

# Multi-Hypergraph Learning for Incomplete Multimodality Data

Mingxia Liu<sup>1</sup>, Member, IEEE, Yue Gao<sup>1</sup>, Senior Member, IEEE, Pew-Thian Yap, Senior Member, IEEE, and Dinggang Shen<sup>2</sup>, Senior Member, IEEE

**Abstract**—Multi-modality data convey complementary information that can be used to improve the accuracy of prediction models in disease diagnosis. However, effectively integrating multi-modality data remains a challenging problem, especially when the data are incomplete. For instance, more than half of the subjects in the Alzheimer’s disease neuroimaging initiative (ADNI) database have no fluorodeoxyglucose positron emission tomography and cerebrospinal fluid data. Currently, there are two commonly used strategies to handle the problem of incomplete data: 1) discard samples having missing features; and 2) impute those missing values via specific techniques. In the first case, a significant amount of useful information is lost and, in the second case, additional noise and artifacts might be introduced into the data. Also, previous studies generally focus on the pairwise relationships among subjects, without considering their underlying complex (e.g., high-order) relationships. To address these issues, in this paper, we propose a multi-hypergraph learning method for dealing with incomplete multimodality data. Specifically, we first construct multiple hypergraphs to represent the high-order relationships among subjects by dividing them into several groups according to the availability of their data modalities. A hypergraph regularized transductive learning method is then applied to these groups for automatic diagnosis of brain diseases. Extensive evaluation of the proposed method using all subjects in the baseline ADNI database indicates that our method achieves promising results in AD/MCI classification, compared with the state-of-the-art methods.

**Index Terms**—Alzheimer’s disease, classification, hypergraph, incomplete data, multi-modality.

## I. INTRODUCTION

Alzheimer’s disease (AD) is characterized by progressive impairment of neurons and the connections between neurons, leading to loss of cognitive/memory function and ultimately

Manuscript received January 14, 2017; revised June 23, 2017; accepted July 24, 2017. Date of publication July 25, 2017; date of current version June 29, 2018. This work was supported by the National Institutes of Health under Grants EB006733, EB008374, EB009634, MH100217, AG041721, AG042599, AG010129, and AG030514. (Corresponding author: Dinggang Shen.)

M. Liu, Y. Gao, and P.-T. Yap are with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: mingxia\_liu@med.unc.edu; kevin.gao@gmail.com; pt yap@med.unc.edu).

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

Digital Object Identifier 10.1109/JBHI.2017.2732287

death. This disease has been regarded as the major challenge to global health care systems [1], and it is estimated that the total prevalence of AD in the United States is about 13.8 million by 2050 [2]. Until now, much effort has been made to investigate computer-aided brain disease diagnosis approaches, aiming to prevent or postpone the onset of AD or its prodrome, i.e., mild cognitive impairment (MCI) [3].

Recent research and clinical studies have investigated extensive candidate biomarkers, and reported that fluorodeoxyglucose positron emission tomography (PET) measurements, structural magnetic resonance imaging (MRI), and cerebrospinal fluid (CSF) are among the best-established biomarkers for AD progression and pathology [3]. Specifically, feature representations (e.g., cortical thickness, connectivity information, and regional volumetric measures) extracted from MRI can be adopted to measure AD-related brain abnormalities [4]–[11]. With PET data, one can detect the abnormality in cerebral metabolic rate for glucose in the brain [12]–[16]. Besides, CSF total-tau (t-tau), CSF tau hyperphosphorylated at threonine 181 (p-tau) and the decrease of CSF amyloid  $\beta$  ( $A\beta$ ) are shown to be related to the cognitive decline in AD/MCI patients [17], [18]. Previous computer-aided disease diagnosis studies have reported that multi-modality data (e.g., MRI, PET, and CSF) can provide complementary information [19]–[22], which can be used for improving the diagnostic results for AD/MCI.

One common challenge in the multi-modality analysis is that there may be missing values in some modalities, which is called *incomplete data problem* [19], [20]. For instance, the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database, which consists of data for AD, MCI, and normal control (NC) subjects with multiple data modalities (e.g., MRI, PET, and CSF), is not complete for all modalities [23]. In the baseline ADNI database, all subjects have MRI data, and only about half subjects have PET and CSF data. Possible reasons for this may include poor data quality, high cost involved with PET scanning and patient dropouts. For instance, CSF data collection requires invasive tests (e.g., lumbar puncture), which might deter patient commitment [19].

In the domain of neuroimaging analysis, researchers have developed various approaches to deal with incomplete data [19], [20]. The first type of methods simply removes subjects with incomplete data. This will unfortunately result in discarding a large amount of the acquired data and may lead to significant reduction of sample size [24]. The second type of methods

aims to impute the missing values via specific data imputation techniques [25], [26]. Some well-known imputation techniques include expectation maximization (EM) [27], singular value decomposition (SVD) [28] and matrix completion [29]. Although these methods may help to process those randomly missing data, they often cannot handle missing data that are in a block-wise manner [19], where subjects have missing values in a whole modality (instead of missing some features in a specific modality). Also, imputation methods may introduce additional imputation artifacts, and thus, degrade the data quality [29]. Different from these two categories, several other methods have been recently developed to directly handle the incomplete data, without having to discard incomplete data or to impute missing data [19], [20]. For instance, Yuan *et al.* [19] proposed a multi-source feature learning method using all subjects in ADNI with at least one available data modality. Xiang *et al.* [20] developed a sparse feature learning method for handling incomplete multi-modality data without imputation. However, the main disadvantage of these methods is that they model only the pairwise relationships between subjects. Intuitively, the higher-order relationships among subjects can be used to improve the learning performance of computer-aided AD/MCI diagnosis.

In this paper, we propose a multi-hypergraph learning (MHL) approach based on incomplete multi-modality data for the automatic diagnosis of AD/MCI, where the high-order relationships among subjects can be modeled explicitly. To be specific, we first divide the dataset into several groups depending on whether the data for a certain modality is available. In this way, each group contains subjects with complete data from a particular combination of different modalities. Then, we construct a hypergraph to model the high-order relationships among subjects in each group. Next, we combine the hyperedges given by multi-hypergraphs associated with those groups and compute a hypergraph Laplacian matrix. Finally, a hypergraph-regularized transductive learning model is used for AD/MCI diagnosis.

The major contributions of this paper are summarized below. *First*, we propose a data grouping strategy for multi-modality data, and each group is corresponding to a particular combination of multiple modalities. *Second*, we develop a multi-hypergraph classification model, by explicitly modeling the inherent high-order relationships among subjects via hypergraphs. *Third*, we conduct extensive experiments on ADNI to empirically analyze our method, including four binary classification tasks (i.e., AD vs. NC, MCI vs. NC, AD vs. MCI, and pMCI vs. sMCI). The experimental results demonstrate the effectiveness of our proposed method.

The work in this paper is different from our previous study in [30]. *First*, while the focus of [30] is to exploit the coherence among different data groups (with each group containing data from a particular combination of different modalities), we focus on modeling the high-order relationships among subjects in this study. *Second*, the strategies for making use of the multi-modality data are different. In [30], we learn optimal weights for different groups from data via a view-aligned hypergraph classification model. In this work, we treat each group equally and propose to learn weights for hyperedges from data automatically. *Third*, the hypergraph construction methods are different.

In [30], we adopt sparse representation for constructing hypergraphs, where the similarity measure is based on the sparse coefficients. In this work, we use the star expansion strategy [31] to generate multiple hyperedges in each group, where Euclidean distance is used as the similarity measure.

The rest of the paper is organized as follows. We first present background information in Section II. In Section III, we describe data pre-processing method and our proposed multi-hypergraph learning approach. In Section IV, we describe experimental settings and report experimental results. In Section V, we compare our method with related approaches and investigate the influences of parameters. In Section VI, we conclude this work and discuss future research directions.

## II. BACKGROUND

### A. Hypergraph Based Transductive Learning

As a typical semi-supervised learning method, transductive learning incorporates not only labeled data but also unlabeled data for improving the performance of supervised learning methods [32], [33]. Generally, transductive learning assumes that the data lie in manifolds and clusters. Based on these two assumptions, many transductive learning methods have been proposed, such as transductive support vector machine [32], graph based learning [34], [35] and hypergraph learning [34], [36]–[38]. Among various transductive learning methods, hypergraph learning achieves promising performance in practice. Here, we denote scalars, vectors, and matrices using normal italic letters, boldface lowercase letters, and boldface uppercase letters, respectively. We further list important notations and definitions used in this paper in Table I.

