

Multimodal Self-Paced Locality-Preserving Learning for Diagnosis of Alzheimer's Disease

Xiaoke Hao¹, Member, IEEE, Ruxue Wang, Yingchun Guo², Yunjia Xiao, Ming Yu³, Member, IEEE, Meiling Wang⁴, Weibin Chen⁵, Member, IEEE, and Daoqiang Zhang⁶, Senior Member, IEEE

Abstract—Alzheimer's disease (AD) is an irreversible neurodegenerative disease that severely impairs human thinking and memory. The accurate diagnosis of AD and its prodromal stages, such as mild cognitive impairment (MCI), is very important for timely treatment or possible interventions of AD. Recent studies have shown that multiple neuroimaging and biological measures contain supplementary information for diagnosis and prognosis. Most existing methods are proposed to simply integrate the multimodal data and train the model using all samples once, which do not fully explore the structural information across the different modalities and ignore the significance of sample learning in the training process. In this article, we propose a multimodal self-paced locality-preserving learning (MSLPL) framework to preserve the inherent structural relationships of the original data and realize the sample selection process from "simple" to "complex." Specifically, the model can project the neuroimaging and genetic data into the label space and learn dimensionality reduction manners with preserving locality structure. Meanwhile, the contributions of each sample are adaptively evaluated by weighting optimization so that the impact of noises can be reduced during model training by self-paced learning (SPL). Finally, a multikernel support vector machine (MK-SVM) is used to fuse the features selected from different modalities for the final prediction. We evaluate MSLPL on 913 subjects from the AD neuroimaging initiative (ADNI) database with imaging and genetic data. The experimental results demonstrate that the proposed method can achieve better classification performances compared with the start-of-the-art multimodality-based methods.

Index Terms—Alzheimer's disease (AD), feature selection, locality preserving projection, multimodal learning, self-paced learning (SPL).

I. INTRODUCTION

ALZHEIMER'S disease (AD) is a common neurodegenerative and irreversible brain disease that affects the health of the elderly. It can cause the death of brain nerve cells and the loss of entire brain tissue in the later stage. The clinical manifestations of AD are mainly memory and cognitive dysfunction, reasoning dysfunction, and may be accompanied by language impairment and motor dysfunction. AD was first described in 1906, and many years later it is considered the main cause of death [1]. According to a survey report in 2017 from the Alzheimer's Association, by 2050, nearly one million new cases of AD will be diagnosed each year, which means that at that time one AD will be diagnosed every 33 s [2]. With the acceleration of population aging, the number of patients will continue to rise. Therefore, effective research on AD has positive significance for human physical and mental health and the sustainable development of society.

According to the development of cognitive models and the severity of functional impairment, the disease status can be divided into normal control (NC), mild cognitive impairment (MCI), and AD. As an intermediary between NC and AD and the early stage of AD [3], MCI refers to patients showing mild symptoms of brain malfunction. According to a report, within six years, 80% of MCI patients will change into AD [4]. AD cannot be cured, while early diagnosis and intervention of MCI may delay the deterioration of the disease, that is, it may reduce the risk of future dementia by managing controllable factors. Therefore, the detection and analysis of sensitive markers for early AD progression can help researchers develop new drugs and treatments to delay disease progression.

The imaging genetics research field integrates both the genetic factors and neuroimaging phenotypic measurements, hoping to reveal the genetic basis of brain structures and functionalities, and to explore the causal relationship between genetic variation and brain diseases such as AD. A large number of studies have shown that the diagnosis of AD and MCI is related to the structural atrophy of the brain, metabolic changes, and pathological amyloid deposition [5]. Neuroimaging technology is a powerful tool for diagnosing neurodegenerative diseases, which provides great potential for

Manuscript received 11 August 2021; revised 4 December 2021; accepted 25 June 2022. Date of publication 11 July 2022; date of current version 12 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61806071 and Grant 62136004; in part by the Natural Science Foundation of Hebei Province of China under Grant F2020202025; in part by the Natural Science Foundation of Zhejiang Province under Grant LY21F020001; in part by the Science and Technology Plan Project of Wenzhou, China, under Grant Y20180232 and Grant ZG2020026. Data Collection and Sharing for this Project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI). (Corresponding authors: Xiaoke Hao; Weibin Chen; Daoqiang Zhang.)

Xiaoke Hao, Ruxue Wang, Yingchun Guo, Yunjia Xiao, and Ming Yu are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China (e-mail: haoxiaoke@hebut.edu.cn).

Meiling Wang and Daoqiang Zhang are with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: dqzhang@nuaa.edu.cn).

Weibin Chen is with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China (e-mail: sun@wzu.edu.cn).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCDS.2022.3189701>.

Digital Object Identifier 10.1109/TCDS.2022.3189701

discovering features related to the early stages of dementia and helps to find the brain regions of interest (ROIs) corresponding to the AD. Among them, commonly used related brain imaging strategies include voxel-based morphometry measures (VBM), fluorodeoxyglucose positron emission tomography (FDG), and F-18 florbetapir PET scans amyloid imaging (AV45). At the same time, with the development of genetics technology, researchers can search for genetic markers related to neurological diseases and mental diseases from a more refined molecular level such as single nucleotide polymorphisms (SNPs). Furthermore, many studies have focused on the AD and MCI based on neuroimaging data and gene data, which play a crucial role in further analysis [6]–[8].

Recently, many machine learning methods have been proposed to analyze the neuroimaging and genetics of AD and MCI, such as sparse learning and classification [9], [10]. Some existing researches are more focused on feature extraction from a single modality, which cannot accurately detect the disease-related characteristic information for diagnosis because of the complexity of brain structure and function. In imaging genetics, data from different modalities can provide necessary and complementary information. Numerous studies have shown that fusion of multimodal data is more conducive to improving the accuracy of AD diagnosis than single modality [6]–[8], [11]–[17]. For example, Liu *et al.* [11] and Zhang *et al.* [17] combined VBM and FDG to analyze and diagnose AD. Zhu *et al.* [13] used VBM, FDG, and cerebrospinal fluid (CSF) to regress and classify AD. Hinrichs *et al.* [6] combined VBM, FDG, CSF, APOE genotype, and cognitive scores to diagnose and predict AD. Liu *et al.* [11] combined VBM, FDG, CSF, and genes to diagnose AD and MCI.

Although many multimodal existing methods have achieved good results, the problems that may limit the performance of diagnostic classification still exist. The features extracted from imaging genetics are often high dimensional and noisy, and there are many redundant or unrelated features. If these features are not handled well, the classification performance will be poor. Feature selection can effectively process features by reducing feature dimensionality and remove irrelevant features. Currently, many feature selection methods to AD classification and to detect the corresponding disease-related brain regions and genetic loci have been applied. For example, Liu *et al.* [18] used the locally linear embedding to convert the high dimensional data of local brain volume and cortical thickness into a local linear space with lower dimensions, which greatly improved the use of brain MRI to predict the performance of AD. Peng *et al.* [19] projected the original data into a new space and used L1p-norm to construct the sparse constraints objective function and for the diagnosis of AD. Shao *et al.* [20] used the hypergraph method to retain the high-order structure information among subjects, and induced a regularization term based on hypergraph to find some potential disease-related connections. Zhu *et al.* [21] proposed a method to embed linear discriminant analysis and locality preserving projections (LPP) into a unified framework to select more discriminative features for multiclass classification in AD diagnosis. Hao *et al.* [22] used the random forest strategy

to construct the similarity matrix among each modality and proposed a new multimodal neuroimaging feature selection method with consistency measurement constraints to analyze AD. Liu *et al.* [11] designed a diagnostic framework with deep learning model named SAE to help AD analysis. Xu *et al.* [23] developed a deep learning architecture named stacked sparse autoencoder (SSAE) to detect nuclei on images of breast cancer. Suk *et al.* [24] proposed a deep sparse multitask learning (DW-S2MTL) to recursively discard uninformative features by performing sparse multitask learning in a hierarchical manner. Shi *et al.* [25] developed a nonlinear metric learning method to improve biomarker identification for AD and MCI.

However, most feature selection methods ignore two important aspects.

- 1) *Intrinsic Structure Information Among Subjects*: In many real-world problems, how to consistently represent structural information among subjects is very important. Due to the types of different modalities that may be different, if pairwise or other simple metrics are used to represent the complex structural relationship among subjects, the topological structure information of the original data will be lost. Therefore, an effective representation of the complex structural relationships among subjects can induce more distinguishing features and further improve subsequent classification performance.
- 2) *Difference in Sample Significance*: There are random noise and systematic biases in different types of imaging genetics data, which not only affect the effectiveness and cost of research but also undermine the precise prediction of diseases [26]. In the process of multimodal fusion, it is unreasonable to regard all samples as uniform. Because these important samples will contribute more to the construction of the decision boundary of the classifier compared with insignificant samples and noise samples. Unfortunately, most multimodality-based methods do not distinguish between specialized and general knowledge, which cannot suppress the interference of noise samples and outliers on research results.

