Contents lists available at ScienceDirect

Medical Image Analysis

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

Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data

Tao Zhou[a,b], Kim-Han Thung[a], Mingxia Liu[a], Feng Shi[c], Changqing Zhang[d], Dinggang Shen[a,e,*]

[a] *Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, NC 27599, USA*
[b] *Inception Institute of Artificial Intelligence, Abu Dhabi 51133, United Arab Emirates*
[c] *United Imaging Intelligence, Shanghai, China*
[d] *School of Computer Science and Technology, Tianjin University, Tianjin 300072, China*
[e] *Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea*

ARTICLE INFO

Article history:
Received 28 November 2018
Revised 15 December 2019
Accepted 19 December 2019
Available online 28 December 2019

Keywords:
Alzheimer's disease (AD)
Multi-modality data
Missing modality
Latent space
Multiple diversified classifiers

ABSTRACT

Fusing multi-modality data is crucial for accurate identification of brain disorder as different modalities can provide complementary perspectives of complex neurodegenerative disease. However, there are at least four common issues associated with the existing fusion methods. *First*, many existing fusion methods simply concatenate features from each modality without considering the correlations among different modalities. *Second*, most existing methods often make prediction based on a single classifier, which might not be able to address the heterogeneity of the Alzheimer's disease (AD) progression. *Third*, many existing methods often employ feature selection (or reduction) and classifier training in two independent steps, without considering the fact that the two pipelined steps are highly related to each other. *Forth*, there are missing neuroimaging data for some of the participants (*e.g.*, missing PET data), due to the participants' "no-show" or dropout. In this paper, to address the above issues, we propose an early AD diagnosis framework via novel multi-modality latent space inducing ensemble SVM classifier. Specifically, we first project the neuroimaging data from different modalities into a latent space, and then map the learned latent representations into the label space to learn multiple diversified classifiers. Finally, we obtain the more reliable classification results by using an ensemble strategy. More importantly, we present a Complete Multi-modality Latent Space (CMLS) learning model for complete multi-modality data and also an Incomplete Multi-modality Latent Space (IMLS) learning model for incomplete multi-modality data. Extensive experiments using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset have demonstrated that our proposed models outperform other state-of-the-art methods.

1. Introduction

Alzheimer's Disease (AD) is the most common form of dementia, which is characterized as a genetically complex and irreversible neurodegenerative disorder and often found in people over 65 years old (Alzheimer's, 2015). Recent studies have demonstrated that there are about 26.6 million AD patients worldwide, and 1 out of 85 people will be affected by AD by 2050 (Palmer, 2011). Since there is no cure for AD, the timely and accurate diagnosis of AD and its prodromal stage (*i.e.*, Mild Cognitive Impairment (MCI)) is clinically important (Zhou et al., 2019a; Lu et al., 2018; Long et al., 2018; Thung et al., 2018; Zhou et al., 2019c; Long , 2016; Wee et al., 2014).

Neuroimaging techniques, such as magnetic resonance imaging (MRI) (Wolz et al., 2012; Chen et al., 2019; Vemuri et al., 2008; Liu et al., 2018; Fan et al., 2019; Lian et al., 2018) and positron emission tomography (PET) (Herholz et al., 2002; Fan et al., 2008), are powerful tools that are able to measure different yet complementary information, and thus provide unprecedented opportunities for dementia study. As neuroimaging data are very high-dimensional, existing methods often use region-of-interest (ROI) based features, instead of the original voxel based features, for analysis (Chaves and et al., 2009; Magnin et al., 2009; Tong et al., 2014). In this context, many machine learning algorithms have been developed to utilize neuroimaging data for AD diagnosis. A conventional machine learning framework concatenates these fea-

[*] Corresponding author at: Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, NC 27599, USA.

E-mail addresses: taozhou.dreams@gmail.com (T. Zhou), henrythung@gmail.com (K.-H. Thung), mxliu@med.unc.edu (M. Liu), feng.shi@united-imaging.com (F. Shi), zhangchangqing@tju.edu.cn (C. Zhang), dgshen@med.unc.edu (D. Shen).

https://doi.org/10.1016/j.media.2019.101630
1361-8415/© 2020 Elsevier B.V. All rights reserved.

tures from multiple modalities into a long vector, and subsequently employs feature selection and classification. While this framework is easy to implement, it ignores many prior knowledge of the data that could be beneficial to the study. To exploit the correlation among multi-modalities, some studies have been developed to fuse the complementary information from multi-modality data for accurate AD diagnosis. For example, Zhu et al. (2016b) use Canonical Correlation Analysis (CCA) to first transform multi-modality data into a common CCA space, and then use the transformed features for classification. Hinrichs et al. (2009) use Multiple Kernel Learning (MKL) to fuse multi-modality data by learning an optimal linearly combined kernels for classification. Zhang et al. (2012) introduce a multi-modality multi-task learning based method by jointly using clinical scores and label information. A multi-task learning based feature selection method with inter-modality relationship preserving constraint is proposed in Liu et al. (2014).

Currently, some multi-modality fusion studies first employ feature selection or multi-modality fusion, and then the selected or fusion features are fed to train a classifier (*e.g.*, Support Vector Machine (SVM)) (Zhu et al., 2016b; Zhang et al., 2012; Lei et al., 2017). It is a popular two-step strategy in AD diagnosis framework. However, the feature selection in the first step may not be the best to the classifier training in the second step, thus this separated pipeline could degrade the final classification performance. Moreover, some existing methods (Zhu et al., 2016b; Lei et al., 2017) often focus on learning a single classifier (*e.g.*, SVM classifier) for AD diagnosis, which is difficult to address the heterogeneity of complex brain disorder. As in many recognition and classification tasks, ensemble approaches can obtain more promising performance (Freund and Schapire, 1997), which can reduce the variance of the base classifiers. More importantly, ensemble classifiers can obtain more promising performance if they have different decision boundaries, allowing for more flexibility through imposing diversity among the all models (Brown et al., 2005). Thus, to deal with this disease heterogeneity issue, it is more reasonable to train a set of diversified classifiers and ensemble them (*i.e.*, instead of training a single classifier), which has been shown effective in previous studies (Freund and Schapire, 1997; Brown et al., 2005; Liu et al., 2016).

In addition, it is unavoidable to have missing data, *i.e.*, some subjects have missing PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, due to the participants' no show-up or dropout. Generally, there are two approaches to deal with missing data issue (Thung et al., 2014), *i.e.*, (1) discard the subjects with missing data, and (2) impute the missing data. Most existing approaches discard subjects with at least one missing modality and perform disease identification based on the remainder of the subjects with complete multi-modalities. In this case, this approach discards a lot of information that is potentially useful, which could not learn a reliable diagnostic model. Besides, imputation methods estimate missing values based on available data using specific imputation techniques, *e.g.*, expectation maximization (EM) (Schneider, 2001), singular value decomposition (SVD) (Hastie et al., 2015), and matrix completion method (Thung et al., 2014). However, the effectiveness of these approaches can be affected by imputation artifacts. Moreover, several recently developed multi-view learning methods (Yuan et al., 2012; Liu et al., 2017) and multi-task learning based methods (Thung et al., 2014) demonstrate greater accuracies in AD diagnosis.

In general, most of the existing approaches suffer from the following challenges. First, some traditional multi-modal fusion methods simply concatenate features from each modality without considering the correlations among different modalities. Second, existing methods often make prediction based on a single classifier, which could not be able to address the heterogeneity of the Alzheimer's disease (AD) progression. Third, most existing methods often employ feature selection (or reduction) and classifier training in two independent steps, without considering the fact that these two pipelined steps are highly related to each other. Forth, the missing data issue is commonly existing in multi-modality setting, thus, how to make use of all available subjects to train a more reliable model is critical for early AD diagnosis.

To this end, we propose a novel AD diagnosis framework via multi-modal latent space inducing ensemble SVM classifier, which can seamlessly perform latent space learning and ensemble of diversified classifiers learning in a unified framework (as shown in Fig. 1). Specifically, we first project neuroimaging features from different modalities (*i.e.*, MRI and PET in our study) into a common latent space, to exploit the cross-modality correlations while learning their latent representations. Besides, to address the heterogeneity of AD progression, we learn multiple diversified classifiers by mapping the latent representations into multiple label spaces, and use an ensemble strategy to obtain a more robust classification result. Furthermore, we integrate latent space learning and classifier training into a unified framework, so that all the components in the framework can work together to achieve a better AD diagnostic model. More importantly, our proposed framework can address missing data issue and make use of all samples to train a reliable prediction model. We conducted experiments on the ADNI database, and the results have demonstrated the superiority of the proposed method over state-of-the-art methods.

Compared with previous AD diagnostic models, the main contributions of our work are four-fold, as described below.

- Our proposed AD diagnosis method seamlessly integrates the latent representations learning and multiple diversified classifiers learning into a unified framework, while previous methods often employ feature selection and classifier training in two separate steps. Thus, the features selected would be optimal to the classifier in our proposed method.
- We learn common latent representation for neuroimaging data from different modalities to better exploit the intrinsic correlations among them. In contrast, previous methods often fuse features from different modalities in their original forms and ignore the correlations among different modalities.
- Unlike existing diagnosis methods that often train a single classifier for model prediction, our proposed method learns multiple diversified classifiers and obtain the final result via an ensemble strategy. In this way, our model can better address the heterogeneity of neurodegenerative disease.
- Last but not least, our proposed method is also able to handle the missing data issue. Specifically, when a subject is associated with one or more missing modalities, instead of excluding this subject from the analysis, our model will project only the available modalities into the latent space. In this way, our proposed method will make use of all available subjects to train a more reliable prediction model. In addition, our method does not need to impute the missing data, which may introduce unnecessary noise into the data that subsequently reduces the classification performance.

A preliminary version of complete multi-modal latent space learning framework was presented in Zhou et al. (2018). This paper extends it and proposes a novel incomplete multi-modal latent space learning framework for incomplete multi-modal datset. We implement disease diagnosis on both complete and incomplete multi-modal dataset in this study.

The rest of this paper is organized as follows. Section 2 describes the materials and neuroimage preprocessing steps. Section 3 gives the details of the proposed approach. The experimental results are reported in Section 4. Finally, Section 5 concludes this paper.

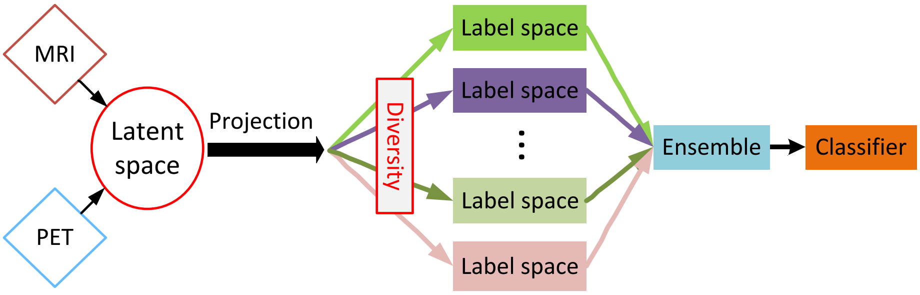


Fig. 1. The flow diagram of our proposed AD diagnosis framework. We project multi-modality data (*i.e.*, MRI and PET in our case) into a common latent space to exploit the correlation among multi-modal neuroimaging data. Then, multiple classifiers with diversity constraint are trained and an ensemble strategy is used to obtain the final classification results.

