This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

Relation-induced Multi-modal Shared Representation Learning for Alzheimer's Disease Diagnosis

Zhenyuan Ning, *Student Member, IEEE,* Qing Xiao, *Student Member, IEEE,* Qianjin Feng, *Member, IEEE,* Wufan Chen, *Senior Member, IEEE,* and Yu Zhang

***Abstract*—The fusion of multi-modal data (e.g., magnetic resonance imaging (MRI) and positron emission tomography (PET)) has been prevalent for accurate identification of Alzheimer's disease (AD) by providing complementary structural and functional information. However, most of the existing methods simply concatenate multi-modal features in the original space and ignore their underlying associations which may provide more discriminant characteristics for AD identification. Meanwhile, how to overcome the overfitting issue caused by high-dimensional multi-modal data remains appealing. To this end, we propose a relation-induced multi-modal shared representation learning method for AD diagnosis. The proposed method integrates representation learning, dimension reduction, and classifier modeling into a unified framework. Specifically, the framework first obtains multi-modal shared representations by learning a bi-directional mapping between original space and shared space. Within this shared space, we utilize several relational regularizers (including feature-feature, feature-label, and sample-sample regularizers) and auxiliary regularizers to encourage learning underlying associations inherent in multi-modal data and alleviate overfitting, respectively. Next, we project the shared representations into the target space for AD diagnosis. To validate the effectiveness of our proposed approach, we conduct extensive experiments on two independent datasets (i.e., ADNI-1 and ADNI-2), and the experimental results demonstrate that our proposed method outperforms several state-of-the-art methods.**

***Index Terms*—Alzheimer's Disease, multi-modal neuroimages, shared representations, relational regularization**

I. Introduction

ALZHEIMER'S disease (AD), as one of the most common neurodegenerative diseases in elderly people, is characterized by irreversible loss of neurons and genetically complex disorder [1]. As the disease progresses, it will result in irreversible brain atrophy and make patients need around-the-clock care which places economic and psychological burdens. Fortunately, early diagnosis of AD is beneficial to patient care and help to slow down progressive deterioration [2]. Thus, accurate identification of AD and its prodromal stage, i.e., mild cognitive impairment (MCI), has drawn extensive attention [3].

Z. Ning, Q. Xiao, Q. Feng, W. Chen, and Y. Zhang are with the School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, 510515, China (E-mail: jonnyning@foxmail.com; felicity_shaw@foxmail.com; 1271992826@qq.com; chenwf@smu.edu.cn; yuzhang@smu.edu.cn). Z. Ning and Q. Xiao contributed equally to this work.

Corresponding author: Yu Zhang (yuzhang@smu.edu.cn). This work was supported in part by the National Natural Science Foundation of China (61971213, 61671230) (Y. Zhang), and in part by the Basic and Applied Basic Research Foundation of Guangdong Province (2019A1515010417) (Y. Zhang).

Neuroimaging techniques have been considered as useful tools for brain disease progression identification [4]. In particular, it has been shown that multi-modal neuroimages (e.g., magnetic resonance imaging (MRI) and positron emission tomography (PET)) can provide complementary structural (e.g., brain atrophy) and functional information (e.g., metabolism) of abnormal brain regions [5]. Therefore, an interesting topic is to develop a robust and accurate model based on multi-modal neuroimages for early diagnosis of AD [6], [7].

Conventional multi-modal based machine learning methods typically concatenate the multi-modal features in the original space for AD identification [8], [9]. However, direct concatenation of multi-modal features cannot take full advantages of the complementary information that exists in multi-modal data. To improve predictive performance, several methods [10]–[13] have been proposed to fuse multi-modal data by exploiting their complementary information for AD diagnosis. For example, Hinrichs *et al*. [10] proposed a multi-kernel learning (MKL) based model to fuse multi-modal features by simultaneously learning kernel weights and a maximum margin classifier. Zhou *et al*. [11], [13] learned a latent space that preserved the specific information of multi-modal data and then projected the features in the latent space into label space for performing prediction. Zhu *et al*. [12] utilized canonical correlation analysis (CCA) to combine multi-modal information by mapping original multi-modal data to a common space and constructed support vector models for joint regression and classification of AD. Although these methods are promising, how to explore the underlying associations inherent in multi-modal data and generate distinguishing representations for AD diagnosis is still challenging.

Moreover, the low sample-to-feature ratio in multi-modal researches brings "dimension curse" issue that may easily result in overfitting [14]. To address this issue, previous studies have conducted feature selection or feature reduction approaches to select informative features for model construction [15]–[18]. For instance, Nie *et al*. [15] applied $L_{2,1}$ norm on the weight of features to remove uninformative features. Similarly, based on $L_{2,1}$ norm, Jie *et al*. [16] constructed a Laplacian matrix to make feature subspace preserve the local structure of original data. To exploit relational information inherent in observations, Zhu *et al*. [17] and Lei *et al*. [18] used relational regularization terms to select features for joint regression and classification in AD diagnosis. Besides, some classical approaches, such as linear discriminant analysis (LDA) and principal component

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

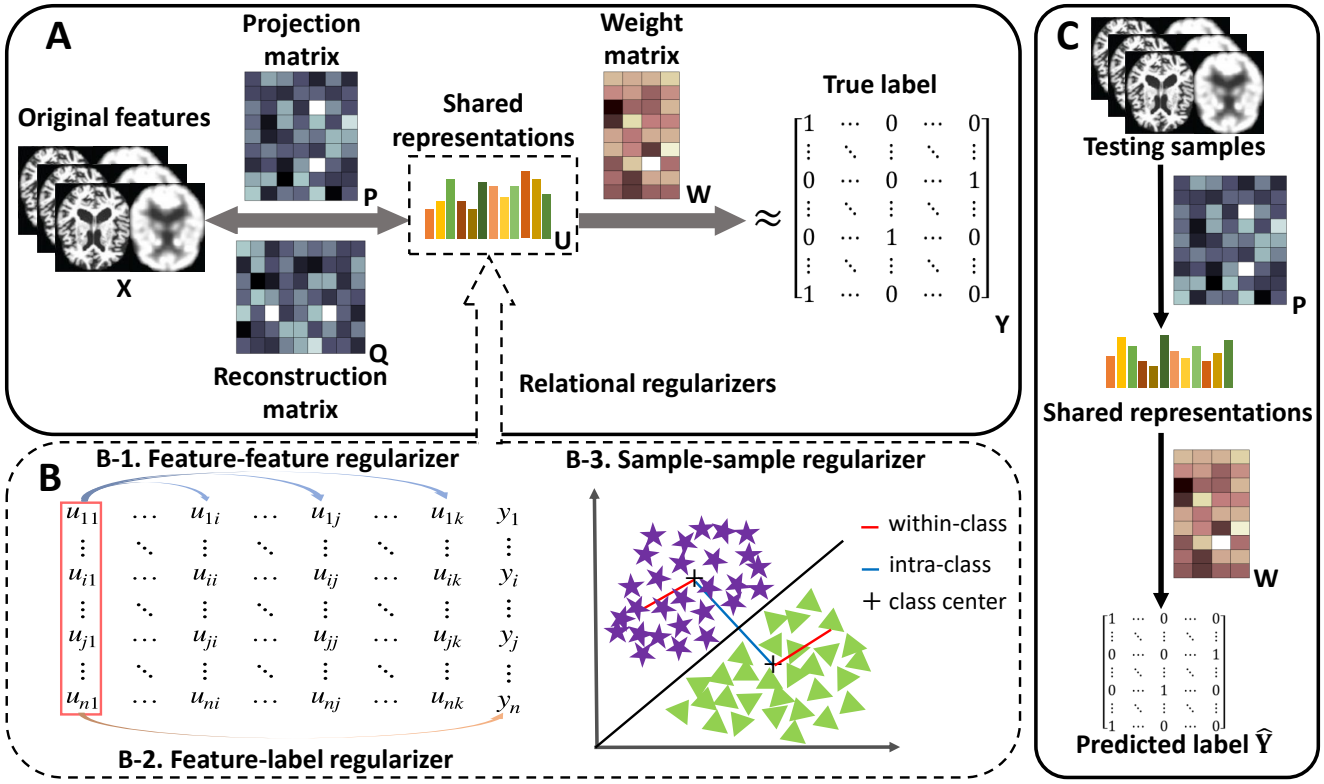

Fig. 1. Flowchart of our proposed framework for Alzheimer's Disease diagnosis. At training stage (A), a shared representations matrix $\mathbf{U}$ is obtained by a bi-directional mapping (including projection matrix $\mathbf{P}$ and reconstruction matrix $\mathbf{Q}$) between original space and shared space. Subsequently, these generated representations are projected to the label space for final prediction. Regularizers devised for representation learning are presented in the dash rectangle B. Specifically, feature-feature and feature-label regularizers are plotted in the B-1 and B-2, respectively. The red solid rectangle denotes the feature vector in shared representations. B-3 gives an illustration of sample-sample (i.e., within-class and intra-class) regularizer, where purple stars and green triangles denote samples of two classes, cross signs represent the mean vectors of each class, and red and blue lines symbolize the within-class and intra-class distances, respectively. For testing (C), a new instance can be directly classified by the projection matrix $\mathbf{P}$ and coefficient matrix $\mathbf{W}$ that are learned at training stage.

analysis (PCA), have also been widely applied to AD-related studies [19], [20]. Although various feature selection/reduction methods have been proposed, there are still two points to further enhance the effectiveness of multi-modal models. *First*, current approaches generally select features in the original feature space. However, with reference to the discussion in [11] and [12], mapping original multi-modal data to a latent space could help to capture the potential characteristics among different modalities. Accordingly, utilizing features in this latent space may boost the diagnostic performance of model. *Second*, previous studies usually perform feature reduction and classifier construction separately, while it has been suggested to further improve the performance of model by jointly training feature reduction and classifier [21].

In this paper, we propose a relation-induced multi-modal shared representation learning framework for AD diagnosis. Fig. 1 illustrates the flowchart of our framework, in which Fig. 1-A, Fig. 1-B, and Fig. 1-C correspond to the training stage, relational regularizers and test stage, respectively. At training stage (Fig. 1-A), the framework first obtains shared representations by learning a bi-directional mapping between original space and shared space. For one thing, we hope to learn latent discriminative representations from multi-modal data by introducing the projection matrix $\mathbf{P}$ which conducts original-to-shared transformation. And for another, we also expect the shared representations can preserve original information as much as possible, and thus the reconstruction matrix $\mathbf{Q}$ is utilized to achieve shared-to-original conversion. We further project the shared representations $\mathbf{U}$ into target space (i.e, label space) by weight matrix $\mathbf{W}$, whose elements stand for the importance of the corresponding feature vectors in $\mathbf{U}$ for final AD diagnosis. Thus, representation learning (from original space to shared space) and classifier modeling (from shared space to label space) are integrated into the unified framework and can be optimized simultaneously (see Section III for details). To encourage learning underlying associations existing in multi-modal data for inducing more distinguishing representations for AD diagnosis, we devise three relational regularizers for the shared space, as depicted in Fig. 1-B, including feature-feature, feature-label, and sample-sample regularizers. The rationales for the devised regularizers are as follows: (1) The learned features should keep low redundancy among themselves. We assume that if a feature can be represented by a linear combination of the rest features in the shared space, it is regarded as the redundant feature when a linear classifier is used for prediction. Thus, redundant features contribute less additional information for accurate classification model construction (as shown in Fig. 1-B1, feature-feature regularizer); (2) The learned features are required to have high relevance with labels. A discriminative feature should have

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

the ability to estimate labels, and thus it's desired to have close connection with labels (as shown in Fig. 1-B2, feature-label regularizer); (3) For enhancing the class separability in the shared space, it's expected that the latent representations of the same-class samples are closed to each other, and the distance of different-class centers is as large as possible (as shown in Fig. 1-B3, sample-sample regularizer). We will describe the regularizers in detail in Section III. Finally, at testing stage (Fig. 1-C), clinical labels of the testing samples can be predicted by two successive projection matrices, i.e., projection matrix $\mathbf{P}$ and weight matrix $\mathbf{W}$. To validate the effectiveness of our proposed method, we conduct extensive experiments on two independent datasets (i.e., ADNI-1 and ADNI-2). It is worthwhile to highlight the main contributions of this work:

- Bi-directional mapping simultaneously considers both data projection and data reconstruction, which helps to learn a latent shared space that preserves original information as much as possible.
- Several regularizers are devised to explore the underlying associations for inducing distinguishing representation learning in the shared space. Feature-wise regularizers make the shared representations more compact and discriminative, while sample-wise regularizers aim to enhance the class separability in the shared space.
- The unification of representation learning, dimension reduction, and classifier modeling makes it possible to optimize these three parts jointly and conduct training and testing process in an efficient way. The experimental results demonstrate our proposed method can not only perform accurate prediction but also discover potential biomarkers for AD identification.