To address the above two problems, we propose a novel heterogeneous multimodal imaging genetics feature analysis method, which uses sample weighting and structured sparsity to analyze brain imaging and genetic data. In terms of estimating sample significance, self-paced learning (SPL) [27]–[31] is adopted to dynamically evaluate the learning difficulties of each sample to realize the automatic growth of samples from simple to complex in the training process. In the process of heterogeneous multimodal fusion, the analysis of the significance of samples helps characterize and explore the information across different modalities. Compared with insignificant samples and noise samples, these important samples contribute more to the establishment of the decision boundary classifier. Therefore, assigning different weights to samples through the SPL mechanism can suppress the interference of noise and outliers. In addition, under the framework of SPL, the LPP method is induced to effectively retain the intrinsic neighborhood structure of the subjects in the sample space. At the same time, the L1-norm constrained projection matrix is used as the regularization term to realize the

feature selection. Finally, we adopt the multikernel support vector machine (MK-SVM) to fuse the features selected from different modalities and make the final prediction. To validate the effectiveness of the proposed method, we perform experimental verification on 913 subjects from the AD neuroimaging initiative (ADNI) with imaging and genetic data. The results demonstrate that our proposed method can achieve better classification performance than the state-of-the-art methods.

II. RELATED WORK

A. Self-Paced Learning

In the research process of multimodal imaging genetics, it is unreasonable to assign all samples the same weight, because the noise samples or outliers contained in the dataset may interfere with the model. SPL is a sample learning mechanism that can effectively control the automatic growth of samples in the training process to suppress the interference of noise samples and make the model more robust. The development of SPL has a certain course. Observing that people usually first learn simple concepts and then learn complex knowledge, Bengio *et al.* [32] proposed curriculum learning (CL) to imitate this learning mechanism of humans. However, properly designed curriculums in CL are usually time consuming. Later, Kumar *et al.* [29] proposed the SPL model to alleviate the deficiency of CL. In contrast, SPL can adaptively select training samples from simple to complex by the feedback of learners without requiring predesigned curriculums. Specifically, SPL embeds CL as a regularization term into the model learning process.

The effectiveness and robustness of the SPL mechanism have been demonstrated in many research fields. Murugesan and Carbonell [33] proposed self-paced multitask learning that can build a shared knowledge base by starting from a set of simpler tasks and gradually introducing more difficult tasks. Gan *et al.* [34] used SPL regularization terms and L21-norm constraints as a new supervised learning method for feature selection to improve efficiency and stability. Yang *et al.* [35] proposed a multimodal SPL method (MSPL) based on multiomics data that can spontaneously identify multiomics data to join the training and improve the accuracy of predicting cancer subtypes. The high-confidence samples of each modality can be used to identify potentially important features, thereby improving the performance of the model. Xu *et al.* [36] proposed a multimodality-based framework by combining SPL and multimodal learning for image classification. Zhu *et al.* [37] proposed a method using self-paced sample weighting and rank minimization (SPMRM) to explore the inherent correlation information across different modalities and estimate the sample importance of AD data for the diagnosis of AD.

In SPL [29], [31], it should learn the model parameter \mathbf{w} and sample variable \mathbf{v} together through the following objective function:

$$\min_{\mathbf{w}, \mathbf{v}} \sum_{i=1}^n v_i l(\mathbf{w}; \mathbf{x}_i; y_i) + f(\mathbf{v}; k) \quad (1)$$

where \mathbf{x}_i denotes the i th sample, and its label is represented by y_i , $l(\mathbf{w}; \mathbf{x}_i; y_i)$ is used to calculate the loss between the truth label y_i and the estimated value. The sample weight variable to describe the importance of the sample is represented by $\mathbf{v} = [v_1, v_2, \dots, v_n]^T \in [0, 1]^n$, $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n]$ is used to represent the model parameter. The age parameter is represented by k , which controls the learning pace. Gradually increasing by k , SPL can automatically obtain more complex samples to joint training according to the learner's self-feedback during the training process.

In the model, the first term of (1) represents the sample loss discounted by weight, as the sample selection is dependent on the weight value of the sample. In the training process, the selected samples with weight parameters greater than 0 are defined as simple samples. Otherwise, treat them as complex samples that weight parameters are equal to 0.

B. Locality Preserving Projection

Imaging genetics data are usually high dimensional, and LPP can preserve the neighborhood structure of the original data by linearly projecting the original high-dimensional data into the low-dimensional space [38]. Therefore, in the research of imaging genetics, the use of LPP can achieve the purpose of linear dimensionality reduction and retain the neighborhood structure of the original data.

The given training set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbf{R}^{m \times n}$ has n samples and m modalities, $\mathbf{x}_i \in \mathbf{R}^m (i = 1, \dots, n)$ represents the i th training data, and the projected data representation $\mathbf{z}_i \in \mathbf{R}^{n \times 1}$ is $\mathbf{z}_i = \mathbf{P}^T \mathbf{x}_i$. The optimization problem of LPP is

$$\begin{aligned} \min_{\mathbf{P}} \quad & \frac{1}{2} \sum_{i,j} \|\mathbf{z}_i - \mathbf{z}_j\|^2 \mathbf{S}_{ij} \\ & = \min_{\mathbf{P}} \frac{1}{2} \sum_{i,j} \|\mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j\|^2 \mathbf{S}_{ij} \\ & = \min_{\mathbf{P}} \sum_i \mathbf{P}^T \mathbf{x}_i \mathbf{D}_{ii} \mathbf{x}_i^T \mathbf{P} - \sum_{i,j} \mathbf{P}^T \mathbf{x}_j \mathbf{S}_{ij} \mathbf{x}_j^T \mathbf{P} \\ & = \min_{\mathbf{P}} \text{tr}(\mathbf{P}^T \mathbf{X} (\mathbf{D} - \mathbf{S}) \mathbf{X}^T \mathbf{P}) \\ & = \min_{\mathbf{P}} \text{tr}(\mathbf{P}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P}) \end{aligned} \quad (2)$$

where the elements of the diagonal matrix \mathbf{D} are: $\mathbf{D}_{ii} = \sum_{j=1}^n \mathbf{S}_{ij}$, and $\mathbf{L} = \mathbf{D} - \mathbf{S}$, which is called the graph Laplacian matrix.

For the form of Gaussian kernel of the element \mathbf{S}_{ij} in similarity matrix \mathbf{S} is defined as follows:

$$\mathbf{S}_{ij} = \begin{cases} \exp\left\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right\}, & \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ in } \mathbf{K} - \text{neighborhood} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

For a given \mathbf{S} , the goal of LPP is to ensure that \mathbf{x}_i and \mathbf{x}_j are similar in the original space, then the projection values $\mathbf{z}_i = \mathbf{P}^T \mathbf{x}_i$ and $\mathbf{z}_j = \mathbf{P}^T \mathbf{x}_j$ are close to each other. After rigorous math transformations, the following generalized eigenvalue problem can easily solve the optimization model (2)

$$\mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P} = \lambda \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{P}. \quad (4)$$

