has only one non-zero element, which takes value from standard norm distribution. In the experiment, the indexes of non-zero elements in coefficient $\boldsymbol{\xi}$ were $\{1, 32, 46, 62, 93\}$, and the values of non-zero elements were $\{0.3591, -0.7943, -0.2273, 1.5938, 0.1552\}$, respectively. The bias was set $\mathbf{b} = -0.8 * \mathbf{1}$. For this simulation study, all of the 5 groups have contribution to the prediction, and for each group, one feature is expected to be selected.

To investigate the effect of structured feature selection, we compared our method ($\ell_{1,p}$ -MKL) with three sparse feature selection methods: 1) the Lasso (ℓ_1 norm penalty) with logistic loss; 2) the Group Lasso ($\ell_{2,1}$ norm penalty) with logistic loss; 3) the ℓ_1 -MKL which is a widely-used model but does not consider structure information. An graphical comparison of Lasso, Group Lasso and the proposed method ($p > 1$) has shown in Fig. 1. Further quantitative studys based on synthetic data are listed in Table 1. The experiment showed that, 1) without taking into account structure information, the Lasso method only selected features from group 2 and 4, and the ℓ_1 -MKL selected features from group 2, 4 and

²<http://adni.loni.usc.edu>

1 with higher chance; 2) the Group Lasso selected three whole groups of features, i.e., group 1, 2, and 3, and the total number of selected features was still large; 3) in contrast, the proposed method with $p > 1$ selected features from all the 5 group, although an alternative feature from group 5 was selected due to high correlation in group 5. Moreover, the classification accuracy (ACC) of our method, especially with $p = 1.5$, was higher than Lasso, Group Lasso and $\ell_{1,p}$ -MKL. For Lasso and Group Lasso, SVM was used as the final classifier; for $\ell_{1,p}$ -MKL and our method, linear base kernel is assigned for each feature to facilitate feature selection. Overall, the results suggest that, when all feature groups are known to contain critical information, the $\ell_{1,p}$ norm penalty is favored. This is the case for AD diagnosis, where crucial and complementary information is encoded in data of multimodalities. For the proposed method, we also investigated different choices of p , i.e., $\ell_{1,1.5}$ -MKL, $\ell_{1,2}$ -MKL and $\ell_{1,6}$ -MKL. With different p , the proposed method selected similar top features with possibly different orders of importance. However, the performance with $p = 1.5$ was better. As noted in Zhao et al. [49], for ℓ_p norm penalty with $1 < p < 2$, estimated coefficients lying in directions closer to the axis are favored; as p ($p > 2$) goes larger, estimated coefficients tend to concentrate along the diagonals, promoting equal coefficients. Roughly speaking, when employed as a group-level regularizer, with $1 < p < 2$ some dominated groups are allowed, and with $p > 2$ dominated groups will be not encouraged. Therefore, when different groups contribute equally, the larger $p > 2$ is favored; otherwise, $1 < p < 2$ is more favorable.

3.2. Real data: ADNI data

The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations as a public-private partnership. Investigators³ within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. The primary goal of ADNI has been to test whether MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers for AD is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

We evaluated our method by applying it on two subsets (named Dataset I and Dataset II in the following) of the ADNI. The Dataset I contains MRI, PET, and SNP data of 189 subjects, including 49 patients with AD, 93 patients with MCI, and 47 Normal Controls (NC); the Dataset II consists of MRI, PET, and SNP data of 360 subjects, including 85 patients with AD, 185 patients with MCI, and 90 Normal Controls (NC). The demographic information of subjects in the two datasets is summarized in Table 2. For inclusion/exclusion of the subjects, we have used the following general criteria: 1) for NC subjects that are non-depressed, non MCI, and non-demented, the MMSE (Mini-Mental State Examination) score is between 24 and 30 with Clinical Dementia Rating (CDR) of 0; 2) For MCI subjects, the MMSE score is between 24 and 30, with CDR of 0.5, and each subject is an absence of

³<http://adni.loni.usc.edu>

significant level of impairment in other cognitive domains; 3) the MMSE score of each Mild AD subject is between 20 and 26, with the CDR of 0.5 or 1.0. The criteria are according to the National Institute of Neurological and Communication Disorders/Alzheimer's Disease and Related Disorders Association for probable AD.

The MRI images acquired from 1.5 T scanners were downloaded from the public ADNI site, and then reviewed for quality. The PET images were 18-fluorodeoxyglucose (FDG) PET images and acquired 30–60 min post-injection. Firstly, they were spatially aligned, interpolated to a standard voxel size, normalized in intensity, and smoothed to a resolution of 8 mm full width at half maximum. Following the same procedures as in Zhang et al. [6], we preprocessed the MRI and PET images by applying anterior commissure-posterior commissure correction using MIPAV software⁴, intensity inhomogeneity correction with the N3 algorithm [56], skull-stripping using both brain surface extractor (BSE) [57] and brain extraction tool (BET) [58], and cerebellum removal. After that, FAST in the FSL package [59] is used to segment structural MR images into three different tissues: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). After registration using HAMMER [60], the MRI and PET images were segmented into 93 regions-of-interest (ROIs) according to the template [61]. The GM volumes of these ROIs in MRI and the average intensity of each ROI in PET were calculated and used as features.

The SNPs [1] were genotyped using the Human 610-Quad Bead-Chip (Illumina, Inc., San Diego, CA, USA). The SNPs, belonging to the top AD candidate genes listed on the AlzGene database⁵ as of June 10, 2010, were selected after the standard quality control (QC) and imputation steps. The QC criteria include (1) gender check, (2) call rate check per subject and per SNP marker, (3) sibling pair identification, (4) the HardyWeinberg equilibrium test, (5) marker removal by the minor allele frequency, and (6) population stratification. Then, the quality-controlled SNPs were imputed using the MaCH software [62] to estimate the missing genotypes. For the Data I, the Illumina annotation information based on the Genome build 36.2 was used to select a subset of SNPs, belonging or proximal to the top 135 AD candidate genes. The above procedure yielded 5677 SNPs. For the Data II, we obtained 3123 SNPs extracted from 153 genes (boundary: 20KB) using the ANNOVAR annotation. Thus, we totally have $93 + 93 + 5677 = 5863$ features on Dataset I, and $93 + 93 + 3123 = 3309$ features on Dataset II from the three modalities for each subject.

3.3. Experimental settings

To evaluate performances of classification methods, we used a 10-fold cross-validation strategy by partitioning the whole dataset into training and testing subsets. The final classification results were obtained by repeating the 10-fold cross-validation 10 times, to avoid any possible bias during dataset partitioning. All parameters tuning were performed by conducting 5-fold inner cross-validation on the training part of the outer cross-validation. Four performance measures including classification accuracy (ACC), sensitivity (SEN), specificity (SPE) and area under receiver operating characteristic (ROC) curve, also known as AUC, were used. Here, the ACC measures the proportion of subjects correctly classified,

⁴<http://mipav.cit.nih.gov/clickwrap.php>

⁵<http://www.alzgene.org>

the SEN represents the proportion of positive class correctly identified, and the SPE denotes the proportion of negative class correctly identified. All datasets are standardized to have zero-mean and unit-variance for each dimension.

On the ADNI data, we compared the proposed method with 1) feature selection based methods, i.e., Fisher Score (FS) [43], and Lasso [15], and 2) MKL based methods, i.e., the multimodal multiple kernel learning method (M-MKL) [10], and l_1 -MKL [46]. FS is one of the most widely used supervised feature selection methods due to its general good performance. In detail, it selects the features such that the feature values of the samples within the same class are small, while the feature values of the samples from different classes are large. The Lasso method can select a small set of discriminative features by directly pooling all features together and enforcing an l_1 norm regularization. M-MKL method represented each modality with a base kernel and further learned a linearly-combined kernel with cross-validation. For FS, Lasso and M-MKL, the support vector machine (SVM) [55] implemented in LibSVM software⁶ was used as the classifier, and they are denoted as FSSVM, Lasso-SVM, and M-MKL respectively. The l_1 -MKL is a special case of our method, but treating features from different modalities equally. *For our proposed model and l_1 -MKL, each base kernel is defined based a single feature to facilitate feature selection.*

For FS method, the proportion of selected feature was determined with cross-validation.⁷ For Lasso-SVM, M-MKL, l_1 -MKL, and our proposed method, we used t -test [63] thresholded by p -value as a feature pre-selection step to reduce feature size and improve computational efficiency. The commonly used p -value < 0.05 was applied for MRI and PET. Considering the large number of SNP features, we selected p -value from $\{0.05, 0.02, 0.01\}$ with inner cross validation. As t -test was used as a feature pre-selection step, t -test-SVM that combined t -test and SVM was designed for comparison with the same p -value setting as well. It should be noted that the t -test based pre-selection was performed on training set in the cross-validation procedures. The parameter in Lasso used to control the contributions of the loss term and l_1 -norm term was selected from the range $\{2^{-10}, 2^{-9}, \dots, 2^1\}$ through an inner cross-validation. The soft margin parameter C' in our method and C in SVM were selected with grid search from $\{2^{-5}, 2^{-4}, \dots, 2^5\}$.