The rest of this paper is organized as follows. In Section II, we describe the data used in this study and introduce the data preprocessing steps. Subsequently, we give a detailed description of our method in Section III. Experimental results are presented in Section IV, followed by discussion and conclusion in Section V and VI, respectively.

II. Data and Image Preprocessing

We obtained data from the public ADNI database[1], which provides various types of data, including neuroimaging, clinical, and genetic information for AD. Totally, we collected 820 subjects that have completely matched MRI and PET images from ADNI-1 and ADNI-2. According to some clinical criteria, such as clinical dementia rating and mini-mental state examination score, these subjects were classified into three categories, namely, normal control (NC), MCI, and AD. Considering that a part of MCI subjects would convert to AD and the others would be stable over time, MCI subjects were further divided into progressive MCI (pMCI) and stable MCI (sMCI). In summary, 93 AD, 99 NC, 121 sMCI, and 79 pMCI from ADNI-1 and 136 AD, 107 NC, 103 sMCI, and 82 pMCI from ADNI-2 were enrolled. More demographic information can be found in TABLE I.

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

TABLE I
Demographic Information of Subjects (AD: Alzheimer's Disease, NC: Normal Control, sMCI: stable MCI, pMCI: progressive MCI).

		AD	NC	sMCI	pMCI
ADNI-1	Female/Male	37/56	39/60	35/86	31/48
	Age	75.4±7.4	75.7±4.8	74.9±7.5	75.0±6.7
	MMSE	23.5±2.1	28.9±1.1	27.4±1.7	26.8±1.7
	Education	14.8±3.0	15.8±3.1	15.8±2.9	15.8±2.7
ADNI-2	Female/Male	55/81	56/51	44/59	36/46
	Age	74.2±8.2	72.9±6.3	71.9±7.2	72.7±7.2
	MMSE	23.1±2.1	29.0±1.3	28.2±1.6	27.2±1.8
	Education	15.9±2.7	16.4±2.6	16.3±2.6	16.2±2.5

We downloaded raw MRI images acquired by 1.5T or 3T scanners with various individualized protocols. All images have been reviewed and corrected by ADNI researchers for spatial distortion caused by B1 field inhomogeneity and gradient nonlinearity. Our image preprocessing contains following procedures: (1) anterior commissure-posterior commissure (AC-PC) correction via MIPAV software[2], (2) intensity inhomogeneity correction using N3 algorithm [22], (3) skull stripping and cerebellum removal with aBEAT[3], (4) three main tissues (i.e. gray matter (GM), white matter, and cerebrospinal fluid) segmentation via FAST algorithm [23], (5) registering images to a template via HAMMER [24], and (6) projecting 90 region of interest (ROI) labels from the template to each subject image. For each subject, we calculated the GM tissue volume of each ROI and regarded it as specific ROI-based feature. For PET images, we aligned them to the corresponding MRI images using affine registration and calculated the average intensity value of each ROI as its feature. Finally, we obtained 90 MRI features and 90 PET features for each subject.

III. Methods

In this paper, matrix, vector, and scalar are symboled by bold uppercase letter, bold lowercase letter, and normal lowercase letter, respectively. For clarity, we list the main notations in TABLE II.

A. Regularized Regression

Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ denote a feature matrix, where n is the number of subjects, m is feature dimension, and the i-th row and j-th column of feature matrix $\mathbf{X}$ are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. $\mathbf{Y} \in \{0,1\}^{n\times c}$ is the corresponding label matrix, where c is the number of classes. $\mathbf{W} \in \mathbb{R}^{m\times c}$ is a weight matrix. Thus, a least square regression model with regularization can be given as:

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \rho R(\mathbf{W}), \tag{1}$$

where $\|\cdot\|_F = \sqrt{\sum_i \|\mathbf{x}_i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}^j\|_2^2}$ denotes Frobenius norm, $R(\mathbf{W})$ is a regularization term (e.g., L_1 norm, L_2 norm,

[2]http://mipav.cit.nih.gov/clickwrap.php
[3]https://www.nitrc.org/projects/abeat

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

TABLE II
MAIN NOTATIONS USED IN FORMULA

Symbol	Size	Description
$\mathbf{X}$	$n \times m$	Original feature matrix with n samples and m features
$\mathbf{X}_M$	$n \times m_1$	Original MRI feature matrix with n samples and m_1 features
$\mathbf{X}_P$	$n \times m_2$	Original PET feature matrix with n samples and m_2 features
$\mathbf{U}$	$n \times k$	Shared representations with n samples and k features
$\mathbf{P}$	$(m_1+m_2)\times k$	Projection matrix (from the original space to the shared space)
$\mathbf{Q}$	$k\times(m_1+m_2)$	Reconstruction matrix(from the shared space to the original space)
$\mathbf{Y}$	$n \times c$	Label matrix with c classes
$\mathbf{W}$	$k \times c$	Weight matrix
$\mathbf{R}(\cdot)$	-	Regularization terms
γ_i, λ_i, ρ_i	-	Regularization parameters

and $L_{2,1}$ norm) and ρ is a nonnegative parameter that balances the importance of the regularization term. In the first part of Eq.(1), label matrix $\mathbf{Y}$ can be estimated by a linear transformation of the feature matrix $\mathbf{X}$. To avoid overfitting issue and improve generalization performance, many regularization terms have been embedded into the least square regression model. For example, ridge regression uses L_2 norm to obtain a trade-off between data fitting and model simplicity. Lasso regression utilizes L_1 norm to build a more sparse model. Additionally, variants of classical regularization terms have also been used to encourage the first part in Eq.(1) to satisfy certain properties [17], [18].

B. Relation-induced Multi-modal Shared Representation Learning

When both MRI and PET data are available, the original feature matrices $\mathbf{X}$ equals to $[\mathbf{X}_M, \mathbf{X}_P] \in \mathbb{R}^{n\times(m_1+m_2)}$, where $\mathbf{X}_M \in \mathbb{R}^{n\times m_1}$ and $\mathbf{X}_P \in \mathbb{R}^{n\times m_2}$ denote the original MRI and PET feature matrix, m_1 and m_2 denote the feature dimension of MRI and PET, respectively. To learn comprehensive information from multi-modal neuroimages, we assume that multi-modal data can be projected into a shared space, whose representations can also reconstruct the original features. Therefore, a multi-modal bi-directional mapping can be defined as follows:

$$\min_{\mathbf{P},\mathbf{Q}} \|\mathbf{U} - [\mathbf{X}_M, \mathbf{X}_P]\mathbf{P}\|_F^2 + \|[\mathbf{X}_M, \mathbf{X}_P] - \mathbf{U}\mathbf{Q}\|_F^2. \quad (2)$$

In the first term, the multi-modal original feature matrix $[\mathbf{X}_M, \mathbf{X}_P]$ is projected into the shared space via a projection matrix $\mathbf{P} \in \mathbb{R}^{(m_1+m_2)\times k}$, where k refers to the feature dimension of latent representations. Meanwhile, the original feature space is reconstructed by the shared repsentations $\mathbf{U} \in \mathbb{R}^{n\times k}$ using a reconstruction matrix (or back projection matrix) $\mathbf{Q} \in \mathbb{R}^{k\times(m_1+m_2)}$ in the second term to ensure the shared representations retain original information as much as possible. Therefore, we replace raw feature matrix in $\|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_F^2$ with shared representations and reformulate Eq. (1) as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P},\mathbf{Q}} \gamma_1\|\mathbf{Y} - \mathbf{U}\mathbf{W}\|_F^2 + \gamma_2\|\mathbf{U} - [\mathbf{X}_M, \mathbf{X}_P]\mathbf{P}\|_F^2 \\ + \gamma_3\|[\mathbf{X}_M, \mathbf{X}_P] - \mathbf{U}\mathbf{Q}\|_F^2 + \rho R(\mathbf{W}, \mathbf{Q}, \mathbf{P}, \mathbf{U}),\end{aligned} \quad (3)$$

where γ_1, γ_2, γ_3, and ρ are the trade-off parameters that are used to balance different terms. In the first term, we utilize the shared representations that considers underlying relationship (under the constraint of $R(\mathbf{U})$) between different modalities data to predict clinical labels. For the last term, we define $\rho R(\mathbf{W}, \mathbf{Q}, \mathbf{P}, \mathbf{U}) = \rho_1\|\mathbf{W}\|_F^2 + \rho_2\|\mathbf{Q}\|_F^2 + \rho_3\|\mathbf{P}\|_F^2 + \rho_4\|\mathbf{U}\|_F^2$, where $\|\cdot\|_F^2 = \sum_i \|\mathbf{x}_i\|_2^2 = \sum_j \|\mathbf{x}^j\|_2^2$ (i.e., squared Frobenius norm), as auxiliary regularization to circumvent the overfitting issue [25]. The middle two terms are for learning latent shared representations. So far, the Eq. (3) has integrated the representation learning and classifier modeling into a unified framework, which encourages learning discriminative multi-modal shared representations in a task-oriented manner.

To explore the potential associations within multi-modal data for learning more compact and discriminative representations in the shared space, we devise several regularization terms based on the feature-feature, feature-label, and sample-sample relation and use them to penalize the target function. Firstly, we assume that if a feature can be represented by a linear combination of the rest features in the shared space, it is regarded as the redundant feature when a linear classifier is used for prediction. We refer this relation as feature-feature relation in this work. Taking $\mathbf{u}_j$ for example, we introduce a new matrix $\tilde{\mathbf{U}}_{(j)} \in \mathbb{R}^{n\times(k-1)}$, where consists of all feature vectors in $\mathbf{U}$ except $\mathbf{u}_j$. If $\mathbf{u}_j$ can be represented by the linear combination of the rest vectors as

$$\mathbf{u}_j = d_{j1}\mathbf{u}_1 + d_{j2}\mathbf{u}_2 + \cdots + d_{jk}\mathbf{u}_k = \tilde{\mathbf{U}}_{(j)}\mathbf{d}_j, \quad (4)$$

where $\mathbf{d}_j \neq \mathbf{0}$. When a linear classifier is used, we can obtain

$$\hat{\mathbf{Y}} = \mathbf{U}\mathbf{W} = \mathbf{u}_j\mathbf{w}^j + \tilde{\mathbf{U}}_{(j)}\tilde{\mathbf{W}} = \tilde{\mathbf{U}}_{(j)}(\mathbf{d}_j\mathbf{w}^j + \tilde{\mathbf{W}}). \quad (5)$$

Therefore, $\hat{\mathbf{Y}}$ can be approached by the rest $(k-1)$ feature vectors without $\mathbf{u}_j$, which suggests that $\mathbf{u}_j$ is redundant among $\mathbf{U}$.