Table 1
Demographic information of the used subjects (MMSE: mini-mental state examination).

	Female / male	Education	Age	MMSE
NC	108 / 118	16.0 ± 2.9	75.8 ± 5.0	29.1 ± 1.0
sMCI	68 / 137	15.7 ± 3.1	75.1 ± 7.6	27.4 ± 1.7
pMCI	62 / 95	15.6 ± 2.9	74.7 ± 6.9	26.6 ± 1.7
AD	87 / 99	14.7 ± 3.1	75.3 ± 7.6	23.3 ± 2.0

2. Materials and neuroimage preprocessing

In this study, we used data from the public ADNI database (Jack , 2008) for performance evaluation. The ADNI dataset had been launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and nonprofit organizations with a five-year public private partnership. The main goal of ADNI is to investigate if MRI, PET and other biological markers, together with clinical and neuropsychological assessments can be combined to measure the progression of AD -.

2.1. Subjects

In this study, we used 774 subjects from ADNI-1, including 226 normal controls (NC), 362 MCI and 186 AD subjects. In addition, there are only 379 subjects with complete MRI and PET data, including 101 NC, 185 MCI, and 93 -AD. Moreover, 362 MCI subjects included 205 stable MCI (sMCI) subjects and 157 progressive MCI (pMCI) subjects. In this study, we defined progressive MCI (pMCI) subjects as MCI subjects that will progress to AD within 24 months, while sMCI subjects as MCI subjects that remain stable at all available time points ($0-96$ months). The demographic and clinical information of all subjects used in this study are summarized in Table 1.

2.2. Neuroimage preprocessing

In this study, we downloaded ADNI preprocessed 1.5T T1-weighted MR images from the ADNI website.[1] All structural MR images were acquired from 1.5T scanners. These MR images were reviewed for quality and automatically corrected for spatial distortion, which were caused by gradient nonlinearity and B1 field inhomogeneity. Further, all PET images (*i.e.*, FDG-PET scans) were collected from a variety of scanners with protocols individualized for each scanner. Following previous works (Zhang et al., 2012; Xue et al., 2006), we further processed the MR images using a standard pipeline including the following steps: (1) anterior commissure-posterior commissure (AC-PC) correction by using MIPAV software,[2] (2) intensity inhomogeneity correction by using N3 algorithm (Sled et al., 1998), (3) brain extraction on all structural MR images by using a robust skull-stripping method (Wang et al., 2014), (4) cerebellum removal based on based on registration and intensity inhomogeneity correction, (5) tissues segmentation by using FAST method in FSL package (Zhang et al., 2001), obtaining three different tissues (*i.e.*, white matter (WM), gray matter (GM), and cerebrospinal fluid), and (6) registration to a template (Kabani, 1998; Wu et al., 2006) by using HAMMER algorithm (Shen and Davatzikos, 2002), and then dissecting images into 93 regions-of-interest (ROIs) by labeling them based on the Jacob template (Kabani, 1998). In detail, we computed the GM tissue volume of each ROI in the labeled image, and then normalized them with the intracranial volume, thus the ROI-based feature was used to represent each subject. Besides, for each subject, we first aligned PET images to their corresponding T1-weighted MR images by using affine registration, and then computed the average PET intensity value of each ROI as PET feature. Thus, in our study, we have 93-dimensional ROI-based features from both the MRI and PET data, respectively.

3. Proposed method

3.1. Preliminary

An SVM (Vapnik, 2013) is a discriminative classifier formally defined by a separating hyperplane, which has been widely used in many fields such as pattern recognition and machine learning. Generally, the primal SVM can be formulated as

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} f\left(y_i, \mathbf{x}_i^\top \mathbf{w} + b\right) + \lambda \Psi(\mathbf{w}), \tag{1}$$

where $f(\cdot)$ is a penalty function, $\mathbf{w}$ and b are the weight vector and bias, respectively, $(\mathbf{x}_i, y_i)$ is the i-th sample of input-output pair, $(\mathbf{x}_i^\top \mathbf{w} + b)$ is the predicted output for the i-th sample, N is the number of samples, and $\Psi(\mathbf{w})$ is a regularizer term imposed on $\mathbf{w}$. Besides, λ is a non-negative parameter used to balance between the data fitting loss term and the regularizer term.

3.2. Common latent space learning for multi-modality data

For a multi-modality data set $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$, where $\mathbf{X}_m \in \mathbb{R}^{d_m \times N}$ denotes the feature matrix for the mth modality with d_m features and N subjects, and M denotes the number of modalities, we exploit the correlations among different modalities by projecting different modalities into a common latent space, as given be-

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

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

low:

$$\min_{\mathbf{V}_m,\mathbf{H}} \sum_{m=1}^{M} \left(\|\mathbf{V}_m^\top \mathbf{X}_m - \mathbf{H}\|_F^2 + \gamma \|\mathbf{V}_m\|_{2,1}\right), \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^{h\times N}$ denotes the common latent representation, $\mathbf{V}_m \in \mathbb{R}^{d_m\times h}$ denotes the projection matrix that project $\mathbf{X}_m$ to $\mathbf{H}$, γ denotes the regularization parameter, and h denotes the dimension of the latent space. The $\ell_{2,1}$-norm regularizer (*i.e.*, $\|\mathbf{V}_m\|_{2,1} = \sum_{i=1}^{d_m}\sqrt{\sum_{j=1}^{h} v_{m,ij}^2}$) has been widely applied to multi-task feature learning (Wang et al., 2017; Zhou et al., 2019b; Zhang et al., 2018; Thung and Wee, 2018), and we use it to collectively penalize the coefficients in each row of $\mathbf{V}_m$, and enforce row-wise sparsity in $\mathbf{V}_m$. Consequently, the $\ell_{2,1}$-norm on $\mathbf{V}_m$ encourages the selection of useful (ROI-based) features from $\mathbf{X}_m$ during the latent space learning in Eq. (2). In the next section, we will give the details of our proposed AD diagnosis model by integrating the latent space learning and classifier training into a unified framework.

3.3. Proposed dementia diagnosis framework

Combining Eqs. (1) and (2), and using the latent features instead of the original features for classifier training, we have a unified framework of latent space learning and classifier training, as given by

$$\min_{\mathbf{V}_m,\mathbf{w},\mathbf{H},b} \sum_{i=1}^{N} f\left(y_i, \mathbf{h}_i^\top \mathbf{w} + b\right) + \lambda\Psi(\mathbf{w}) + \beta\sum_{m=1}^{M}\|\mathbf{V}_m^\top \mathbf{X}_m - \mathbf{H}\|_F^2 + \gamma\sum_{m=1}^{M}\|\mathbf{V}_m\|_{2,1}, \tag{3}$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N] \in \mathbb{R}^{h\times N}$ with $\mathbf{h}_i \in \mathbb{R}^{h\times 1}$ denoting the latent representation of the ith sample (*i.e.*, the ith column of $\mathbf{H}$), and $y_i \in \{-1, 1\}$ is the corresponding label for the ith sample. Besides, λ, β and γ are the regularization parameters that control the contribution of each term in Eq. (3). If we use hinge loss function for $f(\cdot)$, the first term in Eq. (3) is given as

$$\sum_{i=1}^{N} f\left(y_i, \mathbf{h}_i^\top \mathbf{w} + b\right) = \sum_{i=1}^{N}\left(1 - (\mathbf{h}_i^\top \mathbf{w} + b)y_i\right)_+^p, \tag{4}$$

where operation $(\cdot)_+$ is defined as $(x)_+ := \max(x, 0)$, which returns x if it is non-negative, and returns zero otherwise. Besides, p is a constant, which is normally set to value 1 or 2 (Guo et al., 2017).

Eq. (3) consists of only a single SVM classifier with hinge loss function, which may not be able to address the heterogeneity of AD progression. Previous works such as Freund and Schapire (1997), Brown et al. (2005) have indicated that the ensemble of multiple classifiers could be a better and more reliable prediction model. Thus, in this study, following the work in Guo et al. (2017), we extend the single classifier training in Eq. (3) into the following framework of multiple classifiers training, given as

$$\min_{\mathbf{V}_m,\mathbf{W},\mathbf{H},\mathbf{b}} \sum_{c=1}^{C}\sum_{i=1}^{N}\left(1 - (\mathbf{h}_i^\top \mathbf{w}_c + b_c)y_i\right)_+^p + \lambda\Psi(\mathbf{W}) + \beta\sum_{m=1}^{M}\|\mathbf{V}_m^\top \mathbf{X}_m - \mathbf{H}\|_F^2 + \gamma\sum_{m=1}^{M}\|\mathbf{V}_m\|_{2,1}, \tag{5}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C] \in \mathbb{R}^{h\times C}$ is the weight matrix with each column (*i.e.*, $\mathbf{w}_i$) denoting the weight vector for one classifier, $\mathbf{b} = [b_1, b_2, \ldots, b_C] \in \mathbb{R}^{C\times 1}$ is the corresponding bias vector, and C is the number of classifiers. By using the model in Eq. (5), we can learn multiple classifiers, and obtain the final result by devising an ensemble strategy.