To reduce the number of parameters, linear kernel [55] was used as the default kernel type in the experiments, and other kernel types [55] such as Gaussian kernel (also known as Radial function kernel) and Polynomial kernel also have been tested in Section 4. Furthermore, we simply assumed no knowledge on both feature and group weights and thus set $\gamma = \mathbf{1}$ and $\beta = \mathbf{1}$.

3.4. Classification results on Data I from ANDI

On Data I, the classification results of AD vs. NC, MCI vs. NC as well as AD vs. MCI using all the three modalities, i.e., MRI, PET, and SNPs, are listed in Table 3. By taking advantage of both sparse feature selection and dense feature group combination, the proposed method

⁶<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

⁷The parameter set is $\{1\%, 2\%, 3\%, 4\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 70\%\}$.

outperforms all competing methods in classification rate. Specifically, for AD vs. NC classification, our method achieves an ACC of 96.1% with an improvement of 2.1% over the best performance of other methods. For classifying MCI from NC, the improvements by the proposed method is 2.4% in terms of ACC. For classifying AD from MCI, the improvements by the proposed method is 3.9% in terms of ACC. In comparison with t -test-SVM, we obtain 4.2%, 7.6% and 10.6% improvements in terms of ACC for AD vs. NC, MCI vs. NC and AD vs. MCI, respectively. Furthermore, we performed t -test on the ACCs between our proposed method and other compared methods. The asterisk symbols in Table 3 indicate statistically significant improvement of our method compared to each method under comparison.

In Fig. 3, we further plot the ROC curves of different methods under comparison for AD vs. NC, MCI vs. NC and AD vs. MCI classification, respectively. The corresponding AUC values are listed in Table 3. From the ROC curves and AUC values, we can see the superior classification performance by our proposed method. Specifically, the AUC obtained by the proposed method for AD vs. NC, MCI vs. NC and AD vs. MCI classification are 0.992, 0.811 and 0.808, respectively, showing better classification ability than other methods. In summary, the results show that our proposed method can improve the classification results.

To further investigate the benefit of multimodality fusion and also the effect of SNP data, we illustrate the performance of the proposed method w.r.t. different modality combinations. Note that, with single modality and thus single group, the proposed method degenerates to ℓ_1 -MKL. Table 4 summaries all the results. First of all, we can see that the performance of any single modality is much lower than that of their combinations. Among the three modalities, the SNPs show the lowest performance. However, when combined with other modalities, genetic data can obviously help improve predictions. For example, in AD and NC classification, the performances using MRI+SNP and PET+SNP demonstrate 2.7% and 5.7% improvements in terms of ACC over the cases of only using MRI and PET, respectively; the improvement with MRI+PET+SNP over that with MRI+PET is 3.8%. Similar results are obtained for MCI vs. NC and AD vs. MCI classification. All these results show that, the combination of multiple modalities including SNP data can help improve diagnosis performances.

The most frequently selected features in cross-validation are regarded as the most discriminative brain regions or SNPs, which can be potential biomarkers for clinical diagnosis. Top 10 ROIs identified from MRI and PET data for AD classification are illustrated in Fig 4. Specifically, the ROIs selected from MRI include angular gyrus right, amygdala right, uncus right, uncus left, hippocampal formation left, hippocampal formation right, inferior temporal gyrus right, middle temporal gyrus right, temporal pole right, and perirhinal cortex left; the ROIs selected from PET include hippocampal formation left, precuneus left, precuneus right, entorhinal cortex right, entorhinal cortex left, angular gyrus left, angular gyrus right, inferior temporal gyrus left, middle temporal gyrus left, and superior temporal gyrus left. Generally, these identified ROIs are in agreement with other recent AD studies [6,10,11,28,64]. In MRI, hippocampal formation and uncus in parahippocampal gyrus are recognized in AD vs. NC classifications, as well as multiple temporal gyrus regions. This is in line with the findings of the most affected regions in AD in previous neuro-studies [10,11,65]. Amygdala, one of the subcortical regions, is the

integrative center for emotions, and is also identified in AD. In PET, angular gyri, precuneus, and entorhinal cortices are the regions identified, which are also among the altered regions in AD reported in previous studies [10,28,65]. As for the genetic information, the most selected SNPs for AD and NC classification are from ApoE gene (rs429358 and rs769449), VEGFA gene (rs3025035), and SORCS1 gene (rs822097). Generally, our results are consistent with the existing results [1,31]. For instance, the ApoE gene and SORCS1 gene are the well-known top candidate genes related to AD and MCI [1]. VEGFA, the expression of vascular endothelial growth factor, represents a potential mechanism where vascular and AD pathologies are related [66].

3.5. Classification results on Dataset II from ADNI

For further validation of the proposed method, we validate it on Dataset II containing 360 subjects. Fig. 5 illustrates the comparison results. Specifically, for AD vs NC classification, our method achieves an ACC of 94.5%, which is higher than all the other methods, i.e., t -test-SVM 90.6%, FS-SVM 91.3%, M-MKL 91.0%, Lasso-SVM 91.4%, and ℓ_1 -MKL 92.1%. Moreover, the prediction performances (in ACC) of t -test-SVM, FS-SVM, M-MKL, Lasso-SVM, ℓ_1 -MKL, and the proposed method for MCI vs NC classification are respectively (73.1%, 74.5%, 74.3%, 74.8%, 77.8%, 80.2%), and for AD vs MCI classification, are (74.3%, 75.1%, 74.0%, 75.6%, 78.0%, 80.1%), respectively. With statistical significance ($p < 0.05$), our method outperforms the competing methods in terms of the ACC for all of the three tasks.

4. Discussions

To show the effect of different kernel types to our method, we further tested another two widely-used non-linear kernels, i.e., Gaussian kernel and Polynomial kernel. Fig. 6 demonstrates the results for two different classification tasks. In the experiment, all of the methods except of t -test-SVM enable feature selection, resulting in a small subset of features for classification. As shown in Fig. 6, while non-linear kernels performs better than Linear kernel for t -test-SVM, Linear kernel shows competitive performance with well-filtered discriminative features. Specifically, for the proposed method, the ACC performances with Linear, Gaussian and Polynomial kernel for AD vs NC classification are 94.46%, 94.2%, and 94.51%, respectively, and, for MCI vs NC are 80.22%, 79.82%, and 80.36%, respectively. The performance of the proposed method using Linear kernel has no statistically significant difference ($p > 0.05$) with that with other two types of kernel. Moreover, non-linear kernels contains more parameters than their linear counterpart, which requires further cross-validation for parameter selection.

The effects of different choices of p on the $\ell_{1,p}$ penalty are also illustrated in Fig. 8. Taking into account the varied correlations of different modalities with AD, we intuitively selected $p = 1.5$ for a compromise of Lasso and Ridge regularization. This has been further confirmed by Fig. 8, and the proposed model with $p \in [1.5, 2]$ performs better. In Fig. 7, the model performances with the varied trade-off weight C' are demonstrated. Specifically, the ACC performances of our model with parameters in $\{2^{-5}, 2^{-4}, \dots, 2^5\}$ are illustrated. In our experiments, this parameter was selected with cross-validation.