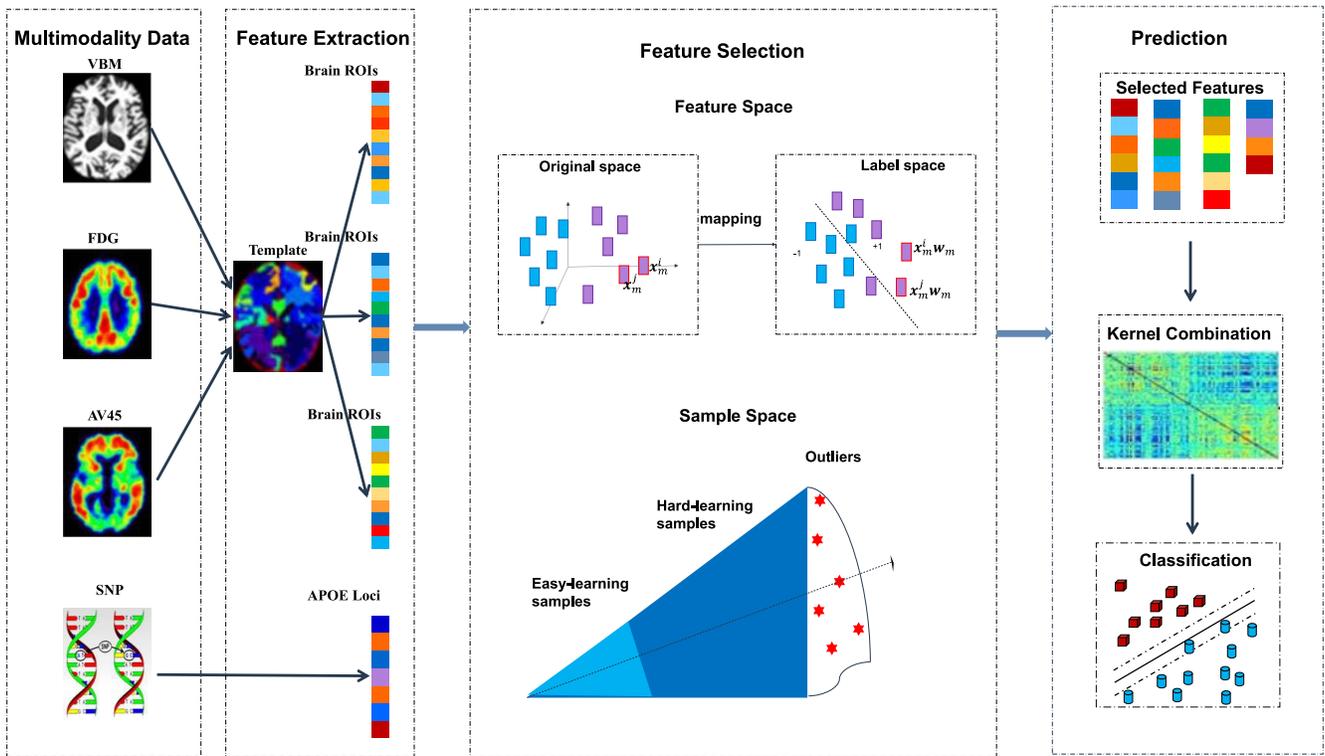


Fig. 1. Flowchart of the proposed method.

III. METHODOLOGY

This section introduces related work and the proposed multimodal self-paced locality-preserving learning (MSLPL). The objective function of the MSLPL model is proposed, and then an effective optimization method is presented to optimize it.

A. Proposed Framework

Fig. 1 illustrates the framework of the proposed method, which mainly includes four steps: 1) data preprocessing; 2) feature extraction; 3) feature selection; and 4) prediction. The innovation of this method is that the inherent structural information of the data and the significance of sample learning in the training process are fully utilized.

B. Subjects and Data Preprocessing

In this study, we perform experimental validation on the ADNI dataset. The ADNI is currently the authoritative data center for the study of AD. It is dedicated to collecting data on AD patients, and testing whether serial VBM, FDG, other biomarkers, and neuropsychological evaluation can be used in combination to explore the pathogenesis of AD and measure the progression of AD.

Imaging data and genetic data from 913 ADNI participants, which have VBM, FDG, AV45, and SNPs data, are collected, including 211 NC subjects, 82 significant memory concern (SMC) subjects, 273 early MCI (EMCI) subjects, 187 late MCI (LMCI) subjects, and 160 AD subjects. Table I lists the clinic and demographic characteristics of the subjects.

TABLE I
CLINIC AND DEMOGRAPHIC CHARACTERISTICS OF THE SUBJECTS

Subjects	NC	SMC	EMCI	LMCI	AD
Number	211	82	273	187	160
Gender(M/F)	190/101	33/49	153/119	108/79	95/65
Age	76.1±6.5	72.5±5.7	71.5±7.1	73.9±8.4	75.18±7.9
Education	16.4±2.6	16.8±2.7	16.1±2.6	16.4±2.8	15.86±2.8
MMSE	29.0±1.2	29.0±1.2	28.4±1.5	27.7±1.7	24.0±2.6
CDR	0.0±0.1	0.0±0.0	0.5±0.1	0.5±0.1	0.7±0.3

The values are expressed as mean ± standard deviation, MMSE = Mini-Mental State Examination, CDR = Clinical Dementia Rating, NC = Normal Control, SMC = Significant Memory Concern, EMCI = Early Mild Cognitive Impairment, LMCI = Late Mild Cognitive Impairment, AD = Alzheimer's Disease.

The diagnosis uses the standard guidelines specified in the ADNI manual. In short, NC participants did not have subjective based on memory decline and normal complaints. SMC participants used the cognitive change index (CCI) for the subjective memory attention, no indication of memory impairment, or decline based on the informant [39]. According to reports provided by subjects, informants, and clinicians, EMCI participants suffer from memory problems and the abnormal memory function is lower than the standard performance. This education level is adjusted according to the Wechsler logical memory delayed recall (LM-delayed), and the sum of the minimal state examination (MMSE) scores is greater than 24. In addition to subjective memory problems clarified in the reports provided by subjects, study partners, or clinicians, LMCI subjects must have memory box (MB) score of at least 0.5 and the clinical dementia rating (CDR) of 0.5;

AD's MMSE score should be in the range of 20–26 and the CDR should be 0.5 or 1.0.

For neuroimaging data, we use the same visit scan to align each participant's preprocessed isomorphic multimodal neuroimaging data (VBM, FDG, and AV45). Then, as $2 \times 2 \times 2 \text{ mm}^3$ voxels, in the standard Montreal Neurological Institute (MNI) space, we create a normalized gray matter density maps from VBM data, and register the FDG and AV45 scans through the SPM software package [40] to the same space. We further extract FDG glucose utilization and AV45 amyloid values based on the MarsBaR anatomical automatic labeling (AAL) atlas [40] of the measured average gray matter densities at 116 ROI levels. In our experiments, the neuroimaging measurements of each modality of 116 ROIs are regarded as quantitative traits to be used.

For the preprocessed genotyping data (SNPs) from the ADNI database, APOE is used as a risk gene located on chromosome 19 and related to neuron development, brain plasticity, and repair based on ANNOVR annotation. The 20-kbp SNPs of the APOE gene boundary are studied, which contains 85 SNPs. Each SNP value is coded as 0, 1, and 2 using an additive coding method to indicate the number of minor alleles as input for this study.

C. MSLPL Model

Multimodal data usually describe data from different perspectives, so compared to single-modal data, they provide supplementary information to each other to make the description more complete. An intuitive method is to select samples based on the inherent relationship among multimodalities. In this article, we propose MSLPL to build a distinguishing and robust model for AD diagnostic classification, which can extract the most important and informative feature information. Here, denote \mathbf{x}_m^i as the feature vector corresponding to the m th modal of the i th sample. Given the training set of the m th modality, $\mathbf{X}_m = [\mathbf{x}_m^1, \dots, \mathbf{x}_m^i, \dots, \mathbf{x}_m^n]^T \in \mathbf{R}^{n \times d}$, n is the number of samples, M is the number of modalities, d represents the dimension of the feature, and \mathbf{y}^i represents the label of each subject i , $\mathbf{y} = [\mathbf{y}^1, \dots, \mathbf{y}^i, \dots, \mathbf{y}^n]^T \in \mathbf{R}^n$ is the label vector of n samples. The objective function of MSLPL can be expressed as

$$\begin{aligned} \min_{\mathbf{w}_m, \mathbf{v}_m} & \sum_{m=1}^M \sum_{i=1}^n \mathbf{v}_m^i \|\mathbf{y}^i - \mathbf{x}_m^i \mathbf{w}_m\|_2^2 + \lambda \sum_{m=1}^M \|\mathbf{w}_m\|_1 \\ & + \frac{1}{2} \sum_{m=1}^M \sum_{i,j} \mu \|\mathbf{x}_m^i \mathbf{w}_m - \mathbf{x}_m^j \mathbf{w}_m\|_2^2 \mathbf{K}_m^{ij} + Mf(\mathbf{v}^i; k) \\ \text{s.t. } & 0 \leq \mathbf{v}_m^i \leq 1, i = 1, 2, \dots, n, m = 1, 2, \dots, M \end{aligned} \quad (5)$$

where \mathbf{w}_m denotes the weight vector of the m th modality, λ is the regularization parameter for constraining feature sparseness, and μ is the regularization parameter for constraining the relationships of samples for each modality. Each element \mathbf{K}_m^{ij} in the matrix represents the proximity relationship between the samples of the m th modality. If \mathbf{K}_m^{ij} is not equal to 0, it means that there is a \mathbf{K} neighbor relationship between the i th sample and j th sample, otherwise, it means that there does exist \mathbf{K}

Algorithm 1 MSLPL

Input: Training data $\{\mathbf{x}_m^1, \dots, \mathbf{x}_m^n\}_{m=1}^M$, label $\mathbf{y}^1, \dots, \mathbf{y}^n$, self-paced parameters k, k' , parameters λ, μ max_iter
Output: $\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{v}_1, \dots, \mathbf{v}_m$

- 1: Initialize $\mathbf{w}(0), \mathbf{v}(0), k, k', \lambda, \mu, \mathbf{P}$
- 2: iter=1
- 3: **while** iter \leq max_iter **do**
- 4: **for** $m \leftarrow \{1, \dots, M\}$ **do**
- 5: $(\mathbf{w}_1, \dots, \mathbf{w}_m)_t \leftarrow$ solution by Eq. (13).
- 6: update \mathbf{P}
- 7: $(\mathbf{v}_1, \dots, \mathbf{v}_m)_t \leftarrow$ solution by Eq. (19).
- 7: **end for**
- 8: iter \leftarrow iter + 1
- 9: **end while**
- 10: Return $\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{v}_1, \dots, \mathbf{v}_m$

neighbor relationship. The proximity relationship is described by the following formula:

$$\mathbf{K}_m^{ij} = \begin{cases} \exp\left\{-\frac{\|\mathbf{x}_m^i - \mathbf{x}_m^j\|_2^2}{2\sigma^2}\right\}, & \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ in } K - \text{neighborhood} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where the parameter σ should be 1 in the general case. \mathbf{K}_m is the weight matrix describing the neighbor relationship of sample points, $\mathbf{K}_m = [\mathbf{K}_m^1, \dots, \mathbf{K}_m^i, \dots, \mathbf{K}_m^n] \in \mathbf{R}^{n \times n}$.

