Latent Representation Learning for Alzheimer's Disease Diagnosis With Incomplete Multi-Modality Neuroimaging and Genetic Data

Tao Zhou, *Member, IEEE*, Mingxia Liu, Kim-Han Thung, and Dinggang Shen, *Fellow, IEEE*

***Abstract*—The fusion of complementary information contained in multi-modality data [*e.g.*, magnetic resonance imaging (MRI), positron emission tomography (PET), and genetic data] has advanced the progress of automated Alzheimer's disease (AD) diagnosis. However, multi-modality based AD diagnostic models are often hindered by the missing data, *i.e.*, not all the subjects have complete multi-modality data. One simple solution used by many previous studies is to discard samples with missing modalities. However, this significantly reduces the number of training samples, thus leading to a sub-optimal classification model. Furthermore, when building the classification model, most existing methods simply concatenate features from different modalities into a single feature vector without considering their underlying associations. As features from different modalities are often closely related (*e.g.*, MRI and PET features are extracted from the same brain region), utilizing their inter-modality associations may improve the robustness of the diagnostic model. To this end, we propose a novel latent representation learning method for multi-modality based AD diagnosis. Specifically, we use all the available samples (including samples with incomplete modality data) to learn a latent representation space. Within this space, we not only use samples with complete multi-modality data to learn a common latent representation, but also use samples with incomplete multi-modality data to learn independent modality-specific latent representations. We then project the latent representations to the label space for AD diagnosis. We perform experiments using 737 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, and the experimental results verify the effectiveness of our proposed method.**

***Index Terms*—Alzheimer's disease, multi-modality data, incomplete multi-modality data, latent representation space**

I. Introduction

As one of the most common neurodegenerative diseases, Alzheimer's disease (AD) often appears in people over 65 years old, and gradually affects their memory and other brain functions. According to a recent research report from the Alzheimer's Association [1], the total prevalence of AD is expected to reach 60 million worldwide over the next 50 years. Thus, a lot of research efforts have been devoted to understand the underlying biological or physiological mechanisms of AD [2], [3]. Although there is no effective cure for AD, studies have demonstrated that the diagnosis of its early stage (*i.e.*, Mild Cognitive Impairment (MCI)) is highly desirable in clinical practice, so that early treatments could be administered to possibly slow down the disease progression [4], [5].

Advanced multi-modality neuroimaging technology, such as Magnetic Resonance Imaging (MRI) [2], [6]–[10] and Positron Emission Tomography (PET) [11], [12], have provided unprecedented opportunities for the early diagnosis of AD. In particular, the fusion of multi-modality data has advanced the progress of automatic AD diagnosis, thanks to the complementary information contained within them [12]–[16]. In addition, several Genome-Wide Association Studies (GWAS) have identified a series of genetic variations (*e.g.*, Single Nucleotide Polymorphism (SNP)) associated with AD [17], which are likely to increase the risk of developing the disease [18], [19]. Therefore, it is of interest to develop a prediction model that fuses both neuroimaging (*e.g.*, MRI and PET) and genetic data (*e.g.*, SNP) to further improve the performance of AD diagnosis [20].

In automated AD diagnosis studies using multi-modality data, the feature dimension is usually very high (*e.g.*, tens of thousands) while the number of training samples is limited, *i.e.*, it is a typical small-sample-size problem. To address this issue, previous studies have applied various dimension reduction or feature selection techniques [21], [22]

Manuscript received January 8, 2019; revised April 14, 2019; accepted April 18, 2019. Date of publication April 25, 2019; date of current version October 1, 2019. This work was supported in part by NIH under Grant EB006733, Grant EB008374, Grant EB009634, Grant MH100217, Grant AG041721, and Grant AG042599. *(Tao Zhou and Mingxia Liu contributed equally to this work.) (Corresponding author: Dinggang Shen.)*

T. Zhou is with the Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA, and also with the Inception Institute of Artificial Intelligence, Abu Dhabi 51133, United Arab Emirates (e-mail: taozhou.dreams@gmail.com).

M. Liu and K.-H. Thung are with the Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: mxliu@med.unc.edu; henrythung@gmail.com).

D. Shen is with the Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea (e-mail: dgshen@med.unc.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2913158

on multi-modality data, to find the most informative feature subset for accurate AD diagnosis. For example, Salas-Gonzalez *et al.*hbox [22] utilize the statistical t-test method to select voxels of interest for AD diagnosis. Sparse learning based feature selection methods have also been widely applied in AD diagnosis [23], [24]. In addition, many classic feature dimensionality reduction algorithms, *e.g.*, Principal Component Analysis (PCA) [25], Linear Discriminant Analysis (LDA) [26], and Laplacian Preserving Projection (LPP) [27], have been applied to AD status prediction.

Although various feature selection and dimension reduction methods have been proposed, there are still two challenges with automatic AD diagnosis systems using multi-modality neuroimaging (*i.e.*, MRI and PET) and genetic data (*i.e.*, SNP). The first challenge is that it is difficult to exploit the inherent association among multi-modality data. To fuse multi-modality data, conventional methods usually first conduct feature selection for each modality separately and then concatenate the selected features for diagnosis or prognosis. However, such approaches ignore the underlying association between different modalities. Although several classic multi-modality fusion approaches (*e.g.*, Multiple Kernel Learning (MKL) [28] and Canonical Correlation Analysis (CCA) [29]) can exploit the relationship between different modalities, they can only be applied to the complete multi-modality data.

As briefly mentioned above, the second challenge in multi-modality based AD diagnosis system is the missing data issue, *i.e.*, not all the samples have complete multi-modality data. Generally, there are two approaches for dealing with this issue, *i.e.*, 1) discarding the samples with missing values, and 2) imputing the missing values. Several current AD diagnostic models follow the first approach, *i.e.*, they simply discard samples with at least one missing modality and perform disease diagnosis using the remaining samples that have complete multi-modality data. However, this simple strategy not only disregards lots of useful information in the discarded samples, it also escalates the small-sample-size issue. The second approach uses various data imputation techniques (*e.g.*, Zero imputation, k-Nearest Neighbor (KNN), expectation maximization, low-rank matrix completion, etc.) to impute the missing data, so that any diagnostic model that works with complete data can be used. However, this strategy could introduce unnecessary noise and thus reduce the classification performance. Several approaches have also been developed to handle incomplete multi-modality data [7], [30], [31]. However, these approaches do not effectively exploit the correlations across multiple modalities. It is expected to boost diagnostic performance by fusing these multi-modality data.

To this end, we propose a novel latent representation learning framework for AD diagnosis (as shown in Fig. 1). Specifically, we assume there exists a latent space for multi-modality data, to which each modality can be projected. The projection from different modalities to this common latent space is expected to model the association among different modalities. That is, we first treat data from each modality (*e.g.*, MRI, PET, and SNP) as explicit features that can reflect different attribute information. For instance, MRI data can provide us with anatomical brain information, while PET and SNP data can provide us with functional brain information and congenital genetic information, respectively. To make full use of all available samples for learning a more reliable prediction model, we utilize samples with complete multi-modality data to learn the common latent feature representation, and utilize samples with incomplete multi-modality data to learn an independent (*i.e.*, modality-specific) latent feature representation for each modality. Furthermore, the learned latent representations are projected to the corresponding label space for AD diagnosis. We evaluate our proposed method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, and the experimental results verify its effectiveness.

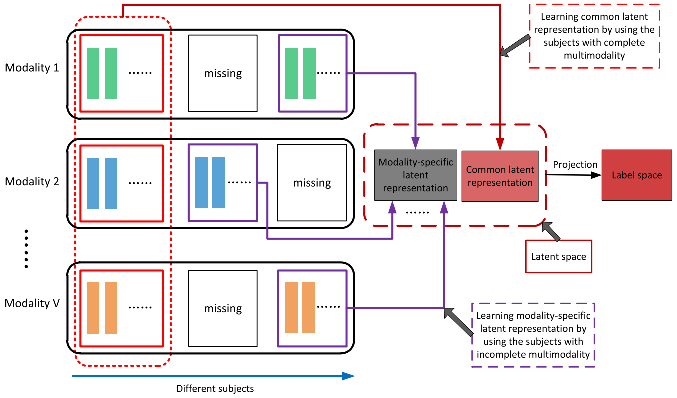

Learning common latent representation by using the subjects with complete multimodality
Modality 1
......
missing
......
Modality 2
......
......
missing
Modality-specific latent representation
......
Common latent representation
Projection
Label space
Latent space
Modality V
......
missing
......
Learning modality-specific latent representation by using the subjects with incomplete multimodality
Different subjects

Fig. 1. Illustration of our proposed AD diagnosis framework. First, we project original features into a latent representation space. Within this space, we utilize samples with complete multi-modality data to learn a common latent representation, and utilize samples with incomplete multi-modality to learn modality-specific representations. Finally, the latent feature representations are projected to the label space for AD diagnosis.