However, if there is no constraint on the weight matrix $\mathbf{W}$ that encourages diversity of classifiers, the performance of ensemble strategy will be limited. For example, if we choose $\Psi(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_F^2$, the weight vectors of the classifiers (*i.e.*, columns in $\mathbf{W}$) may end up similar to each other, thus defeating the purpose of ensembling the results of multiple classifiers. To have a more meaningful ensemble classifier that can better address the heterogeneity of the AD progression, we would like to learn a set of diversified classifiers that have weight vectors (*e.g.*, $\mathbf{w}_i$, $\mathbf{w}_j$) very different from each other. There are many ways to enforce diversity for a set of classifiers (Brown et al., 2005), but we choose method described in Guo et al. (2017) for its good performance. That is, we enforce the diversity of classifiers by minimizing the following function between each pair of classifier weight vectors, *i.e.*, $\{\min\|\mathbf{w}_i \odot \mathbf{w}_j\|_0 = \min\sum_k(\mathbf{w}_i(k)\cdot\mathbf{w}_j(k) \neq 0), i \neq j\}$, where $\odot$ denotes Hadamard product, $\|\cdot\|_0$ denotes ℓ_0-norm, and $\mathbf{w}_i(k)$ denotes the kth element in $\mathbf{w}_i$. This minimization will enforce the column weight vectors (*e.g.*, $\mathbf{w}_i$ and $\mathbf{w}_j$, $i \neq j$) in $\mathbf{W}$ to be as orthogonal as possible, so that the classifiers learnt would be diversified. However, as it is difficult to directly optimize the ℓ_0-norm problem, we choose to minimize the relaxed exclusivity function instead, *i.e.*, $\{\min\|\mathbf{w}_i \odot \mathbf{w}_j\|_1 = \min\sum_k|\mathbf{w}_i(k)|\cdot|\mathbf{w}_j(k)|, i \neq j\}$, where $|\cdot|$ denotes the absolute operator. To guarantee the convexity of the regularizer $\Psi(\mathbf{W})$, we combine the relaxed exclusivity function and the Frobenius norm of $\mathbf{W}$ to obtain the following regularizer (Guo et al., 2017):

$$\begin{aligned}\Psi(\mathbf{W}) &= \frac{1}{2}\|\mathbf{W}\|_F^2 + \sum_{i,j\neq i}\|\mathbf{w}_i \odot \mathbf{w}_j\|_1 \\ &= \frac{1}{2}\sum_{k=1}^{h}\left(\sum_{c=1}^{C}|\mathbf{w}_c(k)|\right)^2 \\ &= \frac{1}{2}\|\mathbf{W}^\top\|_{1,2}^2.\end{aligned} \tag{6}$$

The derivation of Eq. (6) is discussed in details in Guo et al. (2017). In summary, we use the $\min\frac{1}{2}\|\mathbf{W}^\top\|_{1,2}^2$ as a regularizer for $\mathbf{W}$ to encourage diversity of classifiers in our framework.

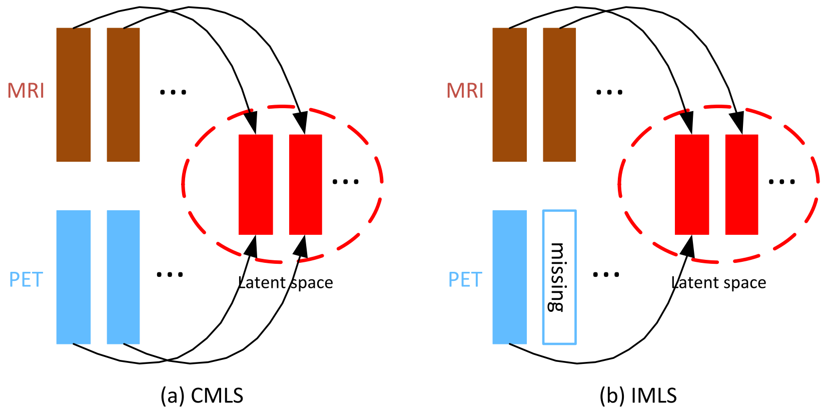
Finally, by substituting Eq. (6) into Eq. (5), we obtain a unified framework of latent space learning and multiple diversified classifiers training, where the final objective function is given as

$$\min_{\mathbf{V}_m,\mathbf{W},\mathbf{H},\mathbf{b}} \sum_{c=1}^{C}\sum_{i=1}^{N}\left(1 - (\mathbf{h}_i^\top \mathbf{w}_c + b_c)y_i\right)_+^p + \frac{\lambda}{2}\|\mathbf{W}^\top\|_{1,2}^2 + \beta\sum_{m=1}^{M}\|\mathbf{V}_m^\top \mathbf{X}_m - \mathbf{H}\|_F^2 + \gamma\sum_{m=1}^{M}\|\mathbf{V}_m\|_{2,1}. \tag{7}$$

Note that this framework assumes that all samples consist of complete multi modality data, thus we call our proposed method in Eq. (7) as Complete Multi-modal Latent Space (CMLS) learning model.

Remarks. In CMLS model, 1) each modality in multi-modality data is projected into a common latent space to exploit the correlations among different modalities; 2) the common latent representations are used to train multiple diversified classifiers to address heterogeneity issue of AD progression, so that an ensemble strategy can be used to improve the classification performance.

However, the applicability of the CMLS model is limited if it can only use samples with complete multi-modality data, as missing data is ubiquitous in multi-modality dataset. To address the missing data issue, we extend our framework in Eq. (7) into an Incom-



(a) CMLS

(b) IMLS

Fig. 2. Comparison of CMLS and IMLS models. Both CMLS and IMLS models project different modalities of multi-modality data into a common latent space, but CMLS model is only applicable if the multi-modality data is complete, while IMLS model is applicable even when the multi-modality data is incomplete. As shown in the figure, IMLS model only projects the existing modality into the common latent space.

Table 2
The main notations used in Eqs. (7) and (8).

Notation	Size	Description
$\mathbf{X}_m$	$d_m \times N$	Feature matrix for the mth modality.
$\mathbf{V}_m$	$d_m \times h$	Projection matrix for the mth modality.
$\mathbf{H}$	$h \times N$	Latent representation matrix.
$\mathbf{W}$	$h \times C$	Weight matrix for SVM classifiers, each column representing a classifier.
$\mathbf{b}$	$C \times 1$	Bias vector for SVM classifiers, each element representing the bias of a classifier.
$\mathbf{O}_m$	$N \times N$	Filter matrix for the mth modality. The value of its ith diagonal element indicates the availability of the mth modality data for the i-th subject.
d_m	–	Number of features for the mth modality data.
C	–	Number of classifiers.
N	–	Number of subjects (or samples).
h	–	Feature dimension in the latent feature representation.
M	–	Number of modalities in multi-modality data.
λ, β, γ	–	Regularization parameters

plete Multi-modal Latent Space (IMLS) learning model as

$$\min_{\mathbf{V}_m,\mathbf{W},\mathbf{H},\mathbf{b}} \quad \sum_{c=1}^{C}\sum_{i=1}^{N}\left(1-(\mathbf{h}_i^\top\mathbf{w}_c+b_c)y_i\right)_+^p+\frac{\lambda}{2}\|\mathbf{W}^\top\|_{1,2}^2$$
$$+\beta\sum_{m=1}^{M}\|\mathbb{P}_{\mathbf{O}_m}\left(\mathbf{H}-\mathbf{V}_m^\top\mathbf{X}_m\right)\|_F^2+\gamma\sum_{m=1}^{M}\|\mathbf{V}_m\|_{2,1}, \tag{8}$$

where $\mathbb{P}_{\mathbf{O}_m}(\mathbf{X})=\mathbf{X}\cdot\mathbf{O}_m$ is a mask operation that filters out missing data component (indicated by $\mathbf{O}_m$) in $\mathbf{X}$, and $\mathbf{O}_m$ is a diagonal matrix with its ith diagonal element indicating the existence of the mth modality data for the ith subject, *i.e.*, value of 1 when we have the data available, and 0 otherwise. For the convenience of our readers to comprehend the proposed models in Eqs. (7) and (8), we have listed the main notations used in Table 2.

Remarks. As shown in Fig. 2, it is worth noting that our IMLS model addresses the missing data issue by making use of all the available samples to train the prediction model. When a sample has the complete multi-modality data, our IMLS model will project all the modality data into a latent space; when one or more modalities are missing, our IMLS model will project only the available modality (or modalities) into the latent space.

3.4. Optimization

The problems in Eqs. (7) and (8) can be solved via the Augmented Lagrange Multiplier (ALM) (Lin et al., 2011) algorithm, which alternatively optimizes the variables, *i.e.*, optimizing one variable at a time with the other variables being fixed. We first provide the details for solving our CMLS model in Eq. (7). With consideration of $1-(\mathbf{h}_i^\top\mathbf{w}_c+b_c)y_i=y_iy_i-(\mathbf{h}_i^\top\mathbf{w}_c+b_c)y_i=y_i(y_i-(\mathbf{h}_i^\top\mathbf{w}_c+b_c))$, we introduce an auxiliary variable $z_i^c:=y_i-(\mathbf{h}_i^\top\mathbf{w}_c+b_c)$. Besides, we also introduce an auxiliary matrix $\mathbf{Q}=\mathbf{W}$ to make the problem separable. Subsequently, we have the following equivalent problem

$$\min_{\mathbf{V}_m,\mathbf{W},\mathbf{H},\mathbf{Q},\mathbf{Z},\mathbf{b}} \quad (\mathbf{Y}\odot\mathbf{Z})_+^p+\frac{\lambda}{2}\|\mathbf{W}^\top\|_{1,2}^2$$
$$+\beta\sum_{m=1}^{M}\|\mathbf{V}_m^\top\mathbf{X}_m-\mathbf{H}\|_F^2+\gamma\sum_{m=1}^{M}\|\mathbf{V}_m\|_{2,1},$$
$$s.t \quad \mathbf{Q}=\mathbf{W},\mathbf{Z}=\mathbf{Y}-\left(\mathbf{H}^\top\mathbf{Q}+\mathbf{1}_N\mathbf{b}^\top\right), \tag{9}$$

where $\mathbf{Y}\in\mathbb{R}^{N\times C}$ is the matrix of ground-truth labels with each of its column equivalent to the vector of ground-truth labels for N samples, $\mathbf{1}_N$ is an all-one column vector of dimension N, and $\mathbf{Z}\in\mathbb{R}^{N\times C}$ is the corresponding matrix of prediction differences between the ground-truth labels and the label predictions from diversified classifiers.

