Information Fusion 66 (2021) 170–183

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion

Yuanpeng Zhang [a,1], Shuihua Wang [b,1], Kaijian Xia [c,1], Yizhang Jiang [d,1], Pengjiang Qian [d,*], For the Alzheimer's Disease Neuroimaging Initiative

[a] Department of Medical Informatics, Nantong University, Nantong, 226001, Jiangsu, PR China
[b] Department of Cardiovascular Science, University of Leicester, Leicester, United Kingdom
[c] Affiliated Changshu Hospital of Soochow University, Changshu, Jiangsu 215500, PR China
[d] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, Jiangsu, PR China

ARTICLE INFO

Dataset link: http://adni.loni.usc.edu/data-samples/access-data/

Keywords:
Alzheimer's disease
Multimodal neuroimaging
Multiple kernel learning
Feature selection
Neuroimaging biomarker
Multiclass classification
Multimodal fusion

ABSTRACT

Alzheimer's disease (AD) will become a global burden in the coming decades according to the latest statistical survey. How to effectively detect AD or MCI (mild cognitive impairment) using reliable biomarkers and robust machine learning methods has become a challenging problem. In this study, we propose a novel AD multiclass classification framework with embedding feature selection and fusion based on multimodal neuroimaging. The framework has three novel aspects: (1) An $l_{2,1}$-norm regularization term combined with the multiclass hinge loss is used to naturally select features across all the classes in each modality. (2) To fuse the complementary information contained in each modality, an l_p-norm ($1 < p < \infty$) regularization term is introduced to combine different kernels to perform multiple kernel learning to avoid a sparse kernel coefficient distribution, thereby effectively exploiting complementary modalities. (3) A theorem that transforms the multiclass hinge loss minimization problem using the $l_{2,1}$-norm and l_p-norm regularizations to a previous solvable optimization problem and its proof are given. Additionally, it is theoretically proved that the optimization process converges to the global optimum. Extensive comparison experiments and analysis support the promising performance of the proposed method.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease with an insidious onset. Clinically, it is characterized by comprehensive dementia such as memory disorder, aphasia, apraxia, agnosia, impairment of visuospatial skills, executive dysfunction, and personality and behavioral changes [1–3]. As the population ages, AD will become a global burden in the coming decades. It was reported that there were approximately 50 million AD cases around the world in 2015 with more than half of them being early cases. That number is expected to triple to 152 million by 2050. Due to the rapid increase in the prevalence of AD, the accurate diagnosis of AD and its early stage, known as mild cognitive impairment (MCI), becomes very crucial for the timely treatment and possible delay of AD [4]. In the past, the diagnosis of AD mainly relied on the evaluation of the patient's medical history, clinical observation or cognitive evaluation. In recent years, research related to AD has shown that finding reliable biomarkers for the automatic detection of AD or MCI is a promising and challenging task [5]. Many research projects looking for biomarkers have already started and achieved landmark results. For example, the Alzheimer's Disease Neuroimaging Initiative (ADNI) [6] is known for collecting different types of candidate biomarkers to further accelerate the development of AD research. Up to now, some biomarkers have been investigated and confirmed to be sensitive to MCI, e.g., protein changes in blood or spinal fluid, brain atrophy detected by imaging, etc. High-quality biomarkers make the diagnosis of MCI more accurate, which can delay and control the further conversion of MCI into AD.

Recent studies have shown that neuroimage analysis is more reliable and sensitive than traditional cognitive assessment in detecting the presence of early AD [7,8]. Thus, many machine learning methods have been developed for automatic neuroimage analysis and further AD classification including traditional methods and deep learning methods. For example, Ahmed et al. [9] used a multikernel-based classifier to fuse complementary information from diffusion tensor imaging (DTI)

* Corresponding author.
E-mail address: qianpjiang@jiangnan.edu.cn (P. Qian).
[1] All authors contributed equally to this study.

https://doi.org/10.1016/j.inffus.2020.09.002
Received 16 May 2020; Received in revised form 15 August 2020; Accepted 12 September 2020
Available online 16 September 2020

and structural magnetic resonance imaging (sMRI) for AD classification. Peng et al. [10] proposed a structured sparse kernel learning method for AD classification by combining MRI, positron emission tomography (PET) and single-nucleotide polymorphism (SNP) features. Additionally, there have also been some representative deep learning-based methods for AD recognition [11]. For example, An et al. [12] proposed a novel method for AD classification based on deep ensemble learning. The core idea was that all base classifiers were taken as surrogates to physicians with different clinical knowledge and their predictions were combined by a deep belief network (DBN) in a stacked way. Suk et al. [13] also proposed a deep ensemble classifier for AD classification. The main difference from An's method was that they used sparse regression models to select different feature subsets. Wang et al. [14] proposed a hierarchical ensemble learning method for AD classification. This was a coarse-to-fine method. First, an MRI image was divided into multiple slices and a multiple pretraining deep neural network was employed to extract the features from the slices. Coarse predictions were obtained based on these features. Then, ensemble learning was conducted using the coarse predictions to generate the refined results for all slices. More studies regarding deep learning methods for AD classification can be found in [15] in which Jo et al. made a comprehensive review. In this study, we focus on traditional machine learning methods. The top-level layout of a generalized traditional machine learning framework for AD classification based on neuroimages is illustrated in Fig. 1. We see that the framework contains several compulsory components, e.g., preprocessing, feature extraction, feature selection (feature reduction) and classification; and one optional component, i.e., feature fusion, which becomes compulsory when combining multimodal neuroimages. Next, we focus on the two components of feature selection and feature fusion and conduct a brief review and analysis of the deficiencies of previous studies as our research motivation.

Regarding feature selection, it aims to select meaningful features and reduce the redundancies from the original feature sets extracted by neuroimaging preprocessing. For instance, Salvatore et al. [16] introduced an unsupervised feature selection method called principal component analysis (PCA) to reduce the redundancies from white matter (WM) and gray matter (GM) density maps. Then, the remaining meaningful features were used for support vector machine (SVM) training. Liu et al. [17] also introduced an unsupervised method called local linear embedding (LLE) to transform multivariate regional brain volume and cortical thickness MRI to a locally low-dimensional linear space while also making use of the global nonlinear data structure. Then, the embedded brain features in the low-dimensional space were employed for classification model training. Beheshti et al. [18] proposed a novel supervised feature selection method. First, feature extraction was performed by using the voxel clusters detected by the voxel-based morphometric (VBM) on sMRI and the voxel values as the volume of interest (VOI). Then, the probability distribution function of the VOI was utilized to represent the statistical patterns of the respective high-dimensional structural MRI sample. Finally, the selected features were used to train SVM classifiers for AD classification tasks. The above-mentioned feature selection (reduction) methods all belong to the filter category, which are independent of classifiers. Nir et al. [19] used DTI-based features and proposed a tractography-based method for AD and MCI classification. They first employed tractography and clustering to locate and partition fibers into 18 fiber bundles. Then, they calculated density maps to quantify the number of fibers passing through each voxel and used the shortest path graph search to reduce the fiber bundles based on maximum density path (MDP) such that the fiber bundles can be represented in a compact and low-dimensional representation. Finally, the diffusivity measures of fractional anisotropy (FA) and MD computed along all the registered across subjects (MDPs) were taken as the features for training an SVM-based classifier. Feature selection methods in this category can be characterized as making use of the global or local statistical information. However, although the features selected by filtering methods can train effective classification models, we do not think that they are straightforward since that they are not specific to classifiers. For example, Martino et al. [20] proposed a mechanism to classify functional MRI spatial patterns through multivariate feature selection. They used recursive feature elimination combined with an SVM (REF-SVM) to reduce the irrelevant voxels recursively. Analogously, Wee et al. [21] proposed a connectivity network measure-based method based on DTI images in which the clustering coefficients of all the anatomical regions, which were computed for all the networks, were taken as the features and the REF-SVM was employed to reduce the feature set for MCI classification. This feature selection strategy was tightly coupled with a specific classifier and had good performance. However, due to the continuous "selection-feedback", this feature selection method has expensive computational costs. Therefore, how to effectively integrate feature selection into a specific classifier in an embedded manner should be further studied.

Many previous works showed that several biomarkers had been demonstrated to be associated with AD patients [10,22]. These different biomarkers contain complementary modalities that can be combined to improve the understanding of the disease pattern over that presented by one modality. Therefore, feature fusion should be considered in the machine learning framework for AD classification. Straightforwardly, one can use each modality to train a classifier and then combine the results using voting or ensemble techniques. Dai et al. [23] used structural and functional MRI images to train MULDA classifiers and combined them as a fusion model using weighted voting to classify AD patients and healthy controls (HCs). Polikar et al. [24] constructed an ensemble classifier based on the multilayer perceptron to combine electroencephalogram (EEG), MRI and PET modalities for AD classification. Furthermore, direct feature concatenation is also often used exploit complementary modalities. Walhovd et al. [25] concatenated MRI, PET and cerebrospinal fluid (CSF) features and performed logistic regression analysis to diagnose and form prognoses for AD. Tang et al. [26] combined the volumetric, shape, and diffusion features of the hippocampus and amygdala and used PCA and the Student's t-test to perform feature selection. Then, the selected features were used to train LDA and SVM for AD classification tasks. Although using straightforward concatenation to exploit complementary modalities sometimes makes sense, it suffers from a major pitfall: straightforward concatenation means that multiple features are treated equally, which makes the complementary modalities become incapable of being effectively utilized.

In addition to the abovementioned fusion methods, another more reasonable fusion method is multiple kernel learning (MKL) [27–29]. Each kernel in MKL is embedded into a feature space of a modality; hence, MKL provides a natural framework for modality fusion. Zhang et al. [22] proposed a very simple MKL model based on the SVM in which kernels were constructed in the MRI, PET, and CSF feature spaces. However, the coefficients of the kernels were not learned automatically but rather were obtained through a grid search. Ahmed et al. [9] selected the simpleMKL [30] as the classification model to recognize AD, MCI and HCs based on local DTI and MRI. However, the simpleMKL can only handle binary classification tasks, thus the authors had to construct AD vs. NC, MCI vs. NC and AD vs. MCI binary classification problems, respectively. Kloft et al. [31] and Kowalski et al. [32] imposed the l1-norm regularization on kernel coefficients to achieve kernel fusion. However, the sparse coefficient distribution is less effective when utilizing complementary modalities.

To naturally integrate feature selection and fusion into a specific classifier for classifying AD, MCI and HCs simultaneously, in this study, taking the multiclass support vector machine (MCSVM) [33] as the basic classifier, we introduce a structured regularization term $l_{2,1}$-norm to yield the group sparsity for all the classes to naturally select features in each kernel space and impose a nonsparsity regularization l_p-norm ($1 < p < \infty$) on the kernel coefficients to perform complementary modality fusion. The main contributions are summarized as follows.

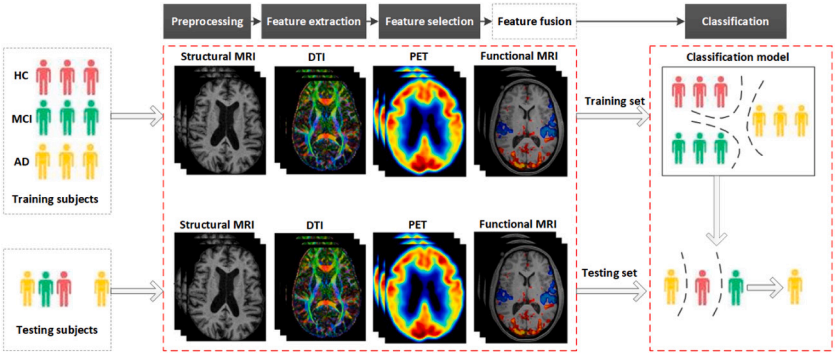


Fig. 1. Machine learning framework for AD classification based on neuroimages.

i Our proposed method uses an embedded feature section strategy to select discriminative features from different modalities. To be specific, an $l_{2,1}$-norm term combined with the hinge loss is used to yield the joint structural sparsity to select the features across all the classes in each modality.

ii For complementary modality fusion, an l_p-norm ($1 < p < \infty$) term is introduced to combine different kernels that correspond to different modalities to avoid a sparse kernel coefficient distribution, thereby effectively exploiting complementary modalities.

iii We give a theorem to transform the minimization problem of the multiclass hinge loss with $l_{2,1}$-norm and l_p-norm regularizations to a previous solvable optimization problem. Additionally, we theoretically demonstrate that our optimization process converges to a global optimum.