The key contribution of this paper is three-fold. *First*, our method performs classification tasks using new feature representations in a latent space, instead of the original features. These new feature representations are expected to be less noisy than the original features, since projecting the original features into the latent space involves sparse feature learning. *Second*, our method makes use of all the available subjects to train a more reliable model, while the conventional methods usually use only subjects with complete multi-modality data or impute missing values via specific algorithms. *Third*, we integrate both the latent feature learning and classifier training into a unified framework, while most conventional methods often conduct the two tasks separately.

The rest of this paper is organized as follows. We describe the materials used in this study and present the data preprocessing steps in Section II. Then, we introduce our proposed latent representation learning method in Section III. We further describe experiments in Section IV and provide the related discussion in Section V. Finally, we conclude this paper in Section VI.

II. Materials and Data Preprocessing

We used the data from the public ADNI database [32] to evaluate our proposed framework. The ADNI dataset was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and

TABLE I
DEMOGRAPHIC INFORMATION OF THE BASELINE SUBJECTS IN THIS STUDY (MMSE: MINI-MENTAL STATE EXAMINATION)

	Female/male	Education	Age	MMSE
NC	92 / 112	15.9 ± 3.0	76.1 ± 4.9	29.1 ± 1.0
sMCI	68 / 137	15.8 ± 3.1	75.1 ± 7.6	27.4 ± 1.6
pMCI	62 / 95	16.1 ± 2.5	74.8 ± 7.0	26.9 ± 1.8
AD	77 / 94	14.5 ± 3.7	75.5 ± 7.7	23.7 ± 1.9

TABLE II
THE MAIN NOTATIONS USED IN THE PROPOSED FORMULATION IN EQ. (6)

Notation	Size	Description
$\mathbf{X}_v^c$	$d_v \times n^c$	The v-th modality feature matrix for samples with complete multi-modality data
$\mathbf{X}_v^{\bar{c}}$	$d_v \times n_v^{\bar{c}}$	The v-th modality feature matrix for samples with incomplete multi-modality data
$\mathbf{W}_v$	$d_v \times h$	Projection matrix for the v-th modality
$\mathbf{H}^c$	$h \times n^c$	Latent feature representation for samples with complete multi-modality data
$\mathbf{H}_v^{\bar{c}}$	$h \times n_v^{\bar{c}}$	Latent feature representation of the v-th modality for samples with incomplete multi-modality data
$\mathbf{Y}^c$	$l \times n^c$	Label matrix for samples with complete multi-modality data
$\mathbf{Y}_v^{\bar{c}}$	$l \times n_v^{\bar{c}}$	Label matrix of the v-th modality for samples with incomplete multi-modality data
$\mathbf{P}$	$l \times h$	Projection matrix from the learned latent feature representation to the label space
d_v	–	Feature dimension of the v-th modality
n_v	–	Number of samples with the v-th modality
n^c	–	Number of samples with complete multi-modality data
$n_v^{\bar{c}}$	–	Number of samples with the v-th modality excluding samples with complete multi-modality data
h	–	Dimension of the latent feature representation
l	–	Number of classes
V	–	Number of multi-modality data
λ, β, γ, η	–	Regularization parameters

Bioengineering, the Food and Drug Administration, private pharmaceutical companies and nonprofit organizations with a five-year public-private partnership. The main aim of ADNI is to research the potential of fusing multi-modality data, including neuroimaging, clinical, biological, and genetic biomarkers, to effectively diagnose AD and its early stage.

A. Studied Subjects

In this study, we used 737 subjects from the ADNI-1 database, including 171 AD, 362 MCI, and 204 normal control (NC) subjects. For MCI subjects, we further labeled those who progressed to AD after a certain period of time as progressive MCI (pMCI) subjects, and those who remained stable as stable MCI (sMCI) subjects. In this study, there are 205 sMCI subjects and 157 pMCI subjects. All the studied subjects have baseline MR images acquired at the first screening time (*i.e.*, the baseline time-point). Among them, only 360 subjects have baseline PET data. Table I shows the demographic information of the studied subjects.

B. Data Preprocessing

In this study, we downloaded 1.5T MR images from the ADNI website.[1] The MR images were collected using a variety of scanners with protocols customized to each scanner. For quality control, ADNI reviewed the MR images and corrected them for spatial distortion caused by B1 field inhomogeneity and gradient nonlinearity. Following previous studies [33], we further processed these MR images and extracted ROI-based features. Specifically, the MR images were processed under the following steps: 1) anterior commissure-posterior commissure (AC-PC) correction using MIPAV software,[2] 2) intensity inhomogeneity correction using the N3 algorithm [34], 3) brain extraction using a robust skull-stripping algorithm [35], 4) cerebellum removal, 5) tissues segmentation using the FAST algorithm in the FSL package [36] to obtain three main tissues (*i.e.*, white matter (WM), gray matter (GM), and cerebrospinal fluid), 6) registration to a template [37] using the HAMMER algorithm [38], and 7) ROI labels projection from the template image to the subject image. Finally, we computed the GM tissue volume of each ROI in the labeled image, and normalized them with the intracranial volume as the ROI-based feature representation for each subject. Moreover, for each subject, we first aligned PET images to their corresponding T1 MR images using affine registration, and then computed the average PET intensity value of each ROI as a feature representation. Using a template with 93 ROIs [37], we obtained a 93-dimensional ROI-based feature vector from a specific type of neuroimaging data (*i.e.*, MRI or PET) for each subject.

Genetic variations can provide us microscopic information about AD. In this study, the SNP data were genotyped using the Human 610-Quad BeadChip. According to the AlzGene database,[3] only SNPs belonging to the top AD gene candidates were selected. The selected SNPs were imputed to estimate the missing genotypes, and Illumina annotation information was used to select a subset of SNPs [39].

III. METHODOLOGY

In this section, we provide the details of the proposed AD diagnosis framework, and then present the optimization algorithm, as well as the model prediction steps.

A. Formulation

Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ denote a feature matrix, where d and n indicate the feature dimension and the number of subjects, respectively, and $\mathbf{Y} \in \mathbb{R}^{l \times n}$ denote the corresponding label matrix, with l being the total number of classes. The least square regression model with multi-class group sparse feature selection method is given as

$$\min_{\mathbf{W}} \; \|\mathbf{W}^{\top}\mathbf{X} - \mathbf{Y}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}, \qquad (1)$$

where $\mathbf{W} \in \mathbb{R}^{d \times l}$ is a regression coefficient matrix, $\top$ denotes the transpose operator, $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$-norm, *i.e.*, $\|\mathbf{W}\|_{2,1} = \sum_i^d \sqrt{\sum_j^l w_{ij}^2}$, and β is a trade-off parameter that is used to balance the reconstruction error (the first term) and the regularizer (the second term). The $\ell_{2,1}$-norm regularizer has been widely applied to

[1] http://www.loni.usc.edu/ADNI

[2] http://mipav.cit.nih.gov/clickwrap.php

[3] www.alzgene.org

multi-task feature learning [40]–[42], and is used to penalize the coefficients in each row of $\mathbf{W}$, so that only certain rows in $\mathbf{W}$ are non-zeros. In Eq. (1), the optimal solution assigns relatively large weights to informative features while zero or small weights to uninformative or less informative features.