It is interesting to compare the proposed method with more existing multi-modality based AD diagnosis methods. Here, we just name some typical works that also used genetic data (which may contain only alleles of ApoE gene). Gray et al. [41] combined MRI, PET, CSF and alleles of ApoE gene; on a data set containing 147 subjects, their graph combination method obtained classification accuracies of 89.0% and 74.6% for AD vs. NC and MCI vs. NC classifications, respectively. Hinrichs et al. [36] integrated MRI, PET, CSF, ApoE genotype data and cognitive data, and achieved an accuracy of 92.4% on 233 subjects for AD and NC classification. Tong et al. [42] used MRI, PET, CSF and ApoE of 147 subjects for features, and their nonlinear graph fusion method achieved an ACC of 91.8% for AD vs. NC and an ACC of 79.5% for MCI vs. NC. Zhang et al. [31] compared several machine learning algorithms for multimodal feature selection using MRI, PET, CSF and SNP data on a dataset of 189 subjects. They reported that the high-order graph matching based method proposed [67] can achieve an accuracy of 92.9% for AD vs. NC classification and an accuracy of 76.4% for MCI vs. NC classification; the sparse multimodal learning proposed by Wang et al. [68] was also tested and obtained an accuracy of 94.8% for AD vs. NC classification and an accuracy of 75.6% for MCI vs. NC classification. Generally speaking, our proposed method showed competing or better performance than these methods, which further validated the efficacy of our proposed method.

5. Conclusions

In this study, we developed a kernel-learning-based method for multimodal feature selection and integration, and further applied it on imaging and genetic data for AD diagnosis. Other than independently selecting features from each modality and then combining them together [6,10] or performing modality selection [22,35], we integrated the feature selection and modal combination in a structured sparsity regularized kernel learning framework, performing both individual-level and group-level feature selection and fusion. Different from the commonly used structured sparsity or group sparsity regularization methods [16,18,47] which focus on sparsely selecting the most relevant feature groups, we proposed to sparsely select features within each modality and densely combine different modalities by taking account of the correlations and interactions between different modalities. The proposed model was formulated into a compact optimization problem with a weighted $\ell_{1,p}$ -norm constraint. A block coordinate descent algorithm applicable to any $p > 1$ was derived to solve the proposed formulation. Comparisons by various experiments on two different datasets have shown competing AD/MCI diagnosis performance by our proposed method.

Despite the sound performance of the proposed model, there are also several limitations that should be considered in future study. First, in our general framework, we just assumed no prior knowledge about feature/group importance. It will be interesting to learn more knowledge about feature/group importance and substructures in each modality from data. Second, kernel weights are fixed for all subjects, we will consider locally weighted kernel weights as [69].

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (11771160, 11401231, 61502182, 61673186), Foundation for the National Institutes of Health (1U01MH110274, EB0 06733, EB008374,

EB009634, MH100217), the Promotion Program in Science and Technology Research of Huaqiao University (ZQN-PY411).

Appendix

In this appendix, we show the derivation of the dual form of Problem (3). According to Theorem 1, we can fold parameter τ and C into a single trade-off weight C' . Further, note that we can rewrite $C'\mathcal{L}(t, y) = \max_{\alpha \in [0, C']} \alpha(1 - ty)$. Thus, the optimization problem in Eq. (3) can be rewritten as follows,

$$\begin{aligned} & \min_{\theta \geq 0, \mathbf{w}, b} \max_{\alpha} \sum_{i=1}^N \alpha_i (1 - y^{(i)} f(\mathbf{x}^{(i)})) + \frac{1}{2} \sum_{l=1}^L \sum_{m \in \mathcal{G}_l} \frac{\|\mathbf{w}_m\|_2^2}{\theta_m} \quad (15) \\ & \text{s.t. } \left\| \boldsymbol{\theta} \right\|_{1, p; \beta, \gamma} \leq 1, \\ & 0 \leq \alpha_i \leq C', \quad i = 1, \dots, N. \end{aligned}$$

where $f(\mathbf{x}^{(i)}) = \sum_{l=1}^L \sum_{m \in \mathcal{G}_l} \mathbf{w}_m^T \boldsymbol{\phi}_m(\mathbf{x}_m^{(i)}) + b$. It is important to note that the minimization and maximization in Eq. (15) are switchable, as the problem is convex in \mathbf{w} and b and concave in α .

For fixed $\boldsymbol{\theta}$ and α , by taking minimization w.r.t \mathbf{w}_m , we have $\mathbf{w}_m = \theta_m \sum_{i=1}^N \alpha_i y^{(i)} \boldsymbol{\phi}_m(\mathbf{x}_m^{(i)})$; by taking minimization w.r.t b , we have $\sum_{i=1}^N \alpha_i y^{(i)} = 0$. Then, plugging back \mathbf{w} , we get the dual form in Eq. (13).

Biography

Jialin Peng is an associate professor in the College of Computer Science and Technology at Huaqiao University in China. He received his Ph.D. in Mathematics from Zhejiang University in 2013. Dr. Peng's research aims to explore intelligent approaches to bridge the data and medical informatics via machine learning. He received the best scientific paper award from the 21st International Conference on Pattern Recognition.

Xiaofeng Zhu received M.Sc. degree (by research in computer science) and Ph.D. degree (in computer science), respectively, from National University of Singapore (NUS), Singapore and The University of Queensland (UQ), Australia. His current research interests include large-scale multimedia retrieval, feature selection, sparse learning, data preprocess, and medical image analysis. He has published in top-tier journals and conferences.

Ye Wang received the M.Sc. degree in mathematical statistics from Zhejiang University in China in 2013. Her current research interest is statistical learning and its applications.

Le An received the B.Eng. degree in telecommunications engineering from Zhejiang University in China in 2006, the M.Sc. degree in electrical engineering from Eindhoven University of Technology in Netherlands in 2008, and the Ph.D. degree in electrical engineering from University of California, Riverside in USA in 2014. His research interests include image processing, computer vision, pattern recognition, and machine learning. He received the best paper award from the 2013 IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS).

Dinggong Shen is a professor of Radiology, Biomedical Research Imaging Center (BRIC), Computer Science, and Biomedical Engineering in the University of North Carolina at Chapel Hill (UNC-CH). He is currently directing the Center for Image Analysis and Informatics, the Image Display, Enhancement, and Analysis (IDEA) Lab in the Department of Radiology, and also the medical image analysis core in the BRIC. He was a tenure-track assistant professor in the University of Pennsylvania (UPenn), and a faculty member in the Johns Hopkins University. Dr. Shen's research interests include medical image analysis, computer vision, and pattern recognition. He has published more than 700 papers in the international journals and conference proceedings. He serves as an editorial board member for six international journals. He also served in the Board of Directors, The Medical Image Computing and Computer Assisted Intervention (MICCAI) Society, in 2012–2015.

References

- [1]. Saykin AJ, Shen L, Foroud TM, Potkin SG, Swaminathan S, Kim S, Risacher SL, Nho K, Huentelman MJ, Craig DW, Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans, *Alzheimer's Dementia* 6 (3) (2010) 265–273.
- [2]. Morris JC, Storandt M, Miller JP, McKeel DW, Price JL, Rubin EH, Berg L, Mild cognitive impairment represents early-stage Alzheimer disease, *Arch. Neurol* 58 (3) (2001) 397–405. [PubMed: 11255443]
- [3]. Guo X, Wang Z, Li K, Li Z, Qi Z, Jin Z, Yao L, Chen K, Voxel-based assessment of gray and white matter volumes in Alzheimer's disease, *Neurosci. Lett* 468 (2) (2010) 146–150. [PubMed: 19879920]
- [4]. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M-O, Chupin M, Benali H, Colliot O, Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the adni database, *Neuroimage* 56 (2) (2011) 766–781. [PubMed: 20542124]
- [5]. Wang H, Nie F, Huang H, Risacher S, Saykin AJ, Shen L, Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression, in: *Medical Image Computing and Computer-Assisted Intervention*, Springer, 2011, pp. 115–123.
- [6]. Zhang D, Shen D, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, *Neuroimage* 59 (2) (2012) 895–907. [PubMed: 21992749]
- [7]. Suk H-I, Lee S-W, Shen D, Hierarchical feature representation and multi-modal fusion with deep learning for AD, MCI diagnosis, *Neuroimage* 101 (2014) 569–582. [PubMed: 25042445]
- [8]. Tong T, Wolz R, Gao Q, Guerrero R, Hajnal JV, Rueckert D, Multiple instance learning for classification of dementia in brain mri, *Med. Image Anal* 18 (5) (2014) 808–818. [PubMed: 24858570]
- [9]. Zhu X, Suk H-I, Lee S-W, Shen D, Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis, *Brain Imaging Behav.* 10 (3) (2015) 818–828.
- [10]. Zhang D, Wang Y, Zhou L, Yuan H, Shen D, Multimodal classification of Alzheimer's disease and mild cognitive impairment, *Neuroimage* 55 (3) (2011) 856–867. [PubMed: 21236349]