To this end, we need to minimize the correlation between $\mathbf{u}_j$ and $\tilde{\mathbf{U}}_{(j)}\mathbf{d}_j$ and make it less possible to use $\tilde{\mathbf{U}}_{(j)}$ to represent $\mathbf{u}_j$ for prediction. Thus, we devise following term to define feature-feature regularizer:

$$\begin{aligned}\mathbf{R_{FF}} &= \sum_{j=1}^{k} r{<}\mathbf{u}_j, d_{j1}\mathbf{u}_1 + d_{j2}\mathbf{u}_2 + \cdots + d_{jk}\mathbf{u}_k{>} \\ &= \sum_{j=1}^{k} r{<}\mathbf{u}_j, \tilde{\mathbf{U}}_j\mathbf{d}_j{>} = \tfrac{1}{n-1}tr(\mathbf{U}^T\tilde{\mathbf{U}}),\end{aligned} \quad (6)$$

where $\tilde{\mathbf{U}} = [\tilde{\mathbf{U}}_1\mathbf{d}_1, \tilde{\mathbf{U}}_2\mathbf{d}_2, \ldots, \tilde{\mathbf{U}}_k\mathbf{d}_k]$, and $r{<}\mathbf{a}, \mathbf{b}{>}$ is computed as follows:

$$r{<}\mathbf{a}, \mathbf{b}{>} = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{a_i - \bar{a}}{\sigma_a})(\frac{b_i - \bar{b}}{\sigma_b}) = \frac{1}{n-1}\mathbf{a}^T\mathbf{b}. \quad (7)$$

For vectors $\mathbf{a}$ and $\mathbf{b}$, n is the vector length, a_i and b_i are the elements, $\bar{a}$ and $\bar{b}$ are the mean values, and σ_a and σ_b are

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

the standard deviations.The right equality holds when $\mathbf{a}$ and $\mathbf{b}$ are normalized by z-score normalization (i.e., $\bar{a}, \bar{b} = 0$ and $\sigma_a, \sigma_b = 1$).

Although feature-feature regularizer can induce learning low-redundant shared representations, it cannot guarantee that these representations are able to estimate clinical labels accurately. Intuitively, a discriminative feature is desired to have close connection (i.e., high relevance) with labels, so we refer this relation as feature-label relation and define the regularizer as follows:

$$\begin{aligned}\mathbf{R_{FY}} =& \sum_{i=1}^{k}\sum_{j=1}^{c}\left|r<\mathbf{u}_i, \mathbf{y}_j>\right| = \|\tfrac{1}{n-1}\mathbf{U}^T\mathbf{Y}\|_F \\ =& \frac{1}{n-1}\sqrt{tr(\mathbf{U}^T\mathbf{Y}\mathbf{Y}^T\mathbf{U})}.\end{aligned} \tag{8}$$

For computational convenience, the constant $\frac{1}{n-1}$ is included into parameters of above regularization terms ($\mathbf{R_{FF}}$ and $\mathbf{R_{FY}}$) and squared Frobenius norm is used to approximate original formula of $\mathbf{R_{FY}}$. The simplified versions of Eq.(6) and Eq.(8) used in the target function are as follows:

$$R_1(\mathbf{U}) = tr(\mathbf{U}^T\tilde{\mathbf{U}}), \tag{9}$$

$$R_2(\mathbf{U}) = -\,tr(\mathbf{U}^T\mathbf{Y}\mathbf{Y}^T\mathbf{U}), \tag{10}$$

where the negative sign is introduced to convert the maximization problem to the minimization one. Different from the filter method maximal relevance and minimal redundancy (mRMR) algorithm [26], which selects features based on mutual information, our proposed method embeds the devised feature-feature and feature-label regularizers into the target function to induce learning low-redundant and high-relevant shared representations for AD diagnosis in a task-oriented manner. It's worth mentioning that the orthogonal constraint has been widely used to subspace learning. However, most previous works only focused on the feature redundancy and commonly ignored the relevance between feature and label. Inspired by the mRMR method, we hope the shared representations with both minimum redundancy and maximum relevance. In this circumstance, orthogonal constraint may be too strict to ensure the close relation between feature and label. Thus, we defined a novel feature-feature regularizer to reduce the correlation among features in a mild way.

Moreover, it's easier to separate different-class samples when the shared representations of the same-class samples are closed to each other and distance of different-class centers are as large as possible. Thus, we introduce the third sample-sample relation to enhance the class separability in the shared space. We make two definitions as follows:

$$\mathbf{R_{SS1}}(\mathbf{U}) = tr\Big\{\sum_{i=1}^{c}\sum_{j=1}^{n_i}\big[(\mathbf{u}_{(i)}^j - \mathbf{m}^i)^T(\mathbf{u}_{(i)}^j - \mathbf{m}^i)\big]\Big\}, \tag{11}$$

$$\mathbf{R_{SS2}}(\mathbf{U}) = tr\Big\{\sum_{i=1}^{c} n_i(\mathbf{m}^i - \bar{\mathbf{m}})^T(\mathbf{m}^i - \bar{\mathbf{m}})\Big\}, \tag{12}$$

where $\mathbf{R_{SS1}}(\mathbf{U})$ and $\mathbf{R_{SS2}}(\mathbf{U})$ are the distance of within-class samples and distance of intra-class centers, respectively, n_i is the number of samples in the i-th class, $\mathbf{u}_{(i)}^j$ is the feature vector of the j-th sample belong to the i-th class, $\mathbf{m}^i$ is the mean vector of the samples in i-th class, and $\bar{\mathbf{m}}$ is the average vector of all samples. Accordingly, the sample-sample relation deduces two sample-sample regularizers. With some algebraic steps, we can get and minimize the modified version of Eq. (11) and Eq. (12) as follows:

$$R_3(\mathbf{U}) = tr(\mathbf{U}^T\mathbf{L}\mathbf{U}), \tag{13}$$

$$R_4(\mathbf{U}) = -tr(-\mathbf{U}^T\mathbf{L}\mathbf{U} + \mathbf{U}^T\mathbf{U}) = tr(-\mathbf{U}^T\mathbf{S}\mathbf{U}), \tag{14}$$

where $\mathbf{s}_{ij} = \frac{1}{n_i}$ if both $\mathbf{u}^i$ and $\mathbf{u}^j$ belong to the i-th class, otherwise $\mathbf{s}_{ij} = 0$. And $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal matrix with its i-th diagonal element being the sum of the i-th row of $\mathbf{S}$. It's worthy noting that normalization is needed to guarantee zero mean value in Eq.(14).

Finally, our target function can be formulated as:

$$\begin{aligned}J =& \gamma_1\|\mathbf{Y} - \mathbf{U}\mathbf{W}\|_F^2 + \gamma_2\|\mathbf{U} - [\mathbf{X}_M, \mathbf{X}_P]\mathbf{P}\|_F^2 \\ &+ \gamma_3\|[\mathbf{X}_M, \mathbf{X}_P] - \mathbf{U}\mathbf{Q}\|_F^2 + \lambda_1 R_1(\mathbf{U}) + \lambda_2 R_2(\mathbf{U}) \\ &+ \lambda_3 R_3(\mathbf{U}) + \lambda_4 R_4(\mathbf{U}) + \rho R(\mathbf{W}, \mathbf{Q}, \mathbf{P}, \mathbf{U}),\end{aligned} \tag{15}$$

where λ_i ($i = 1, 2, 3, 4$) are the trade-off parameters to balance different relational regularizers.

C. Optimization

The objective function in Eq.(15) is non-convex with respect to all variables $\mathbf{U}$, $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{W}$. Fortunately, it is convex with respect to any one of these four variables when the others are fixed. In the following section, we employ an iterative algorithm to solve the problem effectively.

(1) Optimizing $\mathbf{U}$: fixing $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{W}$, we rewrite the target function with respect to $\mathbf{U}$ as follows:

$$\begin{aligned}\min_{\mathbf{U}} &\gamma_1\|\mathbf{Y} - \mathbf{U}\mathbf{W}\|_F^2 + \gamma_2\|\mathbf{U} - [\mathbf{X}_M, \mathbf{X}_P]\mathbf{P}\|_F^2 \\ &+ \gamma_3\|[\mathbf{X}_M, \mathbf{X}_P] - \mathbf{U}\mathbf{Q}\|_F^2 + \lambda_1 R_1(\mathbf{U}) + \lambda_2 R_2(\mathbf{U}) \\ &+ \lambda_3 R_3(\mathbf{U}) + \lambda_4 R_4(\mathbf{U}) + \rho_1\|\mathbf{U}\|_F^2.\end{aligned} \tag{16}$$

Taking the derivative of Eq.(16) with respect to $\mathbf{U}$, we obtain

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{U}} =& 2\gamma_1(\mathbf{U}\mathbf{W}\mathbf{W}^T - \mathbf{Y}\mathbf{W}^T) + 2\gamma_2(\mathbf{U} - [\mathbf{X}_M, \mathbf{X}_P]\mathbf{P}) \\ &+ 2\gamma_3(\mathbf{U}\mathbf{Q}\mathbf{Q}^T - [\mathbf{X}_M, \mathbf{X}_P]\mathbf{Q}^T) + 2\lambda_1\tilde{\mathbf{U}} \\ &- 2\lambda_2\mathbf{Y}\mathbf{Y}^T\mathbf{U} + \lambda_3\mathbf{L}\mathbf{U} - 2\lambda_4\mathbf{S}\mathbf{U} + 2\rho_1\mathbf{U}.\end{aligned} \tag{17}$$

Subsequently, we update $\mathbf{U}$ by

$$\mathbf{U}_{t+1} = \mathbf{U}_t - \frac{\partial J}{\partial \mathbf{U}}. \tag{18}$$

(2) Optimizing $\mathbf{P}$: fixing $\mathbf{U}$, $\mathbf{Q}$, and $\mathbf{W}$, we rewrite the target function with respect to $\mathbf{P}$ as follows:

$$\min_{\mathbf{P}} \gamma_2\|\mathbf{U} - [\mathbf{X}_M, \mathbf{X}_P]\mathbf{P}\|_F^2 + \rho_2\|\mathbf{P}\|_F^2. \tag{19}$$

Taking the derivative of Eq.(19) with respect to $\mathbf{P}$ equal zero, we obtain

$$2\gamma_2[\mathbf{X}_M, \mathbf{X}_P]^T[\mathbf{X}_M, \mathbf{X}_P]\mathbf{P} - 2\gamma_2[\mathbf{X}_M, \mathbf{X}_P]^T\mathbf{U} + 2\rho_2\mathbf{P} = 0. \tag{20}$$

Thus, the closed form solution of $\mathbf{P}$ is

$$\mathbf{P} = (\gamma_2[\mathbf{X}_M, \mathbf{X}_P]^T[\mathbf{X}_M, \mathbf{X}_P] + \rho_2\mathbf{I}_m)^{-1}(\gamma_2[\mathbf{X}_M, \mathbf{X}_P]^T\mathbf{U}), \tag{21}$$

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

where $\mathbf{I}_m$ is the identical matrix.

(3) Optimizing $\mathbf{Q}$: fixing $\mathbf{U}$, $\mathbf{P}$, and $\mathbf{W}$, we rewrite the target function with respect to $\mathbf{Q}$ as follows:

$$\min_{\mathbf{Q}} \gamma_3 \|[\mathbf{X}_M, \mathbf{X}_P] - \mathbf{U}\mathbf{Q}\|_F^2 + \rho_3 \|\mathbf{Q}\|_F^2. \tag{22}$$

Taking the derivative of Eq.(22) with respect to $\mathbf{Q}$ equal zero, we obtain

$$2\gamma_3 \mathbf{U}^T\mathbf{U}\mathbf{Q} - 2\gamma_3 \mathbf{U}^T[\mathbf{X}_M, \mathbf{X}_P] + 2\rho_3 \mathbf{Q} = 0. \tag{23}$$

Thus, the closed form solution of $\mathbf{Q}$ is

$$\mathbf{Q} = (\gamma_3 \mathbf{U}^T\mathbf{U} + \rho_3 \mathbf{I}_k)^{-1}(\gamma_3 \mathbf{U}^T[\mathbf{X}_M, \mathbf{X}_P]), \tag{24}$$

where $\mathbf{I}_k$ is the identical matrix.