For multi-modality data fusion, as in our case, Eq. (1) can be extended to

$$\min_{\mathbf{W}_v} \sum_{v=1}^{V} \|\mathbf{W}_v^\top \mathbf{X}_v - \mathbf{Y}\|_F^2 + \beta \|\mathbf{W}_v\|_{2,1}, \tag{2}$$

where V is the number of modalities, and $\mathbf{X}_v$ denotes feature matrix for the v-th modality. In Eq. (2), the original features from different modalities are projected into the label space without considering the correlation among different modalities. However, we know that both neuroimaging and genetic features are related in some ways, as they are directly and indirectly associated with the AD disease label, respectively, thus giving different views for the AD prediction task. Hence, we hypothesize that there exists a common latent feature space for those modalities, which contains information from different perspectives and gives us a more comprehensive view of the AD prediction task. Thus, we extend Eq. (2) into the following formulation as

$$\min_{\mathbf{W}_v,\mathbf{E}_v,\mathbf{H},\mathbf{P}} \frac{1}{2}\|\mathbf{PH}-\mathbf{Y}\|_F^2 + \gamma \sum_{v=1}^{V} \|\mathbf{W}_v^\top \mathbf{X}_v - \mathbf{H}\|_F^2 + \beta \sum_{v=1}^{V} \|\mathbf{W}_v\|_{2,1} + \frac{\eta}{2}\|\mathbf{P}\|_F^2. \tag{3}$$

Furthermore, to alleviate outlier effect in latent space learning, we adopt the ℓ_1-norm to replace the Frobenius norm. Thus, the objective function in Eq. (3) is transformed to

$$\begin{aligned}\min_{\mathbf{W}_v,\mathbf{E}_v,\mathbf{H},\mathbf{P}} &\frac{1}{2}\|\mathbf{PH}-\mathbf{Y}\|_F^2 + \beta \sum_{v}^{V} \|\mathbf{W}_v\|_{2,1} + \gamma \sum_{v}^{V} \|\mathbf{E}_v\|_1 + \frac{\eta}{2}\|\mathbf{P}\|_F^2,\\ s.t.\ &\mathbf{W}_v^\top \mathbf{X}_v = \mathbf{H} + \mathbf{E}_v, \quad \forall v \in \{1, 2, \cdots, V\},\end{aligned} \tag{4}$$

where β, γ and η are the regularization parameters, $\mathbf{H} \in \mathbb{R}^{h\times n}$ is the common latent feature representation for all modalities, $\mathbf{W}_v \in \mathbb{R}^{d_v \times h}$ is a feature projection matrix for the v-th modality, $\mathbf{E}_v \in \mathbb{R}^{h\times n_v}$ is the sparse error matrix for the v-th modality, and $\mathbf{P} \in \mathbb{R}^{l\times h}$ is a label projection matrix that projects the common latent feature representation to the label space. Note that the dimensions of the coefficient matrix $\mathbf{W}_v$ in Eq. (4) and Eq. (2) are different, where $\mathbf{W}_v$ in Eq. (2) maps the v-th feature matrix to the label space directly, while $\mathbf{W}_v$ in Eq. (4) maps the v-th feature matrix to a common latent space $\mathbf{H}$, which will be subsequently mapped to the label space via $\mathbf{P}$. The equality constraints in Eq. (4) are used to enforce different modality data of the same subject to have a common feature component ($\mathbf{H}$) and a modality specific discrepancy component ($\mathbf{E}_v$) after they are projected by $\mathbf{W}_v$. Furthermore, assuming that the discrepancy component is sparse, we impose ℓ_1-norm on $\mathbf{E}_v$, while Frobenius norm is imposed on $\mathbf{P}$ to limit the magnitudes of its coefficient values.

In addition, to preserve the local structure of the data after the projection (*i.e.*, to preserve the local neighborhood of the samples before and after the projection), we add a Laplacian regularizer in Eq. (4), given as

$$\begin{aligned}\min_{\mathbf{W}_v,\mathbf{E}_v,\mathbf{H},\mathbf{P}} &\frac{1}{2}\|\mathbf{PH}-\mathbf{Y}\|_F^2 + \lambda \sum_{v=1}^{V} \mathrm{tr}\left(\mathbf{W}_v^\top \mathbf{X}_v \mathbf{L}_v (\mathbf{W}_v^\top \mathbf{X}_v)^\top\right)\\ &+ \beta \sum_{v=1}^{V} \|\mathbf{W}_v\|_{2,1} + \gamma \sum_{v=1}^{V} \|\mathbf{E}_v\|_1 + \frac{\eta}{2}\|\mathbf{P}\|_F^2,\\ s.t.\ &\mathbf{W}_v^\top \mathbf{X}_v = \mathbf{H} + \mathbf{E}_v, \quad \forall v \in \{1, 2, \cdots, V\},\end{aligned} \tag{5}$$

where the second term is the Laplacian regularization term, which is added to ensure that similar inputs have similar latent feature representations, and λ is its regularization parameter. Specifically, the Laplacian matrix is given as $\mathbf{L}_v = \mathbf{D}_v - \mathbf{S}_v$, where $\mathbf{D}_v$ is a diagonal matrix with its i-th diagonal element denoting the sum of the i-th row in $\mathbf{S}_v$. $\mathbf{S}_v$ is a similarity matrix for the v-th modality, whose (i, j)-th element is given as $\exp(-\|\mathbf{X}_{v,:i} - \mathbf{X}_{v,:j}\|_2^2/\sigma)$, where $\mathbf{X}_{v,:i}$ and $\mathbf{X}_{v,:j}$ denote the i-th column and the j-th column of $\mathbf{X}_v$, respectively, and $\sigma = 1$ is empirically set in this study.

So far, the proposed method in Eq. (5) can only be applied to subjects with complete multi-modality data, *i.e.*, each modality has the same number of samples, $n = n_v$, $\forall v = \{1, 2, \cdots, V\}$. However, using Eq. (5), subjects with incomplete multi-modality data will be discarded. Intuitively, using more samples for model training should result in a more reliable prediction model. To make the joint latent feature learning model in Eq. (5) applicable to incomplete multi-modality data, we have to make some modifications. First, we decompose $\mathbf{X}_v$ into two parts, one with complete multi-modality data, and the other with incomplete multi-modality data, *i.e.*, $\mathbf{X}_v = [\mathbf{X}_v^c, \mathbf{X}_v^{\bar{c}}] \in \mathbb{R}^{d_v\times(n^c+n_v^{\bar{c}})}$, where $n_v = n^c + n_v^{\bar{c}}$ is the total number of samples in v-modality. Note that the number of complete multi-modality data, denoted by n^c, is the same for each $\mathbf{X}_v$. Then, the corresponding label for $\mathbf{X}_v$ is given by $\mathbf{Y}_v = [\mathbf{Y}^c, \mathbf{Y}_v^{\bar{c}}]$. Similarly, the matrix $\mathbf{H}$ in the constraint of Eq. (5) is also decomposed into two parts, *i.e.*, $\mathbf{H}_v = [\mathbf{H}^c, \mathbf{H}_v^{\bar{c}}] \in \mathbb{R}^{h\times(n^c+n_v^{\bar{c}})}$, where $\mathbf{H}^c$ denotes the latent feature representation for the complete multi-modality data, and $\mathbf{H}_v^{\bar{c}}$ denotes the latent feature representation for the remaning v-th modality data (other than that included in $\mathbf{H}^c$). Using the notations above, we extend Eq. (5) to

$$\begin{aligned}\min_{\mathbf{W}_v,\mathbf{E}_v,\mathbf{H},\mathbf{P}} &\frac{1}{2}\|\mathbf{P}[\mathbf{H}^c, \mathbf{H}_1^{\bar{c}}, \ldots, \mathbf{H}_V^{\bar{c}}] - [\mathbf{Y}^c, \mathbf{Y}_1^{\bar{c}}, \ldots, \mathbf{Y}_V^{\bar{c}}]\|_F^2\\ &+ \lambda \sum_{v}^{V} \mathrm{tr}\left(\mathbf{W}_v^\top [\mathbf{X}_v^c, \mathbf{X}_v^{\bar{c}}] \mathbf{L}_v (\mathbf{W}_v^\top [\mathbf{X}_v^c, \mathbf{X}_v^{\bar{c}}])^\top\right)\\ &+ \beta \sum_{v}^{V} \|\mathbf{W}_v\|_{2,1} + \gamma \sum_{v}^{V} \|\mathbf{E}_v\|_1 + \frac{\eta}{2}\|\mathbf{P}\|_F^2,\\ s.t.\ &\mathbf{W}_v^\top\ [\mathbf{X}_v^c, \mathbf{X}_v^{\bar{c}}] = [\mathbf{H}^c, \mathbf{H}_v^{\bar{c}}] + \mathbf{E}_v,\ \forall v \in \{1, 2, \cdots, V\},\end{aligned} \tag{6}$$

where $\mathbf{H} = [\mathbf{H}^c, \mathbf{H}_1^{\bar{c}}, \ldots, \mathbf{H}_V^{\bar{c}}]$ is the latent representation matrix for all the samples. Note that the latent representation $\mathbf{H}_v^{\bar{c}}$ is specific to the v-th modality, as it is not shared with other modalities like $\mathbf{H}^c$. Thus, we also refer to $\mathbf{H}_v^{\bar{c}}$ as the modality-specific latent representation.