A hypergraph is represented by  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{w})$ , where  $\mathcal{V}$  is a set of vertices (each representing a subject),  $\mathcal{E}$  is a set of hyperedges (each connecting two or more vertices) and  $\mathbf{w} = (w_i) \in \mathcal{R}^{N_e}$  is the weight vector for  $N_e$  hyperedges. Each hyperedge  $e_i$  ( $i = 1, 2, \dots, N_e$ ) is assigned a weight  $w_i$ . In hypergraphs, a hyperedge can connect more than two vertices, through which high-order relationships can be modeled explicitly [34], [36]. In comparison, each edge in a simple graph connects only two vertices. The incidence matrix  $\mathbf{H} = (h_{ij}) \in \mathcal{R}^{N \times N_e}$  of hypergraph  $\mathcal{G}$  encodes the relationships among vertices. The  $(i, j)$ -entry of the incidence matrix  $\mathbf{H}$  indicates whether vertex  $v_i$  is connected with other vertices in the hyperedge  $e_j$ , i.e.,

$$h_{ij} = \begin{cases} 1, & \text{if } v_i \in e_j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The vertex degree and the hyperedge degree are defined, respectively, as

$$d(v_i) = \sum_{e_j \in \mathcal{E}} w_j h_{ij} \quad (2)$$

and

$$\delta(e_j) = \sum_{v_i \in \mathcal{V}} h_{ij} \quad (3)$$

We let  $\mathbf{D}_v$  and  $\mathbf{D}_e$  be diagonal matrices containing the vertex degrees and the hyperedge degrees, respectively.

TABLE I  
NOTATIONS

Notation	Definition
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{w})$	$\mathcal{G}$ denotes a hypergraph, while $\mathcal{V}$ , $\mathcal{E}$ and $\mathbf{w}$ represent the set of vertices, the set of hyperedges and the weights of hyperedges, respectively.
$N$	The number of vertices in the hypergraph $\mathcal{G}$ , i.e., $N =  \mathcal{V} $ .
$N_e$	The total number of hyperedges, i.e., $N_e =  \mathcal{E} $ .
$N_e^g$	The number of hyperedges in the $g$ -th group.
$\mathbf{W}$	The diagonal matrix of the hyperedge weights, $\mathbf{W} \in \mathcal{R}^{N_e \times N_e}$ .
$d(v_i)$	The degree of the vertex $v_i$ .
$\delta(e_j)$	The degree of the hyperedge $e_j$ .
$\mathbf{D}_v$	The diagonal matrix of the vertex degrees, $\mathbf{D}_v \in \mathcal{R}^{N \times N}$ .
$\mathbf{D}_e$	The diagonal matrix of the hyperedge degrees, $\mathbf{D}_e \in \mathcal{R}^{N_e \times N_e}$ .
$\mathbf{H}^g$	The incidence matrix for the hypergraph in the $g$ -th group, and $\mathbf{H}^g \in \mathcal{R}^{N \times N_e}$ .
$\mathbf{y}$	The $N$ -dimensional label vector of $N$ samples, with element $y_i = 1$ if the $i$ -th sample belongs to the positive class, $y_i = -1$ if the $i$ -th sample belongs to the negative class, and otherwise $y_i = 0$ .
$\mathbf{f}$	The relevance score vector for $N$ samples, $\mathbf{f} \in \mathcal{R}^N$ .

Denote  $\mathbf{W} = (W_{ij}) \in \mathcal{R}^{N_e \times N_e}$  as the diagonal matrix of hyperedge weights, with the diagonal element  $W_{ii} = 1$ . Previous approaches for computing the hypergraph Laplacian can be categorized into two classes [34]. The first class aims to construct a simple graph from the initial hypergraph, e.g., star expansion [31], clique expansion [31] and Rodriquez's Laplacian [39]. In the second class, we define a hypergraph Laplacian based on analogs from simple graph Laplacian, e.g., normalized Laplacian [34] and Bolla's Laplacian [40]. Recently, Benson *et al.* [38] proposed a hypergraph construction method to capture the high-level information inside a graph. In this paper, we adopt the method proposed in [34] to compute the hypergraph Laplacian. Letting  $\Theta = \mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^T \mathbf{D}_v^{-1/2}$ , the hypergraph Laplacian is defined as  $\mathbf{L} = \mathbf{I} - \Theta$ . We define the hypergraph Laplacian regularizer  $\Omega(\mathbf{f})$  as in [34]

$$\begin{aligned} \Omega(\mathbf{f}) &= \mathbf{f}^T \mathbf{L} \mathbf{f} \\ &= \frac{1}{2} \sum_{e_i \in \mathcal{E}} \sum_{v_j, v_k \in \mathcal{V}} \frac{w_i h_{ji} h_{ki}}{\delta(e_i)} \left( \frac{f_j}{\sqrt{d(v_j)}} - \frac{f_k}{\sqrt{d(v_k)}} \right)^2 \end{aligned} \quad (4)$$

## B. ADNI Database

This study is based on the incomplete multi-modality data from the baseline ADNI database (ADNI-1) [23]. In ADNI, subjects are divided into three categories (i.e., AD, MCI, and NC) based on particular criteria such as Mini-Mental State Examination (MMSE) scores. To be specific, general inclusion or exclusion criteria used in ADNI are listed in the following: 1) mild AD: MMSE scores between 20–26 (inclusive), CDR of 0.5 or 1.0 and meet NINCDS/ADRDA criteria for probable AD; 2) MCI: MMSE scores between 24–30 (inclusive), a memory complaint, have objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living and an absence of dementia; and 3) NC: Mini-Mental State Examination (MMSE) scores between 24–30 (inclusive),

TABLE II

DEMOGRAPHIC AND CLINICAL INFORMATION OF SUBJECTS IN THE BASELINE ADNI DATABASE

	AD	MCI	NC
Male/Female	99/87	254/141	118/108
Age (Mean $\pm$ SD)	75.40 $\pm$ 7.60	74.90 $\pm$ 7.30	76.00 $\pm$ 5.00
Edu. (years) (Mean $\pm$ SD)	14.70 $\pm$ 3.10	15.70 $\pm$ 3.00	16.00 $\pm$ 2.90
MMSE (Mean $\pm$ SD)	23.30 $\pm$ 2.00	27.00 $\pm$ 1.80	29.10 $\pm$ 1.00
CDR (Mean $\pm$ SD)	0.75 $\pm$ 0.25	0.50 $\pm$ 0.03	0.00 $\pm$ 0.00

Note: Values reported as Mean  $\pm$  standard Deviation (SD); MMSE: mini-mental state examination; CDR: Clinical Dementia Rating.

a Clinical Dementia Rating (CDR) of 0, non-depressed, non MCI and non-demented.

Besides, many MCI subjects could convert to AD within several months after the baseline time, while the others may keep stable over time. Accordingly, we further categorized those MCI subjects into two classes, i.e., progressive MCI (pMCI) and stable MCI (sMCI). Specifically, if the diagnosis for a particular subject was MCI at baseline but this subject converted to AD after baseline within 24 months, this subjects is defined as pMCI. In contrast, an sMCI subject was diagnosed as MCI at all available time points (0–96 months). The detailed description for each category is given online (<http://adni.loni.usc.edu>).

## III. METHOD

### A. Data Pre-Processing

There are 807 subjects in the baseline ADNI database, including 186 AD subjects, 395 MCI subjects (169 pMCI and 226 sMCI), and 226 NCs. All 807 subjects in ADNI have T1-weighted structural MRI data, while only 396 subjects have PET data and 406 subjects have CSF data. That is, in the baseline ADNI database, the missing data are in a block-wise manner, i.e., subjects have missing values in a whole modality, instead of missing some features in a specific modality. In Table II, we report demographic and clinical information of all subjects used in this study.