(4) Optimizing $\mathbf{W}$: fixing $\mathbf{U}$, $\mathbf{Q}$, and $\mathbf{P}$, we rewrite the target function with respect to $\mathbf{W}$ as follows:

$$\min_{\mathbf{W}} \gamma_1 \|\mathbf{Y} - \mathbf{U}\mathbf{W}\|_F^2 + \rho_4 \|\mathbf{W}\|_F^2. \tag{25}$$

Taking the derivative of Eq.(25) with respect to $\mathbf{W}$ equal zero, we obtain

$$2\gamma_1 \mathbf{U}^T\mathbf{U}\mathbf{W} - 2\gamma_1 \mathbf{U}^T\mathbf{Y} + 2\rho_4 \mathbf{W} = 0. \tag{26}$$

Thus, the closed form solution of $\mathbf{W}$ is

$$\mathbf{W} = (\gamma_1 \mathbf{U}^T\mathbf{U} + \rho_4 \mathbf{I}_k)^{-1}(\gamma_1 \mathbf{U}^T\mathbf{Y}). \tag{27}$$

Therefore, the objective function in Eq.(15) can be solved by conducting the above steps iteratively until convergences. Algorithm.1 gives supplementary details.

Algorithm 1 Pseudocode for solving Eq.(15)

Training Stage

Input: original feature matrix $[\mathbf{X}_M, \mathbf{X}_P]$, label matrix $\mathbf{Y}$, pre-determined parameters k, λ_i, ρ_i and t_{max}

Output: projection matrix $\mathbf{P}$, reconstruction matrix $\mathbf{Q}$ and weight matrix $\mathbf{W}$

Initialize: Initialize $\mathbf{U}$, $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{W}$ with random values, normalize $\mathbf{U}$ (mean $= 0$, var $= 1$), $t = 0$

repeat

  update $\mathbf{U}$ by using Eq.(17), Eq.(18);

  update $\mathbf{P}$ by using Eq.(21);

  update $\mathbf{Q}$ by using Eq.(24);

  update $\mathbf{W}$ by using Eq.(27);

  $t = t + 1$;

until $t = t_{max}$

Testing Stage

Input: new sample $[\mathbf{X}_M, \mathbf{X}_P]_{new}$

Output: prediction $\mathbf{Y}_{new}$

Procedure: $\mathbf{Y}_{new} = [\mathbf{X}_M, \mathbf{X}_P]_{new}\mathbf{P}\mathbf{W}$

IV. Experiments and results

In this section, we first introduce experimental settings, including comparison methods, parameter settings, and the evaluation strategy. Then, we present the diagnostic performance of all competing methods. Subsequently, effectiveness of relational regularizers and bi-directional mapping is validated. Finally, we study the influence of parameters to the proposed method.

A. Experiment Settings

In this study, we focus on three classification tasks, namely, AD vs. NC, pMCI vs. sMCI, and MCI vs. NC. First, we compare our proposed framework with several conventional methods, the details of which are briefly introduced below.

- Baseline is defined as a model that simply concatenates original multi-modal features and directly performs classification without using any feature reduction methods.
- Multiple kernel learning (MKL): In [10], the authors used MKL to fuse multi-modal features by simultaneously learning kernel weights and a maximum margin classifier.
- Canonical feature selection (CFS): In [12], the authors utilized CCA to combine multi-modal information by mapping original multi-modal feature space to a common space. Following the literature, we conducted a search for β, γ, and C from $\{10^{-5}, , ..., 10^5\}$, $\{10^{-3}, , ..., 10^8\}$, and $\{2^{-5}, , ..., 2^5\}$, respectively.
- Relational regularization feature selection (RRFS): Zhu *et al*. [17] performed feature selection based on three kinds of relationships for joint regression and classification tasks. We searched the best control parameters of three relation terms in the range of $\{10^{-6}, ..., 10^2\}$ and chose parameter of $L_{2,1}$ term from $\{10^2, , ..., 10^8\}$. And k and σ was set to 3 and 1, respectively.
- Relational-regularized discriminative sparse learning (RrDSL): In [18], Lei *et al*. proposed a discriminative sparse learning method with relational regularization to jointly predict the clinical scores and classify AD stages. For RrDSL, we determined the best control parameters in the range $\{10^{-10}, 10^{-9}, ..., 10^{10}\}$.
- Latent Representation Learning (LRL): To utilize the complementary information of multi-modal data, Zhou et al. [11] used the learned features in a latent space for AD diagnosis. According to the literature, we set $\lambda = 1$ and $h = 30$ and tuned three hyper-parameters in the range of $\{10^{-4}, 10^{-3}, ..., 10^4\}$.
- Complete multi-modality latent space (CMLS): By intergrating latent space learning and ensemble classifier training into a unified work, Zhou *et al*. [13] tried to explore the intrinsic correlations within multi-modal data of Alzheimer's Disease. The regularization parameter values, the dimension of latent space, and the number of classifiers were determined in the range of $\{10^{-6}, 10^{-5}, ..., 10^3\}$, $\{10, 20, ..., 80\}$, and $\{10, 20, ..., 80\}$, respectively.
- Dirty multi-task canonical correlation analysis (SCCA): Du *et al*. [27] utilized the multi-task learning and parameter decomposition to investigate complex multi-SNP-multi-QT associations. Based on this model, we explored not only the shared biomakers but also the modality-specific biomarkers within MRI and PET data. According to the manuscript, we searched λ_s, β_s, λ_w, β_b, and λ_z within $\{10^{-5}, 10^{-4}, ..., 10^5\}$.
- Landmark-based deep multi-instance convolutional network (LDMI): Liu *et al*. [28] developed a patch-based network to perform AD classification task by learning local-to-global structure information. As suggested in

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

[28], the patch size was selected in the range of {24, 36, 48}, and SGD algorithm with the learning rate of 0.01 was used to train the network.

- Hierarchical fully convolutional network (HFCN): In [21], Lian *et al*. proposed to automatically identify discriminant patches and regions using pruning strategy and used multi-scale feature representations for AD diagnosis. The patch size was selected in the range of {25, 35}, and Adam optimizer with the learning rate of 0.001 was used to train the network.

For Baseline method and those approaches which treated feature reduction and classifier training as two standalone procedures (except LRL, CMLS, LDMI, and HFCN), we built a support vector machine model (SVM) for classification via LIBSVM toolbox [29] and selected the optimal margin parameter C in the range of $\{10^{-5}, 10^{-4}, ..., 10^{5}\}$. In our proposed model, parameters can be classified into three categories: feature dimensionality k, regularizer parameter λ_i, and auxiliary parameters (ρ_i and γ_i). We determined the parameters with the spaces of $k \in \{20, 25..., 40\}$, $\lambda_i \in \{10^{-5}, 10^{-4}..., 10^{1}\}$, $\rho_i \in \{10^{-5}, 10^{-4}..., 10^{1}\}$, and $\gamma_i \in \{10^{-3}, 10^{-2}..., 10^{0}\}$. It is time-consuming and inefficient to tune all parameters simultaneously. Therefore, we tuned some parameters each time by fixing other categorical parameters. We evaluated all comparison methods using 10-fold cross-validation strategy. Several metrics were used to evaluate all the comparison methods, including area under curve (AUC), accuracy (ACC), specificity (SPE), and sensitivity (SEN). Finally, considering that AUC is more commonly utilized to evaluate the models based on the imbalance-class dataset when compared with other metrics [30], we conducted the paired *t*-tests (at 95% significance level) on the classification tasks of our method and the competing methods. And $^{\#}$ denotes that it exits significant difference with p-value < 0.05 between our proposed method and other compared methods.

B. Diagnostic Performance

The diagnostic performances of all comparison methods on three tasks are presented in TABLE III, from which we can have following observations. *First*, multi-modal fusion methods can improve the performance of diagnostic model by exploiting complementary multi-modal information. Actually, LRL method obtains the best performance when compared with Baseline, MKL, and CFS methods. A potential reason is that LRL jointly learns latent common space and classifier, which can help to capture useful multi-modal information in a task-driven manner. In addition, we have implemented two mono-modal variants of the proposed method (called Proposed_MRI and Proposed_PET) to demonstrate the superiority of multi-modal fusion. The experimental results also shed light on the benefit of complementary information provided by multi-modal data. *Second*, feature selection/reduction is important when dealing with high-dimensional multi-modal data. Compared with Baseline method without any feature selection/reduction procedures, other methods get better classification performance on all three tasks. It's worthy mentioning that both RRFS and RrDSL yield significant improvement, especially on pMCI vs. sMCI and MCI vs. NC tasks, which may owe to relational regularization utilization. *Third*, compared with conventional learning methods, our proposed method achieves the best performance on all three tasks. Several potential advantages exist in the proposed method: 1) Different common space based learning, such as LRL, CMLS, and SCCA, which just use unidirectional mapping, the proposed method uses bi-directional mapping for simultaneously considering both data projection and reconstruction, which helps to learn an informative shared space that preserves original information as much as possible. 2) Several regularizers are devised to explore the underlying associations of multi-modal data in the shared space (unlike LRL, CMLS, and SCCA). Moreover, the definitions of regularizers in our proposed method also are different from those in RRFS and RrDSL. In this paper, feature-wise regularizers induce the shared representations to be low-redundant among themselves and high-relevant with labels, and sample-wise regularizers aim to enhance the class separability in the shared space. 3) The unification of representation learning and classifier modeling gives the access for each component to interact and supervise with each other during the optimization process, which is different from baseline, MKL, CFS, RrDSL, RRFS, and SCCA. *Forth*, compared with deep learning based methods (LDMI [28], HFCN [21], our method still yields better classification results in terms of most metrics. A potential reason is that the available data is limited for training deep learning models. Similar to deep leaning methods, the proposed method learns latent representations in a task-driven manner. By contrast, our proposed method holds some advantages as follows: 1) The underlying associations of multi-modal data in the shared space are captured, including feature-feature redundancy, feature-label relevancy, and class separability. 2) Just a small number of parameters are contained in the proposed method, which means that it does not refer to time-consuming parameter refining and does not rely on a large amount of training samples for model training. In addition, our experiments also show that the parameters are robust to multi-site datasets (see *E. Influence of Parameters*). 3) Comparing with deep learning methods, our method is more explainable, which is beneficial to discover useful biomarkers for AD diagnosis.

C. Effectiveness of Relational Regularizers

To investigate the effectiveness of proposed relational regularizers, we implemented and compared different variants of our proposed method. For convenience, we denote the proposed method without any relational regularizers as SRL. Accordingly, the SRL with feature-feature, feature-label, sample-sample regularizers are denoted as SRL_$\mathrm{R_{FF}}$, SRL_$\mathrm{R_{FY}}$, SRL_$\mathrm{R_{SS}}$, respectively. The experimental results are shown in TABLE V, and we can observe that: 1) The SRL obtains promising results than Baseline method and some multi-modal fusion methods, which implies that the shared representation learning using bi-directional mapping without relational regularization can also enhance classification performance effectively. 2) All variants of SRL with different relational regularizers outperform SRL, which indicates that these regularizers can help induce learning more distinguishing shared representations via exploring the potential associations

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

TABLE III
RESULTS (MEAN AND STANDARD DEVIATION) OF THREE CLASSIFICATION TASKS ON ADNI-1 (THE BOLDFACE DENOTES THE BEST RESULTS OF EACH METRIC AND $^{\#}$ DENOTES SIGNIFICANT DIFFERENCE WITH P-VALUE < 0.05).