In our model, using sample weighting and structure sparseness to analyze the characteristics of brain imaging and genetic data, we can not only preserve the structural relationship between the sample data but also consider the ‘‘difficulty’’ of the sample during the training process, and realize the automatic growth of the sample by embedding an SPL mechanism.

D. Optimization Algorithm

For all the variables, (5) is not jointly convex, but when the remaining variables are fixed, it is convex for each variable. Then, we adopt the alternative optimization strategy (AOS) algorithm to resolve the proposed MSLPL model, as listed in Algorithm 1.

Update \mathbf{w}_m by Fixing \mathbf{v}_m : For the objective function of (5), fixed \mathbf{v}_m , we can regard the fourth term of (5) as constants, thus we obtain

$$\begin{aligned} \min_{\mathbf{w}_m, \mathbf{v}_m} & \sum_{m=1}^M \sum_{i=1}^n \mathbf{v}_m^i \|\mathbf{y}^i - \mathbf{x}_m^i \mathbf{w}_m\|_2^2 + \lambda \sum_{m=1}^M \|\mathbf{w}_m\|_1 \\ & + \frac{1}{2} \sum_{m=1}^M \sum_{i,j} \mu \|\mathbf{x}_m^i \mathbf{w}_m - \mathbf{x}_m^j \mathbf{w}_m\|_2^2 \mathbf{K}_m^{ij}. \end{aligned} \quad (7)$$

In order to optimize and solve \mathbf{v}_m conveniently, we make the following transformation to the first term of (7)

$$\begin{aligned} \mathbf{Q}_m &= \left[\sqrt{\mathbf{v}_m^1} \mathbf{y}^1, \sqrt{\mathbf{v}_m^2} \mathbf{y}^2, \dots, \sqrt{\mathbf{v}_m^n} \mathbf{y}^n \right] \in \mathbf{R}^n \\ \mathbf{G}_m &= \left[\sqrt{\mathbf{v}_m^1} \mathbf{x}_m^1, \sqrt{\mathbf{v}_m^2} \mathbf{x}_m^2, \dots, \sqrt{\mathbf{v}_m^n} \mathbf{x}_m^n \right] \in \mathbf{R}^{n \times d}. \end{aligned}$$

Then, the first term of (7) can be transformed into

$$\begin{aligned} & \sum_{m=1}^M \sum_{i=1}^n v_m^i \|y^i - x_m^i w_m\|_2^2 \\ &= \sum_{m=1}^M \sum_{i=1}^n \left\| \sqrt{v_m^i} (y^i - x_m^i w_m) \right\|_2^2 = \sum_{m=1}^M \|Q_m - G_m w_m\|_2^2. \end{aligned} \quad (8)$$

For the third term of (7), we make it as:

$$\mathbf{D}_m = \text{diag} \left(\left[\sum_{j=1}^n \mathbf{K}_m^{1j}, \sum_{j=1}^n \mathbf{K}_m^{2j}, \dots, \sum_{j=1}^n \mathbf{K}_m^{nj} \right] \right) \in \mathbf{R}^{n \times n}.$$

Then, it can be transformed into

$$\begin{aligned} & \frac{1}{2} \sum_{m=1}^M \sum_{i,j} \mu \|x_m^i w_m - x_m^j w_m\|_2^2 \mathbf{K}_m^{ij} \\ &= \mu \sum_{m=1}^M \sum_{i,j} \left(w_m^T (x_m^i)^T \mathbf{K}_m^{ij} x_m^j w_m - w_m^T (x_m^i)^T \mathbf{K}_m^{ij} x_m^j w_m \right) \\ &= \mu \sum_{m=1}^M \left(\sum_{i=1}^n w_m^T (x_m^i)^T \mathbf{D}_m^{ii} x_m^i w_m - \sum_{i,j} w_m^T (x_m^i)^T \mathbf{K}_m^{ij} x_m^j w_m \right) \\ &= \mu \sum_{m=1}^M \text{tr} \left(w_m^T \mathbf{X}_m^T \mathbf{D}_m \mathbf{X}_m w_m - w_m^T \mathbf{X}_m^T \mathbf{K}_m \mathbf{X}_m w_m \right) \\ &= \mu \sum_{m=1}^M \text{tr} \left(w_m^T \mathbf{X}_m^T \mathbf{L}_m^h \mathbf{X}_m w_m \right) \end{aligned} \quad (9)$$

where \mathbf{L}_m^h represents the hypergraph Laplacian matrix of the m th modality, $\mathbf{L}_m^h = \mathbf{D}_m - \mathbf{K}_m$. At this time, the target (7) is transformed into

$$\begin{aligned} & \min_{w_m} \sum_{m=1}^M \|Q_m - G_m w_m\|_2^2 + \mu \sum_{m=1}^M \text{tr} \left(w_m^T \mathbf{X}_m^T \mathbf{L}_m^h \mathbf{X}_m w_m \right) \\ & \quad + \lambda \sum_{m=1}^M \|w_m\|_1 \end{aligned} \quad (10)$$

where w_m^i is the i th row of the vector w_m . We define a matrix \mathbf{P}_m

$$\mathbf{P}_m^{ii} = \frac{1}{2|w_m^i|}. \quad (11)$$

Then, we can get $2\text{tr}(w_m^T \mathbf{P}_m w_m) = \|w_m\|_1$. Equation (7) is further transformed into

$$\begin{aligned} & \min_{w_m} \sum_{m=1}^M (Q_m - G_m w_m)^T (Q_m - G_m w_m) \\ & \quad + \mu \sum_{m=1}^M \text{tr} \left(w_m^T \mathbf{X}_m^T \mathbf{L}_m^h \mathbf{X}_m w_m \right) + \lambda \sum_{m=1}^M \text{tr} \left(w_m^T \mathbf{P}_m w_m \right). \end{aligned} \quad (12)$$

Taking the derivative of w_m in (10) and setting the derivative to be 0, we can obtain

$$w_m = \left(G_m^T G_m + \mu \mathbf{X}_m^T \mathbf{L}_m^h \mathbf{X}_m + \lambda \mathbf{P}_m \right)^{-1} G_m^T Q_m. \quad (13)$$

Update v_m by Fixing w_m : In this step, we can obtain the best weight of the current sample in the m th modality. Its physical meaning is to select the confidence samples (v_m^i) to be used for training of the m th modality through the correlation between multiple modalities. For the objective function of (5), fixed w_m , we can regard the second and third term of (5) as constants, so we obtain

$$\begin{aligned} & \min_{v_m} \sum_{m=1}^M \sum_{i=1}^n v_m^i \|y^i - x_m^i w_m\|_2^2 + Mf(v^i; k) \\ & \text{s.t. } 0 \leq v_m^i \leq 1, i = 1, 2, \dots, n, m = 1, 2, \dots, M. \end{aligned} \quad (14)$$

In terms of the self-paced function, the hard weighting method we first used is as follows:

$$f(v^i; k) = -\frac{1}{k} \|v\|_1 = -\frac{1}{k} \sum_{i=1}^n v^i \quad (15)$$

where the elements of the diagonal matrix \mathbf{v} are $\{v^1, v^2, \dots, v^n\}$, $\mathbf{v} \in [0, 1]^n$. Substituting (15) into (14), the solution of v^i can be calculated as

$$v^i = \begin{cases} 1, & \frac{1}{M} \sum_{m=1}^M l_m^i < \frac{1}{k} \\ 0, & \frac{1}{M} \sum_{m=1}^M l_m^i = \frac{1}{k} \end{cases} \quad (16)$$

where l_m^i represents the loss function: $\sum_{m=1}^M \sum_{i=1}^n v_m^i \|y^i - x_m^i w_m\|_2^2$, i represents the i th sample and m represents the m th modality. k is the SPL parameter, which is used to spontaneously select samples participating in training during the learning process. At the beginning, only simple samples with smaller losses are selected. As k decreases, more samples with larger losses will be gradually added to the training. In this way, we can effectively avoid inducing noises or outliers in feature selection.