The following sections are organized as follows. In Section 2, the MCSVM is briefly introduced with the preliminaries for the proposed model. In Section 3, we present the objective and optimization of our model. Furthermore, the ADNI data preprocessing is also given. The experimental results are reported in Section 4. Section 5 concludes the whole study.

2. Preliminaries

Although the classic SVM model can handle multiclass classification tasks by using "one-against-one" or "one-against-all" strategies [34], it only transforms the multiclass problem into several binary ones so that the correlations between multiple classes are omitted. In this study, to naturally classify AD, MCI and HCs into different groups, we take the multiclass support vector machine (MCSVM) [33,35,36] as our basic classifier. Thus, in this section, we give a brief introduction to the MCSVM so that our proposed method can be easily understood in the next part. Suppose we have a training dataset $\{\mathbf{X}, \mathbf{Y}\}$, where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N, \mathbf{x}_i \in \mathbb{R}^D, \mathbf{Y} = \{y_i\}_{i=1}^N, y_i \in \{1, 2, \ldots, C\}$ and C is the number of classes. The MCSVM aims to learn C different decision functions with the projected vector and bias $\{\mathbf{w}_c, b_c\}$. In [37], an efficient method was proposed to train $\{\mathbf{w}_c, b_c\}$ by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w}_c, \xi_i} \frac{1}{2}\sum_{c=1}^{C} \mathbf{w}_c^T \mathbf{w}_c + \zeta \sum_{i=1}^{N}\sum_{c\neq y_i} \xi_{ic} \\ &\text{s.t. } \mathbf{w}_{y_i}^T \mathbf{x}_i + b_{y_i} \geq \mathbf{w}_c^T \mathbf{x}_i + b_c + 2 - \xi_{ic} \\ &\quad \xi_{ic} \geq 0, i = 1, 2, \ldots, N, c = 1, 2, \ldots, C \wedge c \neq y_i \end{aligned} \tag{1}$$

where $\mathbf{w}_c \in \mathbb{R}^{d\times 1}$ is the cth column of $\mathbf{W} \in \mathbb{R}^{D\times C}$. ξ is the slack variable and ζ is the slack regularized parameter. For an unseen sample, its class decision can be determined by

$$f(\mathbf{x}) = \arg\max_{1\leq c\leq C} \left(\mathbf{w}_c^T \mathbf{x} + b_c\right) \tag{2}$$

The decision function in (2) performs more efficiently and is powerful than that in the classic SVM using "one-against-all" or "one-against-one" strategies that ignore the correlations between multiple classes [37]. The solution to the problem in (1) can be obtained using dual variables by finding the saddle point of the following Lagrangian function $L(\mathbf{w}, \mathbf{b}, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta})$:

$$\begin{aligned} L(\mathbf{w}, \mathbf{b}, \xi, \mathbf{a}, \boldsymbol{\beta}) &= \frac{1}{2}\sum_{c=1}^{C} \mathbf{w}_c^T \mathbf{w}_c + \zeta \sum_{i=1}^{N}\sum_{c\neq y_i} \xi_{ic} - \sum_{i=1}^{N}\sum_{c=1}^{c} \beta_{ic}\xi_{ic} \\ &- \sum_{i=1}^{N}\sum_{c=1}^{c} \alpha_{ic}\left(\left(\mathbf{w}_{y_i} - \mathbf{w}_c\right)^T \mathbf{x}_i + b_{y_i} - b_c - 2 + \xi_{ic}\right) \end{aligned} \tag{3}$$

with the dummy variables

$$\alpha_{iy_i} = 0, \beta_{iy_i} = 0, \xi_{iy_i} = 2, i = 1, 2, \ldots, N \tag{4}$$

and constraints

$$\alpha_{ic} \geq 0, \beta_{ic} \geq 0, \xi_{ic} \geq 2, i = 1, 2, \ldots, N, c = 1, 2, \ldots, C \wedge c \neq y_i \tag{5}$$

which should be maximized w.r.t. the two Lagrangian multipliers $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and minimized w.r.t. $\mathbf{w}$ and ξ. Finally, we arrive at the following QP problem via several mathematical manipulations:

$$\begin{aligned} &\max_a 2\sum_{i=1}^{N}\sum_{c=1}^{C} \alpha_{ic} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{c=1}^{c}\left(-\frac{1}{2} m_{jy_i}\sum_{c=1}^{c}\alpha_{ic}\sum_{c=1}^{c}\alpha_{jc} + \alpha_{ic}\alpha_{jy_i} - \frac{1}{2}\alpha_{ic}\alpha_{jc}\right)\left(\mathbf{x}_i \cdot \mathbf{x}_j\right) \\ &\text{s.t. } \sum_{i=1}^{N}\alpha_{ic'} = \sum_{i=1}^{N} m_{ic'}\sum_{c=1}^{c}\alpha_{ic}, c' = 1, 2, \ldots, K \\ &0 \leq \alpha_{ic} \leq \zeta, \alpha_{ij_i} = 0, i = 1, 2, \ldots, N, c = 1, 2, \ldots, C \wedge c \neq y_i \end{aligned} \tag{6}$$

where $m_{ic'} = 1$ if $y_i = c'$ and $m_{ic'} = 0$ otherwise, and $\left(\mathbf{x}_i \cdot \mathbf{x}_j\right)$ represents the inner product between two samples. With $\boldsymbol{\alpha}$ and $\mathbf{b}$, the decision function in (2) can be rewritten as

$$f(\mathbf{x}) = \arg\max_{1\leq c'\leq C}\left(\sum_{i=1}^{N}\left(m_{ic'}\sum_{c=1}^{c}\alpha_{ic} - \alpha_{ic}\right)\left(\mathbf{x}_i \cdot \mathbf{x}\right) + b_{c'}\right) \tag{7}$$

Usually, the inner product $\left(\mathbf{x}_i \cdot \mathbf{x}_j\right)$ is replaced by $K\left(\mathbf{x}_i, \mathbf{x}_j\right)$, where $K(\cdot, \cdot)$ is a kernel function. The commonly used kernel functions include the Gaussian kernel, the linear kernel, the polynomial kernel, the sigmoid kernel, etc.

3. Multimodal neuroimaging feature selection and fusion

In this section, we give the details of the proposed method, including its formulation, optimization, comparison analysis with other methods and convergence analysis.

3.1. Formulation

Suppose we have a multimodal dataset χ having K modalities, where each modality $\mathbf{X}^k = \{\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_N^k\}$ contains N samples and

each sample $\mathbf{x}_i^k = [x_{i1}^k, x_{i2}^k, \ldots, x_{iD_k}^k]^T$ is represented in a D_k-dimensional feature space. When the label information $\mathbf{Y}$ is added into each modality, the training samples are obtained as $\{\mathbf{X}^k, \mathbf{Y}\}$. Since we know that each modality data is scanned/collected from the same subject, label information is shared across all modalities. Thus, in this study, we force $\mathbf{Y}^k$ to be $\mathbf{Y}$. Moreover, we use $\mathbf{W}^k \in \mathbb{R}^{D_k \times C}$ to represent the projected matrix of the kth modality.

Under the MCSVM-based framework, for multimodal neuroimaging fusion and classification, we aim to find the optimal weight vector $\mathbf{h} = [h^1, h^2, \ldots, h^K]^T$ to combine the features of each modality along with $\mathbf{W}^k$. By considering regularized risk minimization, the optimization problem can be formulated as

$$\min_{\mathbf{h},\mathbf{w}^k} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \sum_{k=1}^{K} \left(1 - y_{ic} \sum_{d=1}^{D^k} \sqrt{h^k} w_{dc}^k x_{di}^k\right)_+ + \frac{\alpha}{2} \sum_{k=1}^{K} \left\|\mathbf{W}^k\right\|_2^2 + \mu\Omega(\mathbf{h}) \right) \tag{8}$$

where N denotes the size of multimodal data, C denotes the number of classes, K denotes the number of modalities and D^k denotes the dimension of the k th modality. $(1-z)_+ = \max(1-z, 0)$ is the hinge loss function. represents the regularized risk term and μ is the regularized parameter. Note that the bias b_c for each class is hidden by augmenting the projected vector $\mathbf{w}_c$ and each sample $\mathbf{x}_i$ with an additional dimension: $\mathbf{w}_c^T \leftarrow [\mathbf{w}_c^T, b_c]$ and $\mathbf{x}_i^T \leftarrow [\mathbf{x}_i^T, 1]$, respectively.

In general, we expect that all modalities are able to provide complementary information for model training; hence, nonsparse solutions w.r.t. $\mathbf{h}$ are expected. In this study, we use the l_p -norm $(1 < p < \infty)$ to combine multiple modalities. The nonconvexity of the resulting optimization problem can be resolved by substituting $w_{dc}^k \leftarrow \sqrt{h^k} w_{dc}^k$. Furthermore, by introducing $\zeta = 1/N\alpha$ and adjusting $\mu \leftarrow \mu/\alpha$, we arrive at the following convex optimization problem:

$$\min_{\mathbf{h},\mathbf{w}^k} \left(\zeta \sum_{i=1}^{N} \sum_{c=1}^{C} \sum_{k=1}^{K} \left(1 - y_{ic} \sum_{d=1}^{D^k} w_{dc}^k x_{di}^k\right)_+ + \frac{1}{2} \sum_{k=1}^{K} \frac{1}{h^k} \left\|\mathbf{W}^k\right\|_2^2 + \mu \|\mathbf{h}\|_p^p \right) \tag{9}$$

According to [38, Th. 1], the solution to the problem in (9) can be replaced by optimizing the following problem:

$$\min_{\mathbf{n},\mathbf{w}^k} \left(\zeta \sum_{i=1}^{N} \sum_{c=1}^{c} \sum_{k=1}^{K} \left(1 - y_{ic} \sum_{d=1}^{D^k} w_{dc}^k x_{di}^k\right)_+ + \frac{1}{2} \sum_{k=1}^{K} \frac{1}{h^k} \left\|\mathbf{W}^k\right\|_2^2 \right) \quad \text{s.t.} \sum_{k=1}^{K} (h^k)^p \le 1 \tag{10}$$

The objective function in (10) is always appended with one relatively mild assumption, unless otherwise stated. The assumption is that $\mathbf{W}^k/0 = 0$ if $\mathbf{W}^k = 0$; otherwise, $\mathbf{W}^k/0 = \infty$ [38]. This assumption states that $\mathbf{W}^k = 0$ whenever $h^k = 0$ to reach a finite objective [30].

Additionally, in each modality, we also expect that important and discriminative features should assigned large weights during the modality fusion process. Thus, we introduce a scaling factor vector $\theta^k = \left[\theta_1^k, \theta_2^k, \ldots, \theta_{D^k}^k\right]^T$ for each modality to measure the importance of the features. By plugging θ^k into (10), we have the following optimization problem:

$$\min_{\mathbf{n},\theta^k,\mathbf{w}^k} \left(\zeta \sum_{i=1}^{N} \sum_{c=1}^{c} \sum_{k=1}^{K} \left(1 - y_{ic} \sum_{d=1}^{D^k} \sqrt{\theta_d^k} w_{dc}^k x_{di}^k\right) + \frac{1}{2} \sum_{k=1}^{K} \frac{1}{h^k} \left\|\mathbf{W}^k\right\|_2^2 \right) \quad \text{s.t.} \sum_{k=1}^{K} (h^k)^p \le 1, \theta^k > \mathbf{0}, (\theta^k)^T \mathbf{1} = 1 \tag{11}$$

3.2. Optimization

Directly solving the problem in (11) does not seem to be very easy. In this study, we use a wrapper-based optimization method to train our proposed model based on the following proposition and theorem.