The above objective function can be solved by minimizing the following ALM problem

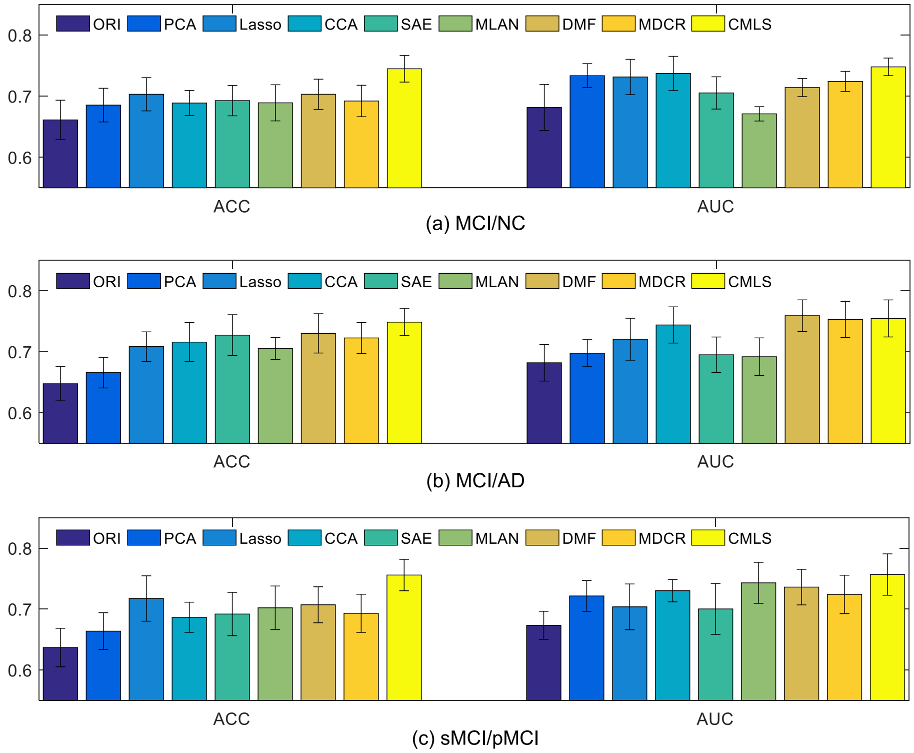
$$\mathcal{L}(\mathbf{V}_m,\ \mathbf{W},\mathbf{H},\mathbf{Q},\mathbf{Z},\mathbf{b},\mathbf{P}_1,\mathbf{P}_2)$$
$$=(\mathbf{Y}\odot\mathbf{Z})_+^p+\frac{\lambda}{2}\|\mathbf{W}^\top\|_{1,2}^2$$
$$+\beta\sum_{m=1}^{M}\|\mathbf{V}_m^\top\mathbf{X}_m-\mathbf{H}\|_F^2+\gamma\sum_{m=1}^{M}\|\mathbf{V}_m\|_{2,1}$$
$$+\Phi(\mathbf{P}_1,\mathbf{Q}-\mathbf{W})+\Phi\left(\mathbf{P}_2,\mathbf{Z}-\mathbf{Y}+\left(\mathbf{H}^\top\mathbf{Q}+\mathbf{1}_N\mathbf{b}^\top\right)\right), \tag{10}$$

where $\Phi(\mathbf{P},\Delta)=\frac{\mu}{2}\|\Delta\|_F^2+\langle\mathbf{P},\Delta\rangle$, with $\langle\ \cdot\ ,\ \cdot\ \rangle$ denoting the matrix inner product, μ is a positive penalty scalar, and $\mathbf{P}_1$ and $\mathbf{P}_2$ are Lagrangian multipliers. To find a minimal point for $\mathcal{L}$, we update one variable while keeping the other variables fixed. Thus, we split the above optimization problem into the following multiple subproblems.

V_m *-subproblem*: The associated optimization problem with respect to $\mathbf{V}_m$ can be written as

$$\min_{\mathbf{V}_m} \ \beta\|\mathbf{V}_m^\top\mathbf{X}_m-\mathbf{H}\|_F^2+\gamma\|\mathbf{V}_m\|_{2,1}. \tag{11}$$

We can solve problem (11) by taking the derivative of the objective function with respect to $\mathbf{V}_m$, and set it to zero (Nie et al., 2010; 2016). We first compute the derivative of the term $\|\mathbf{V}_m\|_{2,1}$ *w.r.t.* $\mathbf{V}_m$, *i.e.*, $\frac{\partial\|\mathbf{V}_m\|_{2,1}}{\partial\mathbf{V}_m}=\mathbf{\Lambda}\mathbf{V}_m$, where $\mathbf{\Lambda}\in\mathbb{R}^{d_m\times d_m}$ is a diagonal matrix with its ith diagonal element given as $\mathbf{\Lambda}_{ii}=\frac{1}{2\|\mathbf{V}_{m,i:}\|_2^2}$, and $\mathbf{V}_{m,i:}$



(a) MCI/NC

(b) MCI/AD

(c) sMCI/pMCI

Fig. 3. Comparison of classification results using two evaluation metrics (*i.e.*, ACC and AUC) for three classification tasks: (a) MCI/NC, (b) MCI/AD, and (c) sMCI/pMCI, using complete multi-modality dataset.

denotes the *i*th row of $\mathbf{V}_m$. Then, taking the derivative of the objective function in Eq. (11) and set it to zero, we have the following close-form solution

$$\mathbf{V}_m = \left(\mathbf{X}_m\mathbf{X}_m^\top + \frac{\gamma}{\beta}\mathbf{\Lambda}\right)^{-1}\mathbf{X}_m\mathbf{H}^\top. \tag{12}$$

Note that, $\mathbf{\Lambda}$ is dependent on $\mathbf{V}_m$, thus we can use an iterative optimization step to update $\mathbf{V}_m$ and $\mathbf{\Lambda}$ until convergence.

W -subproblem: When other variables are fixed except $\mathbf{W}$, the minization of objective function in Eq. (10) is equivalent to

$$\min_{\mathbf{W}} \ \frac{\lambda}{2}\|\mathbf{W}^\top\|_{1,2}^2 + \Phi(\mathbf{P}_1, \mathbf{Q} - \mathbf{W}). \tag{13}$$

The optimization problem in Eq. (13) can be split into a set of subproblems. We optimize each row of $\mathbf{W}$, *i.e.*, $\mathbf{W}_{i:}$, by solving

$$\min_{\mathbf{W}_{i:}} \ \frac{\lambda}{2}\mathbf{W}_{i:}\mathbf{G}\mathbf{W}_{i:}^\top + \Phi(\mathbf{P}_{1,i:}, \mathbf{Q}_{i:} - \mathbf{W}_{i:}), \tag{14}$$

where $\mathbf{G} := \mathrm{diag}([\frac{\|\mathbf{W}_{i:}\|_1}{|\mathbf{W}_{i:}(1)|+\xi}, \ldots, \frac{\|\mathbf{W}_{i:}\|_1}{|\mathbf{w}_{i:}(C)|+\xi}])$, and ξ is a small constant to avoid zero denominator. Then, an iterative strategy can be adopted to update $\mathbf{W}_{i:}$ and $\mathbf{G}$. When $\mathbf{G}$ is fixed, we update $\mathbf{W}_{i:}$ by using

$$\mathbf{W}_{i:} = (\mu\mathbf{Q}_{i:} + \mathbf{P}_{1,i:})(\lambda\mathbf{G} + \mu\mathbf{I})^{-1}. \tag{15}$$

H -subproblem: Dropping all unrelated terms with respect to $\mathbf{H}$ yields

$$\begin{aligned}\min_{\mathbf{H}} \ \beta\sum_{m=1}^{M}\left\|\mathbf{V}_m^\top\mathbf{X}_m - \mathbf{H}\right\|_F^2 \\ + \Phi\Big(\mathbf{P}_2, \mathbf{Z} - \mathbf{Y} + \big(\mathbf{H}^\top\mathbf{Q} + \mathbf{1}_N\mathbf{b}^\top\big)\Big).\end{aligned} \tag{16}$$

Taking the derivative of the above objective with respect to $\mathbf{H}$ and setting it to zero, we get the following close-form solution

$$\begin{aligned}\mathbf{H} = &\left(\frac{2\beta}{\mu}M\mathbf{I} + \mathbf{Q}\mathbf{Q}^\top\right)^{-1} \\ &\times\left(\frac{2\beta}{\mu}\sum_{m=1}^{M}\mathbf{V}_m^\top\mathbf{X}_m - \mathbf{Q}\big(\mathbf{Z} - \mathbf{Y} + \mathbf{1}_N\mathbf{b}^\top + \mathbf{P}_2/\mu\big)^\top\right),\end{aligned} \tag{17}$$

where $\mathbf{I}$ is an identity matrix.

Q -subproblem: The associated optimization problem with respect to $\mathbf{Q}$ can be written as

$$\min_{\mathbf{Q}} \Phi(\mathbf{P}_1, \mathbf{Q} - \mathbf{W}) + \Phi\Big(\mathbf{P}_2, \mathbf{Z} - \mathbf{Y} + \big(\mathbf{H}^\top\mathbf{Q} + \mathbf{1}_N\mathbf{b}^\top\big)\Big). \tag{18}$$

Taking the derivative of the above objective with respect to $\mathbf{Q}$ and setting it to zero, we obtain the following close-form solution

$$\mathbf{Q} = \big(\mathbf{I} + \mathbf{H}\mathbf{H}^\top\big)^{-1}\Big(\mathbf{W} - \mathbf{P}_1/\mu - \mathbf{H}\big(\mathbf{Z} - \mathbf{Y} + \mathbf{1}_N\mathbf{b}^\top + \mathbf{P}_2/\mu\big)\Big), \tag{19}$$

where $\mathbf{I}$ is an identity matrix.

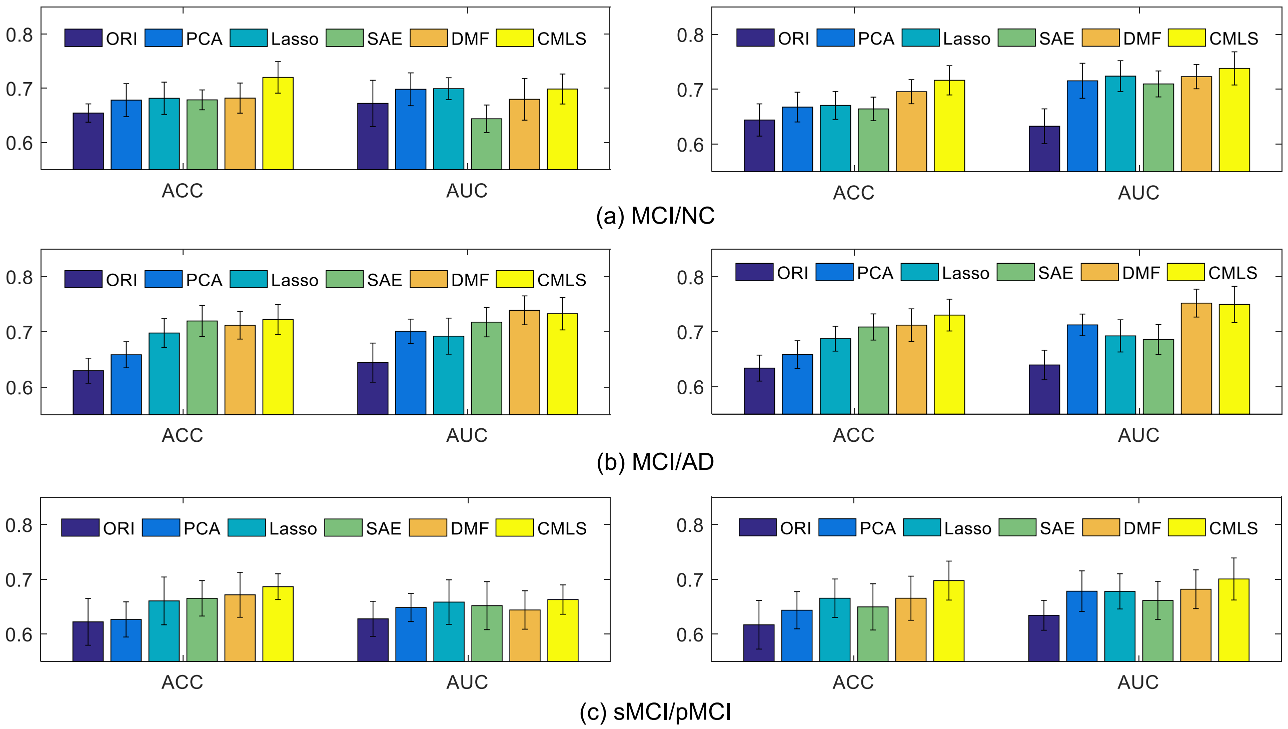
Z -subproblem: When other variables are fixed except $\mathbf{Z}$, the minimization of objective function in Eq. (10) is equivalent to