We process MR and PET images and extract the region of interest (ROI) features from these two data modalities. Specifically, using MIPAV (<http://mipav.cit.nih.gov/index.php>), we first apply anterior commissure (AC)-posterior commissure (PC) correction for MR images. We then re-sample those images to have the resolution of  $256 \times 256 \times 256$  and correct intensity inhomogeneity via N3 algorithm [41]. In the next step, we perform skull stripping, and remove the cerebellum by warping a labeled template to each skull-stripped image. Then, FAST [42] in the FSL software package (<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki>) is adopted to segment the human brain into three tissue types, i.e., gray matter, white matter, and cerebrospinal fluid (CSF). After registration via HAMMER [43], we obtain a subject-labeled image, by using a template with 93 ROIs [44]. For each subject, we finally extract the volumes of GM tissue in 93 ROIs as feature representations, normalized by the total intracranial volume. Here, the total intracranial volume is estimated by the summation of GM, WM and CSF volumes from 93 ROIs. On the other hand, for those PET images, we first align each PET image onto its corresponding MR image via affine transformation. Then, the average intensity of each ROI in the PET image is computed as the feature representation for that subject. Also, we use five CSF biomarkers for representing subjects, i.e., amyloid  $\beta$  ( $A\beta_{42}$ ), CSF total tau (t-tau), CSF tau hyperphosphorylated at threonine 181 (p-tau) and two tau ratios with respect to  $A\beta_{42}$  (i.e., t-tau/ $A\beta_{42}$  and p-tau/ $A\beta_{42}$ ). In this way, we now have a 191-dimensional feature vector for each subject having complete multi-modality data, including 93 MRI features, 93 PET features, and 5 CSF features.

### B. Multi-Hypergraph Construction With Incomplete Multi-Modality Data

Although several methods have been proposed to handle the incomplete data directly without having to discard incomplete data or to impute missing data [19], [20], these methods can only model pairwise relationships between subjects. That is, the high-order relationships among subjects are not captured to further improve the performance of disease diagnosis. For addressing this issue, we propose a multi-hypergraph learning (MHL) method to handle block-wise incomplete multi-modality data, by explicitly incorporating the high-order relationships among subjects into the learning process. As shown in Fig. 1, we first divide the whole dataset into several groups according to the availability of the data associated with a particular combination of modalities. For each group, we construct a hypergraph to model the complex relationships among subjects. We then compute the hypergraph Laplacian matrix based on the hypergraphs associated with different groups. Finally, we perform hypergraph based transductive classification to predict class labels for new testing subjects.

We illustrate the multi-hypergraph construction approach in Fig. 2 by using three modalities (i.e., MRI, PET, and CSF) as an example. In Fig. 2, each column in the data matrix represents a subject with three data modalities, while each row denotes a feature vector among all subjects in a specific modality (some entries may have missing values). We first partition

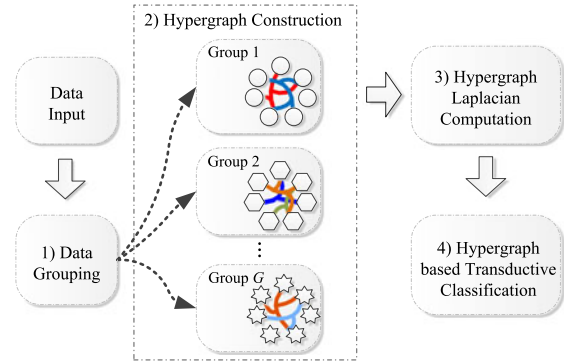
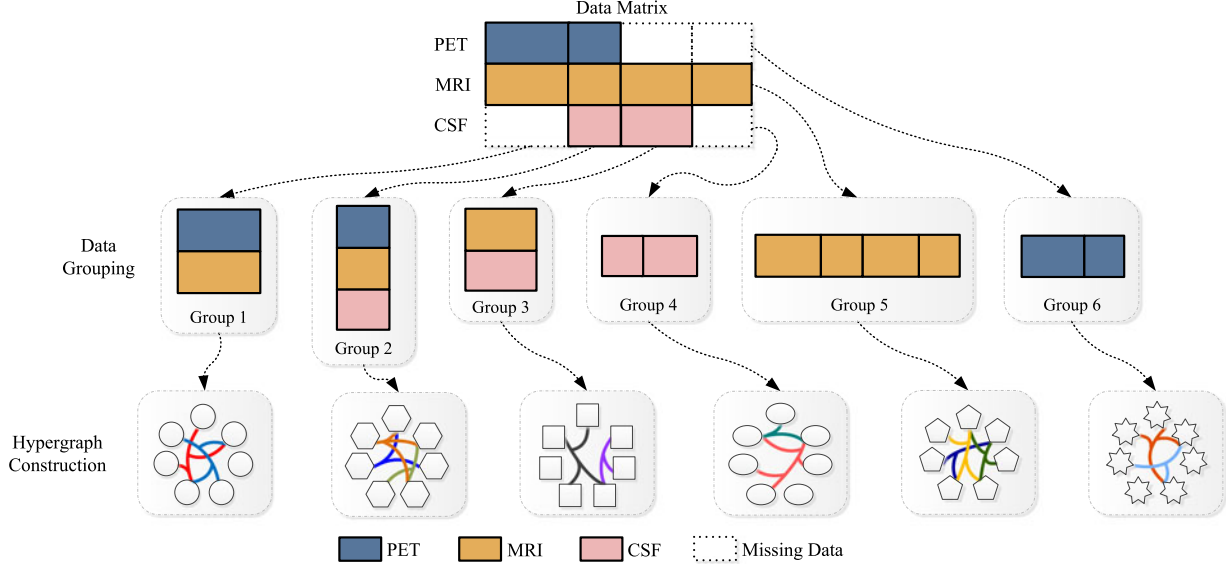


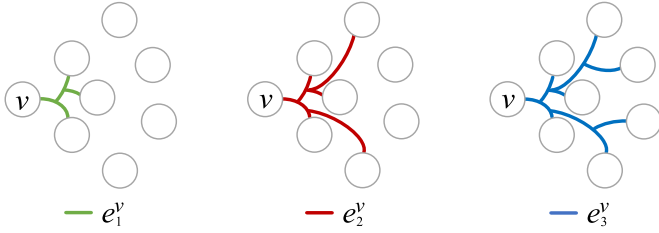
Fig. 1. The proposed multi-hypergraph learning method for incomplete multi-modality data. The main steps include 1) data grouping, 2) hypergraph construction, 3) hypergraph Laplacian computation, and 4) hypergraph based transductive classification.

the whole dataset into six groups, including three *inter-modality* groups (i.e., Group 1, Group 2, and Group 3) and three *single-modality* groups (i.e., Group 4, Group 5, and Group 6). In this study, we denote subjects in *inter-modality* groups as those having features from two or more modalities, while subjects in *single-modality* groups as those having only data from a single modality. As shown in Fig. 2, subjects in Group 1 have PET and MRI features, subjects in Group 2 have features from three modalities (i.e., PET, MRI, and CSF), subjects in Group 3 have MRI and CSF features, while subjects in Groups 4–6 have only CSF, MRI, and PET features, respectively. In each group, we have complete feature representations for each subject. We believe that such data grouping method has at least two advantages. First, we can make full use of subjects with data from at least one modality via such grouping, which is particularly important for dealing with the problem of incomplete multi-modality data. In this way, more available information can be used in the learning process. Second, such data grouping method can largely broaden the feature space by concatenating features from different modalities, through which one can get much richer feature representation. Note that such data grouping method is a general means for problems with complete multi-modality or multi-view data to expand the feature space.

Motivated by [31], [45], we construct the hyperedges via the star expansion strategy, by varying the neighborhood size for each vertex (with a vertex representing a specific subject), where the Euclidean distance is used for computing the neighbors. In Fig. 3, we give an illustration of our hyperedge construction strategy. As shown in Fig. 3, with the vertex  $v$  as the center, we vary the neighborhood size  $s$  (i.e.,  $s = 3, 5, 7$ ) to generate three hyperedges (i.e.,  $e_1^v$ ,  $e_2^v$  and  $e_3^v$ ), with each hyperedge connecting the vertex  $v$  and its  $s$ -nearest neighbors. Since different neighborhood sizes reflect different scales of data structure, the hyperedges constructed by using different neighborhood sizes reflect both local and semi-local structure information of the original data. Finally, we combine the hyperedges generated from multiple groups to compute a hypergraph Laplacian matrix [34], which will be used in a hypergraph based transductive classification model.