Equation (16) implements the ‘‘hard’’ sample selecting method in the form of binary weights. However, it may determine whether to choose these samples overconfidently. In contrast, the soft threshold weight assigns a continuous value from 0 to 1 (including 0 and 1) to each sample, which may reflect the potential importance of the training sample more flexible. Based on this mechanism, the mixed weighting scheme, a combination of the soft weighting and the binary weighting scheme can be adopted. Especially, if the loss is too small or too large, binary weighting is applied, otherwise, soft weighting will be applied. The definition is as follows:

$$f(v^i; k) = -\xi \sum_{i=1}^n \log(v^i + \xi k) \quad (17)$$

where $\xi = [1/(k' - k)]$, k' represents an auxiliary parameter in the range $k' > k > 0$. Then, the derivation of v^i in (17) is

$$\frac{\partial l}{\partial v^i} = \sum_{m=1}^M l_m^i - \frac{M\xi}{v^i + k\xi}. \quad (18)$$

According to (18), the closed-form solution of v^i is as follows:

$$v^i = \begin{cases} 1, & \frac{1}{M} \sum_{m=1}^M l_m^i \leq \frac{1}{k'} \\ 0, & \frac{1}{M} \sum_{m=1}^M l_m^i \geq \frac{1}{k} \\ \frac{M\xi}{\sum_{m=1}^M l_m^i} - k\xi, & \text{otherwise.} \end{cases} \quad (19)$$

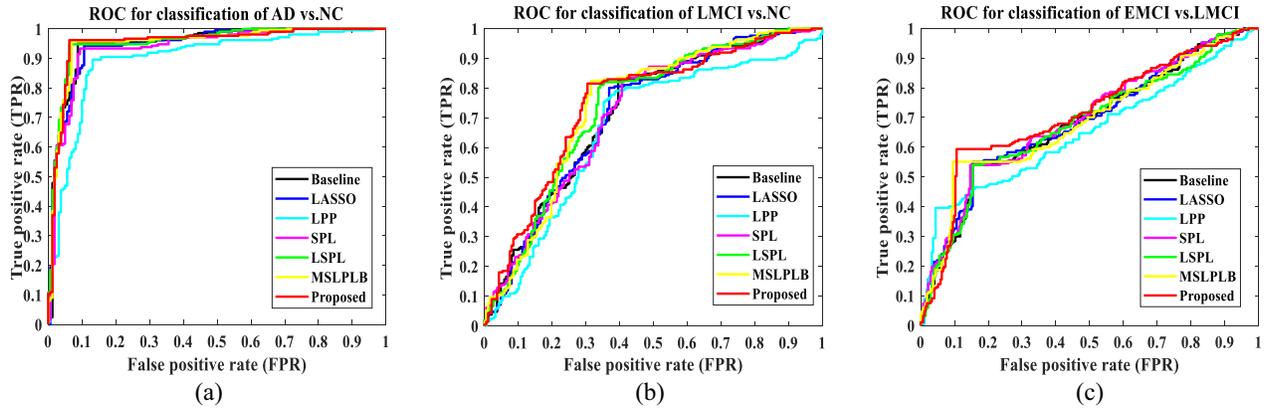


Fig. 2. ROC curves of all comparisons. (a) Classification of AD versus NC. (b) Classification of LMCI versus NC. (c) Classification of EMCI versus LMCI. The horizontal axis represents the false positive rate; the vertical axis represents the true positive rate. The area under the curve (AUC) indicates the diagnosis power.

E. Classification

Following the previous work [17], [41], the MK-SVM is used to make the final prediction after feature selection. For a specific training set, we calculate the kernel matrix of the m th modality of subjects \mathbf{x}_m^i and \mathbf{x}_m^j by $k_m(\mathbf{x}_m^i, \mathbf{x}_m^j) = \phi(\mathbf{x}_m^i)^T \phi(\mathbf{x}_m^j)$. Linear kernel fusion of multimodal data through $k(\mathbf{x}^i, \mathbf{x}^j) = \sum_{m=1}^M \rho_m k_m(\mathbf{x}_m^i, \mathbf{x}_m^j)$, where ρ_m represents the combination coefficient of the m th modality. Therefore, we can get the following dual form of the MK-SVM:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \sum_{m=1}^M \rho_m k_m(\mathbf{x}_m^i, \mathbf{x}_m^j) \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (20)$$

where α_i is the Lagrange multiplier of the i th sample. In this article, with the help of the LIBSVM toolbox [42], the SVM classifier can be readily solved. To find the fusion coefficient ρ_m with the best classification effect, we use a grid search in range $[0, 1]$ by cross-validation on the training set.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

Due to the limited number of subjects, we use 10-fold cross-validation to reduce bias and evaluate the classification performance by averaging the results of different sets of testing. We design three tasks of experiments to verify the effectiveness, including AD versus NC, LMCI versus NC, and EMCI versus LMCI. The dataset used in the experiment is 913 ADNI participants with VBM, FDG, AV45, and SNP modalities. We only use the data of the NC, EMCI, LMCI, and AD patients. In addition, the number of dimensionality of each neuroimaging modality (VBM, FDG, or AV45) is 116, which corresponds to 116 ROIs, and the number of dimensionality of genetic modality is 85, which corresponds to 85 SNP loci. The experimental evaluation indicators are classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and the area under the receiver operating characteristic curve (AUC).

Our proposed MSLPL method is compared with several existing diagnosis methods in AD classification, including: 1) baseline, which is a multikernel method [17] without performing feature selection; 2) LASSO, which is a multi-kernel method with LASSO [43] feature selection; 3) LPP, which is a multikernel method with LPP [44] feature selection; and 4) SPL, which is a multikernel method with SPL mechanism. The feature selection method with the L1-norm under the framework of SPL (denote as LSPL) and the MSLPL feature selection with binary weighting (denote as MSLPLB) are also compared. In the experiments shown above, SVM with a linear kernel is used for classification. In all methods, the regularization parameters λ and μ are chosen in range $\{10^{-5}, 10^{-4}, \dots, 10^0\}$ and $\{10^{-5}, 10^{-4}, \dots, 10^1\}$, respectively.

B. Classification Performance

Table II illustrates the detailed experimental results on the ADNI dataset. The ROC curves of the proposed method and all comparison methods are shown in Fig. 2. As observed, the proposed method consistently outperforms other methods in all tasks. Specifically, the classification accuracy of our proposed method in three tasks is 95.14%, 82.85%, and 76.91%, respectively. In addition, the highest AUC values that our proposed method have obtained are 0.95, 0.75, and 0.71, respectively. These experimental results prove the effectiveness of the proposed method and its better diagnostic performance.

In summary, it is noting that our method can achieve higher sensitivity than other methods in most cases. Relatively speaking, it is very important to obtain high sensitivity in the process of disease diagnosis. The difference between the sensitivity and specificity is large in our experiments, for instance, the sensitivity is relatively high for each method but specificity is low somewhat. The cost of misclassification between patients and NCs is different in medical diagnosis. Obviously, it is more expensive to misclassify patients as NCs and incorrect diagnosis results may delay treatment of patients. Such high sensitivity is beneficial for timely detection of the disease and effective treatment.

TABLE II
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS. (a) AD
VERSUS NC. (b) LMCI VERSUS NC. (c) EMCI VERSUS LMCI

(a)				
Method	ACC	SEN	SPE	AUC
Baseline	93.52±3.66	95.26	91.25	0.95
LASSO	90.84±3.45	94.31	86.25	0.94
LPP	88.15±4.46	89.10	86.88	0.90
SPL	92.18±4.50	93.36	90.63	0.94
LSPL	94.33±3.24	94.79	93.75	0.94
MSLPLB	94.60±2.86	95.73	93.13	0.95
Proposed	95.14±3.07	96.21	93.75	0.95
(b)				
Method	ACC	SEN	SPE	AUC
Baseline	78.33±7.38	87.20	68.45	0.72
LASSO	78.59±7.17	85.31	71.12	0.71
LPP	76.11±3.72	84.83	66.31	0.67
SPL	76.31±7.73	86.26	65.24	0.71
LSPL	79.84±7.19	85.31	73.80	0.73
MSLPLB	82.36±7.52	86.73	77.54	0.74
Proposed	82.85±5.89	86.73	78.61	0.75
(c)				
Method	ACC	SEN	SPE	AUC
Baseline	72.39±5.58	84.98	54.01	0.68
LASSO	72.61±4.57	84.62	55.08	0.69
LPP	72.39±1.70	95.60	38.50	0.65
SPL	72.61±4.40	85.35	54.01	0.69
LSPL	72.17±6.64	84.62	54.01	0.68
MSLPLB	76.06±4.50	90.48	55.08	0.69
Proposed	76.91±3.41	89.01	59.36	0.71

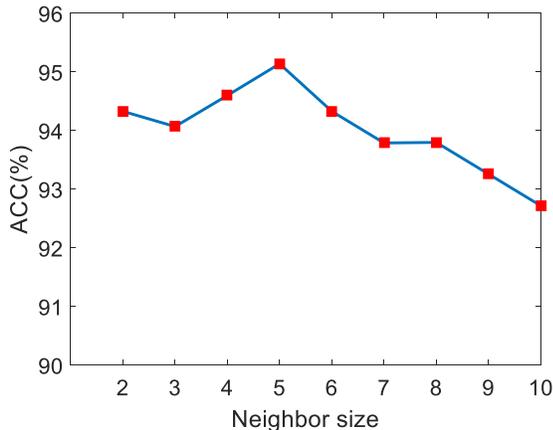


Fig. 3. Classification results with different neighbor size of KNN.