$$\begin{aligned}&\min_{\mathbf{Z}} \ (\mathbf{Y}\odot\mathbf{Z})_+^p + \Phi\Big(\mathbf{P}_2, \mathbf{Z} - \mathbf{Y} + \big(\mathbf{H}^\top\mathbf{Q} + \mathbf{1}_N\mathbf{b}^\top\big)\Big) \\ &\quad\Leftrightarrow \min_{\mathbf{Z}} \ (\mathbf{Y}\odot\mathbf{Z})_+^p + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{S}\|_F^2,\end{aligned} \tag{20}$$

where $\mathbf{S} = \mathbf{Y} - \mathbf{H}^\top\mathbf{Q} - \mathbf{1}_N\mathbf{b}^\top - \mathbf{P}_2/\mu$. In our study, we set $p = 2$. Following Guo et al. (2017), we have the following close-form solution

$$\mathbf{Z} = \Omega\odot\mathbf{S}/\left(1 + \frac{2}{\mu}\right) + \overline{\Omega}\odot\mathbf{S}, \tag{21}$$

where $\Omega := (\mathbf{Y}\odot\mathbf{S} > 0)$ is an indicator matrix, and $\overline{\Omega}$ is the complementary support of Ω.



(a) MCI/NC

(b) MCI/AD

(c) sMCI/pMCI

Fig. 4. Comparison of results for three classification tasks, *i.e.*, (a) MCI/NC, (b) MCI/AD, and (c) sMCI/pMCI, using single modality data: MRI (left) and PET (right).

b -subproblem: The optimization problem associated with **b** can be written as

$$\begin{aligned} &\min_{\mathbf{b}} \ \Phi\left(\mathbf{P}_2, \mathbf{Z}-\mathbf{Y}+\left(\mathbf{H}^\top\mathbf{Q}+\mathbf{1}_N\mathbf{b}^\top\right)\right) \\ &\quad \Leftrightarrow \min_{\mathbf{b}} \ \left\|\left(\mathbf{Z}-\mathbf{Y}+\mathbf{H}^\top\mathbf{Q}+\mathbf{P}_2/\mu\right)+\mathbf{1}_N\mathbf{b}^\top\right\|_F^2. \end{aligned} \tag{22}$$

The close-form solution for the above problem is given as

$$\mathbf{b} = \frac{1}{N}\left(\mathbf{Y}-\mathbf{Z}-\mathbf{H}^\top\mathbf{Q}-\mathbf{P}_2/\mu\right)^\top\mathbf{1}_N. \tag{23}$$

Multipliers: The multipliers $\mathbf{P}_1$ and $\mathbf{P}_2$ can be updated by

$$\begin{cases} \mathbf{P}_1 := \mathbf{P}_1 + \mu(\mathbf{Q}-\mathbf{W}) \\ \mathbf{P}_2 := \mathbf{P}_2 + \mu\left(\mathbf{Z}-\mathbf{Y}+\left(\mathbf{H}^\top\mathbf{Q}+\mathbf{1}_N\mathbf{b}^\top\right)\right). \end{cases} \tag{24}$$

We repeat the above updating steps iteratively until convergence. Similar with solving CMLS model, we can also use ALM algorithm to solve our IMLS model efficiently. Note that, the multi-modality data set $\mathbf{X}$ only consists of these subjects with complete multi-modalities in CMLS, while the multi-modality data set $\mathbf{X}$ consists of the all subjects with complete an incomplete multi-modalities in IMLS.

3.5. Prediction

After training our model, we can obtain the ensemble classifier weight $\mathbf{w}$ and bias b, which are the average of weight vectors and biases of all the diversified classifiers, respectively, *i.e.*, $\mathbf{w} = \frac{1}{C}\sum_{c=1}^{C}\mathbf{w}_c$, and $b = \frac{1}{C}\sum_{c=1}^{C} b_c$. Then, for a testing sample $\mathbf{x}^{te}$ with Θ available modalities, its latent representation is computed by averaging the feature projections from each available modality, *i.e.*, $\mathbf{h}_{te} = \frac{1}{|\Theta|}\sum_{m\in\Theta}\mathbf{V}_m^\top\mathbf{x}_m^{te}$, where $|\Theta|$ ($\le M$) denotes the number of modalities in Θ. Finally, the classification label for this test sample is given as $\mathbf{y}_{te} = \mathrm{sign}(\mathbf{h}_{te}^\top\mathbf{w}+b)$.

4. Experiments

4.1. Experimental setup

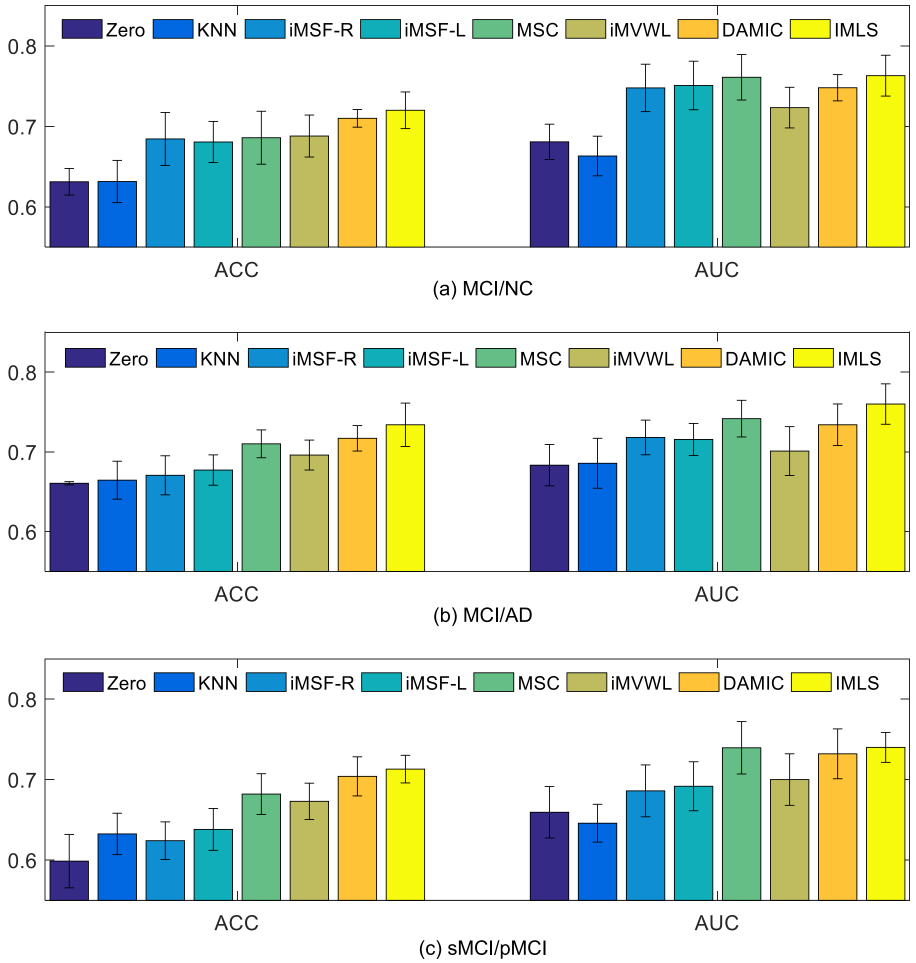
We evaluate the effectiveness of the proposed model by conducting the following three binary classification tasks: *i.e.*, MCI vs. NC, MCI vs. AD, sMCI vs. pMCI classifications. We use classification accuracy (ACC) and Area Under Curve (AUC) as performance metrics to compare our proposed method with the other comparison methods.

We perform 10-fold cross validation for all the methods under comparison, and report the means and standard deviations of the experimental results with ten repetitions. For parameter setting of our method, we determine the regularization parameter values (*i.e.*, $\{\lambda, \beta, \gamma\} \in \{10^{-6}, \ldots, 10^{3}\}$) and the dimension of the latent space (*i.e.*, $h \in \{10, 20, \ldots, 80\}$) via an inner cross-validation search on the training data, and searched the number of classifiers C in the range $\{10, 20, \ldots, 80\}$. We also use inner cross-validation to select hyper-parameter values for all the comparison methods. Besides, the soft margin parameter of SVM is determined via grid search in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{2}\}$.

4.2. Comparison results on complete multi-modality data

Comparison methods. In this subsection, we compare our proposed method (*i.e.*, CMLS) with other classification methods that use complete multi-modality data, as listed below.

- Baseline method. We include the result for the experiment by using the original features to train a SVM classifier, without performing any feature selection or reduction operation (denoted as “ORI”).
- Feature reduction (or selection) methods. Two methods are compared in this category, namely 1) Principal Component Analysis (PCA) (Jolliffe, 2002), and 2) ℓ_1-norm based feature selection method, which is denoted as “Lasso”. For PCA,



(a) MCI/NC

(b) MCI/AD

(c) sMCI/pMCI

Fig. 5. Comparison of classification results using two evaluation metrics (*i.e.*, ACC and AUC) for three classification tasks: (a) MCI/NC, (b) MCI/AD, and (c) sMCI/pMCI) on the incomplete multi-modal dataset.

we determine the optimal dimensionality of the data based on their respective eigenvalues computed by the generalized eigen-decomposition method according to Zhu et al. (2016a). For Lasso method, we optimize its sparsity parameter (λ) by cross-validating its value in the range of $\{10^{-4}, \ldots, 10^{2}\}$.

- Multi-modality fusion methods. Four multi-modality/view learning methods are compared in this category, *i.e.*, (1) CCA (Hardoon et al., 2004); (2) Multi-view Learning with Adaptive Neighbours (MLAN) method (Nie et al., 2018), which performs clustering/semi-supervised classification and local structure learning simultaneously; (3) Deep Matrix Factorization (DMF) method (Zhao et al., 2017), which conducts deep semi-nonnegative matrix factorization (NMF) to seek a common representation for multi-view clustering task; and 4) Multi-view Dimensionality Co-Reduction (MDCR) method (Zhang et al., 2017), which adopts kernel matching to regularize the dependencies across multiple views and projects each view into a low-dimensional space. For CCA method, we optimize its regularization parameter by cross-validating its value in the range of $\{10^{-4}, \ldots, 10^{2}\}$. For MLAN method, the parameter λ is determined in the range of $\{0, 0.02, \ldots, 1\}$. For DMF method, the parameters β and γ are searched in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{2}\}$. For MDCR method, the parameter λ is determined in the range of $\{0, 0.5, \ldots, 4\}$. Note that we use MLAN method to directly perform disease prediction, while resorting to SVM classifier to perform disease prediction for the other methods.
- Deep learning based feature representation method. In this category, we compare our CMLS model with Stacked Auto-Encoder (SAE) (Suk et al., 2015) method. In SAE, the main parameter is the number of hidden units. Following Suk et al. (2015), we use a three-layer network for multi-modality data by using a grid search from $[100, 300, 500] - [50, 100] - [10, 20, 30]$ (bottom-top).