- [11]. Zhu X, Suk H-I, Lee S-W, Shen D, Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification, *IEEE Trans. Biomed. Eng* 63 (3) (2016) 607–618. [PubMed: 26276982]
- [12]. Westman E, Muehlboeck J-S, Simmons A, Combining MRI and CSF measures for classification of Alzheimer’s disease and prediction of mild cognitive impairment conversion, *Neuroimage* 62 (1) (2012) 229–238. [PubMed: 22580170]
- [13]. Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, Green RC, Hu X, Huentelman MJ, Kim S, Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers, *Brain Imaging Behav.* 8 (2) (2014) 183–207. [PubMed: 24092460]
- [14]. Wang H, Nie F, Huang H, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L, Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort, *Bioinformatics* 28 (2) (2012) 229–237. [PubMed: 22155867]
- [15]. Tibshirani R, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc* 58 (1) (1996) 267–288.
- [16]. Yuan M, Lin Y, Model selection and estimation in regression with grouped variables, *J. R. Stat. Soc* 68 (1) (2006) 49–67.
- [17]. Friedman J, Hastie T, Tibshirani R, *The elements of statistical learning*, 1, Springer series in statistics Springer, Berlin, 2001.
- [18]. Jenatton R, Audibert J-Y, Bach F, Structured variable selection with sparsity-inducing norms, *J. Mach. Learn. Res* 12 (2011) 2777–2824.
- [19]. Cao P, Shan X, Zhao D, Huang M, Zaiane O, Sparse shared structure based multi-task learning for mri based cognitive performance prediction of alzheimers disease, *Pattern Recognit.* 72 (2017) 219–235.
- [20]. Cao P, Liu X, Yang J, Zhao D, Huang M, Zaiane O, L2,1-l1 regularized nonlinear multi-task representation learning based cognitive performance prediction of Alzheimer’s disease, *Pattern Recognit.* 79 (2018) 195–215.
- [21]. Guerrero R, Ledig C, Schmidt-Richberg A, Rueckert D, Group-constrained manifold learning: application to ad risk assessment, *Pattern Recognit.* 63 (2017) 570–582.
- [22]. Hinrichs C, Singh V, Xu G, Johnson S, MKL for robust multi-modality AD classification, in: *Medical Image Computing and Computer-Assisted Intervention*, Springer, 2009, pp. 786–794.
- [23]. Liu F, Zhou L, Shen C, Yin J, Multiple kernel learning in the primal for multi-modal Alzheimer’s disease classification, *IEEE J. Biomed. Health Inform* 18 (3) (2014) 984–990. [PubMed: 24132030]
- [24]. Peng J, An L, Zhu X, Jin Y, Shen D, Structured sparse kernel learning for imaging genetics based alzheimers disease diagnosis, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, pp. 70–78.
- [25]. Kowalski M, Sparse regression using mixed norms, *Appl. Comput. Harmon. Anal* 27 (3) (2009) 303–324.
- [26]. Zhou Y, Jin R, Hoi SC, Exclusive lasso for multi-task feature selection, in: *AISTATS*, 9, 2010, pp. 988–995.
- [27]. Kong D, Fujimaki R, Liu J, Nie F, Ding C, Exclusive feature learning on arbitrary structures via L12-norm, in: *Advances in Neural Information Processing Systems*, 2014, pp. 1655–1663.
- [28]. Liu F, Wee C-Y, Chen H, Shen D, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer’s disease and mild cognitive impairment identification, *Neuroimage* 84 (2014) 466–475. [PubMed: 24045077]
- [29]. Kloft M, Brefeld U, Sonnenburg S, Zien A, Lp-norm multiple kernel learning, *J. Mach. Learn. Res* 12 (2011) 953–997.
- [30]. Zhou J, Liu J, Narayan VA, Ye J, Modeling disease progression via multi-task learning, *Neuroimage* 78 (2013) 233–248. [PubMed: 23583359]
- [31]. Zhang Z, Huang H, Shen D, Integrative analysis of multi-dimensional imaging genomics data for Alzheimer’s disease prediction., *Front. Aging Neurosci* 6 (2013). 260–260
- [32]. Xiang S, Yuan L, Fan W, Wang Y, Thompson PM, Ye J, Bi-level multi-source learning for heterogeneous block-wise missing data, *Neuroimage* 102 (2014) 192–206. [PubMed: 23988272]

- [33]. Walhovd K, Fjell A, Brewer J, McEvoy L, Fennema-Notestine C, Hagler D, Jennings R, Karow D, Dale A, Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer's disease, *Am. J. Neuroradiology* 31 (2) (2010) 347–354.
- [34]. Kohannim O, Hua X, Hibar DP, Lee S, Chou Y-Y, Toga AW, Jack CR, Weiner MW, Thompson PM, Boosting power for clinical trials using classifiers based on multiple biomarkers, *Neurobiol. Aging* 31 (8) (2010) 1429–1442. [PubMed: 20541286]
- [35]. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, Feng D, Fulham MJ, Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, *IEEE Trans. Biomed. Eng* 62 (4) (2015) 1132–1140. [PubMed: 25423647]
- [36]. Hinrichs C, Singh V, Xu G, Johnson SC, Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population, *Neuroimage* 55 (2) (2011) 574–589. [PubMed: 21146621]
- [37]. Zu C, Jie B, Liu M, Chen S, Shen D, Zhang D, Label-aligned multi-task feature learning for multimodal classification of alzheimer's disease and mild cognitive impairment, *Brain Imaging Behav.* 10 (4) (2016) 1148–1159. [PubMed: 26572145]
- [38]. Jie B, Zhang D, Cheng B, Shen D, Manifold regularized multitask feature learning for multimodality disease classification, *Hum. Brain Mapp* 36 (2) (2015) 489–507. [PubMed: 25277605]
- [39]. Ye T, Zu C, Jie B, Shen D, Zhang D, Initiative ADN, et al., Discriminative multi-task feature selection for multi-modality classification of alzheimer disease, *Brain Imaging Behav.* 10 (3) (2016) 739–749. [PubMed: 26311394]
- [40]. Ahmed OB, Benois-Pineau J, Allard M, Catheline G, Amar C. Ben, Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning, *Neurocomputing* 220 (2017) 98–110.
- [41]. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, *Neuroimage* 65 (2013) 167–175. [PubMed: 23041336]
- [42]. Tong T, Gray K, Gao Q, Chen L, Rueckert D, Initiative ADN, et al., Multi-modal classification of Alzheimer's disease using nonlinear graph fusion, *Pattern Recognit.* 63 (2017) 171–181.
- [43]. Duda R, Hart P, Stork D, *Pattern classification*, John Wiley & Sons, 2012.
- [44]. Lanckriet GR, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI, Learning the kernel matrix with semidefinite programming, *J. Mach. Learn. Res* 5 (2004) 27–72.
- [45]. Bach FR, Lanckriet GR, Jordan MI, Multiple kernel learning, conic duality, and the SMO algorithm, in: the 21th International Conference on Machine learning, ACM, 2004, pp. 1–8.
- [46]. Rakotomamonjy A, Bach F, Canu S, Simple MKL, *J. Mach. Learn. Res* 9 (2008) 2491–2521.
- [47]. Szafranski M, Grandvalet Y, Rakotomamonjy A, Composite kernel learning, *Mach. Learn* 79 (1–2) (2010) 73–103.
- [48]. Yeh Y-R, Lin T-C, Chung Y-Y, Wang Y-CF, A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection, *IEEE Trans. Multimedia* 14 (3) (2012) 563–574.
- [49]. Zhao P, Rocha G, Yu B, The composite absolute penalties family for grouped and hierarchical variable selection, *Ann. Stat* (2009) 3468–3497.
- [50]. Chen G, Chen Q, Zhang D, Mixed norm regularized discrimination for image steganalysis, *Sens. Imaging* 16 (1) (2015) 17.
- [51]. Rakotomamonjy A, Flamary R, Gasso G, Canu S, Lp-Lq penalty for sparse linear and sparse multiple kernel multitask learning, *IEEE Trans. Neural Netw* 22 (8) (2011) 1307–1320. [PubMed: 21813358]
- [52]. Zhu X, Suk H-I, Wang L, Lee S-W, Shen D, Alzheimer's Disease Neuroimaging Initiative, A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, *Med. Image Anal* 38 (2017) 205–214. [PubMed: 26674971]
- [53]. Szafranski M, Grandvalet Y, Morizet-Mahoudeaux P, Hierarchical penalization, *Advances in Neural Information Processing Systems* 21, 2007 no. 1–8.
- [54]. Nocedal J, Wright S, *Numerical optimization*, Springer Science & Business Media, 2006.