	AD vs. NC				pMCI vs. sMCI				MCI vs. NC			
	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN
Baseline	0.888	0.875	0.889	0.862	0.680	0.720	0.867	0.491	0.705	0.749	0.492	0.875
	(0.064)	(0.070)	(0.099)	(0.127)	(0.073)	(0.054)	(0.131)	(0.243)	(0.086)	(0.055)	(0.193)	(0.089)
MKL [10]	0.893	0.889	0.895	0.877	0.696	0.701	0.782	0.602	0.703	0.711	0.606	0.745
	(0.044)	(0.041)	(0.072)	(0.069)	(0.083)	(0.087)	(0.099)	(0.104)	(0.071)	(0.069)	(0.109)	(0.080)
CFS [12]	0.946	0.932	0.931	0.926	0.791	0.795	0.882	0.633	0.793	0.774	0.619	0.842
	(0.027)	(0.032)	(0.048)	(0.051)	(0.055)	(0.058)	(0.098)	(0.094)	(0.044)	(0.047)	(0.063)	(0.102)
RRFS [17]	0.953	0.953	0.956	0.950	0.813	0.810	0.867	0.725	0.794	0.814	0.633	0.895
	(0.028)	(0.030)	(0.057)	(0.071)	(0.047)	(0.039)	(0.098)	(0.165)	(0.040)	(0.037)	(0.158)	(0.101)
RrDSL [18]	0.942	0.961	0.962	0.948	0.818	0.805	0.878	0.716	0.794	0.804	0.621	0.893
	(0.039)	(0.042)	(0.067)	(0.073)	(0.088)	(0.079)	(0.058)	(0.040)	(0.082)	(0.077)	(0.085)	(0.127)
LRL [11]	0.958	0.942	0.970	0.911	0.794	0.805	0.901	0.661	0.794	0.796	**0.697**	0.845
	(0.039)	(0.052)	(0.048)	(0.115)	(0.044)	(0.050)	(0.085)	(0.129)	(0.050)	(0.045)	(0.125)	(0.076)
CMLS [13]	0.896	0.896	0.879	0.912	0.772	0.760	0.767	0.738	0.754	0.759	0.680	0.795
	(0.068)	(0.068)	(0.113)	(0.087)	(0.062)	(0.077)	(0.161)	(0.279)	(0.045)	(0.051)	(0.253)	(0.114)
SCCA [27]	0.941	0.922	0.960	0.882	0.706	0.755	0.909	0.520	0.713	0.746	0.557	0.840
	(0.033)	(0.036)	(0.052)	(0.080)	(0.077)	(0.044)	(0.062)	(0.126)	(0.100)	(0.057)	(0.248)	(0.084)
LDMI [28]	0.957	0.921	0.963	0.931	0.809	0.806	**0.911**	0.693	0.788	0.743	0.661	0.882
	(0.049)	(0.056)	(0.084)	(0.063)	(0.061)	(0.073)	(0.069)	(0.082)	(0.025)	(0.069)	(0.038)	(0.076)
HFCN [21]	0.949	0.919	0.965	0.928	0.805	0.802	0.894	0.706	0.782	0.754	0.659	0.877
	(0.059)	(0.067)	(0.083)	(0.070)	(0.024)	(0.047)	(0.038)	(0.097)	(0.028)	(0.044)	(0.130)	(0.091)
Proposed_MRI	0.945	0.927	0.940	0.917	0.822	0.825	0.901	0.711	0.789	0.800	0.657	0.870
	(0.056)	(0.050)	(0.084)	(0.095)	(0.040)	(0.042)	(0.094)	(0.165)	(0.086)	(0.066)	(0.245)	(0.114)
Proposed_PET	0.931	0.912	0.907	0.911	0.821	0.810	0.892	0.684	0.782	0.796	0.614	0.885
	(0.077)	(0.073)	(0.117)	(0.102)	(0.075)	(0.046)	(0.097)	(0.188)	(0.080)	(0.059)	(0.234)	(0.145)
Proposed$^{\#}$	**0.976**	**0.969**	**0.978**	**0.956**	**0.840**	**0.845**	0.909	**0.748**	**0.820**	**0.826**	0.657	**0.910**
	(0.033)	(0.035)	(0.070)	(0.057)	(0.073)	(0.055)	(0.127)	(0.143)	(0.053)	(0.034)	(0.183)	(0.061)

of multi-modal data in the shared space. In addition, we can observe that the SRL_$\mathrm{R_{SS}}$ has a relatively marginal improvement on AD vs. NC classification when compared with SRL_$\mathrm{R_{FF}}$ and SRL_$\mathrm{R_{FY}}$, but it still shows an obvious performance improvement on pMCI vs. sMCI and MCI vs. NC especially in terms of ACC and AUC, which demonstrates it might help to learn more subtler information for these two tasks. 3) The proposed method that includes feature-feature, feature-label, and sample-sample regularizers achieves the best performance on all three tasks. All these variants (except SRL) only focus on single perspective of shared representations, which is insufficient for complicated exploration of multi-modal data. Containing both feature-wise and sample-wise regularizers, our complete method can not only induce learning of low-redundant and discriminative shared representations, but also improve the class separability in the shared space.

D. Effectiveness of Bi-directional Mapping

To further validate effectiveness of the bi-directional mapping scheme, especially the effectiveness of reconstruction matrix $\mathbf{Q}$, we conduct experiments by detaching $\mathbf{Q}$-related terms from the proposed framework and its variants. Classification results are reported in TABLE VI, where * refers to detaching matrix Q. Comparing TABLE V with TABLE VI, we can observe that the proposed approach and all its alternative versions (namely SRL, SRL_$\mathrm{R_{FF}}$, SRL_$\mathrm{R_{FY}}$, and SRL_$\mathrm{R_{SS}}$) outperform their corresponding model with unidirectional mapping scheme on three classification tasks in terms of most metrics. A reasonable explanation is that the reconstruction matrix $\mathbf{Q}$ can efficiently regulate the shared representations to preserve the original information and well cooperate with projection matrix $\mathbf{P}$ to capture the complementary information in multi-modal data. In other words, the bi-directional mapping learns matrices $\mathbf{P}$ and $\mathbf{Q}$ simultaneously in order to ensure the shared representations U have both representation and reconstruction abilities. In addition, in testing stage, we would like to directly make a prediction for testing samples using the learned transformation matrices and circumvent any middle processes, e.g. the inverse operation.

E. Influence of Parameters

In this section, we study the influence of interest-of-parameters, that is dimension of shared representations k and regularizer parameters λ_i. For k, we first fix other parameters and determine k in the range of $\{20, 25,..., 40\}$ for each experiment. Subsequently, we tune feature-related regularizer parameters (i.e., λ_1, λ_2) and sample-related regularizer parameters (i.e., λ_3, λ_4) in turn by setting the range of $\{10^{-5}, 10^{-4}, ..., 10^{1}\}$. The experimental results with respect to k are shown in Fig. 2, where the most suitable k value is

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

TABLE IV
RESULTS (MEAN AND STANDARD DEVIATION) OF THREE CLASSIFICATION TASKS ON ADNI-2 (THE BOLDFACE DENOTES THE BEST RESULTS OF EACH METRIC AND $^{\#}$ DENOTES SIGNIFICANT DIFFERENCE WITH P-VALUE < 0.05).

	AD vs. NC				pMCI vs. sMCI				MCI vs. NC			
	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN
Baseline	0.918	0.889	0.940	0.846	0.781	0.790	0.845	0.725	0.739	0.754	0.692	0.790
	(0.067)	(0.062)	(0.107)	(0.062)	(0.100)	(0.083)	(0.114)	(0.193)	(0.044)	(0.033)	(0.173)	(0.114)
MKL [10]	0.909	0.897	0.941	0.853	0.792	0.781	0.849	0.736	0.744	0.760	0.655	0.797
	(0.070)	(0.073)	(0.061)	(0.085)	(0.064)	(0.079)	(0.077)	(0.113)	(0.029)	(0.042)	(0.132)	(0.094)
CFS [12]	0.944	0.931	0.950	0.916	0.815	0.809	0.861	0.762	0.754	0.763	0.659	0.823
	(0.068)	(0.066)	(0.079)	(0.097)	(0.079)	(0.087)	(0.083)	(0.109)	(0.021)	(0.055)	(0.076)	(0.125)
RRFS [17]	0.942	0.922	0.953	0.897	0.818	0.817	0.850	0.775	0.751	0.764	0.641	0.832
	(0.042)	(0.044)	(0.066)	(0.070)	(0.055)	(0.046)	(0.071)	(0.115)	(0.078)	(0.052)	(0.179)	(0.155)
RrDSL [18]	0.947	0.939	0.950	0.911	0.816	0.820	0.866	0.769	0.779	0.766	0.651	0.842
	(0.047)	(0.053)	(0.068)	(0.101)	(0.038)	(0.096)	(0.084)	(0.045)	(0.066)	(0.046)	(0.091)	(0.094)
LRL [11]	0.964	0.940	0.955	0.929	0.823	0.817	0.877	0.743	0.800	0.790	0.745	0.815
	(0.035)	(0.054)	(0.064)	(0.095)	(0.057)	(0.059)	(0.106)	(0.162)	(0.081)	(0.066)	(0.134)	(0.087)
CMLS [13]	0.894	0.897	0.861	0.927	0.790	0.790	0.777	0.803	0.744	0.764	0.675	0.814
	(0.067)	(0.063)	(0.134)	(0.077)	(0.074)	(0.067)	(0.042)	(0.136)	(0.060)	(0.070)	(0.147)	(0.143)
SCCA [27]	0.944	0.922	0.943	0.904	0.797	0.810	0.841	0.767	0.745	0.774	0.700	0.814
	(0.047)	(0.053)	(0.081)	(0.086)	(0.080)	(0.071)	(0.143)	(0.168)	(0.093)	(0.079)	(0.198)	(0.145)
LDMI [28]	0.953	0.920	0.950	0.928	0.799	0.787	0.900	0.701	0.791	0.752	0.673	0.844
	(0.073)	(0.061)	(0.088)	(0.108)	(0.091)	(0.077)	(0.067)	(0.099)	(0.086)	(0.070)	(0.099)	(0.107)
HFCN [21]	0.959	0.931	0.954	0.933	0.815	0.819	**0.902**	0.696	0.788	0.743	0.686	0.851
	(0.037)	(0.044)	(0.061)	(0.069)	(0.081)	(0.066)	(0.093)	(0.117)	(0.091)	(0.111)	(0.089)	(0.083)
Proposed_MRI	0.950	0.922	0.936	0.911	0.809	0.816	0.880	0.735	0.791	0.806	0.708	**0.863**
	(0.038)	(0.047)	(0.096)	(0.081)	(0.054)	(0.026)	(0.114)	(0.140)	(0.089)	(0.066)	(0.181)	(0.136)
Proposed_PET	0.926	0.912	0.964	0.871	0.791	0.784	0.865	0.683	0.783	0.788	0.688	0.847
	(0.058)	(0.062)	(0.064)	(0.105)	(0.058)	(0.055)	(0.095)	(0.178)	(0.077)	(0.056)	(0.151)	(0.072)
Proposed$^{\#}$	**0.977**	**0.968**	**0.982**	**0.957**	**0.856**	**0.859**	0.900	**0.808**	**0.831**	**0.815**	**0.788**	0.831
	(0.035)	(0.045)	(0.038)	(0.069)	(0.049)	(0.059)	(0.125)	(0.191)	(0.071)	(0.061)	(0.175)	(0.128)

TABLE V
CLASSIFICATION RESULTS OF DIFFERENT VARIANTS OF PROPOSED FRAMEWORK (THE BOLDFACE DENOTES THE BEST RESULTS OF EACH METRIC).