Using Eq. (6), we can include all the data from the dataset, regardless of modality completeness when training the model, *i.e.*, to learn $\mathbf{W}_v$ and $\mathbf{P}$. All the modalities are projected to a common latent feature space via $\mathbf{W}_v$, and $\mathbf{P}$ is the classifier coefficient matrix that predicts disease label from the data point in the latent feature space. For clarity, the notations used in Eq. (6) are summarized in Table. II.

B. Optimization and Model Prediction

The objective function in Eq. (6) is not jointly convex with respect to all variables. Therefore, we utilize the Augmented Lagrange Multiplier (ALM) [43] algorithm to solve the problem efficiently and effectively. The detailed optimization steps are provided in the *Supplementary Materials* of this manuscript. After training our model, we can obtain the feature projection matrix $\mathbf{W}_v$ for the v-th modality, and the label projection matrix $\mathbf{P}$. For a given testing sample $\mathbf{x}^{te}$ with $\boldsymbol{\Omega}$ available modalities, the latent feature representation is computed by averaging the feature projections from each available modality, *i.e.*, $\mathbf{h}^{te} = \frac{1}{|\Omega|}\sum_{v\in\Omega}\mathbf{W}_v\mathbf{x}_v^{te}$, where $|\Omega|$ denotes the number of modalities in $\boldsymbol{\Omega}$. Consequently, the final classification label for this test sample is given as $\mathbf{y}^{te} = \mathbf{P}\mathbf{h}^{te}$.

IV. Experiments

In this section, we first describe the parameter settings of our method and the comparison methods. Then we report the classification performance of all the comparison methods, and show the results using different combinations of modalities. In addition, we show classification results using incomplete multi-modality data with different missing rates.

A. Experimental Setup

We evaluate our method using the ADNI dataset for two multi-class classification tasks, including 1) NC vs. MCI vs. AD (*i.e.*, three-class classification task), and 2) NC vs. sMCI vs. pMCI vs. AD (*i.e.*, four-class classification task). In addition, as it is important to distinguish progressive MCI from stable MCI for early diagnosis, we also evaluated our proposed method for the sMCI vs. pMCI classification task. Using only the complete multi-modality data (*i.e.*, subjects with no missing modality), we first compare the proposed method with several conventional methods, the details of which are briefly introduced below.

- Baseline method. We conduct an experiment using only the original features without performing any feature selection (denoted as "Baseline").
- Feature reduction (or selection) methods. Three comparison methods are used in this category, namely 1) Principal Component Analysis (PCA) [25], 2) Locality Preserving Projection (LPP) [27], and 3) the $\ell_{2,1}$-norm based feature selection method as described in Eq. (2), which is denoted as "L21". For PCA and LPP, we determined the optimal dimensionality of the data based on the eigenvalues computed using the generalized eigen-decomposition method described in [44]. For the L21 method, we optimize its sparsity parameter (λ) by cross-validating its value in the range of $\{10^{-4}, \ldots, 10^{2}\}$.
- Modality (or feature) fusion methods. Two comparison methods in this category are used, *i.e.*, 1) CCA [45] and 2) MKL [28]. For CCA, we optimize the regularization parameter by cross-validating its value in the range of $\{10^{-4}, \ldots, 10^{2}\}$. For MKL, we optimize the normalized weight parameter for each modality by cross-validating its value in the range of $\{0, \ldots, 1\}$.
- Deep learning based feature representation method. In this category, we compare our method with the Stacked Auto-Encoder (SAE) [46] method. In SAE, the main parameter to be determined is the number of hidden units. Following [46], we build a three-layer network with multi-modality data input using a grid search from [100, 300, 500]-[50, 100]-[10, 20, 30] (bottom-top).

Furthermore, to verify the advantage of our proposed method in handling the missing data issue in the testing phase, we also compare it with the following state-of-the-art methods.

- Data imputation methods. 1) Zero imputation. In this method, the missing values are filled with zeros. Since all the features are z-normalized (*i.e.*, subtract the mean and divide by the standard deviation) before the imputation process, this method is equivalent to filling the missing feature values with the average observed values. 2) k-Nearest Neighbor (KNN) imputation [47]. In this method, the missing values are filled with the weighted mean of the k nearest-neighbor samples. Following [7], the weights are inversely proportional to the Euclidean distances between the neighboring samples. 3) Low-rank Matrix Completion (LRMC) [48], [49] based imputation method. This was proposed to recover missing data from a limited number of samples. After data imputation, we employe a linear SVM to perform disease classification in these three methods.
- Incomplete Multi-Source Feature learning (iMSF) method [30], [31]. This is a multi-task learning method, which first partitions the data into several groups according to the availability of modalities, and treats the learning of a classifier for each group of data as a task. A joint sparse learning model is then employed to select a common set of features among all these tasks. There are two versions of iMSF, which use different loss functions, *i.e.*, the least square loss (denoted as "iMSF-R") and the logistic loss (denoted as "iMSF-L"). Note that the source code of the iMSF method is designed for binary classification tasks, and here we adopt a one-vs-all strategy for multi-class classification tasks.
- Doubly Aligned Incomplete Multi-view Clustering (DAIMC) method [50]. This is a clustering method that is designed for incomplete multi-modality data using weighted semi-nonnegative matrix factorization [50]. To apply this method for the classification task, we train a SVM classifier using the learned common latent features in DAIMC.

- Matrix Shrinkage and Completion method (MSC) [7]. This method first partitions the combined matrix (*e.g.*, features and targets) into sub-matrices, where each sub-matrix consists of samples with complete features (corresponding to a certain combination of modalities) and target outputs. Then, a multi-task sparse learning framework is employed to select informative features and samples. Subsequently, the shrunk combined matrix with missing features and unknown target outputs is imputed via low-rank matrix completion using a fixed-point continuation method [51].

We use a ten-fold cross-validation strategy to evaluate all comparison methods. Specifically, we first randomly partition the whole dataset into ten subsets (with each subset having a roughly equal number of samples), and then select one subset for testing and use the remaining nine subsets for training. We repeat the whole process fifty times to avoid possible bias in dataset partitioning during cross-validation. Thus, the final results are computed by averaging the fifty repetition results. In our proposed method, it would take a significant amount of time if we simultaneously tune five hyper-parameters. In this case, we first tune λ and h by fixing other hyper-parameters. Here, we empirically set $\lambda = 0.01$ and $h = 30$, and then apply a five-fold inner cross-validation to tune values of the other three hyper-parameters, *i.e.*, $\beta, \gamma, \eta \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$, in Eq. (6). Note that we use the SVM classifier from the LIBSVM toolbox [52] to perform classification for all the comparison methods (except for iMSF and MSC), and determine the margin parameter C in the SVM classifier using a grid search in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{3}\}$. Moreover, for fair comparison, we also conduct five-fold inner cross-validation to conduct parameter selection for all the comparison methods. We use classification accuracy (ACC) and Area Under Curve (AUC) to evaluate all the comparison methods, and follow the AUC computation method in [53] for multi-class classification tasks.

The SNP data are high-dimensional, but only parts of SNPs are related to AD, as shown by previous studies [20]. Thus, in order to reduce the computational complexity for the AD prediction task [17], we employ sparse learning to select 100 SNPs. Specifically, we first randomly partition the whole dataset into 10 subsets (with each subset having a roughly equal number of samples), and select one subset for testing, while the remaining nine subsets are used for training. We then use a five-fold inner cross-validation on the training set to select SNPs via a sparse learning method [54]. Finally, we select the top 100 most related SNPs.

B. Classification Results using Complete Multi-Modality Data

Fig. 2 shows the classification performance of all the competing methods on three classification tasks, using complete multi-modality data (*i.e.*, using only subjects with all three modalities). Here, "Ours_com" denotes our proposed method that uses the samples with only complete multi-modality data to train a model, while "Ours" denotes the proposed method that uses all available samples (including the samples with incomplete multi-modalities). The comparison methods only train models using samples with complete multi-modality data, as they are inapplicable to incomplete dataset.