- [55]. Scholkopf B, Smola AJ, Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press, 2001.
- [56]. Sled JG, Zijdenbos AP, Evans AC, A nonparametric method for automatic correction of intensity nonuniformity in mri data, *IEEE Trans. Med. Imaging* 17 (1) (1998) 87–97. [PubMed: 9617910]
- [57]. Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM, Magnetic resonance image tissue classification using a partial volume model, *Neuroimage* 13 (5) (2001) 856–876. [PubMed: 11304082]
- [58]. Smith SM, Fast robust automated brain extraction, *Hum. Brain Mapp* 17 (3) (2002) 143–155. [PubMed: 12391568]
- [59]. Zhang Y, Brady M, Smith S, Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm, *IEEE Trans. Med. Imaging* 20 (1) (2001) 45–57. [PubMed: 11293691]
- [60]. Shen D, Davatzikos C, Hammer: hierarchical attribute matching mechanism for elastic registration, *IEEE Trans. Med. Imaging* 21 (11) (2002) 1421–1439. [PubMed: 12575879]
- [61]. Kabani NJ, 3d anatomical atlas of the human brain, in: 20th Annual Meeting of the Organization for Human Brain Mapping, 7, 1998, pp. P-0717.
- [62]. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR, Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes, *Genet. Epidemiol* 34 (8) (2010) 816–834. [PubMed: 21058334]
- [63]. Peck R, Devore JL, Statistics: The exploration & analysis of data, Cengage Learning, 2011.
- [64]. Convit A, De Asis J, De Leon M, Tarshish C, De Santi S, Rusinek H, Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer’s disease, *Neurobiol. Aging* 21 (1) (2000) 19–26. [PubMed: 10794844]
- [65]. Hua X, Leow AD, Lee S, Klunder AD, Toga AW, Lepore N, Chou Y-Y, Brun C, Chiang M-C, Barysheva M, 3D characterization of brain atrophy in Alzheimer’s disease and mild cognitive impairment using tensor-based morphometry, *Neuroimage* 41 (1) (2008) 19–34. [PubMed: 18378167]
- [66]. Chiappelli M, Borroni B, Archetti S, Calabrese E, Corsi MM, Franceschi M, Padovani A, Licastro F, VEGF gene and phenotype relation with Alzheimer’s disease and mild cognitive impairment, *Rejuvenation Res.* 9 (4) (2006) 485–493. [PubMed: 17105389]
- [67]. Liu F, Suk H-I, Wee C-Y, Chen H, Shen D, High-order graph matching based feature selection for Alzheimer’s disease identification, in: *Medical Image Computing and Computer-Assisted Intervention*, Springer, 2013, pp. 311–318.
- [68]. Wang H, Nie F, Huang H, Ding C, Heterogeneous visual features fusion via sparse multimodal machine, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3097–3102.
- [69]. Kannao R, Guha P, Success based locally weighted multiple kernel combination, *Pattern Recognit.* 68 (2017) 38–51.

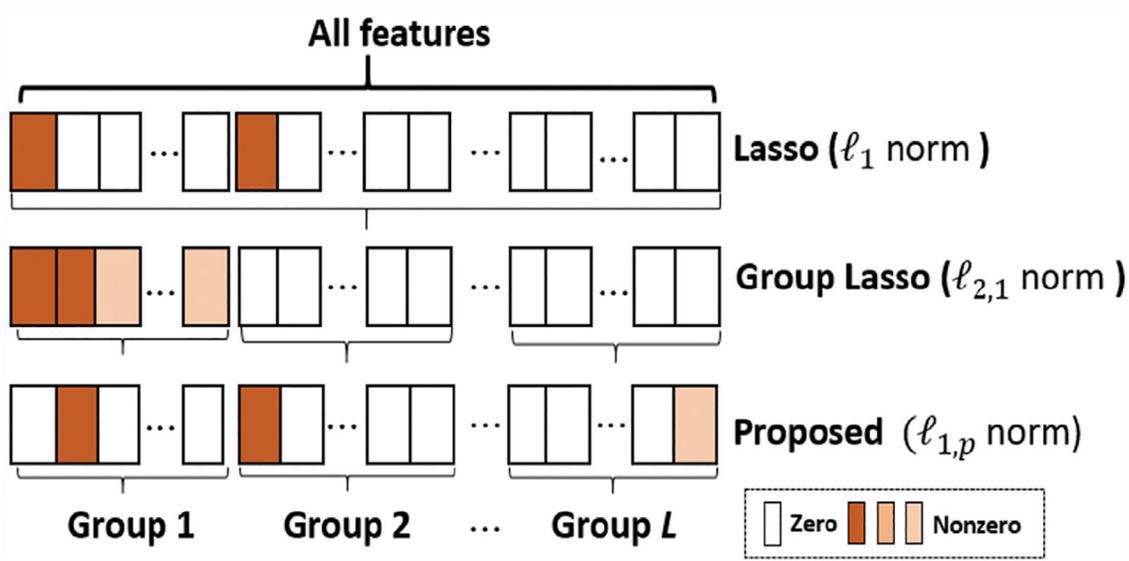


Fig. 1. Illustration of sparsity patterns by different regularizers: a) Lasso (ℓ_1 norm) [15] generates sparse solution but neglects inherent group structures; b) Group Lasso ($\ell_{2,1}$ norm) [16] sparsely selects a few groups of variables with predefined group structure; c) structured sparsity with $\ell_{1,p}$ norm ($p > 1$) in contrast keeps all groups but conducts within-group variable selection. This structured sparsity regularizer is particularly valuable for AD diagnosis, where features are naturally grouped by modalities and each modality is useful. Each box denotes a feature, where box with darker color indicates the feature selected with larger weight, and white box denotes the unselected features.

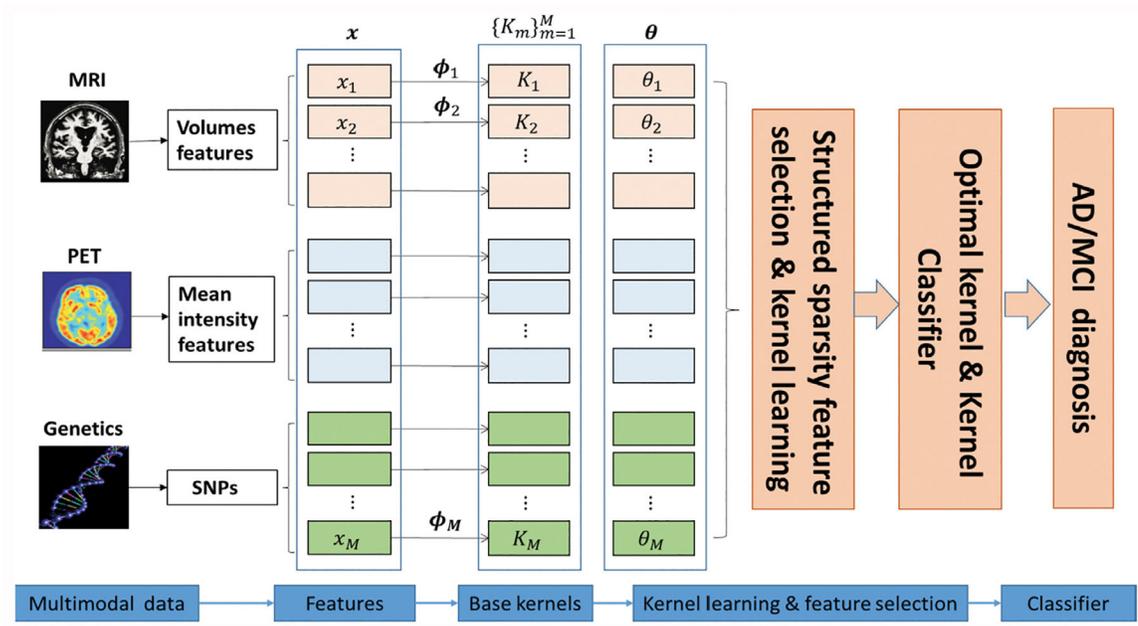


Fig. 2. Schematic illustration of our proposed framework. After representing each feature with a distinct basis kernel, a data-driven kernel representation and an optimal discriminant function are learnt. With a novel structured sparsity regularizer, the finally learnt kernel is a weighted linear combination of the basis kernels, and thus features with zero kernel weights would not be active in the discriminant classifier.