Data	Methods	AD vs. NC				pMCI vs. sMCI				MCI vs. NC			
		AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN
ADNI1	SRL	0.944	0.912	0.939	0.881	0.798	0.795	0.900	0.634	0.778	0.783	0.648	0.850
		(0.053)	(0.061)	(0.085)	(0.134)	(0.071)	(0.060)	(0.095)	(0.147)	(0.086)	(0.039)	(0.231)	(0.094)
	SRL_R_{FF}	0.963	0.938	0.959	0.914	0.822	0.820	0.909	0.686	0.791	0.799	0.653	0.870
		(0.043)	(0.069)	(0.071)	(0.146)	(0.057)	(0.042)	(0.073)	(0.132)	(0.059)	(0.047)	(0.201)	(0.109)
	SRL_R_{FY}	0.955	0.937	0.929	0.944	0.817	0.805	0.883	0.682	0.796	0.799	0.626	0.885
		(0.039)	(0.054)	(0.083)	(0.059)	(0.085)	(0.060)	(0.112)	(0.172)	(0.074)	(0.049)	(0.217)	(0.094)
	SRL_R_{SS}	0.953	0.927	0.959	0.889	0.825	0.820	0.901	0.696	0.808	0.806	0.638	0.890
		(0.041)	(0.062)	(0.053)	(0.139)	(0.088)	(0.082)	(0.095)	(0.244)	(0.068)	(0.044)	(0.22)	(0.084)
	Proposed	**0.976**	**0.969**	**0.978**	**0.956**	**0.840**	**0.845**	**0.909**	**0.748**	**0.820**	**0.826**	**0.657**	**0.910**
		(0.033)	(0.035)	(0.070)	(0.057)	(0.073)	(0.055)	(0.127)	(0.143)	(0.053)	(0.034)	(0.183)	(0.061)
ADNI2	SRL	0.946	0.905	0.936	0.880	0.800	0.800	0.847	0.743	0.764	0.776	0.688	0.828
		(0.028)	(0.027)	(0.075)	(0.056)	(0.079)	(0.065)	(0.088)	(0.151)	(0.046)	(0.047)	(0.157)	(0.106)
	SRL_R_{FF}	0.959	0.936	0.945	0.929	0.825	0.837	0.892	0.768	0.772	0.785	0.690	**0.840**
		(0.034)	(0.047)	(0.064)	(0.089)	(0.043)	(0.039)	(0.058)	(0.093)	(0.089)	(0.075)	(0.183)	(0.134)
	SRL_R_{FY}	0.958	0.924	0.945	0.907	0.828	0.833	0.860	0.798	0.768	0.791	0.742	0.819
		(0.036)	(0.055)	(0.115)	(0.083)	(0.052)	(0.026)	(0.107)	(0.138)	(0.065)	(0.05)	(0.15)	(0.116)
	SRL_R_{SS}	0.956	0.930	0.945	0.919	0.812	0.828	0.847	0.805	0.778	0.778	0.781	0.777
		(0.028)	(0.026)	(0.077)	(0.041)	(0.089)	(0.066)	(0.137)	(0.187)	(0.062)	(0.064)	(0.179)	(0.127)
	Proposed	**0.977**	**0.968**	**0.982**	**0.957**	**0.856**	**0.859**	**0.900**	**0.808**	**0.831**	**0.815**	**0.788**	0.831
		(0.035)	(0.045)	(0.038)	(0.069)	(0.049)	(0.059)	(0.125)	(0.191)	(0.071)	(0.061)	(0.175)	(0.128)

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

TABLE VI
CLASSIFICATION RESULTS OF ONE-DIRECTIONAL VERSION OF VARIANTS OF PROPOSED METHOD (* MEANS DETACHING MATRIX Q AND THE BOLDFACE DENOTES THE BEST RESULTS OF EACH METRIC).

Data	Methods	AD vs. NC				pMCI vs. sMCI				MCI vs. NC			
		AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN
ADNI1	SRL*	0.937	0.901	0.888	0.911	0.795	0.785	0.835	0.709	0.757	0.769	0.622	0.840
		(0.033)	(0.039)	(0.077)	(0.070)	(0.064)	(0.053)	(0.152)	(0.236)	(0.071)	(0.044)	(0.270)	(0.115)
	SRL_R_{FF}*	0.946	0.923	0.938	0.903	0.806	0.805	0.876	0.698	0.774	0.786	0.658	**0.850**
		(0.046)	(0.065)	(0.089)	(0.111)	(0.068)	(0.060)	(0.071)	(0.167)	(0.105)	(0.057)	(0.210)	(0.088)
	SRL_R_{FY}*	0.945	0.916	0.919	0.911	0.802	0.805	0.883	0.680	0.767	0.783	0.677	0.835
		(0.046)	(0.062)	(0.064)	(0.126)	(0.061)	(0.037)	(0.090)	(0.169)	(0.095)	(0.082)	(0.262)	(0.149)
	SRL_R_{SS}*	0.951	0.917	0.950	0.886	0.803	0.815	0.883	0.707	0.773	0.773	0.677	0.820
		(0.043)	(0.06)	(0.071)	(0.095)	(0.054)	(0.063)	(0.112)	(0.224)	(0.051)	(0.041)	(0.204)	(0.101)
	Proposed_Q	**0.957**	**0.933**	**0.950**	**0.919**	**0.809**	**0.825**	**0.884**	**0.736**	**0.790**	**0.803**	**0.724**	0.840
		(0.044)	(0.059)	(0.097)	(0.097)	(0.076)	(0.054)	(0.098)	(0.148)	(0.056)	(0.039)	(0.238)	(0.122)
ADNI2	SRL*	0.940	0.903	0.924	0.887	0.800	0.806	0.835	0.770	0.758	0.769	0.697	0.812
		(0.024)	(0.04)	(0.095)	(0.067)	(0.051)	(0.039)	(0.082)	(0.082)	(0.090)	(0.060)	(0.144)	(0.107)
	SRL_R_{FF}*	0.949	0.923	0.973	0.884	0.820	0.812	0.837	0.780	0.786	0.783	**0.726**	0.816
		(0.039)	(0.045)	(0.061)	(0.099)	(0.056)	(0.059)	(0.098)	(0.129)	(0.043)	(0.035)	(0.143)	(0.112)
	SRL_R_{FY}*	0.943	0.917	0.969	0.876	0.804	0.825	0.862	0.778	0.765	0.777	0.702	0.821
		(0.034)	(0.034)	(0.050)	(0.066)	(0.051)	(0.071)	(0.109)	(0.118)	(0.061)	(0.027)	(0.130)	(0.083)
	SRL_R_{SS}*	0.942	0.916	0.955	0.886	0.811	0.815	0.842	**0.780**	0.765	0.784	0.697	0.835
		(0.054)	(0.072)	(0.077)	(0.122)	(0.067)	(0.077)	(0.119)	(0.194)	(0.090)	(0.057)	(0.167)	(0.080)
	Proposed_Q	**0.952**	**0.933**	**0.973**	**0.901**	**0.826**	**0.828**	**0.885**	0.758	**0.793**	**0.801**	0.711	**0.853**
		(0.043)	(0.049)	(0.044)	(0.075)	(0.078)	(0.061)	(0.089)	(0.126)	(0.059)	(0.061)	(0.148)	(0.108)

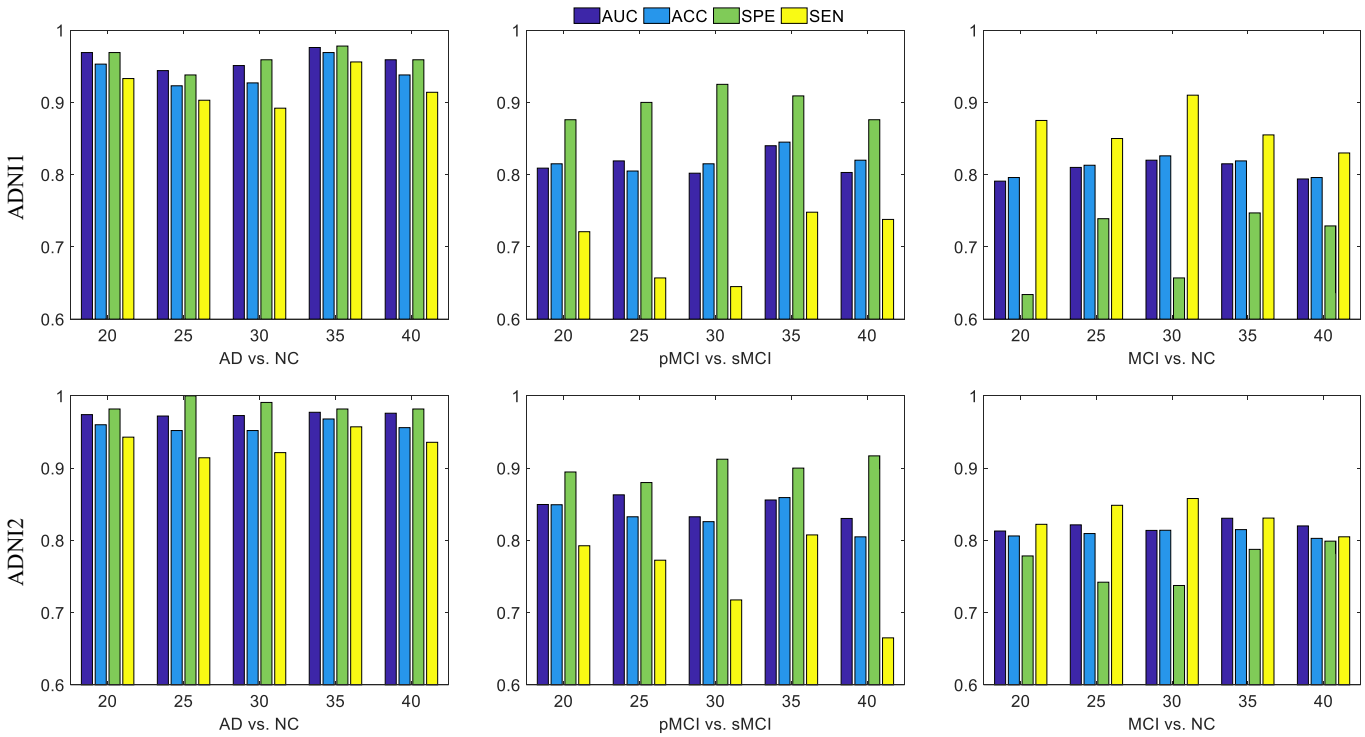

Fig. 2. Classification performance of the proposed method with different dimensions k (i.e., 20, 25, 30, 35, and 40) of shared representations on three binary tasks for ADNI1 and ADNI2 datasets.

located around 35 after overall consideration of four metrics on three classification tasks. Fig. 3 and 4 illustrate the AUC and ACC achieved by grid combinations of (λ_1 and λ_2) and (λ_3 and λ_4). Taking both AUC and ACC into consideration, we can find the most appropriate parameter settings for different classification tasks. Taking pMCI vs. sMCI in Fig. 3 for example, the best λ_1, λ_2, λ_3, and λ_4 falls in the range of $[10^{-2}, 10^{0}]$, $[10^{-2}, 10^{-1}]$, $[10^{-2}, 10^{-1}]$, and $[10^{-4}, 10^{-3}]$,