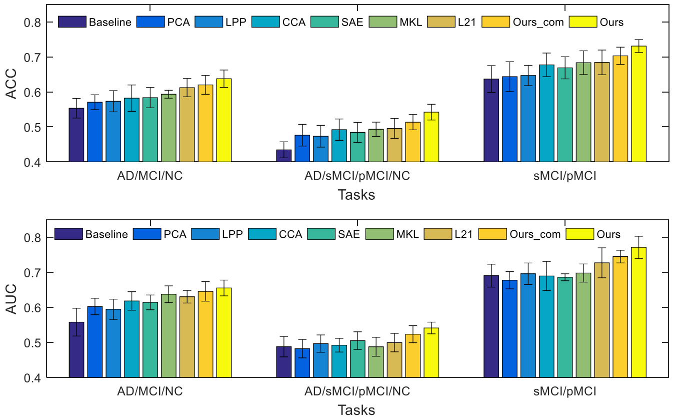


Fig. 2. Classification results in terms of ACC (top) and AUC (bottom) achieved by 9 different methods for three classification tasks, *i.e.*, NC vs. MCI vs. AD (left), NC vs. sMCI vs. pMCI/AD (middle), and sMCI vs. pMCI (right), using complete multi-modality data in training and testing (except "Ours", which denotes our method uses all available samples with incomplete multi-modality data for model training). The error bars denote the standard deviations of the results.

From Fig. 2, we can have following observations. *First*, our proposed method (*i.e.*, Ours_com) outperforms the three feature selection methods (*i.e.*, PCA, LPP and L21), which implies that the new feature representation learned by our proposed method can help improve the classification performance. It also outperforms modality fusion methods (*e.g.*, CCA and MKL) for all three classification tasks. The comparison results indicate that our method is more effective than both the CCA and MKL methods in exploiting the correlations among different modalities. Additionally, it also outperforms the deep learning based method (*i.e.*, SAE) that learns high-level features for classification. Generally, deep learning based methods can learn "good" features for classification, but the diagnosis performance is likely degraded due to the limited number of samples in this study. *Second*, the results reported for multi-class classification tasks (*i.e.*, NC vs. MCI vs. AD, NC vs. sMCI vs. pMCI vs. AD) are lower than the performance on the binary classification task (*i.e.*, sMCI vs. pMCI). A possible reason for this is that multi-class classification tasks are much more challenging than binary classification tasks. Also, our proposed method that uses all available data ("Ours") is better than its degraded version, ("Ours_com"), which only uses complete multi-modality data for training. This shows the advantage of using all available data during the training, rather than discarding samples with incomplete multi-modality data.

C. Classification Results using Incomplete Multi-Modality Data

Based on all available subjects with incomplete multi-modality, we perform three classification tasks and report the ACC and AUC results in Fig. 3. The results show that our method outperforms imputation based methods in terms of ACC and AUC. On one hand, this is probably due to the fact

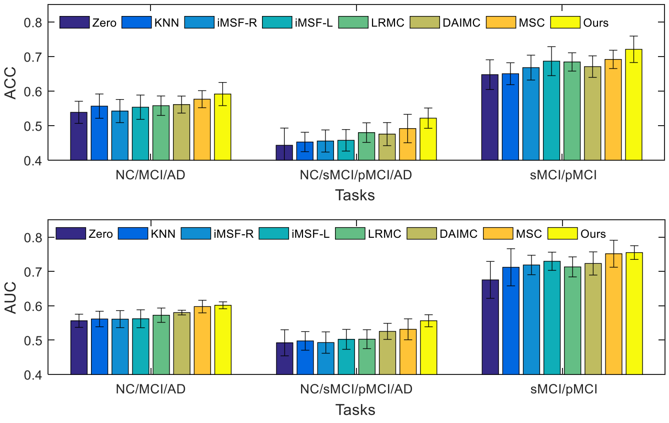


Fig. 3. Classification results in terms of ACC (top) and AUC (bottom) achieved by 8 different methods for three classification tasks, *i.e.*, NC vs. MCI vs. AD (left), NC vs. sMCI vs. pMCI vs. AD (middle), and sMCI vs. pMCI (right), using incomplete multi-modality data in training and testing. The error bars denote the standard deviations of the results.

that our proposed method does not involve any imputation, and thus avoids imputation errors that may affect the classification model like in the imputation based methods. On the other hand, this may also result from the use of latent feature representations when learning the prediction model, which are less noisy (due to the explicit learning of the error matrix), more discriminative (due to feature selection), and more comprehensive (by considering information from different data views). Thus, introducing the latent feature representation not only solves the missing data problem, but also results in a more accurate prediction model. Our method also performs better than two state-of-the-art methods (*i.e.*, DAIMC and MSC). In addition, we perform the McNemar test between our method and other comparison methods. Specifically, we utilize the "testcholdout" function (Matlab code) to conduct the McNemar test, which returns an h value that indicates whether the two comparison classification models have equal predictive accuracies. That is, $h = 0$ (accepting the null hypothesis) indicates that the two models have equal predictive accuracies, while $h = 1$ indicates that the predictive accuracies between the two methods are statistically different. In our experiments, we obtain $h = 1$ for all the McNemar tests, indicating that the predictions between our method and all comparison methods are significantly different statistically.

D. Results Using Different Combinations of Modalities

The main aim of this study is to learn the latent feature representations from multiple modalities by exploiting the correlations among them. To further analyze the benefit of fusing multi-modality neuroimaging and genetic data, we show the performance of our proposed method for different combinations of modalities on different classification tasks in Fig. 4 (*i.e.*, results with only complete multi-modality testing data) and Fig. 5 (*i.e.*, results with incomplete multi-modality testing data). From Fig. 4 and Fig. 5, it can be seen that our model using MRI and PET data outperforms the models using other two modality combinations (*e.g.*, MRI and SNP, or PET and SNP). These results show that SNP data is less discriminative in AD status diagnosis. A possible reason could be that the MRI and PET data are the phenotype features that are closely related to diagnostic labels, while the SNP data are genotype features that are relatively less related to diagnostic labels [17], [20]. On the other hand, when all the three modalities are used, our model performs better than the case of using any combination of two modalities.

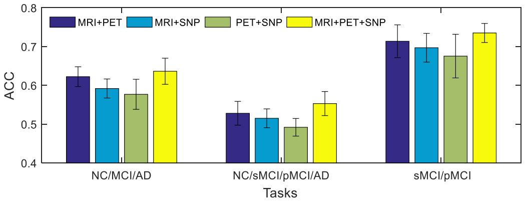


Fig. 4. Classification accuracies of the proposed method trained using all available training data (including incomplete multi-modality data), and tested using only complete multi-modality data. Four different combinations of modalities (*i.e.*, MRI+PET, MRI+SNP, PET+SNP, and MRI+PET+SNP) and three different combinations of disease cohorts (or classification tasks) are used in this group of experiments. The error bars denote the standard deviations of results.

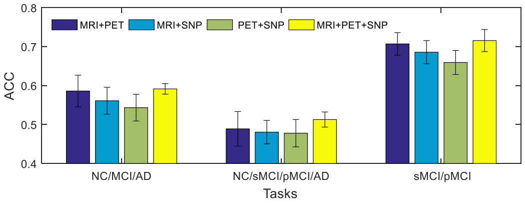
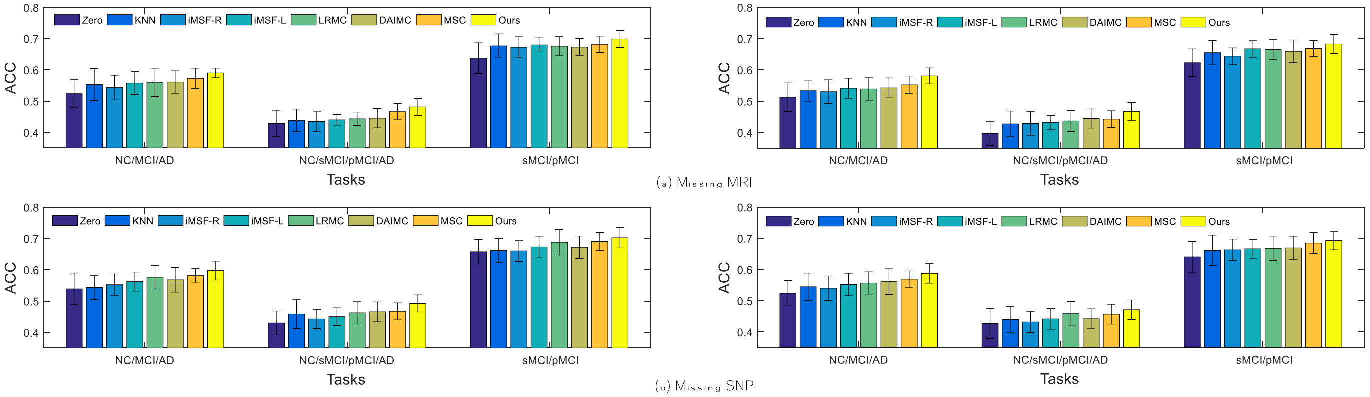


Fig. 5. The classification accuracies of the proposed method, which was trained using all available training data (including incomplete multi-modality data) and tested using all available testing data (including incomplete multi-modality data). Four different combinations of modalities (*i.e.*, MRI+PET, MRI+SNP, PET+SNP, and MRI+PET+SNP) and three different combinations of disease cohorts (or classification tasks) were used in this group of experiments. The error bars denote the standard deviations of results.