**Fig. 2.** Illustration of the proposed group-based multi-hypergraph construction method. First, subjects are divided into  $G$  ( $G = 6$  in this study) groups according to the data availability of a certain combination of modalities. As MRI data are complete in ADNI, there are in total six combinations of modalities, giving six groups, each containing subjects with complete data. Note that we use different shapes to denote different feature representations since they may be different across groups (e.g., circles for Group 1 and squares for Group 3). A hypergraph is constructed for each group, giving in total  $G$  hypergraphs.



**Fig. 3.** Hyperedge construction using different neighborhood sizes. Regarding vertex  $v$  as the center, we vary the neighborhood size  $s$  (i.e.,  $s = 3, 5, 7$ ), and generate three hyperedges (i.e.,  $e_1^v$ ,  $e_2^v$  and  $e_3^v$ ), each connecting the vertex  $v$  and its  $s$ -nearest neighbors.

### C. Multi-Hypergraph Based Classification

Since hypergraphs of the different groups share the same vertices, representing all the subjects, we combine the hypergraphs via the incidence matrix  $\mathbf{H} = [\mathbf{H}^1, \mathbf{H}^2, \dots, \mathbf{H}^G]$ , where  $\mathbf{H}^g \in \mathcal{R}^{N \times N_e^g}$  ( $g = 1, 2, \dots, G$ ) is the incidence matrix corresponding to the  $g$ -th group. It is worth noting that, if a vertex represented by the  $g$ -th group of feature representation has missing values, no hyperedge in  $\mathbf{H}^g$  covers this vertex. Using the hypergraphs constructed for the groups, a binary classification problem can be formulated as estimating the relevance score vector  $\mathbf{f} = (f_i) \in \mathcal{R}^N$  for the  $N$  samples, which is as follows:

$$\arg \min_{\mathbf{f}} \{ \Omega(\mathbf{f}) + \lambda \mathcal{R}_{\text{emp}}(\mathbf{f}) \} \quad (5)$$

where the first term  $\Omega(\mathbf{f}) = \mathbf{f}^T \mathbf{L} \mathbf{f}$  is a hypergraph regularization term defined in (4), where  $\mathbf{L} = \mathbf{I} - \mathbf{D}_v^{-1/2} \tilde{\mathbf{H}} \mathbf{W} \mathbf{D}_e^{-1} \tilde{\mathbf{H}}^T \mathbf{D}_v^{-1/2}$ . The second term is the empirical loss computed based on the training data, and  $\lambda$  is a tuning parameter for trade-off between empirical loss and regularization. Solving problem in (5) allows knowledge learned from the training data to be transferred to

the testing data [32]. In this work, we adopt the square loss function

$$\mathcal{R}_{\text{emp}}(\mathbf{f}) = \|\mathbf{f} - \mathbf{y}\|^2 = \sum_{v_i \in \mathcal{V}} (f_i - y_i)^2 \quad (6)$$

where  $\mathbf{y} = (y_i) \in \mathcal{R}^N$  is the label vector for the  $N$  samples. For the  $i$ -th sample, we let  $y_i = 1$  if it belongs to the positive class,  $y_i = -1$  if it belongs to the negative class, and  $y_i = 0$  if its class is unknown.

Since different hyperedges may play different roles in modeling the structure information of data, there may be different weights for those hyperedges. For jointly learn the hyperedge weights and the estimated class labels automatically from data, we reformulate the problem in (5) as follows:

$$\begin{aligned} \arg \min_{\mathbf{f}, \mathbf{W}} \quad & \mathbf{f}^T \mathbf{L} \mathbf{f} + \lambda \|\mathbf{f} - \mathbf{y}\|^2 + \mu \|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \sum_{i=1}^{N_e} W_{ii} = 1, \quad \forall W_{ii} \geq 0 \end{aligned} \quad (7)$$

where the last term in (7) and those constraints are used to penalize the complexity of weighting values for hyperedges.

The objective function in (7) is not jointly convex with respect to  $\mathbf{f}$  and  $\mathbf{W}$  [46]. Fortunately, it is convex with respect to  $\mathbf{f}$  when  $\mathbf{W}$  is fixed, and also convex with respect to  $\mathbf{W}$  give a fixed  $\mathbf{f}$ . In this study, we adopt an alternating optimization algorithm to solve the problem in (7). That is, we first fix  $\mathbf{W}$  and optimize  $\mathbf{f}$ , and the objective function in (7) can be reformulated as minimizing the following

$$\mathbf{Q} = \mathbf{f}^T \mathbf{L} \mathbf{f} + \lambda \|\mathbf{f} - \mathbf{y}\|^2 \quad (8)$$

By differentiating  $\mathbf{Q}$  with respect to  $\mathbf{f}$ , we obtain

$$\mathbf{f} = (\mathbf{L} + \lambda \mathbf{I})^{-1} \mathbf{y} \quad (9)$$

In the second step, we optimize  $\mathbf{W}$  with the learned  $\mathbf{f}$  in the first step. Denote  $\mathbf{\Lambda} = \mathbf{f}^T \mathbf{D}_v^{-1/2} \tilde{\mathbf{H}}$ , and the objective function in (7) can be rewritten as follows

$$\begin{aligned} \arg \min_{\mathbf{W}} \quad & \mathbf{f}^T \mathbf{L} \mathbf{f} + \mu \|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \sum_{i=1}^{N_e} W_{ii} = 1, \quad \forall W_{ii} \geq 0 \end{aligned} \quad (10)$$

Then, the partial derivative of (10) with respect to  $\mathbf{W}$  is as follows

$$\frac{\partial}{\partial \mathbf{W}} \left\{ \mathbf{f}^T \mathbf{L} \mathbf{f} + \mu \|\mathbf{W}\|_F^2 + \eta \left( \sum_{i=1}^{N_e} W_{ii} - 1 \right) \right\} = 0 \quad (11)$$

$$\begin{aligned} \Rightarrow \mathbf{W} &= \frac{\mathbf{\Lambda}^T \mathbf{\Lambda} \mathbf{D}_e^{-1} - \eta \mathbf{I}}{2\mu} \\ \eta &= \frac{\mathbf{\Lambda} \mathbf{D}_e^{-1} \mathbf{\Lambda}^T - 2\mu}{N_e} \end{aligned} \quad (12)$$

The above-mentioned two steps are performed iteratively until convergence. In the experiments, the iteration number is fixed as 20 empirically.

## IV. EXPERIMENTS

### A. Methods for Comparison

In the experiments, we first compare our MHL method with four baseline imputation-based approaches. These baseline methods utilize various data imputation techniques to impute missing features, which are summarized in the following:

- 1) *Zero*. In this method, those missing values are simply filled with zeros. If we first normalize the original data to have unit standard deviation and zero mean, this method is equal to the mean value imputation method. More specifically, those missing values are filled with the means of values that are available in the same row.
- 2) *k-Nearest Neighbor (KNN)* [47], [48]. In KNN method, we simply fill the missing value with weighted mean of its  $k$ -nearest neighbor rows. Specifically, we first identify the feature rows that are most similar to the one with the missing value via KNN algorithm, and then fill that missing values with the weighted mean of values in neighboring rows. Similar to the study in [49], the weight for a specific neighboring row is inversely proportional to the Euclidean distance between this neighboring row and the row with missing values.
- 3) *Expectation Maximization (EM)* [27]. This method imputes the missing values using the expectation maximization algorithm. To be specific, in the  $E$  step, we first estimate the mean and the covariance matrix from the feature matrix, and then fill those missing values with estimates from the previous  $M$  step (or initialized as zeros). In the  $M$  step, based on the available values, estimated mean and covariance, we fill those missing elements with

conditional expectation values. Then, the mean and the covariance will be re-estimated based on the filled feature matrix. In EM method, the above mentioned two steps will be repeated until convergence.

- 4) *Singular Value Decomposition (SVD)* [28]. The missing values are iteratively filled-in based on matrix completion with the low-rank approximation. Specifically, we first fill those missing values with initial guesses (e.g., zeros), followed by a singular value decomposition (SVD) process to generate a low-rank approximation of a filled-in matrix. Based on the low-rank estimation matrix, we will update the missing elements with their corresponding values in this matrix. Similarly, we will perform SVD again to obtain a new updated matrix, and repeat such process until convergence.