Proposition 1. *Given a fixed $\mathbf{W}^k$ and θ^k, the minimal h^k in optimization problem in (11) can be obtained as*

$$h^k = \left(\left\|\mathbf{W}^k\right\|_2^2\right)^{\frac{1}{p+1}} \left(\sum_{k'=1}^{K} \left(\left\|\mathbf{W}^k\right\|_2^2\right)^{\frac{p}{p+1}} \right)^{-\frac{1}{p}} \tag{12}$$

The proof of Proposition 1 is given in Appendix A. We see that h^k is completely dependent on the value of $\mathbf{W}^k$.

When h^k is fixed, the problem in (11) can be transformed to another solvable optimization problem according to the following theorem, i.e., Theorem 1.

Theorem 1. *Solving the problem in (11) can be completely replaced by solving the problem defined as follows:*

$$\min_{\mathbf{w}^k} \left(\frac{1}{2} \sum_{k=1}^{K} \frac{1}{h^k} \left\|\mathbf{W}^k\right\|_{2,1}^2 + \zeta \sum_{c=1}^{C} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(1 - y_{ic} \sum_{d=1}^{D_k} w_{dc}^k x_{di}^k\right)_+ \right) \tag{13}$$

Proof. In (11), let $\sqrt{\theta_d^k} w_{dc}^k = \tilde{w}_{dc}^k$. Then, we have $w_{dc}^k = \tilde{w}_{dc}^k / \sqrt{\theta_d^k}$. By substituting $\tilde{w}_{dc}^k$ into (11), the optimization problem is updated as

$$\min_{\mathbf{w}^*,\theta^*} \left(\frac{1}{2} \sum_{k=1}^{K} \frac{1}{h^k} \sum_{c=1}^{C} \sum_{d=1}^{D_k} \frac{\left(\tilde{w}_{dc}^k\right)^2}{\theta_d^k} + \zeta \sum_{c=1}^{c} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(1 - y_{ic} \sum_{d=1}^{D_k} \tilde{w}_{dc}^k x_{di}^k\right)_+ \right) \quad \text{s.t.} \ \theta^k > \mathbf{0}, (\theta^k)^T \mathbf{1} = 1 \tag{14}$$

By replacing the notation $\tilde{w}_{dc}^k$ with w_{dc}^k, we arrive at

$$\min_{\mathbf{w}^k,\theta^*} \left(\frac{1}{2} \sum_{k=1}^{K} \frac{1}{h^k} \sum_{c=1}^{c} \sum_{d=1}^{D_k} \frac{\left(w_{dc}^k\right)^2}{\theta_d^k} + \zeta \sum_{c=1}^{c} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(1 - y_{ic} \sum_{d=1}^{D_k} w_{dc}^k x_{di}^k\right)_+ \right) \quad \text{s.t.} \ \theta^k > \mathbf{0}, (\theta^k)^T \mathbf{1} = 1 \tag{15}$$

According to the Cauchy–Schwarz inequality [39]

$$\left(\sum_{d=1}^{D_k} a_d b_d \right)^2 \le \left(\sum_{d=1}^{D_k} a_d \right)^2 \left(\sum_{d=1}^{D_k} b_d \right)^2 \tag{16}$$

where a_d and b_d represent any two real numbers, the first term in (15) can be further simplified by virtue of the constraints of θ^k as

$$\begin{aligned} &\min_{\theta^k>0} \sum_{k=1}^{K} \frac{1}{h^k} \sum_{d=1}^{D_k} \frac{1}{\theta_d^k} \sum_{c=1}^{c} (w_{dc}^k) = \min_{\theta^*>0} \sum_{k=1}^{K} \frac{1}{h^k} \sum_{d=1}^{D_k} \frac{\left\|\mathbf{w}^{d,k}\right\|_2^2}{\theta_d^k} \\ &= \min_{\theta^*>0} \sum_{k=1}^{K} \frac{1}{h^k} \sum_{d=1}^{D_k} \left[\frac{\left\|\mathbf{w}^{d,k}\right\|^2}{\sqrt{\theta_d^k}} \right]^2 \sum_{d=1}^{D_k} \theta_d^k = \sum_{k=1}^{K} \frac{1}{h^k} \left[\sum_{d=1}^{D_k} \left\|\mathbf{w}^{d,k}\right\|_2 \right]^2 \\ &= \sum_{k=1}^{K} \frac{1}{h^k} \left\|\mathbf{W}^k\right\|_{2,1}^2 \end{aligned} \tag{17}$$

Note that $\mathbf{w}^{d,k} \in \mathbb{R}^{C \times 1}$ is the dth column of $\mathbf{W}^k \in \mathbb{R}^{D^* \times C}$. By substituting (17) into (11), we complete the proof of Theorem 1.

When h^k is fixed, according to Theorem 1, we know that solving the problem in (11) can be completely replaced by solving the problem defined in (13) that can be further simplified as

$$\min_{\mathbf{w}^k} \sum_{k=1}^{K} \left\{ tr\left[(\mathbf{W}^k)^T \boldsymbol{\Theta}^k \mathbf{W}^k \right] + \zeta f((\mathbf{W}^k)^T \mathbf{X}^k, \mathbf{Y}) \right\} \tag{18}$$

where

$$tr\left[(\mathbf{W}^k)^T \boldsymbol{\Theta}^k \mathbf{W}^k \right] = \left\|\mathbf{W}^k\right\|_{2,1}^2 \tag{19}$$

$$f\left((\mathbf{W}^k)^T \mathbf{X}^k, \mathbf{Y} \right) = \sum_{c=1}^{C} \sum_{i=1}^{N} \left(1 - y_{ic} \sum_{d=1}^{D_k} w_{dc}^k x_{di}^k\right)_+ \tag{20}$$

and $\boldsymbol{\Theta}^k$ is a D_k by D_k diagonal matrix in which each element is defined as

$$\boldsymbol{\Theta}^k(i,j) = \begin{cases} 0 & \text{if } i \ne j \\ h^k / 2\|\mathbf{w}^{i,k}\|_2 & \text{if } i = j \end{cases} \tag{21}$$

It is obvious that $\boldsymbol{\Theta}^k$ is dependent on $\mathbf{W}^k$ and h^k such that an alternant iteration strategy can be used to search for the optimal solution to the problem in (13) . We should keep in mind that when $\mathbf{w}^{i,k} = \mathbf{0}, \boldsymbol{\Theta}^k(i,i) = 0$ becomes a subgradient of $\|\mathbf{W}^k\|_{2,1}^2$ w.r.t. $\mathbf{w}^{i,k}$. However, we know that $\boldsymbol{\Theta}^k(i,i)$ cannot be set to 0; otherwise, the convergence of the proposed method cannot be guaranteed. To avoid this exceptional case, we introduce a very small constant τ to impose a fine adjustment on $\boldsymbol{\Theta}^k(i,i)$ as $h^k/2\sqrt{\left(\mathbf{w}^{i,k}\right)^T \mathbf{w}^{i,k} + \tau}$.

In each iteration, $\mathbf{W}^k$ is updated by the current $\boldsymbol{\Theta}^k$, and $\boldsymbol{\Theta}^k$ is updated by the current $\mathbf{w}^k$ and h^k. This iteration procedure is repeated until the algorithm converges. The optimization problem in (18) can be considered as solving K subproblems. Let $\tilde{\mathbf{W}}^k = \left(\boldsymbol{\Theta}^k\right)^{\frac{1}{2}} \mathbf{W}^k$ and $\tilde{\mathbf{X}}^k = \left(\boldsymbol{\Theta}^k\right)^{\frac{-1}{2}} \mathbf{X}^k$, that is, $\tilde{w}_{dc}^k = \left(\boldsymbol{\Theta}^k(d,d)\right)^{\frac{1}{2}} w_{dc}^k$ and $\tilde{x}_{di}^k = \left(\boldsymbol{\Theta}^k(d,d)\right)^{\frac{-1}{2}} x_{di}^k$. Then, each subproblem of (18) becomes

$$\begin{aligned}
&\min_{\mathbf{w}^k} tr\left[\left(\mathbf{W}^k\right)^T \boldsymbol{\Theta}^k \mathbf{W}^k\right] + \zeta f\left(\left(\mathbf{W}^k\right)^T \mathbf{X}^k, \mathbf{Y}\right) \\
&\Leftrightarrow \min_{\mathbf{w}^k} \left[\left(\mathbf{W}^k\right)^T \left(\boldsymbol{\Theta}^k\right)^{\frac{1}{2}} \left(\boldsymbol{\Theta}^k\right)^{\frac{1}{2}} \mathbf{W}^k\right] \\
&+\zeta f\left(\left(\mathbf{W}^k\right)^T \left(\boldsymbol{\Theta}^k\right)^{\frac{1}{2}} \left(\boldsymbol{\Theta}^k\right)^{\frac{-1}{2}} \mathbf{X}^k, \mathbf{Y}\right) \\
&\Leftrightarrow \min_{\tilde{\mathbf{W}}^k} tr\left[\left(\tilde{\mathbf{W}}^k\right)^T \tilde{\mathbf{W}}^k\right] + \zeta f\left(\left(\tilde{\mathbf{W}}^k\right)^T \tilde{\mathbf{X}}^k, \mathbf{Y}\right)
\end{aligned} \tag{22}$$

So far, we have a clear idea that the optimization problem in (11) has been transformed into K solvable subproblems in (22). Each subproblem can be solved by using the QP solver used in [38] for the MCSVM. During the problem transformation from (11) to the final (22), we always assume that our problem is linearly separable. In fact, for each subproblem, we can use different kernel functions to replace the inner product in (6) so that our multimodality fusion is actually an MKL method. Algorithm 1 gives the detailed steps of our proposed method. When the training process for each modality's data is finished, for an unseen sample, its predictive result in the kth modality can be computed as

$$f^k\left(\mathbf{x}^k\right) = \left(\mathbf{W}^k\right)^T \mathbf{x}^k \tag{23}$$

where $f^k\left(\mathbf{x}^k\right) \in \mathbb{R}^C$ in which each element indicates the predictive value of the sample in the kth modality w.r.t. the cth class. The predictive values of each modality may be quite different, which means that our proposed classifier performs discriminatively in different modalities. Therefore, for the final decision, it is very meaningful to put the predictive values from all modalities together. To this end, we use the following strategy to normalize the predictive values from different modalities:

$$\tilde{f}^k\left(\mathbf{x}^k\right) = \frac{f^k\left(\mathbf{x}^k\right) - \min\left(f^k\left(\mathbf{x}^k\right)\right)}{\max\left(f^k\left(\mathbf{x}^k\right)\right) - \min\left(f^k\left(\mathbf{x}^k\right)\right)} \tag{24}$$

where $\tilde{f}^k\left(\mathbf{x}^k\right) \in \mathbb{R}^C$. The final predictive value of sample $\mathbf{x}$ via all modalities can be computed as

$$\tilde{f}(\mathbf{x}) = \sum_{k=1}^{K} h^k \tilde{f}^k\left(\mathbf{x}^k\right) \tag{25}$$

Based on $\tilde{f}(\mathbf{x})$, the label information (category) of sample $\mathbf{x}$ can be determined by

$$y^* = \underset{c\in\{1,2,\ldots,C\}}{\arg\max} \left[\tilde{f}(\mathbf{x})\right]_c \tag{26}$$

3.3. Compared with other regularizations

In MKL-based multimodality fusion, different regularization terms, e.g., the l_1-norm, the $l_{2,1}$-norm and the $l_{1,p}$-norm, can yield different modality selection or feature section methods. To be specific, in [30], the l_1-norm was used to select the most discriminative kernels, i.e., modalities. In [33], Liu et al. used the $l_{2,1}$-norm to select the most relevant modalities for AD classification. A similar regularized strategy was also used in [40]. As we stated in the first section, all of the complementary information contained in different modalities is desired for AD classification. The l_1-norm and $l_{2,1}$-norm aim to select the most discriminative features/groups independently so that the features from weak modalities have a lower probability of being selected. Furthermore, they are less effective at exploiting the complementary information among modalities, as shown in Fig. 2. In [10], Peng et al. designed a new regularization $l_{1,p}$-norm that cannot only generate a nonsparse modality coefficient distribution to combine different complementary information among modalities but also generate a sparse feature coefficient distribution to select the discriminative and important features simultaneously from each modality. However, this method is only designed for binary classification tasks. Our method is different from the abovementioned ones. It combines the $l_{2,1}$-norm and l_p-norm to select features across all the classes in each modality and avoid a sparse kernel coefficient distribution, thereby effectively exploiting complementary modalities. These combined regularization terms are particularly valuable for AD, MCI and HC classification, where features are naturally grouped by modalities and each modality is useful.