E. Study on Different Missing Modalities

In the aforementioned experiments, we use 737 subjects from the ADNI-1 database, where all subjects have complete MRI scans and only 360 subjects have PET data. To further verify the effectiveness of our method in handling the missing data issue, we randomly discard partial MRI or SNP data. Specifically, we randomly select $r\%$ (in this study, we set $r = 10, 20$) of subjects to discard their MRI or SNP data to simulate the missing data issue for MRI and SNP modalities. Then, we conduct a set of experiments to investigate the performance of different methods in AD diagnosis based on these subjects. The experimental results are shown in Fig. 6. From Fig. 6, it can be seen that our proposed method outperforms all the comparison methods under different missing modalities and different missing rates.

To verify the effectiveness of our proposed method in handling problems with more missing data (*e.g.*, $r = 50$), we include a second set of experiments with 50% MRI or PET data being missing. That is, half of the studied subjects have missing MRI or SNP data. Fig. 7 shows the comparison results obtained by using different methods for sMCI/pMCI classification. From Fig. 7, it can be seen that our proposed method outperforms all the comparison methods.



(a) Missing MRI

(b) Missing SNP

Fig. 6. Classification accuracy (*i.e.*, ACC) achieved by different methods using data with $r\%$ (Left: $r = 10$; Right: $r = 20$) subjects associated with missing MRI (top) or SNP (bottom) data. The error bar denotes the standard deviation of the results.

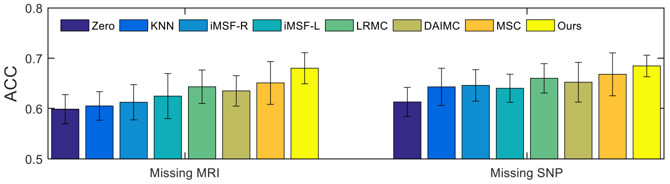


Fig. 7. Classification accuracy (*i.e.*, ACC) achieved by different methods when 50% subjects are with missing MRI (left) or SNP (right) in sMCI/pMCI classification.

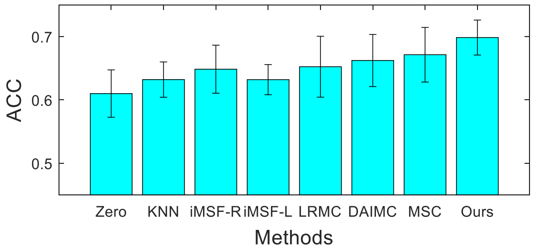


Fig. 8. Classification accuracy (*i.e.*, ACC) achieved by different methods when half of the subjects are missing both MRI and SNP in sMCI/pMCI classification.

Further, we also include a set of experiments when 50% of subjects have no both MRI and SNP data (*i.e.*, only the PET data available) for sMCI/pMCI classification task. Among all 362 subjects in sMCI/pMCI classification, only 185 subjects have the PET data. Therefore, we first select 50% of these 185 subjects and discard both their MRI and SNP data (*i.e.*, these subjects now only have PET data). Fig. 8 shows the comparison results, which further verifies the effectiveness of our method in dealing with the cases when more modalities are missing.

V. Discussion

In this section, we first investigate the influences of different parameters and present the most discriminative ROIs and SNPs identified by our proposed method. Then, we compare our method with several state-of-the-art methods, and provide the differences between our method and several related approaches. Finally, we discuss several limitations of our model.

A. Influences of Parameters

In this section, we study the effects of hyper-parameters (*i.e.*, β, γ, and η). Specifically, we set the values of these parameters in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{2}\}$ for each experiment. We fix the value of one parameter and tune the other two parameters. As an example, Fig. 9 shows the classification accuracy (of the testing samples with complete multi-modality data) achieved by our method for the NC/MCI/AD classification task using different parameter values. From Fig. 9, the experimental results demonstrate that our proposed method obtains better classification results when the values of β, γ, and η fall in [0.1, 10], [1, 10], and [0.1, 10], respectively.

B. Most Related ROIs and SNPs

Based on the formulation in the proposed model (see Eq. (5)), it is expected that when the optimal weight matrices (*i.e.*, $\mathbf{W}_\upsilon$) are found, a small or even zero weight will be assigned to uninformative or less informative feature, while a larger weight will be assigned to more informative feature. For example, for the υ-th modality, the element in the learned $\mathbf{W}_\upsilon$ denotes the contribution of the original feature in the latent space. To study which ROIs/SNPs are selected by our proposed method, we rank the absolute values of all the elements in $\mathbf{W}_\upsilon$, and then report the top elements (with each element corresponding to a specific ROI/SNP) that are frequently selected across all folds.

We show the top ten most related ROIs for MRI and PET data in three classification tasks in Fig. 10 and Fig. 11, respectively. Fig. 10 shows that, hippocampal, amygdala, globus, and lobe WM regions are identified as most related ROIs for AD diagnosis for MRI data. This is consistent with several previous studies which also show that these regions are highly related to AD and MCI classification [44], [55], [56]. For PET data, Fig. 11 shows that precuneus, gyrus, and hippocampal regions are identified as discriminative ROIs for AD and early MCI diagnosis. Again, these identified ROIs are consistent with those reported in previous AD-related studies [20], [44], [57].

In Table S2 of the *Supplementary Materials*, we report those most related ROIs that are frequently selected by our method

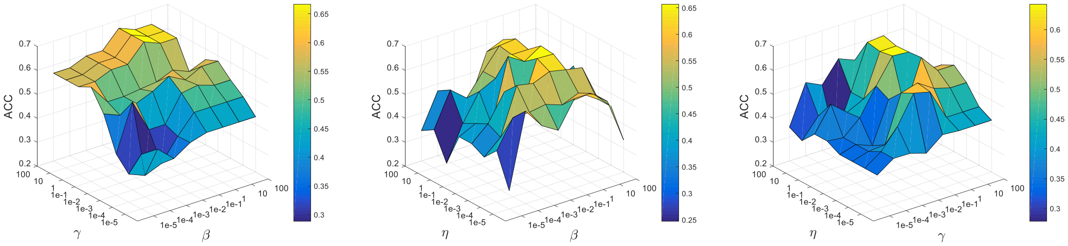


Fig. 9. Classification accuracies of our proposed method for the NC/MCI/AD classification task using different settings of hyper-parameters, *i.e.*, $\beta, \gamma, \eta \in \{10^{-5}, \ldots, 10^{2}\}$.

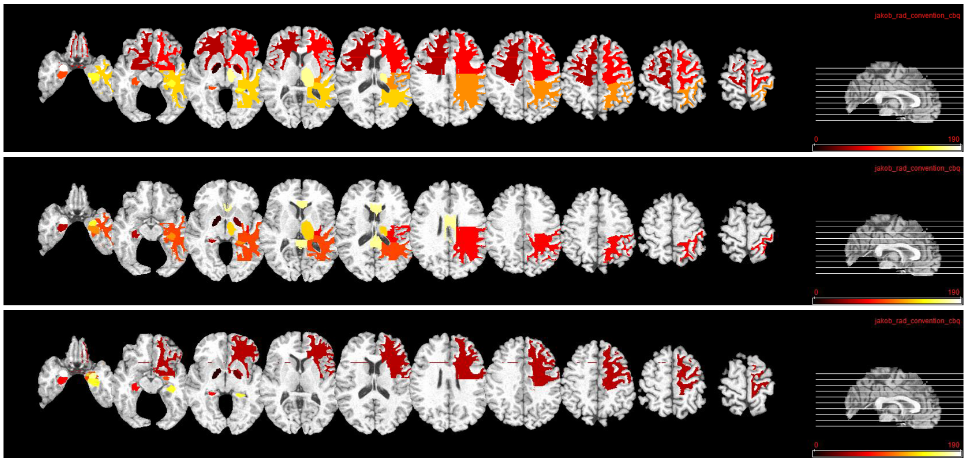


Fig. 10. Top ten selected ROIs for MRI data in three classification tasks. From top to bottom: NC/MCI/AD, NC/sMCI/pMCI/AD, and sMCI/pMCI. Here, different colors denote different ROIs.