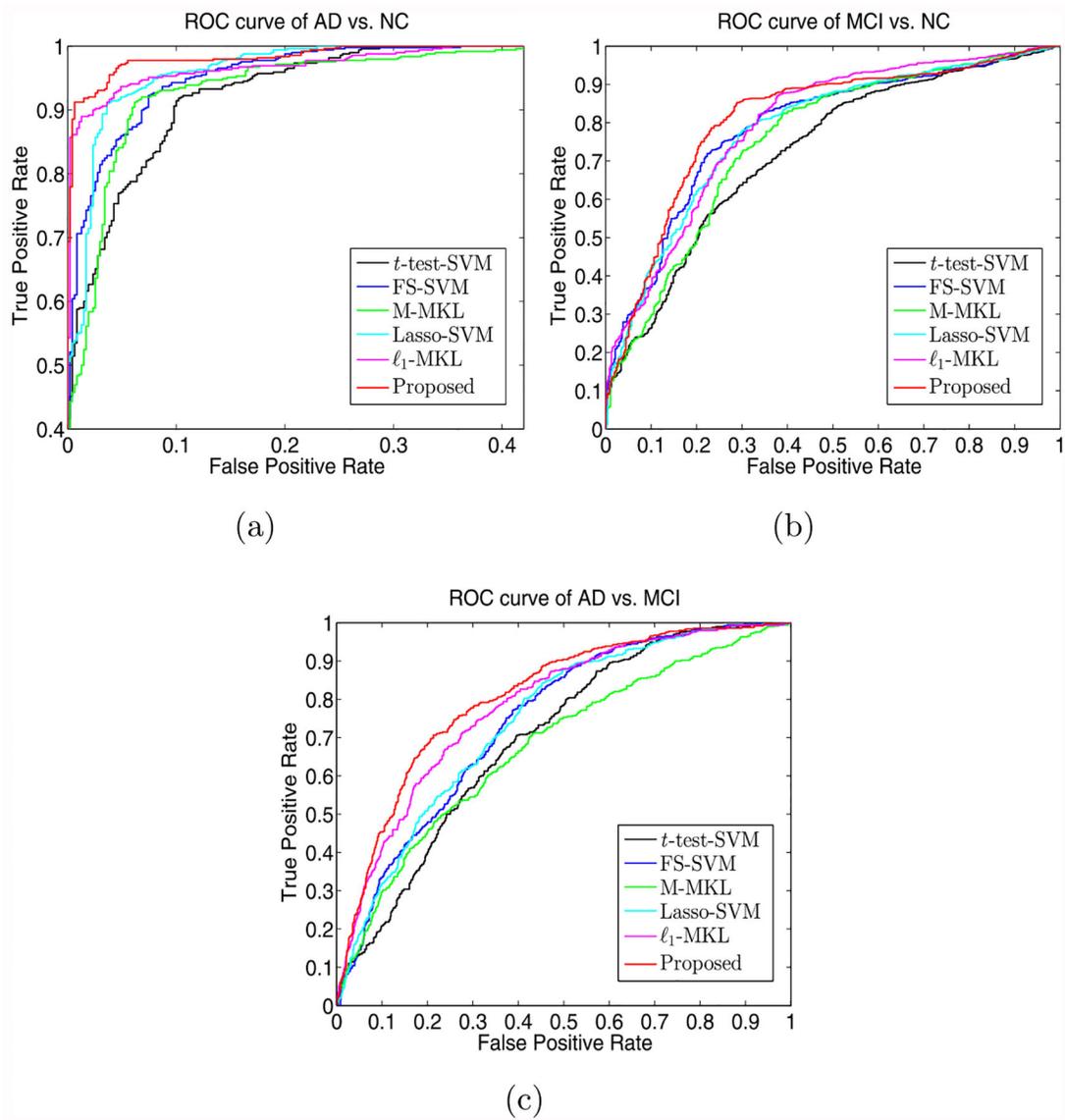


Fig. 3. ROC curves on Dataset I for the methods under comparison, i.e., t -test-SVM, FS-SVM, M-MKL [10], Lasso-SVM, ℓ_1 -MKL [46] and the Proposed.

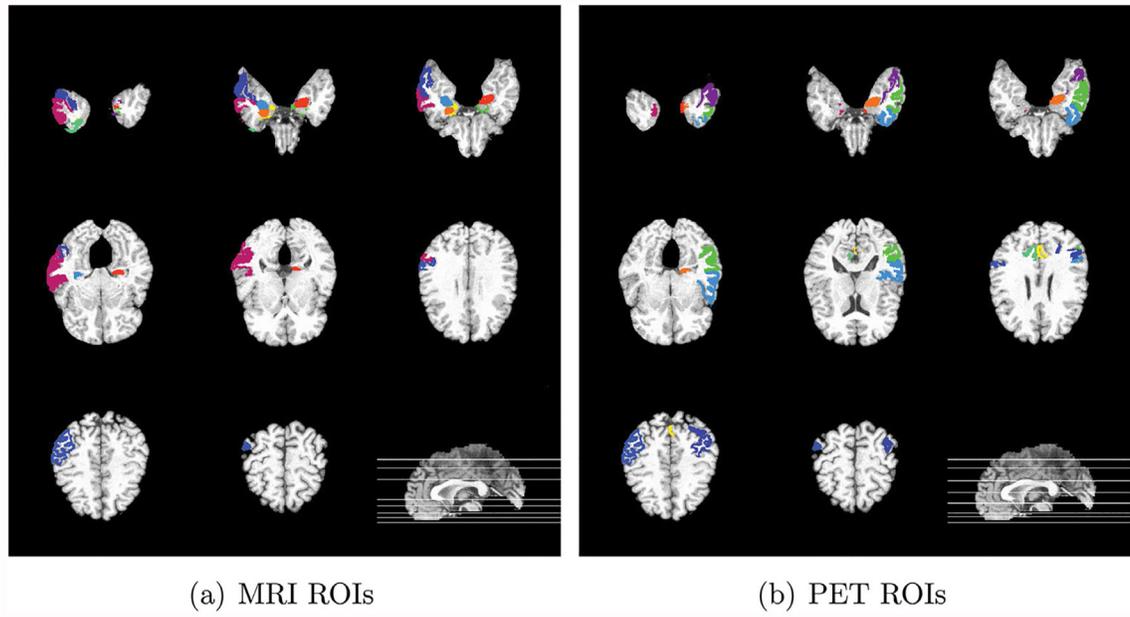


Fig. 4.
Top 10 ROIs detected for AD vs. NC classification in MRI and PET, respectively.