Results. Fig. 3 shows the comparison results between our proposed method and all the comparison methods. From Fig. 3, it can be observed that our proposed method obtains better classification performance in terms of ACC and AUC than all the comparison methods. Specifically, comparing with the Lasso based feature selection method, which fuses multi-modality data without the consideration of correlations between MRI and PET data, our method performs significantly better. For CCA method, we can see that it obtains relatively better AUC performance. This is possibly because the multimodal fusion can better exploit the complementary information and correlation between different modalities, which led to the improved classification performance. Besides, though the SAE method uses high-level features learned from auto-encoder for classification, it has low performance, probably because it is an unsupervised feature learning method that does not consider label information. In addition, when compared with the three state-of-the-art multi-view learning methods, our method still obtains better performance. It is worth noting that the DMF and MDCR methods use two independent steps for dementia diagnosis, while our

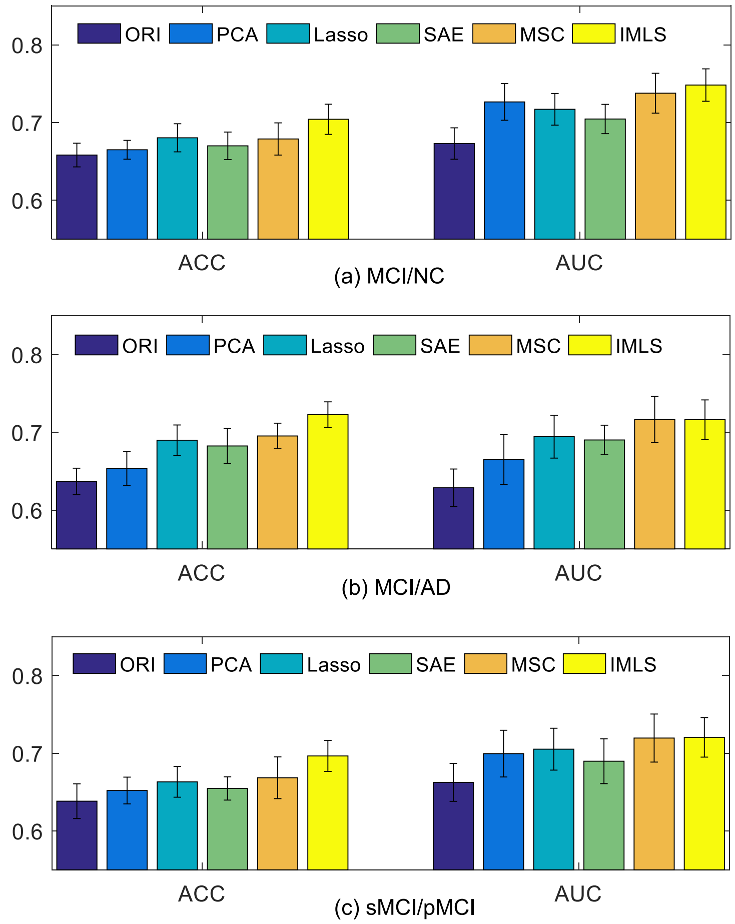


Fig. 6. Comparison of results for three classification tasks (*i.e.*, (a) MCI/NC, (b) MCI/AD, and (c) sMCI/pMCI) using MRI data on the incomplete dataset.

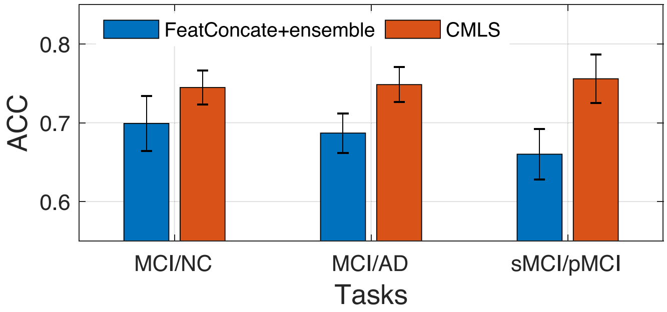


Fig. 7. Comparison between our CMLS model and its degraded counterpart âæFeatConcate+ensembleâğ without using common latent space learning.

method fuses them into a unified framework. Thus, the results also verify the effectiveness of our unified framework.

To investigate the benefit of multi-modality fusion, we also shows classification results that used only single modality. Fig. 4 shows the performance comparison of different methods using single modality (*i.e.*, MRI or PET). Note that, multi-modality fusion methods (*i.e.*, CCA, MLAN and MDCR) are excluded in this experiment. From Fig. 4, it can be observed that the degraded version of our method that uses only single modality (*i.e.*, MRI data or PET data) still outperforms other comparison methods. Besides, comparing Figs. 3 and 4, we can see that all the methods perform better when using multi-modality data, if compared with the case of using only single modality data.

4.3. Comparison results using incomplete multi-modality data

Comparison methods. In this subsection, we compare our proposed method (*i.e.*, IMLS) with some state-of-the-art methods that are applicable to incomplete multi-modality data, as listed below.

- Data imputation methods, including (1) Zero value imputation, and (2) *k*-Nearest Neighbor (KNN) (Hastie et al., 1999; Keller et al., 1985). Specifically, for Zero imputation method, the missing values in all samples are filled with zeros. Since all the features are z-normalized (*i.e.*, minus mean and divide by standard deviation) before the imputation process, thus this imputation method is equivalent to filling the missing feature values with the average observed feature values. For KNN imputation method, the missing values in all samples are filled with the weighted mean of the *k* nearest-neighbor samples. In addition, we search the parameter *k* in the range of {5, 10, 15, 20, 25}.
- iMSF method (Yuan et al., 2012). This method is a multi-view based method, which first partitions subjects into several views, and a specific classifier is constructed for each view. Then, a structural sparse learning model is employed to select a common set of features among these tasks. There are two versions of iMSF that use different loss functions, *i.e.*, the least square loss (denoted as “iMSF-R”) and the logistic loss (denoted as “iMSF-L”).
- Matrix Shrinkage and Completion method (MSC) (Thung et al., 2014). This method can handle classification problem with incomplete multi-modality data.

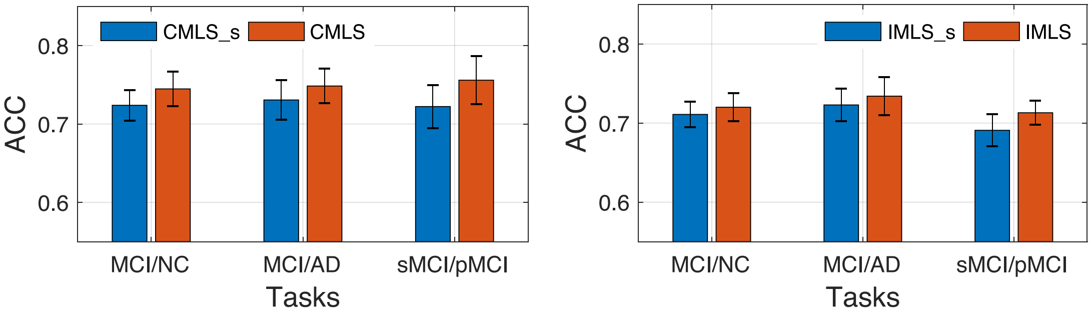


Fig. 8. Comparison between our proposed models (left: CMLS; right: IMLS) and their counterparts using only a single classifier (left: CMLS_s; right: IMLS_s).

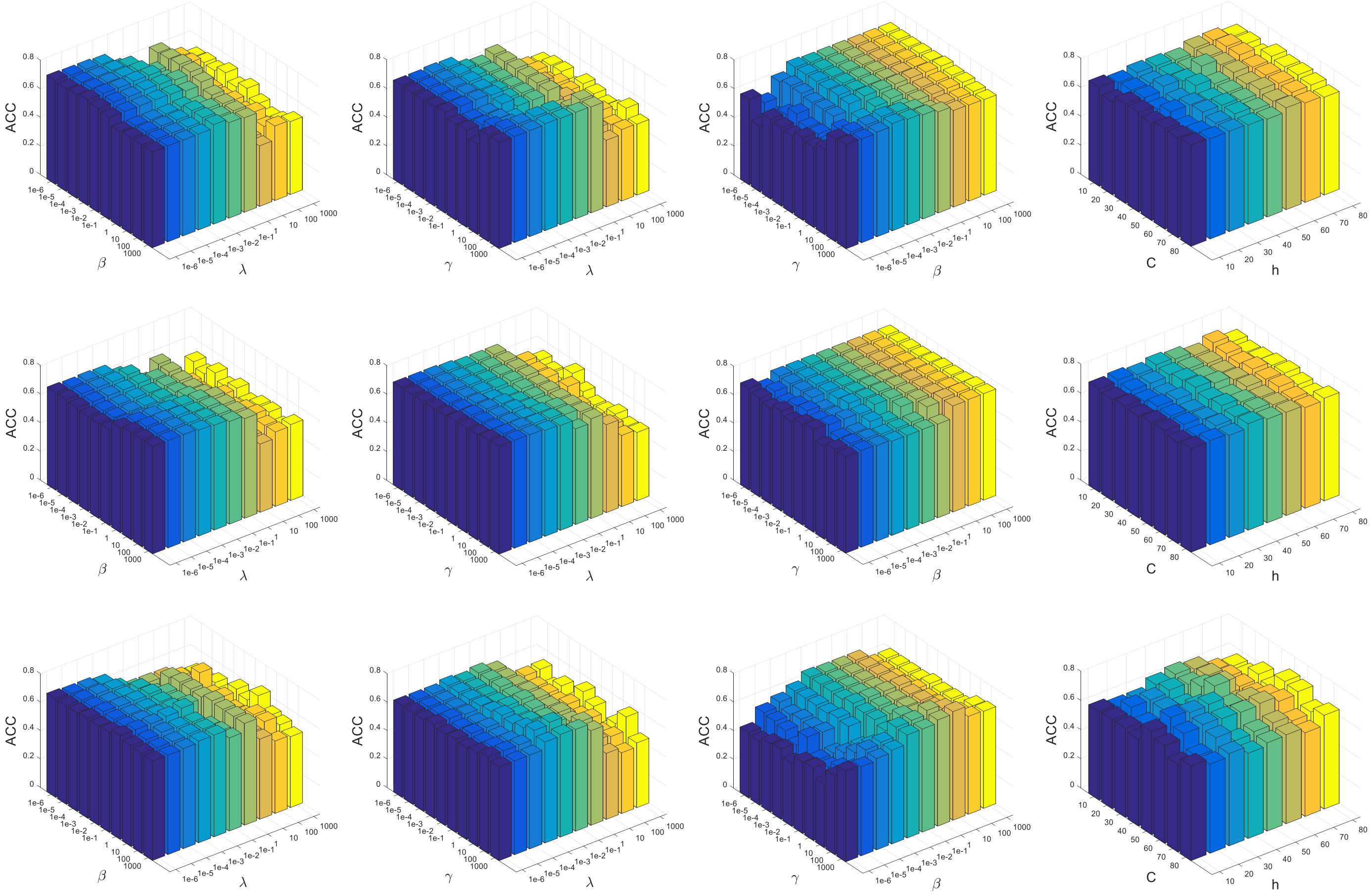


Fig. 9. Effect of the parameter changes for our proposed method (*i.e.*, CMLS) in response to three classification tasks (from top to bottom: MCI/NC, MCI/AD, and sMCI/pMCI) in terms of accuracy (ACC).