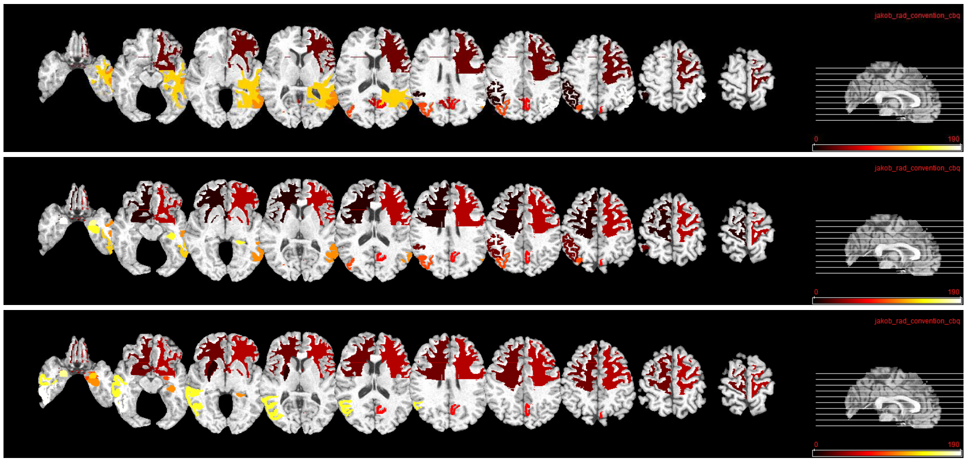


Fig. 11. Top ten selected ROIs for PET data in three classification tasks. From top to bottom: NC/MCI/AD, NC/sMCI/pMCI/AD, and sMCI/pMCI. Here, different colors denote different ROIs.

for the three classification tasks. We also include the color bar to visualize the correspondence between the color and the selected ROI in Fig. S1 and Table S2 of the *Supplementary Materials*. We further report the top five discriminative SNPs that are most frequently identified by our method, which include rs429358, rs10740220, rs2298525, rs7073924, and rs11655156. These genes have been reported to be related to AD in previous studies [17], [58]–[60]. These results suggest that our method is able to identify the most relevant SNPs for AD and early MCI diagnosis. The top 100 selected SNPs can be found in Table S1 of the *Supplementary Materials*.

C. Comparison With State-of-the-Art Methods

We compare the results achieved by our proposed method with those obtained by other state-of-the-art methods that use ADNI subjects. Since very few studies are conducted on multi-class classification tasks, we only report the result of sMCI vs. pMCI classification in Table III. As can be seen, our proposed method generally outperforms other comparison methods [4], [46], [61]–[63] on the sMCI vs. pMCI classification task. The main reason is that our method projects the original features from multi-modality data into a latent space

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON sMCI VS. pMCI CLASSIFICATION

Method	Modality	ACC	AUC
Feature-trans [61]	MRI	0.721	–
Modal-fusion [62]	MRI+CSF	0.685	–
Suk *et al.* [46]	MRI+PET	0.733	–
Thung *et al.* [4]	MRI+PET	0.737	–
Liu *et al.* [63]	MRI+PET	0.733	0.716
Ours	MRI+PET+SNP	0.743	0.755

which provides more comprehensive views for improving diagnosis performance.

D. Comparison with Previous Studies

Our proposed method is similar to but different from several previous studies [64]–[66]. *First*, previous methods [64]–[66] typically conduct the clustering task by first learning common representations and then applying a spectral clustering algorithm on the representation. Different from [64]–[66], our method conducts the latent representation learning and model training in a unified framework, *thus seamlessly integrating them together*. In this way, a “good” latent representation helps to induce a better classification model, and in return a more accurate prediction model will *promote the learning of more discriminative latent representations*. However, previous methods [64]–[66] have ignored the underlying relationship between latent representation learning and model training. *Second*, we utilize the ℓ_1-norm [67]–[71] to alleviate the outlier (*e.g.*, noise) issue in latent space learning, while these methods [64]–[66] do not consider this issue. Intuitively, our method should be more robust to outliers (*e.g.*, noises), resulting in improved classification performance in comparison with previous studies. *Third*, we adopt the $\ell_{2,1}$ constraint on the projection matrices to ensure that the discriminative features have much larger contributions to the latent learning across different modalities. In contrast, the methods in [64]–[66] do not consider this point. We further compare our method with the Partial Multi-View Clustering (PVC) method [64] on the sMCI vs. pMCI classification task, with the comparison results shown in Fig. S2 of the *Supplementary Materials*.

E. Limitations

Although our proposed method achieves promising results in AD and early MCI diagnosis, there are several technical issues to be addressed in future work. First, a linear projection is employed in our model, but it may not be effective to model the complex brain patterns. As such, a non-linear projection can be applied to our formulation in the future. Second, we could extend the proposed model for dealing with the problem of incomplete multi-modality data using a deep learning framework to further improve the classification performance, since deep neural network based features are typically more discriminative than hand-crafted features.

VI. CONCLUSION

In this paper, we propose a novel AD diagnosis framework with latent feature representation learning. Specifically, we first project multi-modality data to a latent feature space, to exploit the underlying association among different modalities. To make use of all available samples to learn an accurate AD prediction model from the incomplete multi-modality dataset, we utilize samples with complete multi-modality data to learn the common latent feature representation, and also utilize samples with the incomplete multi-modality data to learn the modality-specific latent feature representation. Finally, the learned latent feature representations can be linearly projected to the label space for AD diagnosis. Experimental results demonstrate the effectiveness of our proposed method. In future work, our method can be also applied to other diagnosis tasks, such as for schizophrenia [72].

REFERENCES

[1] A. Association, “2013 Alzheimer’s disease facts and figure,” *Alzheimer’s Dementia*, vol. 9, no. 2, pp. 208–245, Mar. 2013.

[2] M. Liu, D. Zhang, and D. Shen, “Relationship induced multi-template learning for diagnosis of Alzheimer’s disease and mild cognitive impairment,” *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1463–1474, Jun. 2016.

[3] T. Zhou, K.-H. Thung, X. Zhu, and D. Shen, “Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis,” *Hum. Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019.

[4] K.-H. Thung, P.-T. Yap, E. Adeli, S.-W. Lee, and D. Shen, “Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion,” *Med. Image Anal.*, vol. 45, pp. 68–82, Apr. 2018.

[5] C. Pennanen *et al.*, “Hippocampus and entorhinal cortex in mild cognitive impairment and early AD,” *Neurobiol. Aging*, vol. 25, no. 3, pp. 303–310, Mar. 2004.

[6] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, “Brain tumor segmentation using convolutional neural networks in MRI images,” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[7] K. H. Thung, C. Y. Wee, P. T. Yap, and D. Shen, “Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion,” *Neuroimage*, vol. 91, pp. 386–400, May 2014.

[8] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen, “Adversarial similarity network for evaluating image alignment in deep learning based registration,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent*. New York, NY, USA: Springer, 2018, pp. 739–746.

[9] C. Lian *et al.*, “Multi-channel multi-scale fully convolutional network for 3D perivascular spaces segmentation in 7T MR images,” *Med. Image Anal.*, vol. 46, pp. 106–117, May 2018.

[10] J. Fan, X. Cao, P.-T. Yap, and D. Shen, “BIRNet: Brain image registration using dual-supervised fully convolutional networks,” *Med. Image Anal.*, vol. 54, pp. 193–206, May 2019.

[11] G. Chetelat, B. Desgranges, V. de la Sayette, F. Viader, F. Eustache, and J. C. Baron, “Mild cognitive impairment: Can FDG-PET predict who is to rapidly convert to Alzheimer’s disease?” *Neurology*, vol. 60, no. 8, pp. 1374–1377, Apr. 2003.

[12] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, “View-aligned hypergraph learning for Alzheimer’s disease diagnosis with incomplete multi-modality data,” *Med. Image Anal.*, vol. 36, pp. 123–134, Feb. 2017.

[13] R. J. Perrin, A. M. Fagan, and D. M. Holtzman, “Multimodal techniques for diagnosis and prognosis of Alzheimer’s disease,” *Nature*, vol. 461, no. 7266, p. 916, 2009.

[14] T. Zhou, K.-H. Thung, M. Liu, F. Shi, C. Zhang, and D. Shen, “Multi-modal neuroimaging data fusion via latent space learning for Alzheimer’s disease diagnosis,” in *Proc. Int. Workshop PRedictive Intell. MEdicine*. Springer, 2018, pp. 76–84.