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

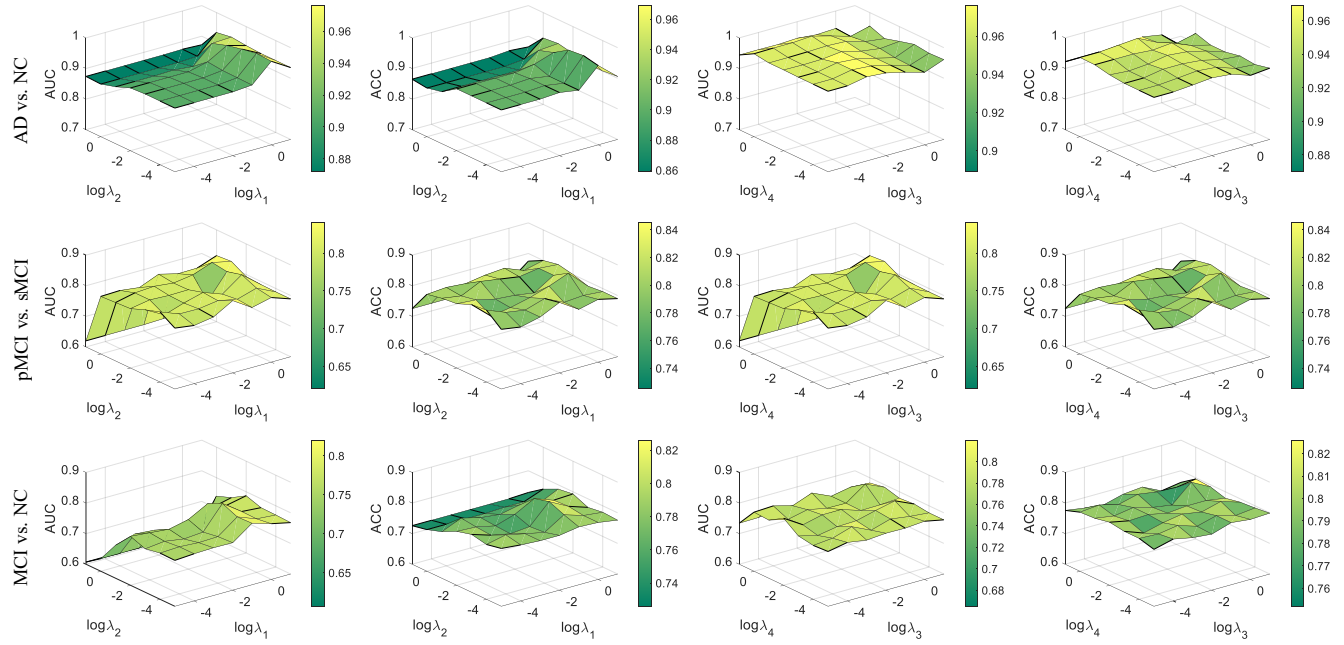

Fig. 3. ACC and AUC results of different parameter settings (i.e., $\lambda_1, \lambda_2, \lambda_3, \lambda_4$) on three tasks for ADNI1. From top to bottom: AD vs. NC, pMCI vs. sMCI, and MCI vs. NC. Taking both ACC and AUC results into consideration, the best λ_1, λ_2, λ_3, and λ_4 falls in the range of $[10^0, 10^1]$, $[10^{-5}, 10^{-3}]$, $[10^{-3}, 10^{-2}]$, and $[10^{-2}, 10^0]$ for AD vs. NC, $[10^{-2}, 10^0]$, $[10^{-3}, 10^{-1}]$, $[10^{-2}, 10^{-1}]$, and $[10^{-4}, 10^{-3}]$ for pMCI vs. sMCI, and $[10^{-1}, 10^0]$, $[10^{-4}, 10^{-2}]$, $[10^{-1}, 10^0]$, and $[10^{-3}, 10^{-1}]$ for MCI vs. NC.

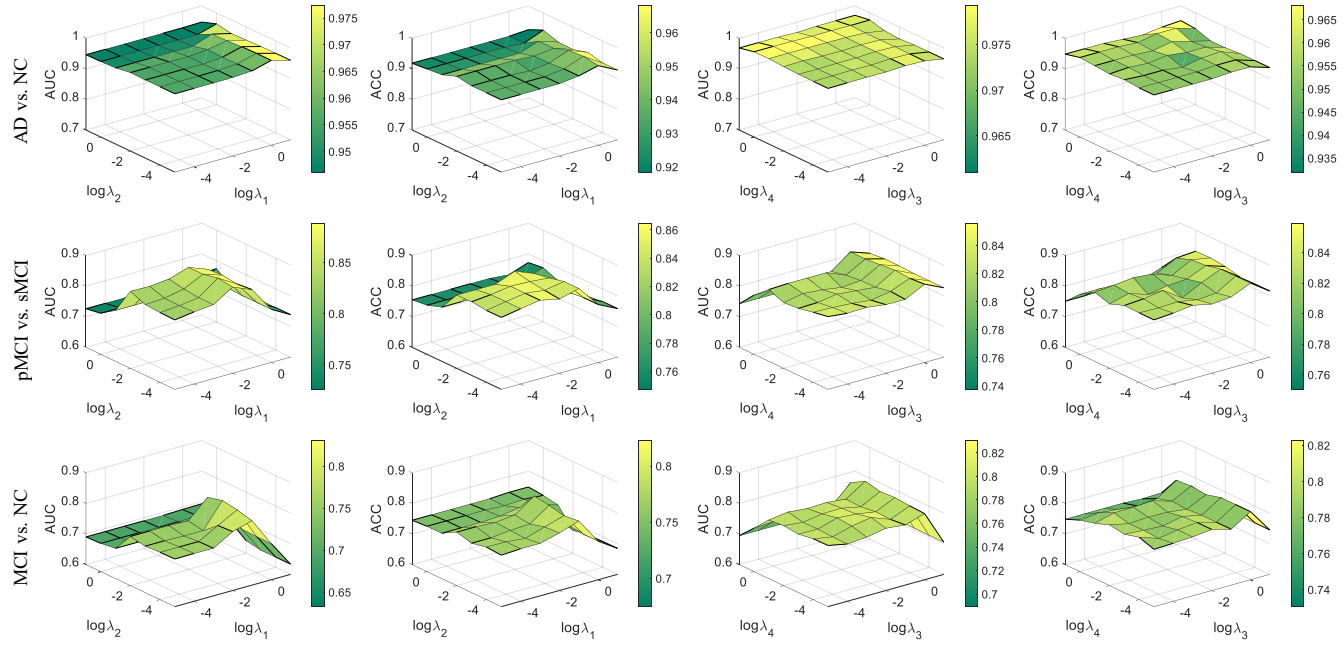

Fig. 4. ACC and AUC results of different parameter settings (i.e., $\lambda_1, \lambda_2, \lambda_3, \lambda_4$) on three tasks for ADNI2. From top to bottom: AD vs. NC, pMCI vs. sMCI, and MCI vs. NC. Taking both ACC and AUC results into consideration, the best λ_1, λ_2, λ_3, and λ_4 falls in the range of $[10^{-1}, 10^0]$, $[10^{-5}, 10^{-3}]$, $[10^{-2}, 10^{-1}]$, and $[10^{-2}, 10^0]$ for AD vs. NC, $[10^{-2}, 10^{-1}]$, $[10^{-4}, 10^{-2}]$, $[10^{-1}, 10^0]$, and $[10^{-4}, 10^{-3}]$ for pMCI vs. sMCI, and $[10^{-2}, 10^{-1}]$, $[10^{-4}, 10^{-2}]$, $[10^{-1}, 10^0]$, and $[10^{-3}, 10^{-1}]$ for MCI vs. NC.

respectively. We also observe that with the parameters used for ADNI1, the suitable parameters can be easily determined for ADNI2. Specifically, we have performed fine-tuning according to the range on which best parameters falls for ADNI1. As the Fig. 3 and 4 shown, the best parameter combination for ADNI1 and ADNI2 is basically consistent, which might mean that it is efficient to optimize parameters when our method is applied to a new dataset. In addition, in our experiments, we found the influence of ρ_i and γ_i to the model is marginal when ρ_i and γ_i fall in the range of $[10^{-2}, 10^{-1}]$ and $[10^{-5}, 10^0]$, respectively. And the optimal average performance with AUC of 0.965 ± 0.011, 0.824 ± 0.027 and 0.799 ± 0.021 is

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

TABLE VII
A BRIEF SUMMARY OF THE STATE-OF-THE-ART STUDIES BASED ON ADNI DATASET FOR AD DIAGNOSIS (TOP: TRADITIONAL MACHINE LEARNING METHODS; BOTTOM: DEEP LEARNING APPROACHES).

Methods	Modalities	Subjects	AD vs. NC				pMCI vs. sMCI				MCI vs. NC			
			AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN
Shao *et al*. [31]	MRI+PET	160 AD + 160 NC + 187 pMCI +273 sMCI	0.950	0.925	0.904	0.941	0.710	0.755	0.633	0.838	0.800	0.825	0.786	0.861
Hao *et al*. [32]	MRI+PET	51 AD + 52 NC + 43 pMCI +56 sMCI	0.980	0.976	0.967	0.984	0.760	0.778	0.855	0.674	0.810	0.845	0.662	0.940
Tong *et al*. [33]	MRI+PET+CSF+Gene	37 AD + 35 NC +75 MCI	0.983	0.918	0.947	0.889	-	-	-	-	0.812	0.795	0.671	0.851
Jie *et al*. [34]	MRI+PET+CSF	51 AD +52 NC + 43 pMCI + 56 sMCI	0.970	0.950	0.950	0.949	0.700	0.689	0.718	0.647	0.820	0.793	0.665	0.859
Zhu *et al*. [35]	MRI+PET	51 AD + 52 NC + 43 pMCI +56 sMCI	-	0.955	-	-	-	0.712	-	-	-	0.797	-	-
Shi *et al*. [36]	MRI+PET+CSF	51 AD + 99 MCI + 52 NC	0.937	0.949	0.944	0.954	-	-	-	-	0.766	0.799	0.710	0.846
Suk *et al*. [37]	MRI+PET+CSF	51 AD + 52 NC + 43 pMCI +56 sMCI	-	0.951	0.980	0.920	-	0.730	0.890	0.530	-	0.788	0.560	0.908
Feng *et al*. [38]	MRI+PET	93 AD + 100 NC + 76 pMCI +128 sMCI	0.968	0.948	0.925	0.977	-	-	-	-	-	-	-	-
Pan *et al*. [39]	MRI+PET	267 AD + 440 NC + 254 pMCI + 405 sMCI	0.958	0.905	0.898	0.910	0.827	0.762	0.763	0.761	-	-	-	-
Proposed	MRI+PET	229 AD + 206 NC + 224 pMCI + 161 sMCI	0.977	0.969	0.980	0.957	0.848	0.852	0.905	0.778	0.826	0.821	0.723	0.871

obtained by the model for AD vs. NC, pMCI vs. sMCI, and MCI vs. NC, respectively.

V. DISCUSSION

In this section, we first summarize the main differences between our proposed method and previous studies on AD-related diagnosis. We then present the most discriminative ROIs identified by our proposed method.

A. Comparison with Previous Studies

In TABLE VII, we roughly summarize and compare our results with those of several state-of-the-art methods [31]–[39] (including traditional machine learning methods [31]–[36] and deep learning approaches [37]–[39]) reported in the literature for AD diagnosis using baseline multi-modal data from ADNI. As we can observe from TABLE VII, our proposed method get comparable performance with the state-of-the-art methods in most situations. Notably, although direct comparison among these methods is impossible and unfair due to varying subject numbers and inconsistent dataset partitions, we can still draw some conjectures: 1) The multi-modal fusion methods [32], [33] can learn more discriminative information via exploring the comprehensive characteristics inherent in multi-modal data for AD diagnosis. 2) A larger dataset would further improve predictive performance of classifier, which can be implemented by collecting more data and some technologies that deal with incomplete data (e.g., generative adversarial network [39]). 3) The increase of data modalities would boost the classification performance via providing informative specific views for AD [33], [37]. Compared with most of conventional machine learning methods, the proposed method adopts the bi-directional mapping for simultaneously considering both data projection and reconstruction. In this way, informative shared space that preserves original information can be found. Also, the underlying associations (i.e., redundancy, relevance, and class separability) of multi-modal data in the shared space are captured by several relational regularizers. Moreover, the proposed method learns latent discriminant representations in a task-driven manner by integrating representation learning and classifier into a unified framework. Although deep learning methods are also end-to-end frameworks to learn discriminant features, rare studies focus on multi-modal underlying associations. Additionally, different from deep learning methods, our proposed method is more explainable for discovering useful biomarkers, and has less parameters which are robust to multi-site datasets.

B. Related ROIs

Apart from the learned shared representations, we are still interested in ROIs that make great contributions to the construction of shared space. Based on Eq.(15), each row vector of the projection matrix $\mathbf{P}$ corresponds to a column vector (or feature vector) of original feature matrix $[\mathbf{X}_M, \mathbf{X}_P]$. The higher L_2 norm of each row vector of projection matrix $\mathbf{P}$, the more important the corresponding feature of $[\mathbf{X}_M, \mathbf{X}_P]$ is, and vice versa. Thus, we rank the L_2 norm of row vectors of projection matrix $\mathbf{P}$ and select top 10 ROIs for each modality on the basis of the frequency across all folds. Fig. 5 visualizes the selection results on all three tasks based on ADNI1 and ADNI2. Results show that selected regions refer to hippocampus, putamen, insula, pallidum, and different gyri (such as parahippocampal gyrus, middle temporal gyrus, and postcentral gyrus) in MRI and orbitofrontal cortex, temporal pole, superior frontal gyrus, supplementary motor area, and hippocampus in PET. Previous studies [40]–[43] have also demonstrated these regions are more helpful for the AD-related diagnosis.