C. Influence of Neighbor Size

In this article, the KNN method is used to find the K -nearest neighbor relationship among the sample points. The number of neighbors in KNN will have an impact on the classification results of the proposed method. Fig. 3 shows the experimental result, where the neighbor size range is $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. As we can see from Fig. 3, the classification accuracy reaches the highest when the number of neighbors is 5. However, the classification accuracy drops

sharply when the number of neighbors exceeds 5. The potential cause may be that when the number of neighbors is large, the sketched K -nearest neighbor relationship will contain subjects of different classes and the description is more inclined to the global structure, which may not reflect the real data structure.

D. Classification Based on Multimodal Data

A large number of studies shows that different modalities in imaging genetics can provide supplementary information for assisting the identification of AD [8], [45], [46]. It is reported that different modalities of information fusion can enhance diagnostic performance. Different methods of fusing biomarkers have been proposed to produce more powerful classifiers [6], [17], [47]. There are many ways to combine multimodal data. For example, Kohannim *et al.* [48] directly concatenated features of multimodal data into a long vector and then used SVM for AD classification. Gray *et al.* [7] used multiple random forests to fuse and classify AD-related multimodal data. In addition, the multimodal classification method that uses multiple classifiers to vote to obtain the final result is a common ensemble learning strategy, but it may introduce bias due to using multiple modalities. The kernel-based method is also a common way to merge different modalities such as multikernel learning [17]. In the multikernel learning, we first calculate a separate kernel matrix for each modality, and merge the kernel functions through their linear combination. A large number of results in the previous studies indicate that multikernel learning can achieve better performance.

Single modality and multiple modalities for a fair comparison are used to evaluate the effectiveness of multimodal classification. We perform experiments on imaging data (including VBM, FDG, and AV45) or gene data (i.e., SNPs), or their combination. Table III and Fig. 4 show the corresponding results of multimodal classification. As we can see, the classification performance of the proposed method is getting better with the increase of the number of modalities. The comparison results show that multimodal data can provide supplementary information and the proposed method has better classification performance in multiple modalities compared with the single modality.

E. Parameter settings

We set two regularization parameters (i.e., λ, μ) in the objective function to balance the relative contribution of the locality preserving regularization and sparsity regularization. In this section, the impact of the two regularization parameters involved in the proposed method on the classification performance is discussed. Specifically, we vary λ in range $\{10^{-5}, 10^{-4}, \dots, 10^0\}$ and μ in range $\{10^{-5}, 10^{-4}, \dots, 10^1\}$. Fig. 5 shows the corresponding results. It can be seen from the Fig. 5 that our proposed method fluctuates slightly when varying the parameters λ and μ , which shows that our proposed method is not particularly sensitive to parameter values.

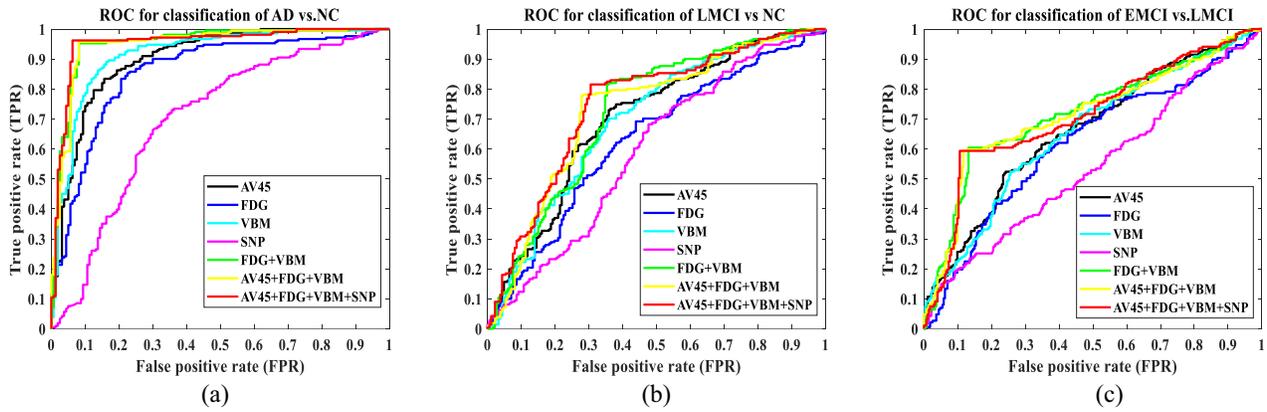


Fig. 4. ROC curves with different modalities. (a) Classification of AD versus NC. (b) Classification of LMCI versus NC. (c) Classification of EMCI versus LMCI. The horizontal axis represents the false positive rate; the vertical axis represents the true positive rate. AUC indicates the diagnosis power.

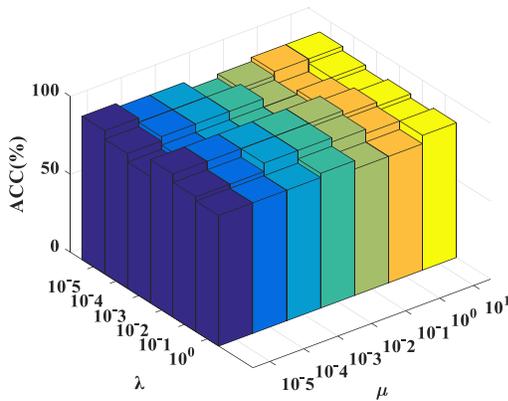


Fig. 5. Accuracy of AD versus NC classification with respect to different parameter values (λ in range $\{10^{-5}, 10^{-4}, \dots, 10^0\}$ and μ in range $\{10^{-5}, 10^{-4}, \dots, 10^1\}$). The x -axis and y -axis represent the diverse value of parameters and the z -axis represents the classification accuracy for AD diagnosis.

F. Algorithm Comparison

To further verify the superiority of the proposed method, some of the latest results reported in the literature based on the multimodal imaging data for AD classification are compared. Specifically, as the dimensions of the genetic data and the imaging data are inconsistent, the dimension alignment is required when solving the group sparsity. The optimal weight matrix consists of the feature weighted column vectors for each modality. Therefore, we set the hypergraph method [20] and MFCC method [22] only on multimodal imaging data (i.e., including VBM, FDG, and AV45) as fair comparisons. The comparative experimental results based on VBM, FDG, and AV45 are shown in Table IV.

As we can see, our model achieves the best results on the evaluation indicators of the classification accuracy and AUC value. In particular, our proposed method achieves an accuracy of 94.06% and an AUC value of 0.95. The hypergraph method induces a regularization term based on hypergraph to preserve the high-order relationship among subjects. Although this method takes into account the inherent high-order structural relationship among samples, due to the “difficulty” of the

TABLE III
CLASSIFICATION PERFORMANCE WITH DIFFERENT MODALITIES. (a) AD VERSUS NC. (b) LMCI VERSUS NC. (c) EMCI VERSUS LMCI

(a)				
Method	ACC	SEN	SPE	AUC
AV45	83.28±4.76	87.20	78.13	0.90
FDG	81.68±3.66	87.68	73.75	0.86
VBM	85.97±6.69	89.57	81.25	0.92
SNP	68.74±4.73	74.41	61.25	0.70
VBM+FDG	93.80±3.14	95.26	91.88	0.95
VBM+FDG+AV45	94.06±3.08	95.73	91.88	0.95
VBM+FDG+AV45+SNP	95.14±3.07	96.21	93.75	0.95
(b)				
Method	ACC	SEN	SPE	AUC
AV45	70.57±4.76	75.83	64.71	0.70
FDG	65.02±8.13	71.56	57.75	0.63
VBM	69.79±6.55	75.83	63.10	0.70
SNP	60.33±7.20	76.78	41.71	0.59
VBM+FDG	82.36±7.17	86.73	77.54	0.74
VBM+FDG+AV45	82.58±6.22	83.89	81.28	0.73
VBM+FDG+AV45+SNP	82.85±5.89	86.73	78.61	0.75
(c)				
Method	ACC	SEN	SPE	AUC
AV45	64.70±6.87	77.29	46.52	0.66
FDG	62.16±3.91	74.73	43.85	0.62
VBM	63.25±7.37	76.92	43.32	0.66
SNP	62.15±2.35	95.60	13.37	0.54
VBM+FDG	75.83±4.55	86.81	59.89	0.72
VBM+FDG+AV45	76.51±3.49	88.28	59.36	0.72
VBM+FDG+AV45+SNP	76.91±3.41	89.01	59.36	0.71

samples is different during the training process, all samples are regarded as uniform and easy to learn, which may increase the impact of the noise value on the model, thus affecting the classification accuracy. The MFCC method uses random forest strategy to calculate the similarity for each modality separately to extract pairwise similarity measures for multiple modalities. While this method only takes into account the paired similarity relationship rather than the similarity among

TABLE IV

CLASSIFICATION PERFORMANCE WITH DIFFERENT METHODS BASED ON VBM, FDG, AND AV45. (a) AD VERSUS NC. (b) LMCI VERSUS NC. (c) EMCI VERSUS LMCI

(a)				
Method	ACC	SEN	SPE	AUC
hypergraph	92.45±3.66	94.31	90.00	0.95
MFCC	93.53±3.18	95.21	90.00	0.95
Proposed	94.06±3.08	95.73	91.88	0.95
(b)				
Method	ACC	SEN	SPE	AUC
hypergraph	78.83±8.82	85.78	71.12	0.73
MFCC	78.57±8.23	83.41	73.26	0.72
Proposed	82.58±5.89	83.89	81.28	0.73
(c)				
Method	ACC	SEN	SPE	AUC
hypergraph	72.81±5.65	87.55	51.34	0.71
MFCC	73.45±6.21	88.65	51.34	0.70
Proposed	76.51±3.49	88.28	59.36	0.72

many samples, which can also provide effective information for disease diagnosis.