We further compare our method with state-of-the-art methods for dealing with incomplete multi-modality data in the field of neuroimaging analysis. These methods include: 1) incomplete Multi-Source Feature (iMSF) learning [19], 2) Ingahlalikar's Algorithm [50], 3) incomplete Source-Feature Selection (iSFS) [20] method, and 4) a matrix shrunk and completion (MSC) method [49]. In the following, we briefly introduce these methods.

- 1) *Incomplete Multi-Source Feature (iMSF) Learning* [19]. Using the similar data grouping technique as in our method, iMSF regards the classification problem with incomplete multi-modality data as a sparse multi-task learning problem, without discarding or imputing incomplete data. As shown in [19], iMSF is effective in finding informative features from incomplete multi-modality data. Two versions of iMSF are available based on two particular loss functions (i.e., least squares loss and logistic loss), including least-squares (iMSF-Least) and logistic loss (iMSF-Logistic).
- 2) *Ingahlalikar's Algorithm* [50]. In this method, an ensemble classification technique is used to fuse outputs of multiple classifiers, where these classifiers are built based on different subsets of subjects with complete feature representations. Specifically, this method first groups the original incomplete multi-modality data into multiple subsets, and then adopts the signal-to-noise ratio coefficient filter algorithm to perform feature selection. Using those selected features, it constructs a linear discriminant analysis (LDA) [51] classifier in each individual subset. Finally, the classification results from all subjects are fused by a particular ensemble strategy for making a final decision for a new testing subject. Two versions of this method are available based on different ensemble strategies. The first one, denoted as Ingahlalikar-Weighted, adopts the weighted averaging strategy, where each classifier is assigned a particular weight based on its classification error on the training data. The second one, denoted as Ingahlalikar-Average, is based on the averaging strategy, where all classifiers are assigned equal weights.
- 3) *Incomplete Source-Feature Selection (iSFS) method* [20]. In iSFS, subjects with incomplete multi-modality data are first partitioned into several groups according to the

availability of data modalities. Then, a feature learning model is developed to find the most informative features from different data groups. That is, all subjects can be used in this method, without any discarding or imputation operation.

- 4) *Matrix Shrinkage and Completion (MSC) method* [49]. In MSC, the input features and the output label vector are first combined into an incomplete matrix. Then, this incomplete matrix is further partitioned into several sub-matrices, with each one containing subjects with complete feature representation (*w.r.t.* certain combinations of multi-modalities). A multi-task learning model is adopted to select both discriminative features and informative samples in each individual sub-matrix, leading to a shrunk matrix. Based on an EM imputation method [27], MSC finally completes those missing feature values and unknown target outputs of the shrunk matrix.

In this study, the proposed MHL method, Ingahlalikar’s algorithm, and MSC can directly perform classification tasks based on incomplete multi-modality data. The other methods need to either impute the missing data (e.g., Zero, KNN, EM, and SVD) or select a subset of features (e.g., iMSF, and iSFS). Motivated by [19], [20], in the experiments, we adopt support vector machine (SVM) for classification [52] after data imputation using Zero/KNN/EM/SVD and feature selection using iMSF/iSFS. A linear SVM is used in the experiments, since the max-margin classification nature of the linear SVM results in good generalizability [20].

## B. Experimental Settings

We adopt a 10-fold cross-validation strategy [24] for performance evaluation. The subjects are partitioned into 10 subsets, and each subset has roughly equal number of subjects. Each time one subset is used as the testing set, while the remaining 9 subsets are adopted as the training set. In order to avoid any bias introduced by random partitioning of the data, such process is repeated 10 times, and we record the average classification results. To optimize parameters for different methods, we further perform an inner 10-fold cross-validation using each training set. Specifically, we further partition each training set into 10 subsets for cross-validation parameter selection [20]. Similar to [19], the neighborhood size  $k$  for KNN is selected from  $\{3, 5, 7, 9, 11, 15, 20\}$ . For the SVD-based imputation method, the rank parameter is chosen from  $\{5, 10, 15, 20, 25, 30\}$ . The regularization parameter  $\lambda$  for iMSF is chosen from  $\{10^{-5}, 10^{-4}, \dots, 10^1\}$ . The parameters  $\lambda$  and  $\mu$  for MHL and  $C$  for SVM are chosen from  $\{10^{-3}, 10^{-2}, \dots, 10^4\}$ . The neighborhood sizes of  $\{3, 5, 7, 9, 11, 15, 20\}$  are used to construct multiple hyperedges in MHL. The influence of the neighborhood size is discussed in Section V-D.

Seven metrics are used for performance evaluation: classification accuracy (ACC), sensitivity (SEN), specificity (SPE), balanced accuracy (BAC), positive predictive value (PPV), negative predictive value (NPV) and the area under the receiver operating characteristic curve (AUC) [53]. Let TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and

**TABLE III**  
COMPARISON WITH BASELINE IMPUTATION-BASED METHODS: AD VS. NC CLASSIFICATION (%)

	Zero	KNN	EM	SVD	MHL (Ours)
ACC	84.48	85.20	84.23	86.37	<b>90.29</b>
SEN	77.96	70.98	67.76	80.78	<b>84.40</b>
SPE	89.84	96.91	<b>97.78</b>	90.87	95.13
BAC	83.90	83.94	82.77	85.82	<b>89.77</b>
PPV	87.01	94.91	<b>96.14</b>	87.78	93.45
NPV	83.36	80.31	78.82	85.64	<b>88.11</b>
AUC	83.90	83.94	82.77	85.82	<b>89.77</b>
<i>p</i> -value	0.0017*	0.0014*	0.0038*	0.0037*	–

**TABLE IV**  
COMPARISON WITH BASELINE IMPUTATION-BASED METHODS: MCI VS. NC CLASSIFICATION (%)

	Zero	KNN	EM	SVD	MHL(ours)
ACC	67.74	67.25	68.06	71.32	<b>74.35</b>
SEN	81.20	79.12	80.96	83.46	<b>86.25</b>
SPE	44.49	46.69	45.93	50.16	<b>53.74</b>
BAC	62.84	62.91	63.44	66.80	<b>70.01</b>
PPV	71.71	72.17	72.87	74.44	<b>76.35</b>
NPV	58.07	56.76	64.24	63.82	<b>69.31</b>
AUC	62.84	62.91	63.44	64.49	<b>70.01</b>
<i>p</i> -value	0.0013*	0.0013*	0.0043*	0.0054*	–

**TABLE V**  
COMPARISON WITH BASELINE IMPUTATION-BASED METHODS: AD VS. MCI CLASSIFICATION (%)

	Zero	KNN	EM	SVD	MHL(ours)
ACC	72.41	72.75	73.79	72.93	<b>79.65</b>
SEN	29.08	30.68	29.96	26.35	<b>39.24</b>
SPE	92.89	92.64	95.94	94.93	<b>98.73</b>
BAC	60.99	61.66	62.95	60.64	<b>68.98</b>
PPV	72.25	72.99	74.94	77.39	<b>93.58</b>
NPV	74.16	74.06	73.71	73.40	<b>77.49</b>
AUC	60.99	61.66	61.45	60.64	<b>69.98</b>
<i>p</i> -value	0.0051*	0.0056*	0.0024*	0.0013*	–

False Negative, respectively. And the evaluation metrics are defined as:  $ACC = \frac{TP+TN}{TP+TN+FP+FN}$ ;  $SEN = \frac{TP}{TP+FN}$ ;  $SPE = \frac{TN}{TN+FP}$ ;  $BAC = \frac{SEN+SPE}{2}$ ;  $PPV = \frac{TP}{TP+FP}$ ;  $NPV = \frac{TN}{TN+FN}$ . The McNemar test [54] is used to evaluate the statistical significance of the difference between classification accuracies of two methods. We report the  $p$ -values in **Tables III–VI** and mark statistically significant differences ( $p < 0.05$ ) with the asterisk (\*).