Algorithm 1

Input: Multimodal data $\mathbf{X}^k = \left\{\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_N^k\right\}_{k-1}^K$ having K modalities; label information $\mathbf{Y}$; penalty parameter ζ; p in l_p -norm

Output: Projection matrix $\mathbf{W}^k$ of each modality

1: Initialize iteration counter $t \leftarrow 0$

2: Initialize the coefficient vector

$$\mathbf{h}(t) = \left\{h^k = 1/K\right\}, k = 1, 2, \ldots, K$$

3: Initialize the projection matrix

$$\mathbf{W}^k(t) = \left\{w_{dk}^k\right\}, d = 1, 2, \ldots, D^k, c = 1, 2, \ldots, C$$

4: Initialize the diagonal matrix $\boldsymbol{\Theta}^k$ in which each diagonal element is computed by

$$\boldsymbol{\Theta}^k(i,i)(t) = h^k(t) \left\|\mathbf{W}^k(t)\right\|_{2,1} / \left\|\mathbf{w}^{j,k}(t)\right\|_2$$

5: Initialize $\tilde{\mathbf{W}}^k(t) = \left(\boldsymbol{\Theta}^k(t)\right)^{\frac{1}{2}} \mathbf{W}^k(t)$ and $\tilde{\mathbf{X}}^k(t) = \left(\boldsymbol{\Theta}^k(t)\right)^{\frac{-1}{2}} \mathbf{X}^k$

6: **Repeat**

7: Update $\tilde{\mathbf{W}}^k(t)$ using the QP solver in [37]:

$$\min_{\tilde{\mathbf{W}}^k} \mathrm{tr}\left[\left(\tilde{\mathbf{W}}^k\right)^T \tilde{\mathbf{W}}^k\right] + \zeta f\left(\left(\tilde{\mathbf{W}}^k\right)^T \tilde{\mathbf{X}}^k, \mathbf{Y}\right)$$

8: Update the projection matrix

$$\mathbf{W}^k(t+1) = \left(\boldsymbol{\Theta}^k(t)\right)^{\frac{-1}{2}} \tilde{\mathbf{W}}^k(t+1)$$

9: Update the kernel coefficient vector

$$h^k(t+1) = \left(\left\|\mathbf{W}^k(t+1)\right\|_2^2\right)^{\frac{1}{p+1}} \left(\sum_{k'=1}^{K} \left(\left\|\mathbf{W}^k(t+1)\right\|_2^2\right)^{\frac{p}{p+1}}\right)^{\frac{-1}{p}}$$

10: Update the diagonal matrix $\boldsymbol{\Theta}^k$ in which each diagonal element is computed by

$$\boldsymbol{\Theta}^k(i,i)(t+1) = h^k(t+1) \left\|\mathbf{W}^k(t)\right\|_{2,1} / \left\|\mathbf{w}^{j,k}(t+1)\right\|_2$$

11: Update iteration counter $t \leftarrow t+1$

12: **Until** algorithm converges

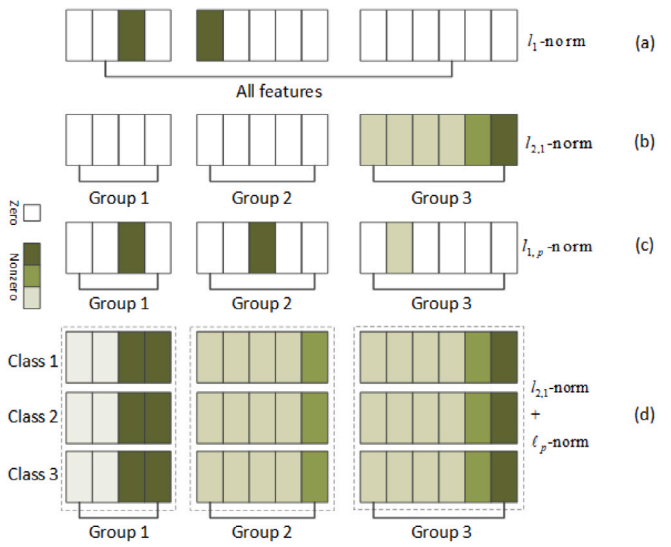


Fig. 2. Schematic illustration of different sparse regularizations. (a) The l_1-norm generates a sparse distribution but loses sight of the inherit group structures. (b) The $l_{2,1}$-norm sparsely chooses a few groups of features with predefined group structures. (c) The $l_{1,p}$-norm ($p > 1$) retains all groups and conducts within-group feature selection. (d) The $l_{2,1}$-norm with the multiclass hinge loss conducts natural feature selection for all the classes and the l_p-norm ($p > 1$) combines different groups to exploit complementary information. Each box represents a feature where a box with a darker color indicates that the corresponding feature has a larger weight.

3.4. Convergence

For an alternating iteration algorithm, it is very important to analyze its convergence. In this section, we make use of the following theorem, i.e., Theorem 2, to prove the convergence of our proposed algorithm.

Theorem 2. *The returned value of the objective function in (15) will be monotonically decreased and finally converge to the global optimum via each iteration in Algorithm 1.*

Proof. According to the abovementioned analysis, the solver applied in [37] can also be used to solve the following subproblem by changing the variable:

$$\min_{w} \frac{1}{2} \operatorname{tr}\left(\mathbf{W}^T \boldsymbol{\Theta} \mathbf{W}\right) + \zeta f\left(\mathbf{W}^T \mathbf{X}, \mathbf{Y}\right) \tag{27}$$

Therefore, when the tth iteration is coming, the next status of $\mathbf{W}$ can be formulated as

$$\mathbf{W}^{(t+1)} = \arg\min_{\mathbf{w}} \frac{1}{2} \operatorname{tr}\left(\mathbf{W}^T(t)\boldsymbol{\Theta}(t)\mathbf{W}(t)\right) + \zeta f\left(\mathbf{W}^T(t)\mathbf{X}, \mathbf{Y}\right) \tag{28}$$

which means that

$$\begin{aligned} &\tfrac{1}{2} tr\left(\mathbf{W}^T(t+1)\boldsymbol{\Theta}(t)\mathbf{W}(t+1)\right) + \zeta f\left(\mathbf{W}^T(t+1)\mathbf{X}, \mathbf{Y}\right) \\ &\leq \tfrac{1}{2} tr\left(\mathbf{W}^T(t)\boldsymbol{\Theta}(t)\mathbf{W}(t)\right) + \zeta f\left(\mathbf{W}^T(t)\mathbf{X}, \mathbf{Y}\right) \end{aligned} \tag{29}$$

Furthermore, the inequality in (29) can be extended as

$$\begin{aligned} &\frac{1}{2}\sum_{d=1}^{D} \frac{\left\|\mathbf{w}^d(t+1)\right\|_2^2}{2\left\|\mathbf{w}^d(t)\right\|_2} + \zeta f\left(\mathbf{W}^T(t+1)\mathbf{X}, \mathbf{Y}\right) \\ &\leq \frac{1}{2}\sum_{d=1}^{D} \frac{\left\|\mathbf{w}^d(t)\right\|_2^2}{2\left\|\mathbf{w}^d(t)\right\|_2} + \zeta f\left(\mathbf{W}^T(t)\mathbf{X}, \mathbf{Y}\right) \end{aligned} \tag{30}$$

Due to $\left(\left\|\mathbf{w}^d(t+1)\right\|_2 - \left\|\mathbf{w}^d(t)\right\|_2\right)^2 \geq 0$, the following inequality holds:

$$\left\|\mathbf{w}^d(t+1)\right\|_2 - \frac{\left\|\mathbf{w}^d(t+1)\right\|_2^2}{2\left\|\mathbf{w}^d(t)\right\|_2} \leq \left\|\mathbf{w}^d(t)\right\|_2 - \frac{\left\|\mathbf{w}^d(t)\right\|_2^2}{2\left\|\mathbf{w}^d(t)\right\|_2} \tag{31}$$

Furthermore, the following inequality also holds:

$$\begin{aligned} &\sum_{d=1}^{D} \left\|\mathbf{w}^d(t+1)\right\|_2 - \frac{\left\|\mathbf{w}^d(t+1)\right\|_2^2}{2\left\|\mathbf{w}^d(t)\right\|_2} \\ &\leq \sum_{d=1}^{D} \left\|\mathbf{w}^d(t)\right\|_2 - \frac{\left\|\mathbf{w}^d(t)\right\|_2^2}{2\left\|\mathbf{w}^d(t)\right\|_2} \end{aligned} \tag{32}$$

By putting (32) and (30) together, we have

$$\begin{aligned} &\tfrac{1}{2}\textstyle\sum_{d=1}^{D} \left\|\mathbf{w}^d(t+1)\right\|_2 + \zeta f\left(\mathbf{W}^T(t+1)\mathbf{X}, \mathbf{Y}\right) \\ &\leq \tfrac{1}{2}\textstyle\sum_{d=1}^{D} \left\|\mathbf{w}^d(t)\right\|_2 + \zeta f\left(\mathbf{W}^T(t)\mathbf{X}, \mathbf{Y}\right) \end{aligned} \tag{33}$$

According to the definition of the $l_{2,1}$-norm, we arrive at

$$\begin{aligned} &\tfrac{1}{2}\|\mathbf{W}(t+1)\|_{2,1} + \zeta f\left(\mathbf{W}^T(t+1)\mathbf{X}, \mathbf{Y}\right) \\ &\leq \tfrac{1}{2}\|\mathbf{W}(t)\|_{2,1} + \zeta f\left(\mathbf{W}^T(t)\mathbf{X}, \mathbf{Y}\right) \end{aligned} \tag{34}$$

Therefore, we complete the proof. Since our optimization subproblem is convex, a global optimum solution will be obtained. The following experimental results show that the convergence is always very fast. In our classification tasks, the proposed method converges within a few iterations.

3.5. Multimodal neuroimaging data preprocessing

In this sub section, data acquisition and preparation, experimental settings are given for reproducibility.

i. Data acquisition

The neuroimaging data we used in this study are collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data repository (http://adni.loni.usc.edu/about/). ADNI is a five-year public partnership that is sponsored by serval institutes, companies and nonprofit organizations. The main purpose of this project is to discover MCI and early AD by combining different/multiple biomarkers such as MRI, PET and CSF. The subjects in the ADNI were enrolled from at least 50 areas of the United States and Canada. Approximately 200 normal subjects and 400 MCI subjects were followed for two years and 200 AD patients were followed within two years. Measurements of the sensitivity and specificity of early biomarkers can help researchers and clinicians determine the progress of AD, which in turn can help physicians develop new treatments to monitor their effectiveness, reduce the time and cost of clinical trials, and improve the safety of drug development. Therefore, the ADNI data repertory has become the preferred database for AD research. In this study, we combine three biomarkers using MRI, PET and CSF as the three modalities to construct a classification task for classifying the selected subjects into three groups, i.e., AD, MCI and HC. Table 1 gives detailed information of the selected subjects, where the inclusion criteria of the subjects are listed as:

(1) HC group: mini-mental state examination score (MMSE) belongs to [29,40] and clinical dement rating (CDR) is equal to 0;
(2) MCI group: MMSE belongs to [29,40] and CDR is equal to 0.5. The subjects in this group have memory impairment and memory loss, but they do not have any serious impairment in other aspects of cognition. They can basically maintain normal daily activities.
(3) AD group: MMSE belongs to [26,31] and CDR is equal to 0.5 or 1.

ii. Data preprocessing

We use the SPM (statistical parametric mapping) tool to preprocess the MRI and PET neuroimaging data. Fig. 4 illustrates the preprocessing pipeline of MRI data, which can be divided into three brief steps.

Step 1: The TPM (tissue probability map) template is used to segment the original MRI data into WM (white matter), GM (gray matter) and other tissues. During the segmentation process, WM and GM tissues

Table 1
Detailed information of the selected subjects.