The MSLPL algorithm is compared with the state-of-the-art methods using multimodal data as shown in Table V, including self-paced related methods and machine learning methods. To verify the reliability of the results, our experimental dataset and processing framework are consistent with previous work [39], [49] in the literature.

In this comparison, the accuracy of our method is higher than other SPL methods [38], [49] in AD versus NC classification. An important reason may be that our proposed method can make full use of the local structure information in the data. Due to the objective function that introduces local similarity constraints and specificity constraints between different samples, the features selected using our proposed method are more informative and discriminative. In particular, the proposed method achieves better classification accuracy than some deep learning methods [11], [24], [25], [50], [51], such as DW-S2MTL [24] and Dropout-DL [50]. One essential reason may be that our model considers the difference of sample significances and suppress the interference of noise samples and outliers on the model. As the objective function is induced by the SPL mechanism, our proposed method will automatically select training samples during each iteration to reduce the impact of noise samples and improve the robustness of the model. In addition, the ability of the traditional hand-draft features from the candidate brain regions of pathogen may be stronger than deep feature representation when the number of training samples is very limited.

V. DISCUSSION AND LIMITATIONS

The novel approach proposed in this article is to solve two issues, including: 1) feature selections of brain ROIs and SNPs related to AD and 2) diagnosis of AD. We perform all classification experiments on the ADNI dataset to prove the validity and reliability of the MSLPL method. The results show that

the model cannot only use the supplementary information in the multimodal imaging genetics data to classify AD but also help to discover biomarkers related to disease for studying the pathogenesis of AD.

It is very important to determine the ROIs and SNPs related to the disease. According to the experimental results of AD versus NC classification, we count that the top ten regions, which are most frequently selected, and regard them as the most relevant markers to the pathogenesis of AD. Table VI show the top ten regions detected from the VBM, FDG, and AV45 in the template space. As we can see, most selected ROIs, such as *Hippocampus*, *Temporal Pole*, and *Amygdala*, have proved to be relevant to AD. According to reports, the *Hippocampus* is closely related to declarative memory [52] and *Amygdala* plays an important role in remembering emotionally significant experiences [53]. Besides, we count the top ten SNPs that are most relevant to the pathogenesis of AD, as shown in Table VII. As we can see, the notable AD risk markers, such as *rs66626994*, *rs10119*, *rs111789331*, *rs10414043*, and *rs7256200*, have been reported to be associated with the pathogenesis of AD in previous studies [54]–[56]. This indicates the MSLPL has the ability to identify meaningful SNPs from massive genetic markers.

Despite its promising performance, our proposed method still has some limitations. First, we only considered the specificity of the different modalities of imaging data and genetic data and ignored the consistent ROI feature selections from the three modalities of imaging data, including VBM, FDG, and AV45. We can try to add the consistency constraints of imaging data in the later work to ensure that the same brain regions from different modalities are selected simultaneously. Second, due to the limited number of subjects, we used 10-fold cross-validation to verify the effectiveness of our proposed method. We hope to collect massive amounts of data from other sites in the future, and then do some independent site validations in future work to verify the generalization and scalability of our proposed model. Finally, we did not test the multiclass classification performance and only studied AD-related binary classification problems, which is very valuable for accurately diagnosing different stages of the disease.

VI. CONCLUSION

This study proposes a new self-paced local preservation projection feature selection method for AD diagnosis. The proposed method can explore the supplementary information provided by multimodal imaging genetics data to select discriminative features related to disease and make the further prediction. In summary, our model induces three novel aspects: 1) using the locality preserving projection to retain the inherent structural relationship of AD data; 2) using the SPL mechanism to adaptively evaluate the significance of samples in the feature selection model; and 3) using sparse regularization terms to constrain the specificity of different modalities of imaging genetics. The experimental results on the ADNI dataset demonstrate that benefiting from the above three aspects, our proposed method achieves better performances compared with the existing state-of-the-art methods.

TABLE V
COMPARISON OF THE PERFORMANCE OF DIFFERENT MULTIMODAL CLASSIFICATION ALGORITHMS ON AD VERSUS NC

Algorithms	Subjects	Modalities	ACC	SEN	SPE	AUC
MKL [17]	51AD, 43MCI-C, 56MCI-NC, 52NC	FDG+ VBM +CSF	93.20	93.00	93.30	0.97
RFSM [7]	37AD, 75MCI, 35NC	FDG + VBM +CSF+genetic	89.00	87.9	90.00	--
SAE [11]	85AD, 67MCI-C, 102 MCI-NC, 77 NC	FDG + VBM	91.35	92.32	90.42	--
SPMRM [37]	160AD, 272EMCI, 187LMCI, 210NC	FDG+VBM+AV45	88.02	94.14	80.00	0.97
DW-S ² MTL [24]	51AD, 43MCI-C, 56MCI-NC, 52NC	FDG +VBM +CSF	95.09	92.00	98.00	--
Dropout-DL [50]	51AD, 43MCI-C, 56MCI-NC, 52NC	FDG +VBM+CSF	91.40	--	--	--
SDSAE [25]	94AD, 121MCI, 123NC	Longitudinal MRI	91.95	89.49	93.82	--
SPLRR [49]	160AD, 272EMCI, 187LMCI, 210NC	FDG+VBM+AV45	89.20	95.56	81.27	0.99
MMDM [6]	48AD, 119MCI, 66NC	FDG +VBM +CSF + APOE+Cognitive scores	92.40	86.70	96.60	0.97
MM-SDPN-SVM [51]	51AD, 43MCI-C, 56MCI-NC, 52NC	MRI + PET	97.13	95.93	98.53	0.97

TABLE VI
TOP 10 REGIONS SELECTED BY THE PROPOSED METHOD FOR AD DIAGNOSIS

VBM		FDG		AV45	
ROI index	ROI names	ROI index	ROI names	ROI index	ROI names
105	Temporal_Sup_Left	62	Lingual_Right	85	Precuneus_Left
86	Precuneus_Right	50	Frontal_Sup_Medial_Right	95	SupraMarginal_Left
85	Precuneus_Left	7	Caudate_Left	5	Calcarine_Left
6	Occipital_Mid_Right	90	Rectus_Right	104	Temporal_Pole_Sup_Right
89	Rectus_Left	82	Postcentral_Right	96	SupraMarginal_Right
82	Postcentral_Right	81	Postcentral_Left	94	Supp_Motor_Area_Right
81	Postcentral_Left	61	Lingual_Left	74	Supp_Motor_Area_Right
50	Frontal_Sup_Medial_Righ	2	Amygdala_Right	6	Calcarine_Right
27	Cingulum_Ant_Left	1	Amygdala_Left	57	Hippocampus_Left
43	Frontal_Mid_Left	57	Hippocampus_Left	90	Rectus_Right

TABLE VII
TOP 10 SNPs SELECTED BY THE PROPOSED METHOD FOR AD DIAGNOSIS

SNP index	SNP names	SNP index	SNP names
22	rs10119_A	47	rs66626994_A
57	rs157585_C	45	rs111789331_A
74	rs1081106_C	4	rs71352238_C
30	rs10414043_A	55	rs77301115_A
31	rs7256200_T	58	rs116881820_C

In the future, we would like to extend this proposed framework on biospecimens (e.g., including blood, urine, and CSF) to improve AD or MCI classification performance based on more rich multimodality data.