## C. Comparison With Baseline Methods

We first compare MHL with four baseline imputation-based methods, including Zero, KNN [47], [48], EM [27] and SVD [28]. In **Tables III–VI**, we report results for four classification tasks: AD vs. NC, MCI vs. NC, AD vs. MCI, and pMCI vs. sMCI, respectively, where the best results are marked in boldface. We also show the receiver operating characteristic (ROC) curves achieved by different methods in **Fig. 4**. From

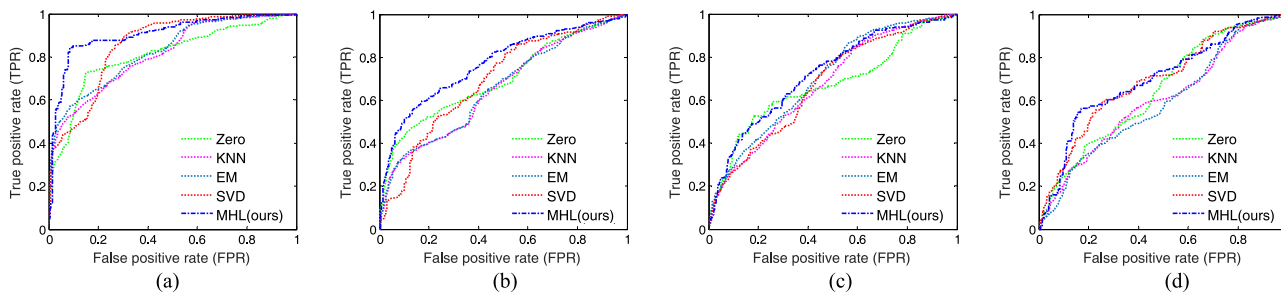


Fig. 4. ROC curves achieved by five different methods in (a) AD vs. NC, (b) MCI vs. NC, (c) AD vs. MCI, and (d) pMCI vs. sMCI classification tasks. Here, different colors denote the ROC curves achieved by five different methods.

TABLE VI

COMPARISON WITH BASELINE IMPUTATION-BASED METHODS: pMCI VS. sMCI CLASSIFICATION (%)

	Zero	KNN	EM	SVD	MHL(ours)
ACC	66.83	65.82	65.82	66.58	<b>74.68</b>
SEN	51.49	53.85	56.07	51.37	<b>68.49</b>
SPE	78.13	74.61	73.41	78.36	<b>79.47</b>
BAC	64.81	64.23	64.74	64.86	<b>73.98</b>
PPV	63.81	60.88	62.04	63.72	<b>79.13</b>
NPV	68.79	68.87	70.09	67.70	<b>72.85</b>
AUC	64.81	64.23	64.74	67.54	<b>71.98</b>
<i>p</i> -value	0.0032*	0.0035*	0.0033*	0.0037*	—

Tables III–VI and Fig. 4, we can observe that MHL consistently outperforms the competing methods regarding all seven evaluation criteria in four classification tasks. For instance, MHL achieves an accuracy of 90.29% and an AUC of 89.77% for AD vs. NC classification, outperforming other methods. The superiority of MHL is confirmed by the fact that the differences are all statistically significant as shown in the tables. On the other hand, it is interesting to observe from Table V and Fig. 4(c) that, for AD vs. MCI classification, the sensitivity achieved by MHL and other methods are low. This implies that it is difficult to distinguish AD from MCI, because MCI (the prodrome of AD) might manifest abnormalities similar to AD.

#### D. Comparison With State-of-the-Art Methods

We further compare MHL with six state-of-the-art methods for AD/MCI classification, including two versions of Ingalhalikar’s algorithm (i.e., Ingalhalikar-Weighted, and Ingalhalikar-Average) [50], two versions of iMSF (i.e., iMSF-Least, and iMSF-Logistic) [19], iSFS [20], and MSC [49]. The results for different classification tasks are reported in Tables VII–X. Note that in these tables, we directly report the results of iSFS [20] and MSC [49] in their respective reference papers. From Tables VII–X, we can observe that the proposed MHL method achieves the best ACC values in four classification tasks, and outperforms six competing methods regarding AUC in both AD vs. MCI and pMCI vs. sMCI classification tasks.

#### E. Computational Complexity

According to Section III-B, the computational complexity for hypergraph construction is  $\mathcal{O}(GN^2)$ , where  $G$  is the number of

TABLE VII

COMPARISON WITH STATE-OF-THE-ART METHODS: AD VS. NC CLASSIFICATION (%)

	ACC	SEN	SPE	AUC
Ingalhalikar-Weighted	83.03	78.54	86.72	89.82
Ingalhalikar-Average	81.07	76.37	84.94	87.39
iMSF-Least	86.41	76.91	94.24	85.57
iMSF-Logistic	86.97	75.78	93.90	86.34
iSFS	88.48	<b>88.95</b>	88.16	88.56
MSC	88.50	83.70	92.70	<b>94.40</b>
MHL(ours)	<b>90.29</b>	84.40	<b>95.13</b>	89.77

TABLE VIII

COMPARISON WITH STATE-OF-THE-ART METHODS: MCI VS. NC CLASSIFICATION (%)

	ACC	SEN	SPE	AUC
Ingalhalikar-Weighted	62.58	65.42	<b>57.73</b>	64.40
Ingalhalikar-Average	61.61	64.16	57.28	62.07
iMSF-Least	70.64	81.62	54.42	63.02
iMSF-Logistic	71.61	82.83	54.73	63.78
MSC	71.50	75.30	64.90	<b>77.30</b>
MHL(ours)	<b>74.35</b>	<b>86.25</b>	53.74	70.01

TABLE IX

COMPARISON WITH STATE-OF-THE-ART METHODS: AD VS. MCI CLASSIFICATION (%)

	ACC	SEN	SPE	AUC
Ingalhalikar-Weighted	63.44	<b>47.51</b>	66.22	63.24
Ingalhalikar-Average	63.10	45.46	65.71	61.69
iMSF-Least	73.44	23.68	96.95	60.31
iMSF-Logistic	73.96	25.36	96.95	61.16
MHL(ours)	<b>79.65</b>	39.24	<b>98.73</b>	<b>69.98</b>

TABLE X

COMPARISON WITH STATE-OF-THE-ART METHODS: pMCI VS. sMCI CLASSIFICATION (%)

	ACC	SEN	SPE	AUC
Ingalhalikar-Weighted	62.58	65.42	57.73	64.40
Ingalhalikar-Average	61.61	64.16	57.28	62.07
iMSF-Least	70.64	71.62	54.42	63.02
iMSF-Logistic	71.61	<b>72.83</b>	54.73	63.78
MHL(ours)	<b>74.68</b>	68.49	<b>79.47</b>	<b>71.98</b>



**TABLE XI**  
RUNTIME COMPARISON OF DIFFERENT METHODS IN AD VS. NC CLASSIFICATION

Method	Time (s)
Zero	0.48
KNN	1.55
EM	1.93
SVD	2.92
Ingalhalikar-Weighted	2.55
Ingalhalikar-Average	2.35
iMSF-Least	2.71
iMSF-Logistic	4.16
MHL(ours)	4.01

data grouping (e.g.,  $G = 6$  for ADNI with MRI, PET and CSF data modalities), and  $N$  is the number of subjects in the database. Besides, the computational complexity for the hypergraph regularized transductive classification is  $\mathcal{O}(N^2T)$ , where  $T$  is the iteration number for solving the optimization problem in (7). Hence, the overall computational complexity of the proposed MHL method is  $\mathcal{O}(N^2)$ .

We further empirically compare the computational time cost between MHL and 8 competing methods. Table XI reports the computational time costs of different methods in AD vs. NC classification. From Tables VII and XI, we can see that the computational time cost of MHL is similar to that of iMSF-Logistic, while the classification results achieved by MHL are much higher than those of iMSF-Logistic.

## V. DISCUSSION

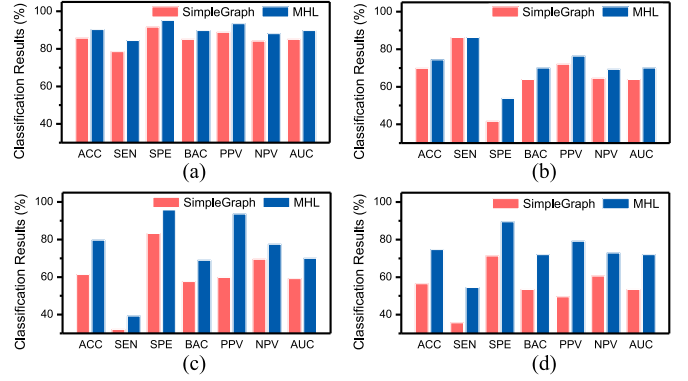
We compare MHL with a simple graph method and a single-modality method in Sections V-A and V-B, respectively. In Section V-C, MHL is further compared with the commonly used classifier, i.e., SVM. Sections V-D and V-E investigate the influences of neighborhood size in hyperedge construction and regularization parameters.