[15] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert, and A. D. N. Initiative, “Random forest-based similarity measures for multi-modal classification of Alzheimer’s disease,” *NeuroImage*, vol. 65, pp. 167–175, Jan. 2013.

[16] A.-R. Mohammadi-Nejad, G.-A. Hossein-Zadeh, and H. Soltanian-Zadeh, “Structured and sparse canonical correlation analysis as a brain-wide multi-modal data fusion approach,” *IEEE Tran. Med. Imag.*, vol. 36, no. 7, pp. 1438–1448, Jul. 2017.

[17] L. An, E. Adeli, M. Liu, J. Zhang, S.-W. Lee, and D. Shen, "A hierarchical feature and sample selection framework and its application for Alzheimer's disease diagnosis," *Sci. Rep.*, vol. 7, Mar. 2017, Art. no. 45269.

[18] J. Chung, L. A. Farrer, G. Jun, and A. D. N. Initiative, "Genome-wide association study in different clinical stages of Alzheimer's disease," *Alzheimer's Dementia, J. Alzheimer's Assoc.*, vol. 11, no. 7, pp. P357, Jul. 2015.

[19] Q. Li *et al.*, "Genetic interactions explain variance in cingulate amyloid burden: An AV-45 PET genome-wide association and interaction study in the ADNI cohort," *BioMed. Res. Int.*, vol. 2015, Mar. 2015, Art. no. 647389.

[20] J. Peng, L. An, X. Zhu, Y. Jin, and D. Shen, "Structured sparse kernel learning for imaging genetics based Alzheimer's disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervent.*, 2016, pp. 70–78.

[21] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images," *NeuroImage*, vol. 60, no. 1, pp. 59–70, Mar. 2012.

[22] D. Salas-Gonzalez *et al.*, "Feature selection using factor analysis for Alzheimer's diagnosis using F18-FDG PET images," *Med. Phys.*, vol. 37, no. 11, pp. 6084–6095, Nov. 2010.

[23] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, Jun. 2013.

[24] P. Cao, X. Shan, D. Zhao, M. Huang, and O. Zaiane, "Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease," *Pattern Recognit.*, vol. 72, pp. 219–235, Dec. 2017.

[25] I. Jolliffe, *Principal Components Analysis*. Hoboken, NJ, USA: Wiley, 2002.

[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[27] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, Canada, 2005, pp. 507–514.

[28] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal Alzheimer's disease classification," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 3, pp. 984–990, May 2014.

[29] B. B. Avants, P. A. Cook, L. Ungar, J. C. Gee, and M. Grossman, "Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis," *NeuroImage*, vol. 50, no. 3, pp. 1004–1016, Apr. 2010.

[30] L. Yuan *et al.*, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.

[31] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, J. Ye, and A. D. N. Initiative, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, Nov. 2014.

[32] C. R. Jack, jr., *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, 2008.

[33] Z. Xue, D. Shen, and C. Davatzikos, "CLASSIC: Consistent longitudinal alignment and segmentation for serial image computing," *NeuroImage*, vol. 30, no. 2, pp. 388–399, 2006.

[34] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998.

[35] Y. Wang *et al.*, "Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates," *PloS One*, vol. 9, no. 1, 2014, Art. no. e77810.

[36] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[37] N. J. Kabani, "3D anatomical atlas of the human brain," *NeuroImage*, vol. 7, p. 717, Feb. 1998.

[38] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Aug. 2002.

[39] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi, "Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database," *Nature Genet.*, vol. 39, no. 1, pp. 17–23, Jan. 2007.

[40] Y. Zhang, C. S. Nam, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Temporally constrained sparse group spatial patterns for motor imagery BCI," *IEEE Trans. Cybern.*, to be published.

[41] J. Wang *et al.*, "Multi-task diagnosis for autism spectrum disorders using multi-modality features: A multi-center study," *Human Brain Mapping*, vol. 38, no. 6, pp. 3081–3097, 2017.

[42] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *Proc. Int. Conf. Mach. Learn.*, Feb. 2013, pp. 352–360.

[43] Z. Lin, M. Chen, and Y. Ma. (2010). "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices." [Online]. Available: https://arxiv.org/abs/1009.5055

[44] X. Zhu, H. Suk, S. 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. 2015.

[45] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[46] H. Suk, S. Lee, and D. Shen, "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis," *Brain Struct. Function*, vol. 220, no. 2, pp. 841–859, 2015.

[47] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[48] P. Liu, J. P. Lewis, and T. Rhee, "Low-rank matrix completion to reconstruct incomplete rendering images," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 8, pp. 2353–2365, Aug. 2017.

[49] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, Jun. 2013, pp. 665–674.

[50] M. Hu and S. Chen, "Doubly aligned incomplete multi-view clustering," in *Proc. IJCAI*, Jul. 2018, pp. 2262–2268.

[51] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and Bregman iterative methods for matrix rank minimization," *Math. Program.*, vol. 128, nos. 1–2, pp. 321–353, 2011.

[52] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, Apr. 2011.

[53] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, 2001.

[54] J. Liu, S. Ji, and J. Ye, "SLEP: Sparse learning with efficient projections," *Arizona State Univ.*, vol. 6, no. 491, p. 7, 2009.

[55] T. M. Nir *et al.*, "Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging," *NeuroImage*, vol. 3, pp. 180–195, May 2013.

[56] A. Convit, J. de Asis, M. de Leon, C. Tarshish, S. de Santi, and H. Rusinek, "Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease," *Neurobiol. Aging*, vol. 21, no. 1, pp. 19–26, Feb. 2000.

[57] M. Liu, D. Zhang, S. Chen, and H. Xue, "Joint binary classifier learning for ECOC-based multi-class classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2335–2341, Nov. 2016.

[58] A. J. Saykin *et al.*, "Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans," *Alzheimer's Dementia*, vol. 6, no. 3, pp. 265–273, 2010.

[59] H. A. Wishart *et al.*"Increased brain activation during working memory in cognitively intact adults with the APOE $\varepsilon 4$ allele," *Amer. J. Psychiatry*, vol. 163, no. 9, pp. 1603–1610, Jul. 2006.

[60] T. Zhou, K.-H. Thung, M. Liu, and D. Shen, "Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 165–175, Jan. 2019.

[61] X. Zhu, H.-I. Suk, Y. Zhu, K.-H. Thung, G. Wu, and D. Shen, "Multi-view classification for identification of Alzheimer's disease," in *Proc. Int. Workshop Mach. Learn. Med. Imaging*. New York, NY, USA: Springer, 2015, pp. 255–262.

[62] E. Westman, J.-S. Muehlboeck, and A. Simmons, "Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion," *NeuroImage*, vol. 62, no. 1, pp. 229–238, 2012.

[63] M. Liu, D. Zhang, and D. Shen, "Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis," *Hum. Brain Mapping*, vol. 35, no. 4, pp. 1305–1319, Apr. 2014.

[64] S.-Y. Li, Y. Jiang, and Z.-H. Zhou, "Partial multi-view clustering," in *Proc. AAAI*, Jun. 2014, pp. 1968–1974.

[65] H. Zhao, H. Liu, and Y. Fu, “Incomplete multi-modal visual data grouping,” in *Proc. Int. Joint Conf. Artif. Intell.*, May 2016, pp. 2392–2398.

[66] J. Wen, Z. Zhang, Y. Xu, and Z. Zhong. (2018). “Incomplete multiview clustering via graph regularized matrix factorization.” [Online]. Available: https://arxiv.org/abs/1809.05998

[67] K.-H. Thung, P.-T. Yap, and D. Shen, “Joint robust imputation and classification for early dementia detection using incomplete multi modality data,” in *Proc. Int. Workshop PRedictive Intell. MEdicine*. Springer, 2018, pp. 51–59.

[68] B. Jie, M. Liu, J. Liu, D. Zhang, and D. Shen, “Temporally constrained group sparse learning for longitudinal data analysis in Alzheimer’s disease,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 238–249, Jan. 2017.

[69] T. Zhou, F. Liu, H. Bhaskar, and J. Yang, “Robust visual tracking via online discriminative and low-rank dictionary learning,” *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2643–2655,

[70] M. Liu and D. Zhang, “Feature selection with effective distance,” *Neurocomputing*, vol. 215, pp. 100–109, Nov. 2016.

[71] 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, 2015.

[72] Y. Fan *et al.*, “Unaffected family members and schizophrenia patients share brain structure patterns: A high-dimensional pattern classification study,” *Biol. Psychiatry*, vol. 63, no. 1, pp. 118–124, 2008.