	AD	MCI	HC
#subjects	38	42	40
Age	77.2 ± 7.7	74.1 ± 6.3	77.2 ± 5.2
CDR	0.96 ± 0.41	0.49 ± 0.17	0.05 ± 0.15
MMSE	21.82 ± 5.04	27.10 ± 2.74	29.25 ± 0.90

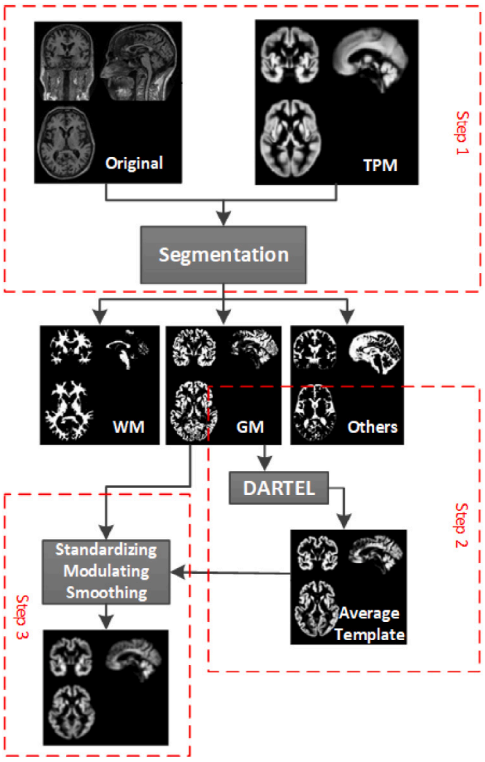


Fig. 3. Preprocessing pipeline of MRI data.

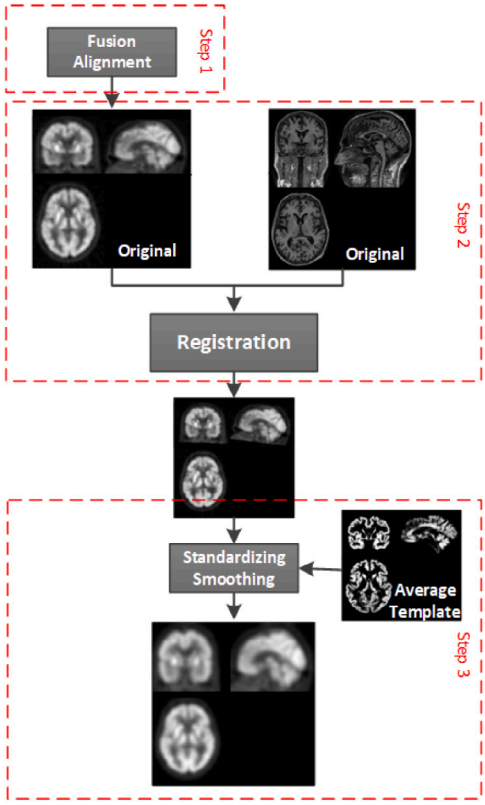


Fig. 4. Preprocessing pipeline of PET data.

Table 2
Classification tasks.

Tasks	Groups	Classes
T1	AD vs. MCI	2
T2	AD vs. HC	2
T3	MCI vs. HC	2
T4	AD vs. MCI vs. HC	3

are mapped into the MNI (Montreal neurological institute) space for the following preprocessing steps.

Step 2: DARTEL (diffeomorphic anatomical registration through exponentiated lie algebra) is used to create average templates for the obtained WM and GM tissues. To be specific, this method models the shape of each brain image to increase the registration accuracy between samples. Compared with the classic MNI template, DARTEL can reduce the influence of the errors caused by race and the proportion of normal people to patients on subsequent studies. Simultaneously, the process also calculates the deformation of each region, i.e., the Jacobian matrix, for subsequent gray matter space normalization.

Step 3: The original GM images are spatially normalized to the standard MNI space and Jacobian scaled based on the obtained average template data and the deformation field. Before normalization, all GM images are modulated to transform the density information to volume information. Additionally, they are smoothed (8 mm Gaussian) to avoid the influences caused by noises.

Fig. 4 illustrates the preprocessing pipeline of PET data, which can also be divided into three brief steps.

Step 1: There are 96 PET images for each subject in the ADNI. SPM 12 is used to fuse these PET images to construct a 3-D one that has brain spatial information and the feature information between tissue structures is also retained. Additionally, motion correction is performed due to head motion.

Step 2: The MRI image and PET image of each subject are registered and affinely aligned.

Step 3: The average template data generated in Fig. 3 is used to spatially normalize all PET images to the standard MNI space. PET images are also smoothed (8 mm Gaussian) to avoid the influences caused by noises.

iii. Feature extraction

The AAL (automated anatomical atlas), which is available as a toolbox (http://www.gin.cnrs.fr/AAL) for SPM, is used as a template to extract original features from MRI and PET images. Based on the AAL, the brain is segmented into 116 regions and we select 90 regions from the cerebrum for feature extraction. To be specific, first, the MRI images and PET images are resampled to the same size as the AAL template so that each region spatially corresponds. The size of the AAL template is 61 × 73 × 61. Then, we extract the volume values from all regions of the MRI images and the average intensity values from all regions of the PET images as the original features for our proposed classification model.

After data preprocessing and feature extraction, we finally obtain 90 features of each modality. Based on these features, we construct several classification tasks shown in Table 2 to verify the proposed method.

4. Experimental studies

4.1. Experimental settings

To fairly evaluate the performance of the proposed model, we introduce eight benchmarking models, which belong to four different

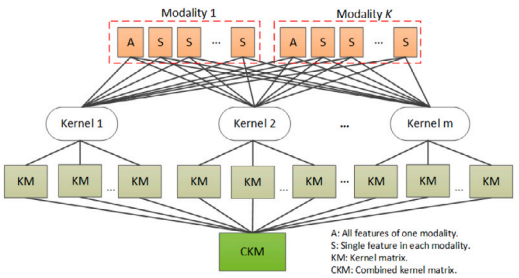


Fig. 5. Combinations between modalities and kernels.

types, i.e., the feature section-based type, the MKL-based type, the feature selection & MKL-based type and the deep learning-based type, for comparison studies. Table 3 gives the benchmarking models and their corresponding parameter settings.

4.2. Experimental results

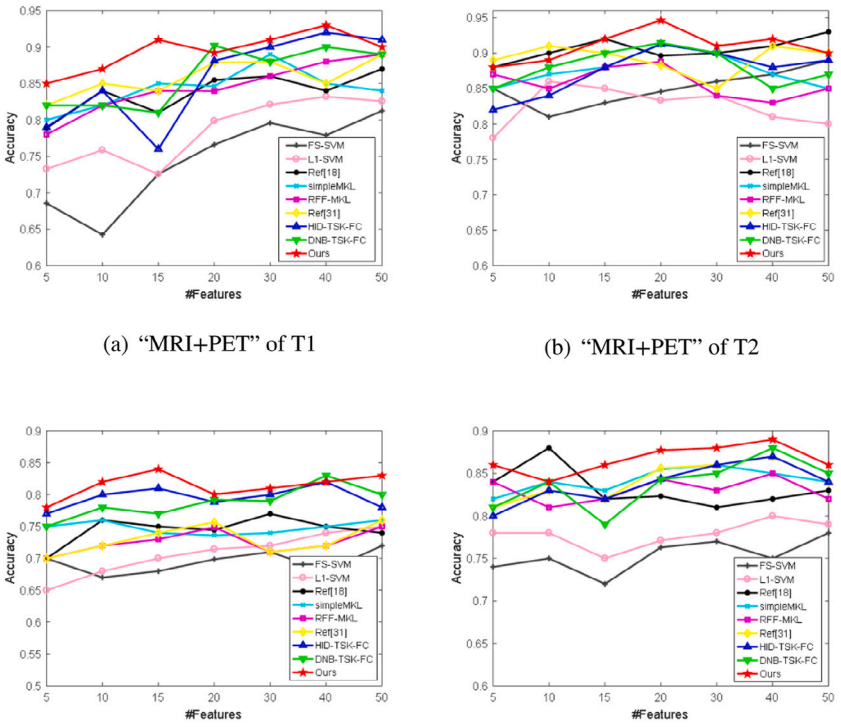
Tables 4 and 5 show the training and testing classification performance of all models with the top 20 selected features in terms of ACC and MAP, respectively. From Table 4, we see that except for the training accuracy on PET of T2, our model performs better than the two feature selection-based models, i.e., FS-SVM and l_1 -SVM. Compared with the two MKL-based models, i.e., Ref. [22] and simpleMKL, our model has absolute advantages. Additionally, except for the training accuracy on MRI of T2, our model also achieves better performance than the two feature selection & MKL-based models, RFF-MKL and Ref. [10]. By observing the results from HID-TSK-FC and DBN-TSK-FC, which are two deep learning-based models, we find that overfitting problems often occur. The training accuracies of some modalities are better than that of our model. However, except for "MRI+PET" of T1 and MRI of T4, the testing accuracy of our model is better than those of the two deep learning-based models. Very similar results can also be observed from Table 5 according to MAP. Additionally, we also tune the number of selected features in each modality instead of only picking 20 features and report the corresponding results in Fig. 6 in terms of ACC. We observe that on the top 5, 10, 15, 30 and 40 selected features, our model performs better than all benchmarking methods on the T1 task. Such advantages can also be observed from T2 on the top 20, 30 and 40 selected features; T3 on the top 5, 10, 15, 20, 30, and 50 selected features and T4 on the top 5, 15, 20, 30, 40 and 50 selected features.

Furthermore, our proposed model can also be used for feature selection. It can actually be considered as an embedded feature selection model. To be specific, the obtained projection matrix $\mathbf{W}^k$ of each modality can be used as a selection guideline, where each row corresponds to each original feature and each column corresponds to each class. To independently evaluate the feature selection ability, several feature selection methods, i.e., mRMR [46], REF-SVM [20] and relief [47], are introduced for comparison and SVM with the Gaussian kernel function is adopted as the classifier. Fig. 8 shows the comparison results on MRI of T4. We see that the top 15 to 40 features selected by our model are more discriminating to the classifier than mRMR and relief belonging to the filter type and the REF-SVM belonging to the wrapped type. This is because our feature selection is embedded into the multiclass hinge loss function during the optimization procedure so that the topper features become more discriminating to the classifier (see Fig. 7).

We also visualize the projection matrix $\mathbf{W}^k$ obtained from MRI of T4 in Fig. 8, where a brighter pixel means a higher value. From Fig. 8, we can clearly see the structural sparsity across all classes generated by the $l_{2,1}$-norm combined with the multiclass hinge loss function. This promising structure provides a very simple way to select discriminative features.

Therefore, compared to all the benchmark models, ours has some advantages. This is mainly due to the following reasons.

(1) Our feature selection is embedded into the multiclass hinge loss function so that the selected features are more discriminative for



(a) "MRI+PET" of T1 (b) "MRI+PET" of T2

(c) "MRI+PET" of T3 (d) "MRI+PET" of T4

Fig. 6. Comparison results of "MRI+PET" in T1, T2, T3 and T4 in terms of the testing ACC under different numbers of selected features.

Table 3
Parameter settings.

Groups	Algorithms	Description and settings
Feature-selection-based type	FS-SVM l1-SVM	The Fisher score [41] and Lasso [42] are used for feature ranking and selection, respectively; and finally, the SVM is used as the classifier.
MKL-based type	Ref. [22]	The coefficient of each kernel is determined by cross-validation.
	simpleMKL [30]	The simpleMKL toolbox for MATLAB is used.
Deep-learning-based type	HID-TSK-FC [43]	HID-TSK-FC is a deep ensemble learning method in which components are integrated in a stacked manner. We use the parameter settings recommended by [43].
	DBN-TSK-FC [44]	DBN-TSK-FC is a DBN (deep belief network)-based method. It builds its DBN-based neural representation in a hierarchical way. We use the parameter settings recommended by [44].
Feature-selection & MKL-basedtype	Ref. [10] RFF-MKL [45]	We use the parameter settings recommended by the original references.
	Ours	is set to 1.