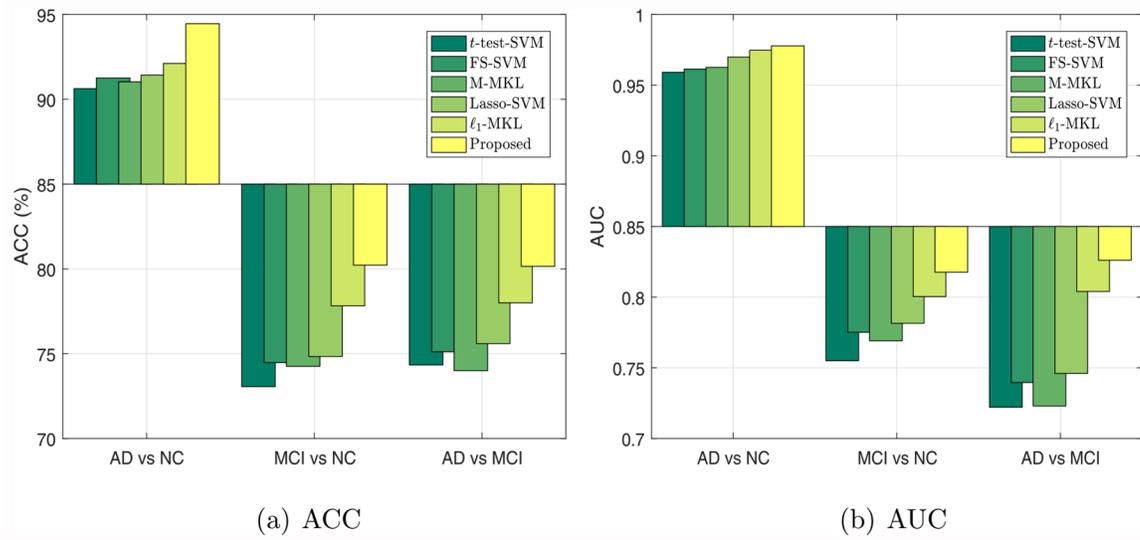


Fig. 5. Comparison of classification results on Dataset II.

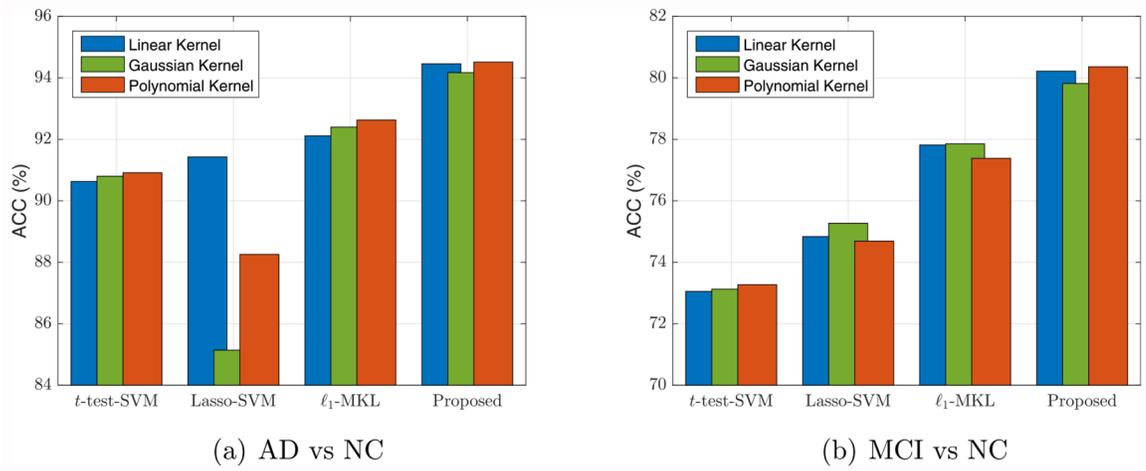


Fig. 6. Classification results on Dataset II with different types of kernels. Linear kernel, Gaussian kernel, and Polynomial kernel are tested.

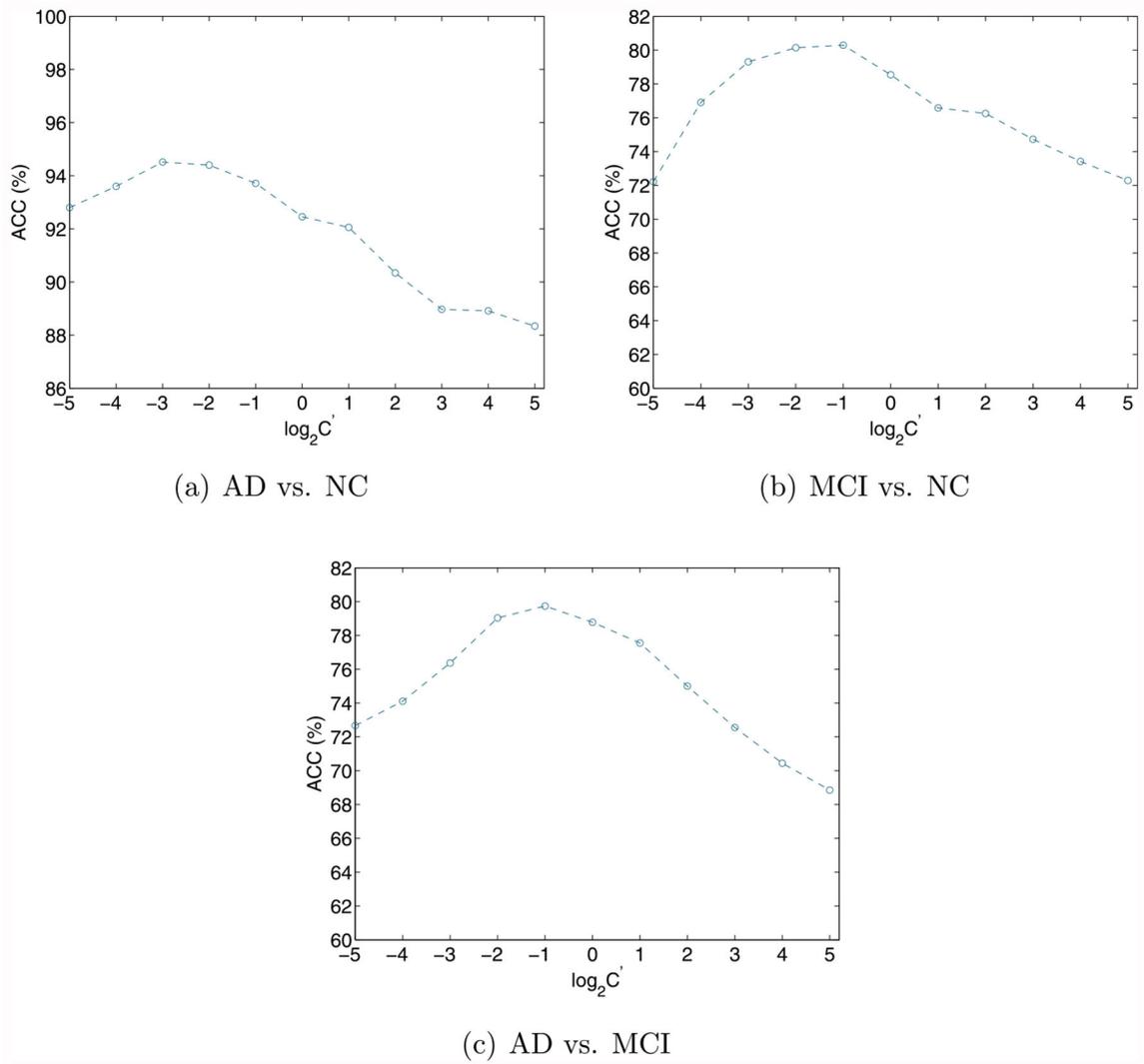


Fig. 7. Performance changes on Dataset II with the changes of the trade-off weight C' .