### A. Comparison With Simple Graph

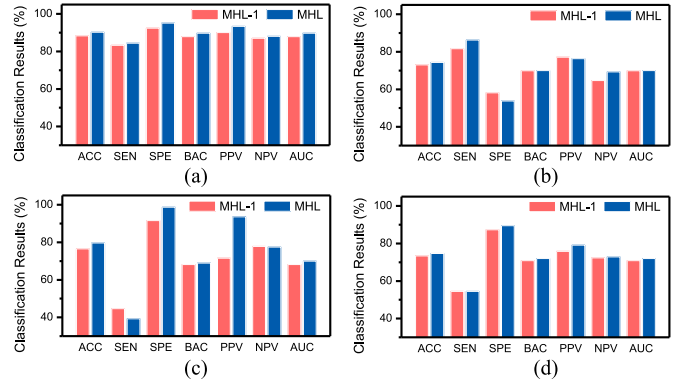
We compare MHL with a simple graph method (SimpleGraph), which is based on [34] and uses the normalized graph Laplacian for transductive classification. Note that a simple graph can only model pairwise relationships among subjects. Similar to MHL, the regularization parameter in SimpleGraph is selected from  $\{10^{-3}, 10^{-2}, \dots, 10^4\}$  via cross validation. Fig. 5 shows that MHL achieves better results in most cases. For instance, MHL is superior to SimpleGraph in AD vs. MCI and pMCI vs. sMCI classification. These results support the fact that explicitly modeling complex relationships among subjects can boost the learning performance.

### B. Comparison With Single-Modality

In the proposed MHL method, we use not only *single-modality* data, but also *inter-modality* data (with features from more than one modalities). We now investigate whether using inter-modality data can improve diagnosis performance. We run



**Fig. 5.** Comparison of MHL with a simple graph (SimpleGraph) method. (a) AD vs. NC classification. (b) MCI vs. NC classification. (c) AD vs. MCI classification. (d) pMCI vs. sMCI classification.



**Fig. 6.** Comparison of MHL with MHL-1, where MHL-1 only uses single-modality data (i.e., MRI, PET, and CSF) to construct hypergraphs. (a) AD vs. NC classification. (b) MCI vs. NC classification. (c) AD vs. MCI classification. (d) pMCI vs. sMCI classification.

MHL with data from only one single-modality (i.e., data in Groups 4–6 in Fig. 2) that is called MHL-1 in this paper. Fig. 6 shows that MHL performs better than MHL-1 in all cases, indicating that *inter-modality* data improve the diagnostic accuracy.

### C. Comparison With SVM

We compare MHL with the commonly used linear SVM classifier, using only the complete MRI data. For a fair comparison, the parameter  $C$  for SVM is chosen from  $\{10^{-3}, 10^{-2}, \dots, 10^4\}$ , which is similar to MHL. Fig. 7 shows that MHL outperforms SVM in most cases. In particular, MHL achieves better results for pMCI vs. sMCI classification in seven evaluation criteria. In addition to the improved performance, another advantage of MHL over SVM is that MHL can perform classification directly using incomplete multi-modality data, while SVM is only suitable for problems with complete data.

### D. Influence of Neighbor Size

As mentioned in Section III, in each group, we construct multiple hyperedges by varying the neighborhood size. To investigate the effectiveness of such strategy, we compare MHL with its variant that uses only one fixed neighborhood size

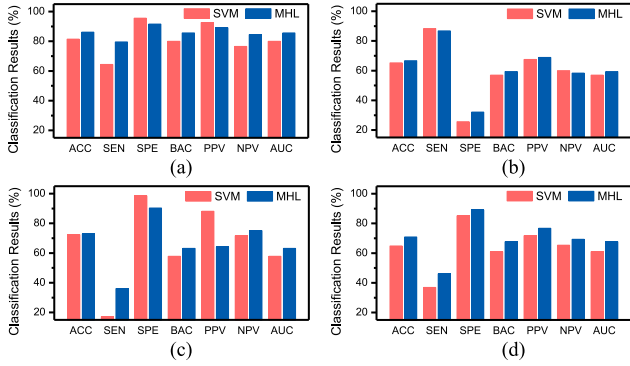


Fig. 7. Comparison between SVM and the proposed MHL method using complete MRI data from the baseline ADNI database. (a) AD vs. NC classification. (b) MCI vs. NC classification. (c) AD vs. MCI classification. (d) pMCI vs. sMCI classification.

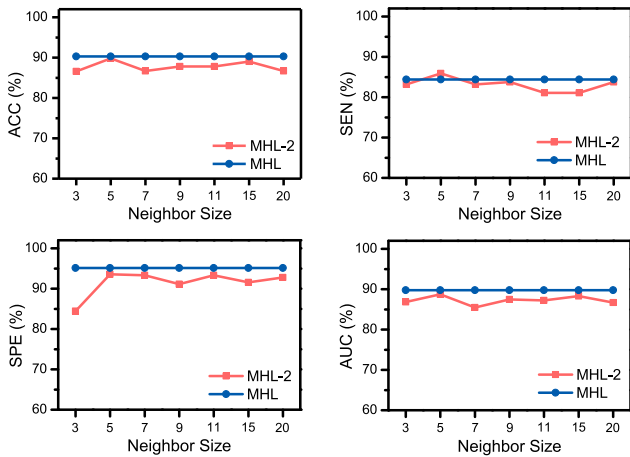


Fig. 8. Comparison between MHL and MHL-2 in AD vs. NC classification, where MHL-2 constructs hyperedges using one fixed neighborhood size.

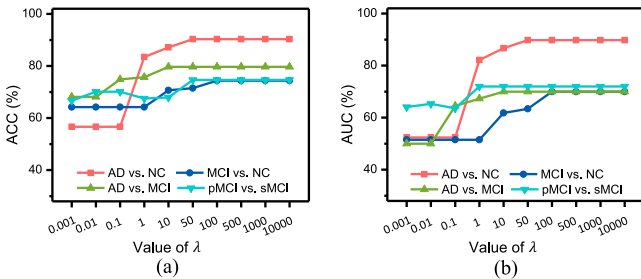


Fig. 9. Results achieved by MHL using different values of  $\lambda$ .

(denoted as MHL-2) in AD-NC classification. Fig. 8 shows that, in most cases, MHL achieves better performance than MHL-2. The superiority of MHL can be attributed to its ability to capture both local and semi-local structure information captured by our hyperedge construction method.

### E. Influence of Parameters

Here we evaluate the influence of two regularization parameters (i.e.,  $\lambda$  and  $\mu$ ) in (7) on the performance of MHL, with results shown in Figs. 9 and 10, respectively. The value of  $\lambda$  and  $\mu$  are varied within  $\{10^{-3}, 10^{-2}, \dots, 10^4\}$ . Fig. 9 indicates that

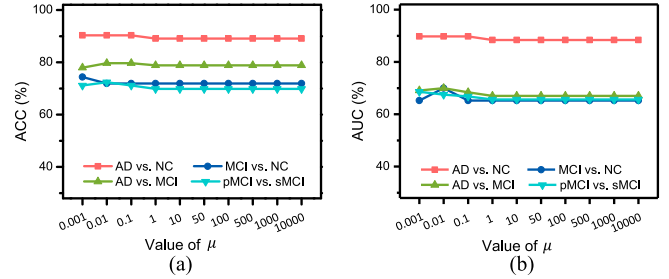


Fig. 10. Results achieved by MHL using different values of  $\mu$ .

when the value of  $\lambda$  is larger than 50, MHL generally achieves better results in both accuracy and AUC. This trend is stable with the increase of  $\lambda$ . From Fig. 10, we can see that the performance of the proposed MHL method is overall stable using different  $\mu$ , and the best ACC and AUC are achieved when  $\mu \leq 1$ .