Remark 1: For the FS-SVM and l1-SVM, the linear kernel, Gaussian kernel and polynomial kernel are introduced as the kernel functions, respectively. The average results of the different kernels are reported.
Remark 2: For the benchmarking models in the MKL-based type and the feature selection & MKL-based type, we use the following "all-single" schema (see Fig. 5) to combine modalities and kernels.
Remark 3: To be fair, we use FS to select features for Ref. [22] and simpleMKL since both have no feature selection mechanisms.
Remark 4: The classification performance is evaluated by the accuracy (termed as ACC) and mean average precision (termed as MAP) criteria. Notably, the MAP is defined as the arithmetic mean of the average pre-cision values for an information retrieval system over a set of N query topics. The ACC and MAP are defined as follows:

$$ACC = \frac{\#correctly\ \ classified\ \ subjects}{N}, \quad (35)$$

$$MAP = \frac{1}{N}\sum_{N} AP_N, \quad (36)$$

where AP (average precision) is a criterion that combines Recall and Precision.
Remark 5: 5-fold CV is used to search for the kernel parameters and other parameters for each method. Using the optimal parameters, each method is performed 30 times and the average performance is recorded.

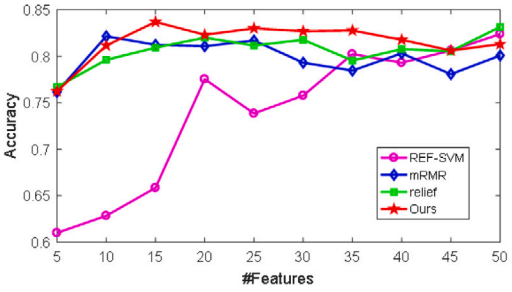


Fig. 7. Classification results on MRI of T1 using different feature section methods.

Fig. 8. Projection matrix $\mathbf{W}^k$ of MRI on T4.

the corresponding classification tasks. Although FS is employed for Ref. [22] and simpleMKL to select the features, it belongs to the filter type, which does not consider the discriminability to classifiers.

(2) In Ref. [22], the cross-validation strategy is used to search for the kernel coefficient. Although it is straightforward, it seems not to generate an optimal kernel combination to minimize the objective. Our method employs the l_p-norm combined with the convex multiclass loss function to combine kernels so that an updated rule of the kernel coefficient can be derived. Our optimal kernel combination can guarantee a global optimum objective.

(3) Unlike RFF-MKL that uses random Fourier features (RFFs) to approximate Gaussian kernels, the proposed model transforms the optimization problem to a solvable one that can be solved in its dual form so that the kernel mapping ability is not damaged.

(4) Compared with simpleMKL and RFF-MKL that generate sparse solutions to kernel coefficients, in our model, the l_p-norm regularization term is introduced to combine different kernels to perform multiple kernel learning to avoid a sparse kernel coefficient distribution, thereby effectively exploiting complementary modalities.

4.3. Efficiency analysis

To observe the convergence of the proposed model, Fig. 9 shows the convergence behavior curves of "MRI+PET" on classification tasks T1, T2, T3 and T4. We observe that within 20 iterations, our model converges in a fast manner.

Fig. 10 shows the average CPU computation time in seconds of 30 tries for each model on "MRI+PET" of T1. We see that our model consumes less time than most of the MKL models (comparable with Ref. [22]) and deep learning models. However, our model has a higher CPU computation time in seconds than FS-SVM and l_1-SVM. In general, multiple kernel-based models are more time-consuming and space-consuming than single kernel-based models due to the multikernel storage and combination. This is also the limitation we point out in the conclusion.

4.4. Nonparametric statistical analysis

To detect whether significant differences exist among the 9 models on the T1, T2, T3 and T4 tasks, two test methods, the Friedman ranking test and the Holm post-hoc test [48,49], are introduced to statistically analyze the results from the top selected 20 features in terms of ACC.

Table 4
Classification performance of all models on top 20 selected features in terms of the training *ACC* (left) and the testing *ACC* (right). The numbers in the parentheses are the standard deviations of the testing.

Task	Modality	FS-SVM	l1-SVM	Ref. [22]	simpleMKL	RFF-MKL	Ref. [10]	HID-TSK-FC	DBN-TSK-FC	Ours
T1	MRI	0.7598/0.7563 (0.0021)	0.7621/0.7467 (0.0019)	0.8099/0.8019 (0.0016)	0.8120/0.8002 (0.0020)	0.7924/0.7832 (0.0027)	0.8026/0.8034 (0.0018)	**0.8478**/0.7963 (0.0029)	0.8423/0.8025 (0.0054)	0.8405/**0.8362** (0.0025)
	PET	0.7532/0.7445 (0.0011)	0.7796/0.7789 (0.0023)	0.8247/0.8150 (0.0010)	0.8025/0.7963 (0.0027)	0.8257/0.7990 (0.0042)	0.8372/0.8136 (0.0028)	0.8863/0.8354 (0.0014)	**0.8922**/0.8369 (0.0040)	0.8784/**0.8467** (0.0014)
	MRI + PET	0.8012/0.7661 (0.0015)	0.8125/0.7987 (0.0028)	0.8747/0.8499 (0.0023)	0.8263/0.8359 (0.0056)	0.8632/0.8401 (0.0023)	0.9012/0.8790 (0.0041)	0.9148/0.8812 (0.0020)	**0.9358**/**0.9021** (0.0025)	0.9257/0.8963 (0.0020)
T2	MRI	0.8258/0.8179 (0.0021)	0.8364/0.8256 (0.0022)	0.8500/0.8478 (0.0015)	0.8678/0.8603 (0.0028)	**0.8805**/0.8400 (0.0027)	0.8570/0.8451 (0.0032)	0.8669/0.8056 (0.0016)	0.8711/0.8563 (0.0071)	0.8800/**0.8765** (0.0032)
	PET	**0.8789**/0.8224 (0.0008)	0.8256/0.8058 (0.0011)	0.8598/0.8547 (0.0021)	0.8701/0.8487 (0.0036)	0.8605/0.8502 (0.0007)	0.8480/0.8372 (0.0022)	0.8569/0.8214 (0.0034)	0.8598/0.8254 (0.0035)	0.8702/**0.8652** (0.0027)
	MRI + PET	0.8745/0.8457 (0.0023)	0.8527/0.8333 (0.0016)	0.9240/0.8957 (0.0023)	0.9365/0.9087 (0.0026)	0.9060/0.8881 (0.0017)	0.9000/0.8787 (0.0011)	0.9563/0.9021 (0.0014)	0.9487/0.9025 (0.0022)	**0.9674**/**0.9458** (0.0005)
T3	MRI	0.7147/0.6869 (0.0008)	0.7215/0.6982 (0.0020)	0.7398/0.7251 (0.0018)	0.7365/0.7256 (0.0018)	0.7350/0.7287 (0.0005)	0.7524/0.7208 (0.0011)	0.7666/0.7124 (0.0030)	0.7752/0.7469 (0.0015)	**0.7902**/**0.7720** (0.0021)
	PET	0.6743/0.6563 (0.0011)	0.7001/0.6759 (0.0026)	0.7301/0.7109 (0.0023)	0.6975/0.6789 (0.0030)	0.7324/0.7000 (0.0016)	0.7263/0.7182 (0.0017)	**0.7965**/0.7361 (0.0058)	0.7785/0.7360 (0.0032)	0.7541/**0.7523** (0.0028)
	MRI + PET	0.7325/0.6987 (0.0020)	0.7251/0.7145 (0.0012)	0.7678/0.7469 (0.0025)	0.7630/0.7362 (0.0024)	0.8025/0.7489 (0.0027)	0.8241/0.7578 (0.0023)	**0.8486**/0.7855 (0.0074)	0.8365/0.7921 (0.0057)	0.8321/**0.8102** (0.0030)
T4	MRI	0.7532/0.7311 (0.0022)	0.7741/0.7474 (0.0022)	0.8210/0.7785 (0.0023)	0.8205/0.8021 (0.0015)	0.8423/0.8141 (0.0020)	0.8378/0.8167 (0.0032)	0.8555/0.8001 (0.0025)	**0.8652**/0.8452 (0.0028)	0.8601/0.8378 (0.0012)
	PET	0.7638/0.7458 (0.0018)	0.7543/0.7400 (0.0030)	0.8251/0.8024 (0.0010)	0.8578/0.8229 (0.0021)	0.8242/0.7989 (0.0027)	0.8410/0.8109 (0.0019)	0.8569/0.8021 (0.0008)	0.8478/0.7983 (0.0036)	**0.8687**/**0.8527** (0.0032)
	MRI + PET	0.7854/0.7630 (0.0030)	0.8002/0.7711 (0.0028)	0.8602/0.8301 (0.0024)	0.8901/0.8587 (0.0021)	0.8760/0.8442 (0.0020)	0.8860/0.8572 (0.0022)	**0.9006**/0.8436 (0.0059)	0.8897/0.8430 (0.0030)	**0.8997**/**0.8854** (0.0027)

Table 5
Classification performance of all models on top 20 selected features in terms of the training *MAP* (left) and the testing *MAP* (right). The numbers in the parentheses are the standard deviations of the testing.

Task	Modality	FS-SVM	l1-SVM	Ref. [22]	simpleMKL	RFF-MKL	Ref. [10]	HID-TSK-FC	DBN-TSK-FC	Ours
T1	MRI	0.7736/0.7664 (0.0022)	0.7801/0.7554 (0.0023)	0.8320/0.8198 (0.0023)	0.8021/0.7856 (0.0021)	0.8007/0.8001 (0.0030)	0.8107/0.8132 (0.0017)	0.8424/0.8122 (0.0011)	**0.8521**/0.8132 (0.0043)	0.8485/**0.8430** (0.0033)
	PET	0.7638/0.7553 (0.0012)	0.7892/0.7801 (0.0021)	0.8387/0.8156 (0.0032)	0.8210/0.7830 (0.0024)	0.8402/0.8102 (0.0019)	0.8502/0.8239 (0.0022)	0.8665/0.8154 (0.0014)	**0.9021**/0.8552 (0.0027)	0.8698/**0.8630** (0.0014)
	MRI + PET	0.8157/0.7754 (0.0020)	0.8253/0.8036 (0.0021)	0.8855/0.8671 (0.0036)	0.8511/0.8365 (0.0021)	0.8691/0.8479 (0.0023)	0.9207/0.8902 (0.0028)	0.9478/0.8992 (0.0040)	**0.9487**/0.8992 (0.0026)	0.9268/**0.9109** (0.0020)
T2	MRI	0.8362/0.8235 (0.0022)	0.8459/0.8361 (0.0018)	0.8741/0.8498 (0.0011)	0.8802/0.8486 (0.0005)	**0.8897**/0.8441 (0.0015)	0.8714/0.8578 (0.0028)	0.8741/0.8420 (0.0026)	0.8871/0.8496 (0.0058)	0.8780/**0.8681** (0.0041)
	PET	**0.8912**/0.8361 (0.0017)	0.8367/0.8136 (0.0020)	0.8760/0.8666 (0.0024)	0.8802/0.8563 (0.0014)	0.8820/0.8574 (0.0009)	0.8630/0.8452 (0.0023)	0.8774/0.8114 (0.0028)	0.8698/0.8420 (0.0036)	0.8808/**0.8678** (0.0023)
	MRI + PET	0.8821/0.8559 (0.0008)	0.8678/0.8452 (0.0011)	0.9370/0.9111 (0.0024)	0.9547/0.9360 (0.0020)	0.9231/0.8990 (0.0016)	0.9127/0.8953 (0.0015)	0.9744/0.9125 (0.0021)	0.9587/0.9146 (0.0057)	**0.9710**/**0.9568** (0.0032)
T3	MRI	0.7265/0.6972 (0.0011)	0.7363/0.7083 (0.0019)	0.7674/0.7291 (0.0009)	0.7523/0.7187 (0.0031)	0.7358/0.7453 (0.0017)	0.7427/0.7361 (0.0016)	0.7669/0.7258 (0.0060)	0.7620/0.7198 (0.0025)	**0.7974**/**0.7782** (0.0022)
	PET	0.6868/0.6671 (0.0022)	0.7201/0.6863 (0.0030)	0.7196/0.7201 (0.0014)	0.7160/0.6698 (0.0021)	0.7360/0.7063 (0.0063)	0.7328/0.7258 (0.0022)	**0.8025**/0.7489 (0.0053)	0.7896/0.7256 (0.0027)	0.7698/**0.7487** (0.0037)
	MRI + PET	0.7485/0.7089 (0.0022)	0.7362/0.7258 (0.0021)	0.7902/0.7778 (0.0022)	0.7712/0.7452 (0.0025)	0.8125/0.7591 (0.0028)	0.8359/0.7639 (0.0018)	**0.8456**/0.7892 (0.0037)	0.8362/0.7743 (0.0050)	0.8370/**0.8220** (0.0028)
T4	MRI	0.7687/0.7412 (0.0009)	0.7801/0.7587 (0.0030)	0.8170/0.7933 (0.0030)	0.8257/0.8063 (0.0011)	0.8456/0.8145 (0.0027)	0.8467/0.8239 (0.0018)	0.8621/0.8147 (0.0063)	0.8685/0.8145 (0.0029)	**0.8784**/**0.8469** (0.0022)
	PET	0.7756/0.7587 (0.0020)	0.7610/0.7530 (0.0021)	0.8234/0.8034 (0.0026)	0.8631/0.8365 (0.0022)	0.8347/0.8082 (0.0013)	0.8412/0.8147 (0.0023)	0.8698/0.8009 (0.0051)	0.8510/0.7998 (0.0057)	**0.8759**/**0.8489** (0.0014)
	MRI + PET	0.7920/0.7752 (0.0021)	0.8210/0.7903 (0.0017)	0.8657/0.8370 (0.0022)	0.8971/0.8632 (0.0014)	0.8870/0.8569 (0.0017)	0.8930/0.8664 (0.0012)	0.9010/0.8369 (0.0027)	0.8894/0.8391 (0.0043)	**0.9112**/**0.8781** (0.0041)