Specifically, MSC method first partitions the combined matrix (*e.g.*, features and targets) into sub-matrices, and each sub-matrix consists of samples with complete features from multi-modality data (*i.e.*, corresponding to a certain combination of modalities) and target outputs. Following that, a multi-task sparse learning framework is used to select informative features and samples. Subsequently, the shrunk combined matrix with missing features and unknown target outputs is imputed, which is realized via low-rank matrix completion algorithm by using a fixed-point continuation method (Ma et al., 2011).

- State-of-the-art incomplete multi-view learning methods, including (1) Doubly Aligned Incomplete Multi-view Clustering (DAIMC) method (Hu and Chen, 2018), and (2) Incomplete Multi-View Weak-label Learning (iMVWL) method (Tan et al., 2018). DAIMC is a clustering method that is designed for incomplete multi-modality data using weighted semi-nonnegative matrix factorization. To apply this method for classification task, we train a SVM classifier using the learned common latent features in DAIMC. On the other hand, iMVWL learns a shared subspace from incomplete views, local label correlations, and a predictor in this subspace, simultaneously.

Results. Fig. 5 shows the comparison results between our proposed method and all the comparison methods by using incomplete multi-modality data. From Fig. 5, we have the following observations: (1) Our proposed method performs better than all the comparison methods. (2) The MSC method obtains relatively better AUC performance in the three classification tasks. (3) The DAMIC method also obtains relatively better performance in terms of ACC

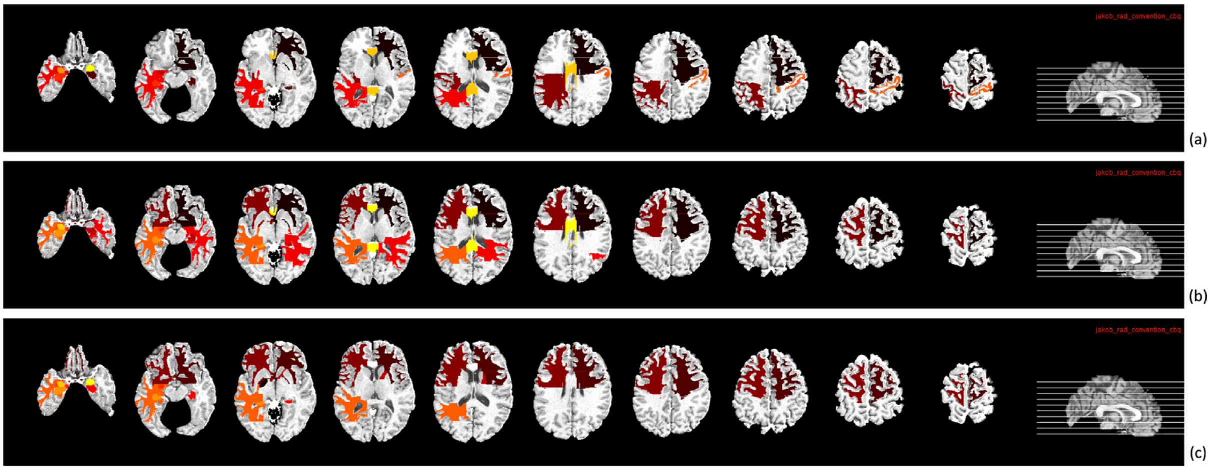


Fig. 10. Top selected regions from MRI data for three classification tasks: (a) MCI/NC, (b) MCI/AD, and (c) sMCI/pMCI.

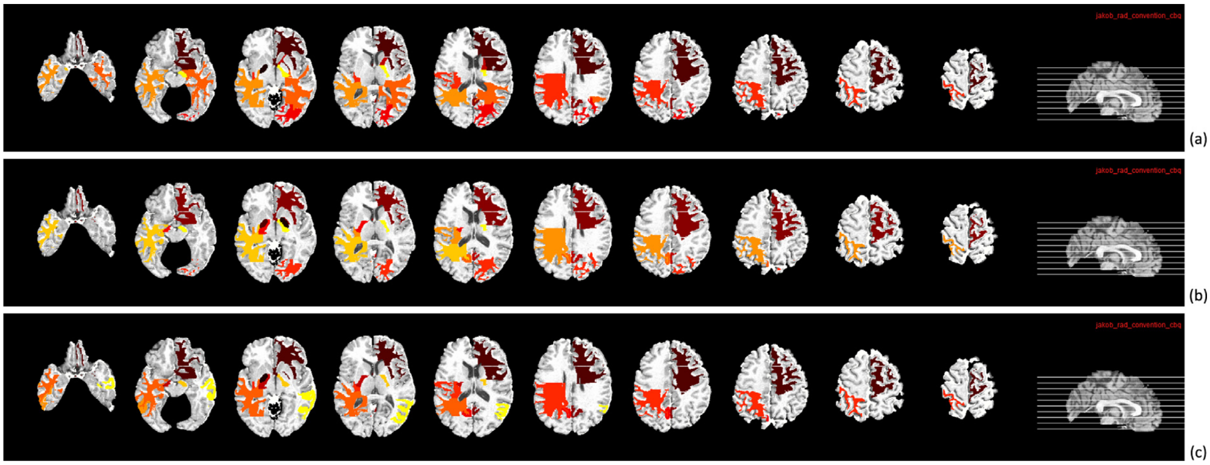


Fig. 11. Top selected regions from PET data for three classification tasks: (a) MCI/NC, (b) MCI/AD, and (c) sMCI/pMCI.

and AUC than other comparison methods. Overall, our method obtains some improvements over other existing methods and several state-of-the-art methods.

Further, to verify the benefit of multi-modality data fusion, Fig. 6 shows the classification results of different methods. Note that, in this comparison experiment, the subjects have complete MRI data but with incomplete PET data, thus, we compare the methods (*i.e.*, ORI, PCA, Lasso, SAE) using complete MRI data with the methods (*i.e.*, MSC and Ours) using complete MRI data and partial PET data. From Fig. 6, it can be observed that our proposed method performs better than other methods in term of ACC on three classification tasks, and also in term of AUC on MCI/NC and sMCI/pMCI task. Besides, it is also worth noting that the MSC obtains relative classification performance than the remaining methods. More importantly, from Figs. 5 to 6, we can see that all methods perform better by combing MRI and PET data than the case of only using MRI data. Thus, these results have verified the effectiveness of our proposed method by fusing incomplete multi-modalities.

4.4. Ablation study

To validate the effectiveness of the common latent space learning, we conduct an experiment by comparing our method with our degraded method that does not use latent space learning. Specifically, we denote our degraded method as FeatConcate+ensemble, which directly concatenates the original features from the two modalities (MRI and PET), and feed them into the learning framework of multiple diversified classifiers. The comparison results are shown in Fig. 7. From the results, it can be observed that our method with using latent space learning performs better than our model without using it.

To verify the effectiveness of ensemble of diversified classifiers, we compare the performance of our proposed methods for the cases of using single classifier (*i.e.*, $C = 1$) and multi-classifiers, where “CMLS_s” denotes our complete multi-modality latent space learning model using a single classifier and “IMLS_s” denotes our incomplete multi-modality latent space learning model using a single classifier. The comparison results are shown in Fig. 8. From the results, it can be seen that our proposed two models using the ensemble of diversified classifiers perform better than the case of using only a single classifier. Thus, these results validate the superiority of ensembling multiple diversified models for classification.

4.5. Parameter analysis

It is critical to select a set of robust parameters for our proposed models, so that our proposed methods can work well in most situations. Thus, in this section, we study the effects of different hyper-parameter (*i.e.*, λ, β, γ, C, and h) values to the performance metrics. First, for the three regularization parameters (*i.e.*, λ, β, and γ), we fix the value for one parameter and tune the other two parameters in the range of $\{10^{-6}, 10^{-5}, \dots, 10^{3}\}$. Then, we tune the parameters C and h in the range of $\{10, 20, \dots, 80\}$.

Fig. 9 shows the effect of the parameter changes by using our proposed CMLS for three classification tasks (from top to bottom: MCI/NC, MCI/AD, and sMCI/pMCI) in terms of classification accuracy (*i.e.*, ACC). Specifically, for MCI/NC classification task, when λ and β are within $[10^{-3}, 10^{-1}]$ and $\gamma \geq 1$, our proposed method has better classification performance. For MCI/AD classification task, when λ is within $[10^{-2}, 1]$, and $\beta \geq 0.1$ and $\gamma \geq 1$, our proposed method achieves better classification results. For sMCI/pMCI classification task, when λ is within $[10^{-2}, 1]$ and $\gamma \geq 1$, our proposed method achieves reasonably good performance. Besides, for the parameters C and h, our method obtains reasonably betetr classification results when $C \in [30, 60]$ and $h \in [40, 50]$.

4.6. Most related brain regions

Furthermore, we also identified the potential brain regions that can be used as biomarkers in AD and its early stage diagnosis. We computed the frequency of the ROIs and reported the ten most related ROIs in Figs. 10 and 11 for three classification tasks. Generally, these identified ROIs are in agreement with many previous studies about AD and its early stage diagnosis (Thung et al., 2014; Zhu et al., 2016a; Misra et al., 2009). Specifically, in MRI data, the top selected ROIs common to three classification tasks are the frontal lobe WM right, hippocampal formation right, uncus left, temporal lobe WM left, hippocampal formation left, and amygdala left. In PET data, the top selected ROIs common to three classification tasks are the globus palladus left, frontal lobe WM right, precuneus right, parietal lobe WM left, temporal lobe WM left, and precuneus left, which are the altered regions in AD reported in some previous studies (Jie et al., 2013; Zhu et al., 2016a; Hua et al., 2008; Chételat , 2005). Thus, in future work, these brain regions can be used as potential biomarkers for AD diagnosis.