### F. Limitations

In this work, we validate the effectiveness of our proposed MHL method on the baseline ADNI database with MRI, PET, and CSF incomplete multi-modality data. Although MHL yields promising results in four classification tasks, there are still several limitations in our method. First, we partition data into multiple groups (corresponding to combinations of modalities), and simply treat different groups equally. Actually, different groups could play different roles in modeling the structure information of data. It is reasonable to further learn the optimal weights for multiple groups to improve the performance of the proposed method, which is one of our future works. On the other hand, the proposed method can only deal with problems of block-wise incomplete data. For more general problems (e.g., with missing values in a specific modality for some subjects), we can first adopt a simple technique (e.g., EM or SVD) to impute these missing values, and then partition subjects into different groups, followed by multi-hypergraph construction and hypergraph based classification using our proposed model. Besides, due to the computational cost in hypergraph construction, our method cannot efficiently deal with large-scale samples. As a future work, we will investigate the deep learning based methods for dealing with the problem of incomplete multi-modality data, by using our proposed multi-modal data grouping strategy.

## VI. CONCLUSION

In this paper, we propose a multi-hypergraph learning (MHL) method to effectively utilize incomplete multi-modality data for AD/MCI diagnosis. We first divide the whole dataset into several groups according to data availability for different combinations of modalities. Each of these groups contains subjects with complete data. We then construct a hypergraph to explicitly model the complex relationships among subjects in each group. With the combination of hyperedges in multiple hypergraphs, we compute the hypergraph Laplacian matrix that is finally utilized for hypergraph regularized transductive classification. Experimental results on the baseline ADNI database demonstrate that MHL makes effective use of incomplete multi-modality data and

improves AD/MCI diagnostic accuracy. MHL is general and can be extended to other problems with incomplete multi-modality data, such as those involved in longitudinal studies.

### ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this study. A complete listing of ADNI investigators can be found online [https://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf).

### REFERENCES

- [1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimer's Dementia*, vol. 3, no. 3, pp. 186–191, 2007.
- [2] A. Association *et al.*, "2013 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 9, no. 2, pp. 208–245, 2013.
- [3] E. M. Reiman, J. B. Langbaum, and P. N. Tariot, "Alzheimer's prevention initiative: A proposal to evaluate presymptomatic treatments as quickly as possible," *Biomarkers Med.*, vol. 4, no. 1, pp. 3–14, 2010.
- [4] R. Cuingnet *et al.*, "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *NeuroImage*, vol. 56, no. 2, pp. 766–781, 2011.
- [5] R. Wolz *et al.*, "Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease," *PLoS ONE*, vol. 6, no. 10, 2011, Art. no. e25446.
- [6] J. Zhang, Y. Gao, Y. Gao, B. C. Munsell, and D. Shen, "Detecting anatomical landmarks for fast Alzheimer's disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2524–2533, Dec. 2016.
- [7] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal Alzheimer's disease classification," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 3, pp. 984–990, May 2014.
- [8] J. Zhang, M. Liu, L. An, Y. Gao, and D. Shen, "Alzheimer's disease diagnosis using landmark-based features from longitudinal structural MR images," *IEEE J. Biomed. Health Informat.*, doi: 10.1119/JBHI.2017.2704614, 2017.
- [9] E. E. Bron, M. Smits, W. J. Niessen, and S. Klein, "Feature selection based on the SVM weight vector for classification of dementia," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 5, pp. 1617–1626, Sep. 2015.
- [10] J. Zhang, M. Liu, and D. Shen, "Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4753–4764, Oct. 2017.
- [11] M. Liu, D. Zhang, and D. Shen, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1463–1474, Jun. 2016.
- [12] G. Chetelat, B. Desgranges, V. De La Sayette, F. Viader, F. Eustache, and J.-C. Baron, "Mild cognitive impairment can FDG-PET predict who is to rapidly convert to Alzheimer's disease?" *Neurology*, vol. 60, no. 8, pp. 1374–1377, 2003.
- [13] K. Herholz *et al.*, "Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET," *NeuroImage*, vol. 17, no. 1, pp. 302–316, 2002.
- [14] N. L. Foster *et al.*, "FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease," *Brain*, vol. 130, no. 10, pp. 2616–2635, 2007.
- [15] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, and J. Li, "A robust deep model for improved classification of AD/MCI patients," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 5, pp. 1610–1616, Sep. 2015.
- [16] C. Lian, S. Ruan, T. Denœux, F. Jardin, and P. Vera, "Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction," *Med. Image Anal.*, vol. 32, pp. 257–268, 2016.
- [17] O. Hansson, H. Zetterberg, P. Buchhave, E. Londos, K. Blennow, and L. Minthon, "Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: A follow-up study," *Lancet Neurol.*, vol. 5, no. 3, pp. 228–234, 2006.
- [18] T. Kawarabayashi, L. H. Younkin, T. C. Saido, M. Shoji, K. H. Ashe, and S. G. Younkin, "Age-dependent changes in brain, CSF, and plasma amyloid  $\beta$  protein in the tg2576 transgenic mouse model of Alzheimer's disease," *J. Neurosci.*, vol. 21, no. 2, pp. 372–381, 2001.
- [19] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.
- [20] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, 2014.
- [21] M. Liu, D. Zhang, and D. Shen, "View-centralized multi-atlas classification for Alzheimer's disease diagnosis," *Human Brain Mapping*, vol. 36, no. 5, pp. 1847–1865, 2015.
- [22] X. Cao, J. Yang, Y. Gao, Y. Guo, G. Wu, and D. Shen, "Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis," *Med. Image Anal.*, doi: 10.1016/j.media.2017.05.004, 2017.
- [23] C. R. Jack *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, 2008.
- [24] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: Data mining, inference and prediction," *Math. Intell.*, vol. 27, no. 2, pp. 83–85, 2005.
- [25] K. Lakshminarayan, S. A. Harp, R. P. Goldman, and T. Samad, "Imputation of missing data using machine learning techniques," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 1996, pp. 140–145.
- [26] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2014.
- [27] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *J. Climate*, vol. 14, no. 5, pp. 853–871, 2001.
- [28] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, no. 5, pp. 403–420, 1970.
- [29] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
- [30] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data," *Med. Image Anal.*, vol. 36, pp. 123–134, 2017.
- [31] J. Y. Zien, M. D. Schlag, and P. K. Chan, "Multilevel spectral hypergraph partitioning with arbitrary vertex sizes," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 18, no. 9, pp. 1389–1399, Sep. 1999.
- [32] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. 16th Int. Conf. Mach. Learn.*, 1999, pp. 200–209.
- [33] S. Agarwal, K. Branson, and S. Belongie, "Higher order learning with graphs," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 17–24.
- [34] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1601–1608.
- [35] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 55–67, Jan. 2008.
- [36] C. Berge and E. Minieka, *Graphs and Hypergraphs*, vol. 7. Amsterdam, The Netherlands: North Holland, 1973.
- [37] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.
- [38] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.
- [39] J. Rodriguez, "On the Laplacian spectrum and walk-regular hypergraphs," *Linear Multilinear Algebra*, vol. 51, no. 3, pp. 285–297, 2003.
- [40] M. Bolla, "Spectra, Euclidean representations and clusterings of hypergraphs," *Discrete Math.*, vol. 117, no. 1, pp. 19–39, 1993.
- [41] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998.
- [42] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.
- [43] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002.

- [44] N. Kabani, "A 3D neuroanatomical atlas of the human brain," *NeuroImage*, vol. 7, no. 4, p. 717, 1998.
- [45] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.
- [46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [47] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," Tech. Rep., 1999.
- [48] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
- [49] K.-H. Thung, C.-Y. Wee, P.-T. Yap, and D. Shen, "Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion," *NeuroImage*, vol. 91, pp. 386–400, 2014.
- [50] M. Ingalhalikar, W. A. Parker, L. Bloy, T. P. Roberts, and R. Verma, "Using multiparametric data with missing features for learning patterns of pathology," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012*. New York, NY, USA: Springer-Verlag, 2012, pp. 468–475.
- [51] S. Mikat, G. ftscht, J. Weston, B. Scholkopf, and K.-R. Mullert, "Fisher discriminant analysis with kernels," *Neural Netw. Signal Process.*, vol. 9, pp. 41–48, 1999.
- [52] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.
- [53] R. H. Fletcher, S. W. Fletcher, and G. S. Fletcher, *Clinical Epidemiology: The Essentials*. Baltimore, MD, USA: Williams & Wilkins, 2012.
- [54] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.