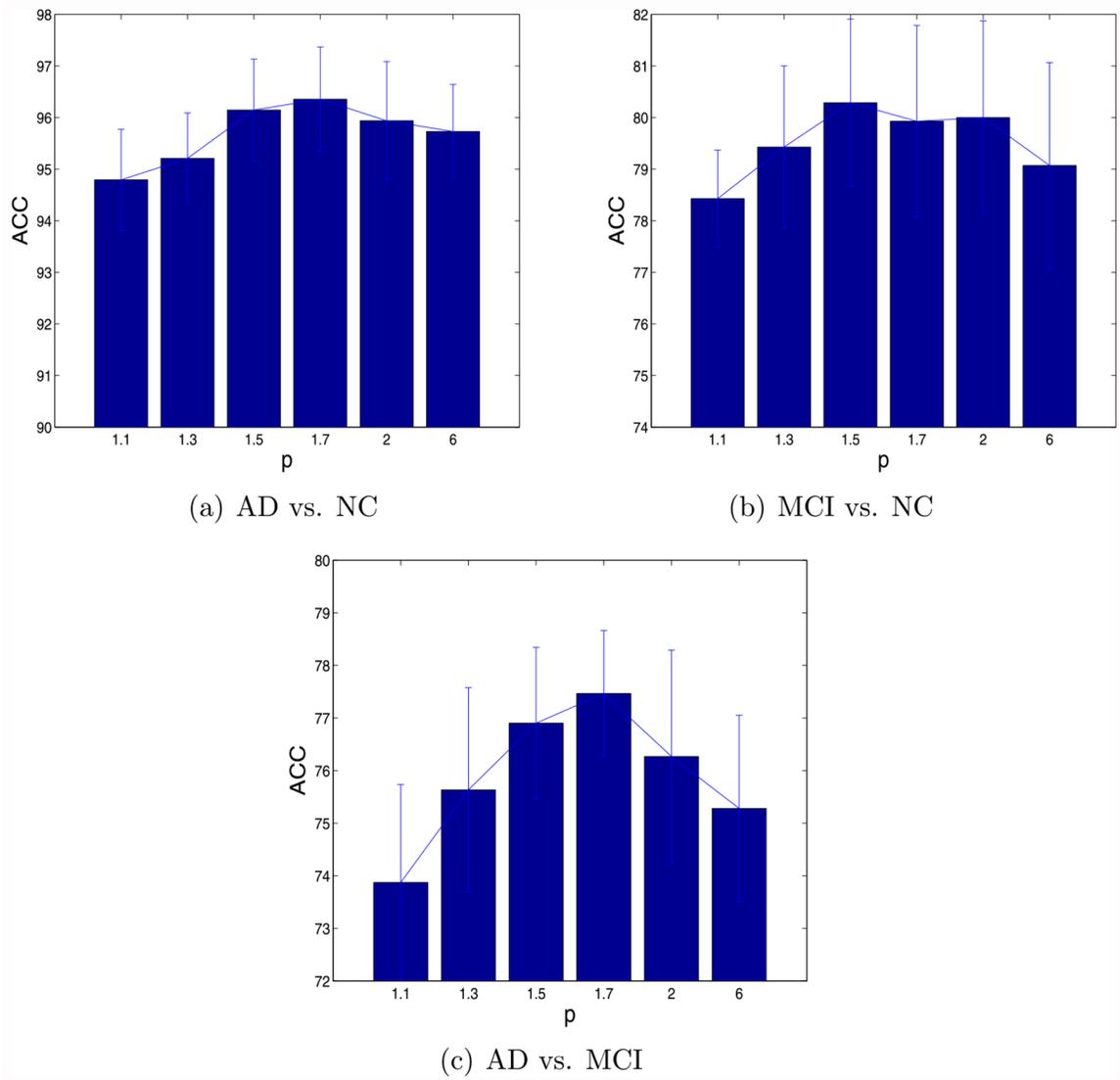


Fig. 8. The effect of different p in the proposed weighted $l_{1,p}$ norm constraint in terms of ACC (mean and standard deviation).

Algorithm 1

Block Coordinate Descent Algorithm for the Proposed Model.

-
- 1: **initial input:** $C', \{K_m\}, \beta, \gamma$, feasible θ , such as $\theta_m = \left(\sum_{l=1}^L \gamma_l \left(\sum_{m' \in \mathcal{G}_l} \beta_{m'} \right)^p \right)^{-\frac{1}{p}}$
- 2: **while** optimality condition is not satisfied **do**
- 3: Compute α in Eq. (12) and b using SVM solver [54]
- 4: Compute $\|w_m\|_2$ for all $m = 1, \dots, M$ according to Eq. (13)
- 5: Update θ_m for all $m = 1, \dots, M$ according to Eq. (4), i.e.,

$$\theta_m = \frac{\|w_m\|_2}{\beta_m^{\frac{1}{2}} \gamma_l^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1; \beta}^{\frac{p-1}{p+1}}} \cdot \frac{1}{\left(\sum_{l=1}^L \gamma_l^{\frac{1}{p+1}} \left\| \mathbf{W}_{\mathcal{G}_l} \right\|_{1; \beta}^{\frac{2p}{p+1}} \right)^{\frac{1}{p}}}$$

6: **end while**

Table 1

Prediction and feature selection results on synthetic data. The indexes of true features are {1, 32, 46, 62, 93} from the 5 groups. Top 5 selected features with their group indexes are listed. With 10 times 10 fold cross-validation, only features selected over 20 times are illustrated. The symbol # means all features of the whole group are selected.

Methods	ACC (%)	Top selected features (Group index)				
Lasso	79.5	32(2)	62(4)	61(4)	26(2)	-
Group Lasso	74.7	#(1)	#(2)	#(3)	-	-
l_1 -MKL	80.6	32(2)	62(4)	1(1)	61(4)	13(1)
$l_{1,5}$ -MKL	84.3	32(2)	62(4)	1(1)	46(3)	85(5)
$l_{1,2}$ -MKL	83.5	32(2)	62(4)	1(1)	46(3)	85(5)
$l_{1,6}$ -MKL	81.5	32(2)	62(4)	85(5)	1(1)	46(3)

Demographic information (Mean \pm SD) of the used subjects. (MMSE: Mini-Mental State Examination; SD: standard deviation).

Table 2

Attributes	Dataset I (189 subjects)			Dataset II (360 subjects)		
	AD	NC	MCI	AD	NC	MCI
Age	75.4 \pm 7.2	75.2 \pm 5.2	75.2 \pm 7.1	75.6 \pm 7.4	75.8 \pm 4.8	75.5 \pm 6.9
Education	14.7 \pm 3.7	15.9 \pm 3.1	16.0 \pm 3.0	14.9 \pm 3.0	16.0 \pm 2.9	15.7 \pm 2.9
MMSE	23.8 \pm 1.9	29.1 \pm 1.1	27.2 \pm 1.6	23.6 \pm 2.1	29.0 \pm 1.1	27.3 \pm 1.7

Table 3

Performance comparison of different methods on Data I for AD vs. NC, MCI vs. NC, and AD vs. MCI classifications, respectively. The asterisk indicates statistically significant difference ($p < 0.05$) compared with our method. The best result of each column is denoted in bold face.

Methods	AD vs. NC				MCI vs. NC				AD vs. MCI			
	ACC (%)	SEN (%)	SPE (%)	AUC	ACC (%)	SEN (%)	SPE (%)	AUC	ACC (%)	SEN (%)	SPE (%)	AUC
t-test-SVM	91.9*	92.7	91.1	0.965	72.7*	85.4	47.7	0.725	66.3*	43.1	78.5	0.700
FS-SVM	92.4*	93.5	91.3	0.979	76.1*	84.3	59.8	0.789	70.1*	40.6	85.7	0.740
M-MKL	92.6*	92.7	92.6	0.968	75.1*	82.8	59.8	0.756	67.4*	50.6	76.2	0.679
Lasso-SVM	93.5*	94.9	92.1	0.980	76.3*	85.2	58.7	0.783	70.3*	49.4	81.3	0.745
ℓ-MKL	94.0*	94.3	93.6	0.986	77.9*	85.7	62.6	0.795	73.0*	61.4	79.0	0.781
Proposed	96.1	97.3	94.9	0.992	80.3	85.6	69.8	0.811	76.9	65.9	82.7	0.808

Comparison of our proposed method on Dataset II in the cases of using different modality combinations. The asterisk indicates statistically significant difference ($p < 0.05$) compared with the performance based on all the three modalities, i.e., MRI, PET, and SNPs. The best result of each column is denoted in bold face.

Table 4

Modalities	AD vs. NC (%)			CI vs. NC (%)			AD vs. MCI (%)		
	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
MRI	88.4*	84.1	93.0	71.6*	83.9	47.2	65.4*	41.2	78.1
PET	86.3*	84.5	88.1	68.8*	85.5	35.7	65.1*	36.7	80.0
SNP	76.0*	69.8	82.6	66.2*	75.4	48.1	63.4*	25.9	83.1
MRI+PET	92.3*	91.9	91.7	76.4*	83.9	61.5	73.6*	55.1	83.3
MRI+SNP	91.1*	89.8	92.6	74.9*	84.5	55.7	70.0*	54.3	78.3
PET+SNP	92.0*	90.8	93.2	71.3*	81.4	51.3	68.7*	43.5	82.0
MRI+PET+SNP	96.1	97.3	94.9	80.3	85.6	69.8	76.9	65.9	82.7