5. Conclusion

In this paper, we propose a multi-modality latent space inducing ensemble SVM classifier for early AD diagnosis framework. Specifically, we first project the original ROIs-based features into a latent space to effectively exploit the correlations among modalities in multi-modality data. Then, by using the learnt latent representations, we learn multiple diversified classifiers and use an ensemble strategy to obtain the final result, so that our proposed model is more robust to disease heterogeneity. Furthermore, we extend our AD diagnosis framework to address the missing data issue which is common in multi-modality dataset. Experimental results using the ADNI dataset demonstrate that our proposed method outperforms other state-of-the-art methods in early AD diagnosis. In future work, we could extend the proposed model using a deep learning framework to improve the classification performance, since features learned from deep networks are typically more discriminative than the hand-crafted features (Fan et al., 2018; Bernard et al., 2018; Shen et al., 2017).

Declaration of Competing Interest

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work entitled Multimodal Latent Space Inducing Ensemble SVM Classifier for Early Dementia Diagnosis with Neuroimaging Data".

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We further confirm that any aspect of the work covered in this manuscript that has involved either experimental animals or human patients has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript.

We understand that the corresponding author (Prof. Dinggang Shen) is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from dgshen@med.unc.edu.

Acknowledgments

This research was supported in part by NIH grant (No. AG041721).

References

Alzheimer's, A., 2015. 2015 Alzheimer's disease facts and figures. Alzheimer's & Dement. 11 (3), 332. The journal of the Alzheimer's Association

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F.A.A., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11), 2514–2525.

Brown, G., Wyatt, J., Harris, R., Yao, X., 2005. Diversity creation methods: a survey and categorisation. Information Fusion 6 (1), 5–20.

Chaves, R., et al., 2009. SVM-Based computer-aided diagnosis of the Alzheimer's disease using *t*-test NMSE feature selection with feature correlation weighting. Neurosci. Lett. 461 (3), 293–297.

Chen, G., Wu, Y., Shen, D., Yap, P.-T., 2019. Noise reduction in diffusion MRI using non-local self-similar information in joint x- q space. Med. Image Anal. 53, 79–94.

Chételat, G., 2005. FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. Neurocase 11 (1), 14–25.

Fan, J., Cao, X., Xue, Z., Yap, P.-T., Shen, D., 2018. Adversarial similarity network for evaluating image alignment in deep learning based registration. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 739–746.

Fan, J., Cao, X., Yap, P.-T., Shen, D., 2019. Birnet: brain image registration using dual-supervised fully convolutional networks. Med. Image Anal. 54, 193–206.

Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C., 2008. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. NeuroImage 41 (2), 277–285.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139.

Guo, X., Wang, X., Ling, H., 2017. Exclusivity regularized machine: a new ensemble SVM classifier. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, pp. 1739–1745.

Hardoon, D.R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Comput. 16 (12), 2639–2664.

Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R., 2015. Matrix completion and low-rank SVD via fast alternating least squares. J. Mach. Learn. Res. 16 (1), 3367–3402.

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., Botstein, D., 1999. Imputing missing data for gene expression arrays. Technical Report, Division of Biostatistics. Stanford University.

Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frolich, L., Schonknecht, P., Ito, K., Mielke, R., Kalbe, E., 2002. Discrimination between alzheimer dementia and controls by automated analysis of multicenter FDG PET. NeuroImage 17 (1), 302–316.

Hinrichs, C., Singh, V., Xu, G., Johnson, S., 2009. MKL for Robust Multi-Modality AD Classification. Springer, pp. 786–794.

Hu, M., Chen, S., 2018. Doubly aligned incomplete multi-view clustering. In: Proceedings of the IJCAI, pp. 2262–2268.

Hua, X., Leow, A.D., Lee, S., Klunder, A.D., Toga, A.W., Lepore, N., Chou, Y.-Y., Brun, C., Chiang, M.-C., Barysheva, M., 2008. 3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry. NeuroImage 41 (1), 19–34.

Jack, C.R., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Jie, B., Zhang, D., Gao, W., Wang, Q., Wee, C.Y., Shen, D., 2013. Integration of network topological and connectivity properties for neuroimaging classification. IEEE Trans. Biomed. Eng. 61 (2), 576–589.

Jolliffe, I., 2002. Principal Component Analysis. Wiley Online Library.

Kabani, N.J., 1998. 3D anatomical atlas of the human brain. NeuroImage 7, P–0717.

Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy k-nearest neighbor algorithm. IEEE Trans. Syst. Man Cybern. (4) 580–585.

Lei, B., Yang, P., Wang, T., Chen, S., Ni, D., 2017. Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis. IEEE Trans. Cybern. 47 (4), 1102–1113.

Lian, C., Liu, M., Zhang, J., Shen, D., 2018. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans. Pattern Anal. Mach. Intell. in press.

Lin, Z., Liu, R., Su, Z., 2011. Linearized alternating direction method with adaptive penalty for low-rank representation. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 612–620.

Liu, F., Wee, C.-Y., Chen, H., Shen, D., 2014. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. NeuroImage 84, 466–475.

Liu, M., Zhang, D., Shen, D., 2016. Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans. Med. Imaging 35 (6), 1463–1474.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2018. Landmark-based deep multi-instance learning for brain disease diagnosis. Med. Image Anal. 43, 157–168.

Liu, M., Zhang, J., Yap, P.-T., Shen, D., 2017. View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. Med. Image Anal. 36, 123–134.

Long, Z., Huang, J., Li, B., Li, Z., Li, Z., Chen, H., Jing, B., 2018. A comparative atlas-based recognition of mild cognitive impairment with voxel-based morphometry. Front. Neurosci. 12, 916.

Long, Z., 2016. A support vector machine-based method to identify mild cognitive impairment with multi-level characteristics of magnetic resonance imaging. Neuroscience 331, 169–176.

Lu, D., Popuri, K., Ding, G., Balachandar, R., Beg, M.F., ADNI, 2018. Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. Med. Image Anal. 46, 26–34.

Ma, S., Goldfarb, D., Chen, L., 2011. Fixed point and Bregman iterative methods for matrix rank minimization. Math. Program. 128 (1), 321–353.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51 (2), 73–83.

Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. NeuroImage 44 (4), 1415–1422.

Nie, F., Cai, G., Li, J., Li, X., 2018. Auto-weighted multi-view learning for image clustering and semi-supervised classification. IEEE Trans. Image Process. 27 (3), 1501–1511.

Nie, F., Huang, H., Cai, X., Ding, C.H., 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1813–1821.

Nie, F., Zhu, W., Li, X., 2016. Unsupervised feature selection with structured graph optimization.. In: Proceedings of the Association for the Advancement of Artificial Intelligence, pp. 1302–1308.

Palmer, A.M., 2011. Neuroprotective therapeutics for Alzheimer's disease: progress and prospects. Trends Pharmacol. Sci. 32 (3), 141–147.

Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J. Clim. 14 (5), 853–871.

Shen, D., Davatzikos, C., 2002. HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21 (11), 1421–1439.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Suk, H., Lee, S., Shen, D., 2015. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220 (2), 841–859.

Tan, Q., Yu, G., Domeniconi, C., Wang, J., Zhang, Z., 2018. Incomplete multi-view weak-label learning.. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2703–2709.

Thung, K.-H., Wee, C.-Y., 2018. A brief review on multi-task learning. Multimed. Tools Appl. 77 (22), 29705–29725.

Thung, K.-H., Wee, C.-Y., Yap, P.-T., Shen, D., Initiative, A.D.N., et al., 2014. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. NeuroImage 91, 386–400.

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

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D., ADNI, 2014. Multiple instance learning for classification of dementia in brain MRI. Med. Image Anal. 18 (5), 808–818.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr, C.R., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39 (3), 1186–1197.

Wang, J., Wang, Q., Peng, J., Nie, D., Zhao, F., et al., 2017. Multi-task diagnosis for autism spectrum disorders using multi-modality features: a multi-center study. Hum. Brain Mapp. 38 (6), 3081–3097.

Wang, Y., Nie, J., Yap, P.-T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., ADNI, 2014. Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PloS One 9 (1), e77810.

Wee, C.Y., Yap, P.T., Zhang, D., Wang, L., Shen, D., 2014. Group-constrained sparse fMRI connectivity modeling for mild cognitive impairment identification. Brain Struct. Funct. 219 (2), 641–656.

Wolz, R., Aljabar, P., Hajnal, J.V., Lötjönen, J., Rueckert, D., ADNI, 2012. Nonlinear dimensionality reduction combining mr imaging with non-imaging information. Med. Image Anal. 16 (4), 819–830.

Wu, G., Qi, F., Shen, D., 2006. Learning-based deformable registration of MR brain images. IEEE Trans. Med. Imaging 25 (9), 1145–1157.

Xue, Z., Shen, D., Davatzikos, C., 2006. CLASSIC: consistent longitudinal alignment and segmentation for serial image computing. NeuroImage 30 (2), 388–399.

Yuan, L., Wang, Y., Thompson, P.M., et al., 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage 61 (3), 622–632.

Zhang, C., Fu, H., Hu, Q., Zhu, P., Cao, X., 2017. Flexible multi-view dimensionality co-reduction. IEEE Trans. Image Process. 26 (2), 648–659.

Zhang, D., Shen, D., ADNI, 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59 (2), 895–907.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57.

Zhang, Y., Nam, C.S., Zhou, G., Jin, J., Wang, X., Cichocki, A., 2018. Temporally constrained sparse group spatial patterns for motor imagery BCI. IEEE Trans. Cybern. 49 (9), 3322–3332.

Zhao, H., Ding, Z., Fu, Y., 2017. Multi-view clustering via deep matrix factorization. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.

Zhou, T., Liu, M., Thung, K.-H., Shen, D., 2019a. Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Trans. Med. Image 38, 2411–2422.

Zhou, T., Thung, K.-H., Liu, M., Shen, D., 2019b. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. IEEE Trans. Biomed. Eng. 66 (1), 165–175.

Zhou, T., Thung, K.-H., Liu, M., Shi, F., Zhang, C., Shen, D., 2018. Multi-modal neuroimaging data fusion via latent space learning for Alzheimerâs Disease diagnosis. In: Proceedings of the International Workshop on PRedictive Intelligence In MEdicine. Springer, pp. 76–84.

Zhou, T., Thung, K.-H., Zhu, X., Shen, D., 2019c. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum. Brain Mapp. 40 (3), 1001–1016.

Zhu, X., Suk, H., Lee, S.-W., Shen, D., 2016a. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. IEEE Trans. Biomed. Eng. 63 (3), 607–618.

Zhu, X., Suk, H.-I., Lee, S.-W., Shen, D., 2016b. Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. Brain Imaging Behav. 10 (3), 818–828.