Fig. 11 shows the average rankings of the Friedman test. We observe that the proposed method has the highest average ranking. The p-value is 0.001555, which indicates that there are significant differences among the 8 methods at the 5% significance level.

Furthermore, we use the Holm post-hoc test to further confirm which benchmarking methods have significant differences with the proposed method. The statistical results with the significance level $\alpha = 0.05$ are shown in Table 6.

We observe that Holm's test rejects those hypotheses that have an unadjusted p-value that is smaller or equal to 0.00833 3. Hence, there are significant differences between the FS-SVM, l1-SVM and RFF-MKL. Although Holm's test does not reject the hypothesis of Ref. [22], simpleMKL, Ref. [10], HID-TSK-FC and DBN-TSK-FC, the smaller p-value also indicates the competitiveness of the proposed model.

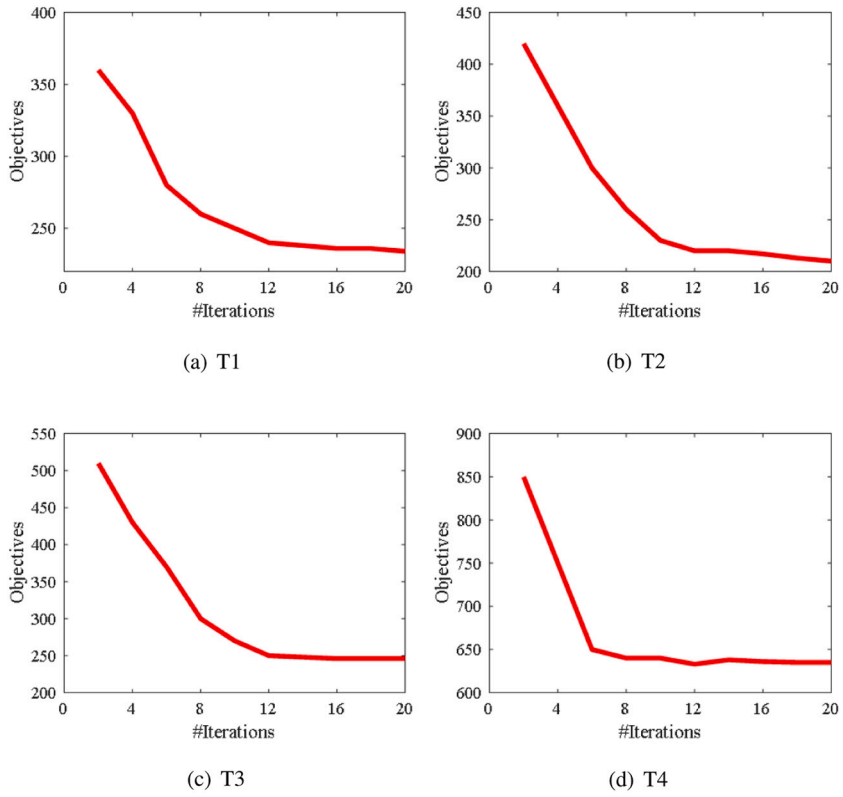


Fig. 9. Convergence behaviors of the proposed model on the top 20 selected features for all classification tasks.

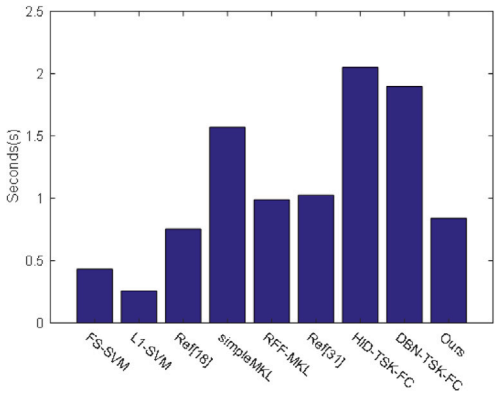


Fig. 10. Average CPU computation time in seconds for all models on "MRI+PET" in T1.

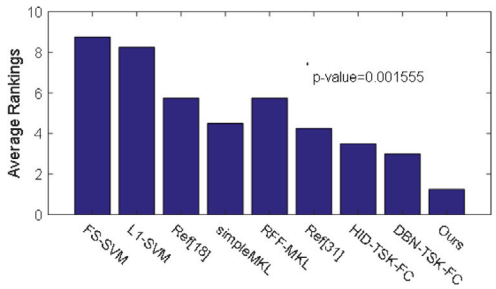


Fig. 11. Average rankings of the Friedman test.

Table 6
Holm post-hoc test results for our model vs. the benchmarking methods.

i	Algorithms	z	p	$Holm = \alpha/i$	Hypothesis
8	FS-SVM	3.872983	0.000108	0.00625	Rejected
7	l1-SVM	3.614784	0.000301	0.007143	Rejected
6	RFF-MKL	2.32379	0.020137	0.008333	Rejected
5	Ref. [22]	2.32379	0.020137	0.01	Not Rejected
4	simpleMKL	1.678293	0.09329	0.0125	Not Rejected
3	HID-TSK-FC	1.549193	0.121335	0.016667	Not Rejected
2	DBN-TSK-FC	1.161895	0.245278	0.025	Not Rejected
1	Ref. [10]	0.903696	0.366157	0.05	Not Rejected

5. Conclusions

In this study, by taking the multiclass support vector machine as the basic classifier, we proposed a novel multiclass classifier for Alzheimer's disease multiclass diagnosis. Our proposed method uses an embedding feature section strategy, i.e., a $l_{2,1}$-norm term combined with the multiclass hinge loss, to select discriminative features from different modalities. To perform complementary modality fusion, an l_p-norm ($1 < p < \infty$) term is introduced to combine different kernels that correspond to different modalities to avoid a sparse kernel coefficient distribution. We give a theorem to transform the minimization problem of the multiclass hinge loss with the $l_{2,1}$-norm and l_p-norm regularizations to a previous solvable optimization problem. Furthermore, we theoretically demonstrate that our optimization process converges to a global optimum. Our method is different from the similar studies in [30,33,40] and [10] in that it combines the $l_{2,1}$-norm and l_p-norm to yield joint structural sparsity to select features across all the classes in each modality and avoid a sparse kernel coefficient distribution, thereby effectively exploiting complementary modalities.

This study is not without limitations. For example, when the number of subjects is large, storing a kernel matrix requires a large amount of space. This limitation also exists in other kernel-based methods. As we know, the stochastic variance reduced gradient [50,51] is effective in MKL optimization. Therefore, in our future works, we will consider how to optimize the proposed method more efficiently.

CRediT authorship contribution statement

Yuanpeng Zhang: Methodology, Writing - original draft. **Shuihua Wang:** Investigation, Coding. **Kaijian Xia:** Data preprocessing, Analysis. **Yizhang Jiang:** Data preprocessing, Analysis. **Pengjiang Qian:** Supervision, Writing - review & editing.

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.

Data availability statement

The data used in this work are from public datasets: ADNI (http://adni.loni.usc.edu/). The access to these datasets is managed through secure LONI image and data archive (https://ida.loni.usc.edu/login.jsp) and contingent on adherence to the ADNI data use agreement and the publications' policies. To apply for the access to data please visit: http://adni.loni.usc.edu/data-samples/access-data/. As such, the investigators 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. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNIAcknowledgementList.pdf.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI), USA (National Institutes of Health Grant U01 AG024904) and DOD ADNI, USA (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A

The corresponding Lagrangian function of the optimization problem in (11) is defined as follows:

$$\begin{aligned} L = & \frac{1}{2}\sum_{k=1}^{K}\frac{1}{h^k}\left\|\mathbf{W}^k\right\|_2^2 \\ & + \zeta\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(1-y_{ic}\sum_{d=1}^{D^k}\sqrt{\theta_d^k}w_{dc}^k x_{di}^k\right)_+ \\ & + \frac{\rho}{p}\left(\sum_{k=1}^{K}\left(h^k\right)^p-1\right)-\sum_{k=1}^{K}v^k h^k \end{aligned} \tag{A.1}$$

where ρ and v^k are the multipliers that act on the constraints $\sum_{k=1}^{K}\left(h^k\right)^p \le 1$ and $h^k \ge 0$, respectively. To simplify the following mathematical derivations, we use a factor $1/p$ to rescale the multiplier ρ. At stationary points of the Lagrangian function L, we have the following KKT conditions:

$$-\frac{1}{2\left(h^k\right)^2}\left\|\mathbf{W}^k\right\|_2^2+\frac{\rho}{p}p\left(h^k\right)^{p-1}-v^k=0 \text{ from } \partial L/\partial h^k=0 \tag{A.2}$$

$$v^k h^k = 0 \text{ complementary slackness.} \tag{A.3}$$

From (A.3), we know that when $h^k \neq 0, v^k$ should be set to 0; additionally, due to (A.2) and the mild assumption we stated below (10), when $h^k = 0, v^k$ also should be set to 0. Therefore, we arrive at

$$\frac{1}{2}\left\|\mathbf{W}^k\right\|_2^2=\rho\left(h^k\right)^{p+1}, \text{ and hence} \tag{A.4}$$

$$h^k=\left(\frac{1}{2}\left\|\mathbf{W}^k\right\|_2^2\frac{1}{\rho}\right)^{\frac{1}{p+1}} \tag{A.5}$$

To find ρ, we further substitute (A.4) back to (A.1) and then arrive at

$$\begin{aligned} L = & \sum_{k=1}^{K}\rho\left(h^k\right)^p+\zeta\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(1-y_{ic}\sum_{d=1}^{D^k}\sqrt{\theta_d^k}w_{dc}^k x_{di}^k\right)_+ \\ & + \frac{\rho}{p}\left(\sum_{k=1}^{K}\left(h^k\right)^p\right)-\frac{\rho}{p} \\ = & \left(\frac{p+1}{p}\right)\sum_{k=1}^{K}\left(h^k\right)^p-\frac{\rho}{p} \\ & + \zeta\sum_{i=1}^{N}\sum_{c=1}^{c}\sum_{k=1}^{K}\left(1-y_{ic}\sum_{d=1}^{D^k}\sqrt{\theta_d^k}w_{dc}^k x_{di}^k\right)_+ \end{aligned} \tag{A.6}$$

$$\begin{aligned} L = & \left(\frac{p+1}{p}\right)\sum_{k=1}^{K}\left(\frac{1}{2}\left\|\mathbf{W}^k\right\|_2^2\right)^{\frac{p}{p+1}}\left(\frac{1}{\rho}\right)^{\frac{p}{p+1}} \\ & + \zeta\sum_{i=1}^{N}\sum_{c=1}^{c}\sum_{k=1}^{K}\left(1-y_{ic}\sum_{d=1}^{D^k}\sqrt{\theta_d^k}w_{dc}^k x_{di}^k\right)_+-\frac{\rho}{p} \\ = & \left(\frac{p+1}{p}\right)\rho^{\left(1-\frac{p}{p+1}\right)}\sum_{k=1}^{K}\left(\frac{1}{2}\left\|\mathbf{W}^k\right\|_2^2\right)^{\frac{p}{p+1}} \\ & + \zeta\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(1-y_{ic}\sum_{d=1}^{D^k}\sqrt{\theta_d^k}w_{dc}^k x_{di}^k\right)_+-\frac{\rho}{p}. \end{aligned} \tag{A.7}$$

By setting the partial derivative of L w.r.t. ρ to 0, i.e., $\partial L/\partial\rho = 0$, we have

$$\rho = \left[\sum_{k=1}^{K}\left(\frac{1}{2}\left\|\mathbf{w}^k\right\|_2^2\right)^{\frac{p}{p+1}}\right]^{\frac{p+1}{p}}. \tag{A.8}$$

By substituting (A.8) into (A.5), we complete the proof for Proposition 1.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.inffus.2020.09.002.