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

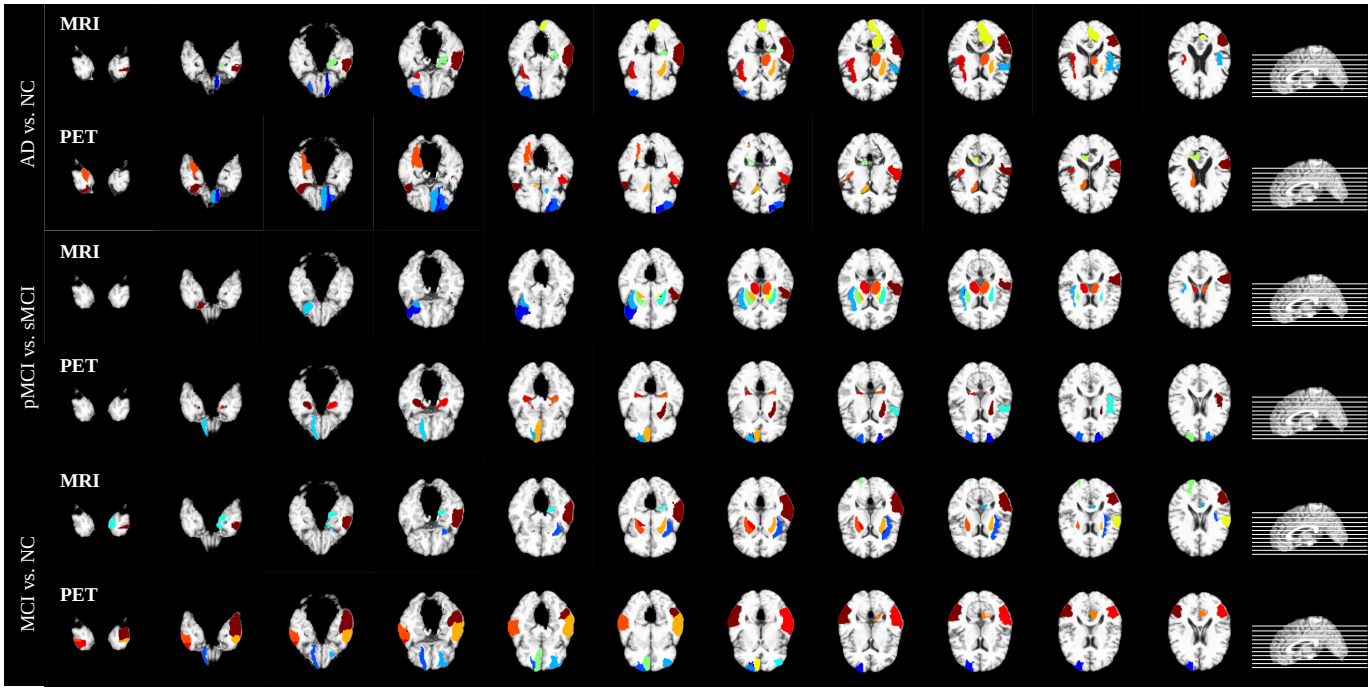

Fig. 5. The selected top 10 ROIs on three tasks. For each task, the top row relates to the MRI data and the bottom one refers to the PET data. Different colors denote different ROIs.

VI. Conclusion

In this paper, we propose a relation-induced multi-modal shared representation learning framework for AD diagnosis. The proposed method integrates representation learning, dimension reduction, and classifier modeling into a unified framework. Within this shared space, we utilize several relational regularizers (including feature-feature, feature-label, and sample-sample regularizers) and auxiliary regularizers to induce learning potential associations inherent in multi-modal data and alleviate overfitting, respectively. Then we project the shared representations into target space for AD diagnosis. The experimental results demonstrate that our proposed method not only outperforms several state-of-the-art methods, but also identifies some potential biomarkers for AD diagnosis. In the future work, we will investigate the feasibility of using our proposed method in the diagnosis of other brain diseases.

Acknowledgement

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The investigators within the ADNI contributed to the design and implementation of ADNI but did not participate in the analysis or writing of this paper. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how to apply/ADNI Acknowledgement List.pdf.

References

[1] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[2] C. Andrade and R. Radhakrishnan, "The prevention and treatment of cognitive decline and dementia: An overview of recent research on experimental treatments," *Indian Journal of Psychiatry*, vol. 51, no. 1, p. 12, 2009.

[3] M. W. Weiner, D. P. Veitch, P. S. Aisen *et al.*, "The alzheimer's disease neuroimaging initiative: a review of papers published since its inception," *Alzheimer's & Dementia*, vol. 9, no. 5, pp. e111–e194, 2013.

[4] L. M. Shaw, H. Vanderstichele, M. Knapik-Czajka *et al.*, "Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects," *Annals of Neurology*, vol. 65, no. 4, pp. 403–413, 2009.

[5] L. K. Ferreira and G. F. Busatto, "Neuroimaging in alzheimer's disease: current role in clinical practice and potential future applications," *Clinics*, vol. 66, pp. 19–24, 2011.

[6] S. Liu, S. Liu, W. Cai *et al.*, "Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1132–1140, 2014.

[7] 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 Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 173–183, 2017.

[8] K. Ritter, J. Schumacher, M. Weygandt *et al.*, "Multimodal prediction of conversion to alzheimer's disease based on incomplete biomarkers," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 2, pp. 206–215, 2015.

[9] J. L. Shaffer, J. R. Petrella, F. C. Sheldon *et al.*, "Predicting cognitive decline in subjects at risk for alzheimer disease by using combined cerebrospinal fluid, mr imaging, and pet biomarkers," *Radiology*, vol. 266, no. 2, pp. 583–591, 2013.

[10] C. Hinrichs, V. Singh, G. Xu, and S. Johnson, "Mkl for robust multimodality ad classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2009, pp. 786–794.

[11] T. Zhou, M. Liu, K.-H. Thung, and D. Shen, "Latent representation learning for alzheimer's disease diagnosis with incomplete multimodality neuroimaging and genetic data," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2411–2422, 2019.

[12] X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen, "Canonical feature selection for joint regression and multi-class identification in alzheimer's disease diagnosis," *Brain Imaging and Behavior*, vol. 10, no. 3, pp. 818–828, 2016.

[13] T. Zhou, K.-H. Thung, M. Liu, F. Shi, C. Zhang, and D. Shen, "Multimodal latent space inducing ensemble svm classifier for early dementia

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3063150, IEEE Transactions on Medical Imaging

diagnosis with neuroimaging data," *Medical Image Analysis*, vol. 60, p. 101630, 2020.

[14] D. L. Donoho *et al.*, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, vol. 1, no. 2000, p. 32, 2000.

[15] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Advances in Neural Information Processing Systems*, 2010, pp. 1813–1821.

[16] B. Jie, D. Zhang, B. Cheng, and D. Shen, "Manifold regularized multi-task feature selection for multi-modality classification in alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 275–283.

[17] X. Zhu, H.-I. Suk, L. Wang *et al.*, "A novel relational regularization feature selection method for joint regression and classification in ad diagnosis," *Medical Image Analysis*, vol. 38, pp. 205–214, 2017.

[18] B. Lei, P. Yang, T. Wang, S. Chen, and D. Ni, "Relational-regularized discriminative sparse learning for alzheimer's disease diagnosis," *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 1102–1113, 2017.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[20] I. Jolliffe, *Principal Component Analysis*. American Cancer Society, 2005.

[21] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 880–893, 2020.

[22] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in mri data," *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998.

[23] 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 Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[24] D. Shen and C. Davatzikos, "Hammer: hierarchical attribute matching mechanism for elastic registration," *IEEE Transactions on Medical Imaging*, vol. 21, no. 11, pp. 1421–1439, 2002.

[25] B. Shi, Y. Chen, P. Zhang *et al.*, "Nonlinear feature transformation and deep fusion for alzheimer's disease staging analysis," *Pattern Recognition*, vol. 63, pp. 487–498, 2017.

[26] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[27] L. Du, F. Liu, K. Liu *et al.*, "Associating multi-modal brain imaging phenotypes and genetic risk factors via a dirty multi-task learning method," *IEEE Transactions on Medical Imaging*, 2020.

[28] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Medical Image Analysis*, vol. 43, pp. 157 – 168, 2018.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[30] Z. Ning, J. Luo, Y. Li, S. Han, Q. Feng, Y. Xu, W. Chen, T. Chen, and Y. Zhang, "Pattern classification for gastrointestinal stromal tumors by integration of radiomics and deep convolutional features," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1181–1191, 2018.

[31] W. Shao, Y. Peng, C. Zu *et al.*, "Hypergraph based multi-task feature selection for multimodal classification of alzheimer's disease," *Computerized Medical Imaging and Graphics*, vol. 80, p. 101663, 2020.

[32] X. Hao, Y. Bao, Y. Guo *et al.*, "Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of alzheimer's disease," *Medical Image Analysis*, vol. 60, p. 101625, 2020.

[33] T. Tong, K. Gray, Q. Gao *et al.*, "Multi-modal classification of alzheimer's disease using nonlinear graph fusion," *Pattern Recognition*, vol. 63, pp. 171–181, 2017.

[34] B. Jie, D. Zhang, B. Cheng, D. Shen, and A. D. N. Initiative, "Manifold regularized multitask feature learning for multimodality disease classification," *Human Brain Mapping*, vol. 36, no. 2, pp. 489–507, 2015.

[35] X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen, "Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 607–618, 2015.

[36] Y. Shi, H.-I. Suk, Y. Gao, S.-W. Lee, and D. Shen, "Leveraging coupled interaction for multimodal alzheimer's disease diagnosis," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[37] H.-I. Suk, S.-W. Lee, D. Shen *et al.*, "Deep sparse multi-task learning for feature selection in alzheimer's disease diagnosis," *Brain Structure and Function*, vol. 221, no. 5, pp. 2569–2587, 2016.

[38] C. Feng, A. Elazab, P. Yang *et al.*, "Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm," *IEEE Access*, vol. 7, pp. 63 605–63 618, 2019.

[39] Y. Pan, M. Liu, C. Lian, Y. Xia, and D. Shen, "Disease-image specific generative adversarial network for brain disease diagnosis with incomplete multi-modal neuroimages," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 137–145.

[40] Y. Han, J. Wang, Z. Zhao *et al.*, "Frequency-dependent changes in the amplitude of low-frequency fluctuations in amnestic mild cognitive impairment: a resting-state fmri study," *Neuroimage*, vol. 55, no. 1, pp. 287–295, 2011.

[41] N. S. Ryan, S. Keihaninejad, T. J. Shakespeare *et al.*, "Magnetic resonance imaging evidence for presymptomatic change in thalamus and caudate in familial alzheimer's disease," *Brain*, vol. 136, no. 5, pp. 1399–1414, 2013.

[42] E. D. Vidoni, G. P. Thomas, R. A. Honea, N. Loskutova, and J. M. Burns, "Evidence of altered corticomotor system connectivity in early-stage alzheimer's disease," *Journal of Neurologic Physical Therapy*, vol. 36, no. 1, p. 8, 2012.

[43] C. T. Watson, P. Roussos, P. Garg *et al.*, "Genome-wide dna methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with alzheimer's disease," *Genome Medicine*, vol. 8, no. 1, p. 5, 2016.

0278-0062 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Southern California. Downloaded on March 23,2021 at 21:20:34 UTC from IEEE Xplore. Restrictions apply.