REFERENCES

- [1] R. Katzman, "The prevalence and malignancy of Alzheimer disease: A major killer," *Arch. Neurol.*, vol. 33, no. 4, pp. 217–218, 1976.
- [2] Alzheimer's Association, "2017 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 13, no. 4, pp. 325–373, 2017.
- [3] E. M. Reiman, J. B. Langbaum, and P. N. Tariot, "Alzheimer's prevention initiative: A proposal to evaluate presymptomatic treatments as quickly as possible," *Biomark. Med.*, vol. 4, no. 1, pp. 3–14, 2010.
- [4] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, and E. G. Tangalos, "Mild cognitive impairment: Clinical characterization and outcome," *Arch. Neurol.*, vol. 56, no. 3, pp. 303–308, 1999.
- [5] S. Takeda and N. Sato, "Systemic inflammation, blood-brain barrier vulnerability and cognitive/non-cognitive symptoms in Alzheimer disease: Relevance to pathogenesis and therapy," *Front. Aging Neurosci.*, vol. 6, p. 171, Jul. 2014.
- [6] C. Hinrichs, V. Singh, G. Xu, and S. C. Johnson, "Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population," *Neuroimage*, vol. 55, no. 2, pp. 574–589, 2011.
- [7] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," *Neuroimage*, vol. 65, pp. 167–175, Jan. 2013.
- [8] T. Tong, K. Gray, Q. Gao, L. Chen, and D. Rueckert, "Multi-modal classification of Alzheimer's disease using nonlinear graph fusion," *Pattern Recognit.*, vol. 63, pp. 171–181, Mar. 2017.

- [9] G. Orru, W. Pettersson-Yeo, A. F. Marquand, and G. Sartori, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review," *Neurosci. Behav. Rev.*, vol. 36, no. 4, pp. 1140–1152, 2012.
- [10] J. Ye, T. Wu, and J. Li, "Machine learning approaches for the neuroimaging study of Alzheimer's disease," *Computer*, vol. 44, no. 4, pp. 99–101, Apr. 2011.
- [11] S. Liu *et al.*, "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132–1140, Apr. 2015.
- [12] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *Neuroimage*, vol. 101, pp. 569–582, Jul. 2014.
- [13] X. Zhu, H.-I. Suk, and D. Shen, "A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis," *Neuroimage*, vol. 100, pp. 91–105, Oct. 2014.
- [14] O. B. Ahmed, J. Benois-Pineau, M. Allard, G. Catheline, and C. B. Amar, "Recognition of alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning," *Neurocomputing*, vol. 220, pp. 98–110, Jan. 2017.
- [15] M. Liu, D. Zhang, and D. Shen, "Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis," *Human Brain Map.*, vol. 35, no. 4, pp. 1305–1319, 2014.
- [16] B. Lei, P. Yang, T. Wang, S. Chen, and D. Ni, "Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1102–1113, Apr. 2017.
- [17] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.
- [18] X. Liu, D. Tosun, M. W. Weiner, and N. Schuff, "Locally linear embedding (LLE) for MRI based Alzheimer's disease classification," *Neuroimage*, vol. 83, pp. 148–157, Dec. 2013.
- [19] J. Peng, X. Zhu, Y. Wang, L. An, and D. Shen, "Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis," *Pattern Recognit.*, vol. 88, pp. 370–382, Apr. 2019.
- [20] W. Shao *et al.*, "Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease," *Comput. Med. Imag. Graph.*, vol. 80, Mar. 2020, Art. no. 101663.
- [21] X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen, "Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 607–618, Mar. 2016.
- [22] X. Hao *et al.*, "Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101625.
- [23] J. Xu *et al.*, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 3, pp. 119–130, Jan. 2016.
- [24] H.-I. Suk, S.-W. Lee, and D. Shen, "Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis," *Brain Struct. Funct.*, vol. 221, no. 5, pp. 2569–2587, 2016.
- [25] B. Shi, Y. Chen, P. Zhang, C. D. Smith, and J. Liu, "Nonlinear feature transformation and deep fusion for Alzheimer's Disease staging analysis," *Pattern Recognit.*, vol. 63, pp. 487–498, Mar. 2017.
- [26] R. R. Kitchen, V. S. Sabine, A. A. Simen, J. M. Dixon, J. M. S. Bartlett, and A. H. Sims, "Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments," *BMC Genom.*, vol. 12, no. 1, p. 589, 2011.
- [27] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. MM*, Orlando, FL, USA, 2014, pp. 547–556.
- [28] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 7–19, Jan. 2018.
- [29] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. NIPS*, vol. 23, Whistler, BC, CA, 2010, pp. 1189–1197.
- [30] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann, "Self-paced learning for matrix factorization," in *Proc. AAAI*, vol. 29, Austin, TX, USA, 2015, pp. 3196–3202.
- [31] D. Meng, Q. Zhao, and L. Jiang, "A theoretical understanding of self-paced learning," *Inf. Sci.*, vol. 414, pp. 319–328, Nov. 2017.
- [32] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, Montreal, QC, Canada, 2009, pp. 41–48.
- [33] K. Murugesan and J. Carbonell, "Self-paced multitask learning with shared knowledge," in *Proc. IJCAI*, Melbourne, VIC, Australia, 2017, pp. 2522–2528.
- [34] J. Gan, G. Wen, H. Yu, W. Zheng, and C. Lei, "Supervised feature selection by self-paced learning regression," *Pattern Recognit. Lett.*, vol. 132, pp. 30–37, Apr. 2020.
- [35] Z.-Y. Yang, L.-Y. Xia, H. Zhang, and Y. Liang, "MSPL: Multimodal self-paced learning for multi-omics feature selection and data integration," *IEEE Access*, vol. 7, pp. 170513–170524, 2019.
- [36] W. Xu, W. Liu, X. Huang, J. Yang, and S. J. N. Qiu, "Multi-modal self-paced learning for image classification," *Neurocomputing*, vol. 309, pp. 134–144, Oct. 2018.
- [37] Q. Zhu, N. Yuan, J. Huang, X. Hao, and D. Zhang, "Multi-modal AD classification via self-paced latent correlation analysis," *Neurocomputing*, vol. 355, pp. 143–154, Aug. 2019.
- [38] X. He and P. Niyogi, "Locality preserving projections," in *Proc. NIPS*, vol. 16, Vancouver, BC, Canada, 2003, pp. 153–160.
- [39] S. L. Risacher *et al.*, "APOE effect on Alzheimer's disease biomarkers in older adults with significant memory concern," *Alzheimer's Dementia*, vol. 11, no. 12, pp. 1417–1429, 2015.
- [40] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, Jan. 2002.
- [41] W. Shao, M. Liu, and D. Zhang, "Human cell structure-driven model construction for predicting protein subcellular location from biological images," *Bioinformatics*, vol. 32, no. 1, pp. 114–121, 2016.
- [42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, Apr. 2011.
- [43] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 73, no. 3, pp. 273–282, 2011.
- [44] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, pp. 1921–1932, 2010.
- [45] S. M. Landau *et al.*, "Comparing predictors of conversion and decline in mild cognitive impairment," *Neurology*, vol. 75, no. 3, pp. 230–238, 2010.
- [46] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun, "A review of multivariate methods for multimodal fusion of brain imaging data," *J. Neurosci. Methods*, vol. 204, no. 1, pp. 68–81, 2012.
- [47] J. Young *et al.*, "Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment," *Neuroimage Clin.*, vol. 2, pp. 735–745, May 2013.
- [48] O. Kohannim *et al.*, "Boosting power for clinical trials using classifiers based on multiple biomarkers," *Neurobiol. Aging*, vol. 31, no. 8, pp. 1429–1442, 2010.
- [49] N. Yuan, D. Guan, Q. Zhu, and W. Yuan, "Self-paced learning for multi-modal fusion for Alzheimer's disease diagnosis," in *Proc. ICSPAC*, 2017, pp. 1–6.
- [50] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, and J. Li, "A robust deep model for improved classification of AD/MCI patients," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1610–1616, Sep. 2015.
- [51] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 173–183, Jan. 2018.
- [52] H. Wolf, V. Jelic, H.-J. Gertz, A. Nordberg, P. Julin, and L.-O. Wahlund, "A critical discussion of the role of neuroimaging in mild cognitive impairment," *Acta Neurologica Scandinavica*, vol. 107, pp. 52–76, Feb. 2003.
- [53] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, and B. C. Dickerson, "Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity," *Psychiatry Res. Neuroimag.*, vol. 194, no. 1, pp. 7–13, 2011.
- [54] L. Gao, Z. Cui, L. Shen, and H.-F. Ji, "Shared genetic etiology between type 2 diabetes and Alzheimer's disease identified by bioinformatics analysis," *J. Alzheimer's Dis.*, vol. 50, no. 1, pp. 13–17, 2016.
- [55] X. Zhou *et al.*, "Non-coding variability at the APOE locus contributes to the Alzheimer's risk," *Nat. Commun.*, vol. 10, no. 1, pp. 1–16, 2019.
- [56] G. Davies *et al.*, "Genetic contributions to variation in general cognitive function: A meta-analysis of genome-wide association studies in the CHARGE consortium ($N = 53949$)," *Mol. Psychiatry*, vol. 20, no. 2, pp. 183–192, 2015.