References

[1] F.J. Martinez-Murcia, J.M. Górriz, J. Ramírez, F. Segovia, D. Salas-Gonzalez, D. Castillo-Barnes, I.A. Illán, A. Ortiz, Alzheimer's Disease Neuroimaging Initiative, et al., Evaluating alzheimer's disease diagnosis using texture analysis, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2017, pp. 470–481.

[2] J.M. Górriz, J. Ramírez, F. Segovia, F.J. Martínez, M.-C. Lai, M.V. Lombardo, S. Baron-Cohen, M.A. Consortium, J. Suckling, A machine learning approach to reveal the neurophenotypes of autisms, Int. J. Neural Syst. 29 (07) (2019) 1850058.

[3] D. Castillo-Barnes, L. Su, J. Ramírez, D. Salas-Gonzalez, F.J. Martinez-Murcia, I.A. Illan, F. Segovia, A. Ortiz, C. Cruchaga, M.R. Farlow, et al., Autosomal dominantly inherited alzheimer disease: Analysis of genetic subgroups by machine learning, Inf. Fusion 58 (2020) 153–167.

[4] F.J. Martinez-Murcia, J.M. Górriz, J. Ramírez, A. Ortiz, A structural parametrization of the brain using hidden markov models-based paths in alzheimer's disease, Int. J. Neural Syst. 26 (07) (2016) 1650024.

[5] J. Munilla, A. Ortiz, J.M. Górriz, J. Ramírez, M.W. Weiner, P. Aisen, M. Weiner, R. Petersen, C.R. Jack Jr, W. Jagust, et al., Construction and analysis of weighted brain networks from sice for the study of alzheimer's disease, Front. Neuroinform. 11 (2017) 19.

[6] R.C. Petersen, P. Aisen, L.A. Beckett, M. Donohue, A. Gamst, D.J. Harvey, C. Jack, W. Jagust, L. Shaw, A. Toga, et al., Alzheimer's disease neuroimaging initiative (adni): clinical characterization, Neurology 74 (3) (2010) 201–209.

[7] P. Qian, Y. Chen, J.-W. Kuo, Y.-D. Zhang, Y. Jiang, K. Zhao, R. Al Helo, H. Friel, A. Baydoun, F. Zhou, et al., Mdixon-based synthetic ct generation for pet attenuation correction on abdomen and pelvis jointly using transfer fuzzy clustering and active learning-based classification, IEEE Trans. Med. Imaging 39 (4) (2019) 819–832.

[8] J.M. Górriz, J. Ramirez, J. Suckling, M.A. Consortium, et al., On the computation of distribution-free performance bounds: Application to small sample sizes in neuroimaging, Pattern Recognit. 93 (2019) 1–13.

[9] O.B. Ahmed, J. Benois-Pineau, M. Allard, G. Catheline, C.B. Amar, A.D.N. Initiative, et al., Recognition of alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning, Neurocomputing 220 (2017) 98–110.

[10] J. Peng, L. An, X. Zhu, Y. Jin, D. Shen, Structured sparse kernel learning for imaging genetics based alzheimer's disease diagnosis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 70–78.

[11] F.J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, D. Castillo-Barnes, Studying the manifold structure of alzheimer's disease: a deep learning approach using convolutional autoencoders, IEEE J. Biomed. Health Inform. 24 (1) (2019) 17–26.

[12] N. An, H. Ding, J. Yang, R. Au, T.F. Ang, Deep ensemble learning for alzheimer's disease classification, J. Biomed. Inform. (2020) 103411.

[13] H.-I. Suk, S.-W. Lee, D. Shen, A.D.N. Initiative, et al., Deep ensemble learning of sparse regression models for brain disease diagnosis, Med. Image Anal. 37 (2017) 101–113.

[14] R. Wang, H. Li, R. Lan, S. Luo, X. Luo, Hierarchical ensemble learning for alzheimer's disease classification, in: 2018 7th International Conference on Digital Home (ICDH), IEEE, 2018, pp. 224–229.

[15] T. Jo, K. Nho, A.J. Saykin, Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data, Front. Aging Neurosci. 11 (2019) 220.

[16] C. Salvatore, A. Cerasa, P. Battista, M.C. Gilardi, A. Quattrone, I. Castiglioni, Magnetic resonance imaging biomarkers for the early diagnosis of alzheimer's disease: a machine learning approach, Front. Neurosci. 9 (2015) 307.

[17] X. Liu, D. Tosun, M.W. Weiner, N. Schuff, A.D.N. Initiative, et al., Locally linear embedding (lle) for mri based alzheimer's disease classification, Neuroimage 83 (2013) 148–157.

[18] I. Beheshti, H. Demirel, A.D.N. Initiative, et al., Probability distribution function-based classification of structural mri for the detection of alzheimer's disease, Comput. Biol. Med. 64 (2015) 208–216.

[19] T.M. Nir, J.E. Villalon-Reina, G. Prasad, N. Jahanshad, S.H. Joshi, A.W. Toga, M.A. Bernstein, C.R. Jack Jr, M.W. Weiner, P.M. Thompson, et al., Diffusion weighted imaging-based maximum density path analysis and classification of alzheimer's disease, Neurobiol. Aging 36 (2015) S132–S140.

[20] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, E. Formisano, Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns, Neuroimage 43 (1) (2008) 44–58.

[21] C.-Y. Wee, P.-T. Yap, W. Li, K. Denny, J.N. Browndyke, G.G. Potter, K.A. Welsh-Bohmer, L. Wang, D. Shen, Enriched white matter connectivity networks for accurate identification of mci patients, Neuroimage 54 (3) (2011) 1812–1822.

[22] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A.D.N. Initiative, et al., Multimodal classification of alzheimer's disease and mild cognitive impairment, Neuroimage 55 (3) (2011) 856–867.

[23] Z. Dai, C. Yan, Z. Wang, J. Wang, M. Xia, K. Li, Y. He, Discriminative analysis of early alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (m3), Neuroimage 59 (3) (2012) 2187–2195.

[24] R. Polikar, C. Tilley, B. Hillis, C.M. Clark, Multimodal eeg, mri and pet data fusion for alzheimer's disease diagnosis, in: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, IEEE, 2010, pp. 6058–6061.

[25] K. Walhovd, A. Fjell, J. Brewer, L. McEvoy, C. Fennema-Notestine, D. Hagler, R. Jennings, D. Karow, A. Dale, A.D.N. Initiative, et al., Combining mr imaging, positron-emission tomography, and csf biomarkers in the diagnosis and prognosis of alzheimer disease, Amer. J. Neuroradiol. 31 (2) (2010) 347–354.

[26] X. Tang, Y. Qin, J. Wu, M. Zhang, W. Zhu, M.I. Miller, Shape and diffusion tensor imaging based integrative analysis of the hippocampus and the amygdala in alzheimer's disease, Magn. Reson. Imaging 34 (8) (2016) 1087–1099.

[27] F. Orabona, L. Jie, B. Caputo, Online-batch strongly convex multi kernel learning, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 787–794.

[28] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, J. Mach. Learn. Res. 12 (2011) 2211–2268.

[29] M. Varma, B.R. Babu, More generality in efficient multiple kernel learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 1065–1072.

[30] A. Rakotomamonjy, F.R. Bach, S. Canu, Y. Grandvalet, Simplemkl, J. Mach. Learn. Res. 9 (Nov) (2008) 2491–2521.

[31] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien, Lp-norm multiple kernel learning, J. Mach. Learn. Res. 12 (2011) 953–997.

[32] M. Kowalski, Sparse regression using mixed norms, Appl. Comput. Harmon. Anal. 27 (3) (2009) 303–324.

[33] K. Crammer, Y. Singer, On the learnability and design of output codes for multiclass problems, Mach. Learn. 47 (2–3) (2002) 201–233.

[34] P. Schiilkop, C. Burgest, V. Vapnik, Extracting support data for a given task, in: Proceedings, First International Conference on Knowledge Discovery & Data Mining, AAAI Press, Menlo Park, CA, 1995, pp. 252–257.

[35] X. Cai, F. Nie, H. Huang, C. Ding, Multi-class l2, 1-norm support vector machine, in: 2011 IEEE 11th International Conference on Data Mining, IEEE, 2011, pp. 91–100.

[36] J. Xu, F. Nie, J. Han, Feature selection via scaling factor integrated multi-class support vector machines, in: IJCAI, 2017, pp. 3168–3174.

[37] K. Crammer, Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines, J. Mach. Learn. Res. 2 (Dec) (2001) 265–292.

[38] M. Kloft, U. Brefeld, P. Laskov, K.-R. Müller, A. Zien, S. Sonnenburg, Efficient and accurate lp-norm multiple kernel learning, in: Advances in Neural Information Processing Systems, 2009, pp. 997–1005.

[39] L. Di Stefano, S. Mattoccia, A sufficient condition based on the cauchy-schwarz inequality for efficient template matching, in: Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429), Vol. 1, IEEE, 2003, pp. I–269.

[40] M. Szafranski, Y. Grandvalet, A. Rakotomamonjy, Composite kernel learning, Mach. Learn. 79 (1–2) (2010) 73–103.

[41] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.

[42] L. Meier, S. Van De Geer, P. Bühlmann, The group lasso for logistic regression, J. R. Stat. Soc. Ser. B-Stat. Methodol. 70 (1) (2008) 53–71.

[43] Y. Zhang, H. Ishibuchi, S. Wang, Deep takagi–sugeno–kang fuzzy classifier with shared linguistic fuzzy rules, IEEE Trans. Fuzzy Syst. 26 (3) (2017) 1535–1549.

[44] X. Zhang, F.-L. Chung, S. Wang, An interpretable fuzzy dbn-based classifier for indoor user movement prediction in ambient assisted living applications, IEEE Trans. Ind. Inform. 16 (1) (2019) 42–53.

[45] F. Liu, L. Zhou, C. Shen, J. Yin, Multiple kernel learning in the primal for multimodal alzheimer's disease classification, IEEE J. Biomed. Health Inform. 18 (3) (2013) 984–990.

[46] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[47] I. Kononenko, Estimating attributes: analysis and extensions of relief, in: European Conference on Machine Learning, Springer, 1994, pp. 171–182.
[48] J. Hodges, E.L. Lehmann, Rank methods for combination of independent experiments in analysis of variance, in: Selected Works of EL Lehmann, Springer, 2012, pp. 403–418.
[49] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. (1979) 65–70.
[50] F. Shang, Y. Liu, J. Cheng, J. Zhuo, Fast stochastic variance reduced gradient method with momentum acceleration for machine learning, 2017, arXiv preprint arXiv:1703.07948.
[51] M. Alioscha-Perez, M.C. Oveneke, J. Dongmei, H. Sahli, Multiple kernel learning via multi-epochs svrg, in: 9th NIPS Workshop on Optimization for Machine Learning, Vol. 12, 2016, pp. 1–5.