Published in final edited form as:

Methods. 2021 May ; 189: 86–94. doi:10.1016/j.ymeth.2020.04.008.

Hyper-graph based Sparse Canonical Correlation Analysis for the Diagnosis of Alzheimer's Disease from Multi-dimensional Genomic Data

Wei Shao[1], **Shunian Xiang**[2,3], **Zuoyi Zhang**[1,4], **Kun Huang**[1,4], **Jie Zhang**[3]

[1.]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, 46202 USA.

[2.]Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University, Shenzhen, 518060, China

[3.]Department of Medical & Molecular Genetics, Indiana University, Indianapolis, IN, 46202, USA

[4.] Regenstrief Institute, Indianapolis, IN, ,46202, USA

Abstract

The effective and accurate diagnosis of Alzheimer's disease (AD), especially in the early stage (i.e., mild cognitive impairment (MCI)) remains a big challenge in AD research. So far, multiple biomarkers have been associated with AD diagnosis and progression. However, most of the existing research only utilized single modality data for diagnostic biomarker identification, which did not take the advantages of multi-modal data that provide comprehensive and complementary information at multiple levels into consideration. In this paper, we integrate multi-modal genomic data from postmortem AD brains (i.e., mRNA, miRNA and epigenomic data) and propose a hyper-graph based sparse canonical correlation analysis (HGSCCA) method to extract the most correlated multi-modal biomarkers associated with AD and MCI. Specifically, our model utilizes the sparse canonical correlation analysis framework (SCCA), which aims at finding the best linear projections for each input modality so that the strongest correlation within the selected features of multi-dimensional genomic data can be captured. In addition, with the consideration of high-order relationships among different subjects, we also introduce a hyper-graph-based regularization term that will lead to the selection of more discriminative biomarkers. To evaluate the effectiveness of the proposed method, we conduct the experiments on the well-known AD cohort study, The Religious Orders Study and Memory and Aging Project (ROSMAP) dataset, and the results show that our method can not only identify meaningful biomarkers for the diagnosis AD disease, but also achieve superior classification performance than the comparing methods.

In this paper, Kun Huang and Jie Zhang conceived the project. Wei Shao, Shunian Xiang conceived and designed the experiments. Wei Shao performed the experiments and analyzed the data. Zuoyi Zhang preprocess the original data. Wei Shao wrote the manuscript, and Kun Huang and Jie Zhang edited and revised the manuscript.

Keywords

Hyper-graph; Multi-modal; Alzheimer's Disease; Sparse Canonical Correlation Analysis; Diagnostic Biomarker

1. Introduction:

Alzheimer's disease (AD) is the most common form of dementia in the elderlies that is characterized by the progressive loss of neurons and their connections [1] . It is reported that the number of affected AD patients will double in the next 20 years, and 1 in 85 people will be affected by 2050 [2]. AD is a slowly developing dementia, with the initial pathological changes happens decades before the cognitive impairment finally show, therefore, an effective and accurate diagnosis of AD, especially at its early stage, is very important for prompt interventions. Just as most diseases, the molecular architecture of AD is manifested at multiple levels including genomic transcriptional, post-transcriptional and epigenomic events [3–5]. At present, many biomarkers have been proven to be sensitive to the diagnosis of AD and mild cognitive impairment (MCI), Specifically, Hansson *et al.* [6] found that the major neuropathological hallmarks of AD are the presence of extracellular amyloid plaques that are composed of Aβ_{40}, Aβ_{42}, and intracellular neurofibrillary tangles (NFT). For RNA-level analysis, many studies have focused on transcriptomic changes in AD [7], and other studies showed that miRNA acts as a post-transcriptional regulator and the dysregulation of miRNA expression has also been associated with AD disease [8, 9]. In addition, DNA methylation and hydroxymethylation [10, 11] are also found to be linked to the development and potential treatment of AD.

However, most of the existing studies focused on utilizing single modality of data for the identification of diagnostic biomarkers for AD or MCI. For example, many classification and statistical models classified AD (MCI) and normal control (NC) patients based on mRNA or protein data alone. Agarwal et al. [12] developed a neural-network-based classification algorithm, which identified nine plasma proteins associated with the occurrence of AD . Li et al [13] applied the least absolute shrinkage and selection operator (LASSO) regression feature selection method, followed by the Support Vector Machine (SVM) classifier to discriminate AD patients from control subjects using protein data . Mukherjee et al [14] proposed a ranking algorithm to identify driving genes from postmortem brain tissue in patient cohorts. Zhang [15] et al. used the Support Vector Machine based recursive feature elimination (SVM-RFE) algorithm to extract gene biomarkers with low redundancy from large amount of gene expression data. In addition, for transcriptomic features, gene co-expression network is a great way to reduce gene features from 22000 to less than a hundred [16–18], and Wang et.al [19] proposed a Bayes-based method to identify potential diagnostic biomarkers of AD based on gene co-expression network. For miRNA data, Leidinger et al [20] found that a panel of 12 miRNAs can predict cognitive decline or the conversion to AD, Fransquet et al [21] identified that microRNA works as a potential blood-based epigenetic biomarker for AD progression. Other works such as Dong et al [22] identified four miRNAs decreased in serum of patients with AD pathology and vascular dementia. Besides gene and miRNA biomarkers, there are also

some DNA methylation markers developed for the diagnosis of AD or MCI. For example, Li et al [23] examined 1.2 million CpG and CpH sites and identified 1224 AD-related methylated enhancer regions by applying a network analysis algorithm, and Vasanthakumar et al. [24] investigated the DNA methylation dynamics that are associated with the AD disease progression.

Recently, several studies combined multiple data types for AD disease mechanism study. For instance, Bihmeyer et al [25] integrated microarray and RNA-seq data to identify important genes in the pathophysiology of AD. Humphries et al [26] proposed an integrative method that combined the transcriptome and DNA methylation data to identify gene networks specific to late-onset Alzheimer's disease. Jiang et al. [27] systematically analyzed AD-related mRNA and miRNA expression profiles to identify active transcription factor and miRNA regulatory pathways in AD. Although the above methods indicated that the combinational analysis of multimodality biomarkers can uncover the hidden mechanism of AD that cannot be found using a single type of data, most of them are focused on characterizing AD pathology and look for disease mechanisms, instead of diagnostic biomarker identification for disease classification. In clinical practice, multiple biomarkers are necessary for a sensitive and precise assessment of disease status and progression stages [28–30]. In this paper, our goal is to apply machine learning algorithms on multi-modal AD data to identify biomarkers for highly accurate individual classification of AD (MCI) patients from normal age-matched people.

To analyze multi-modal data, the major challenge is to distinguish and remove redundant or irrelevant features. Thus, feature selection is the key step in the biomarker identification. Most existing multi-modal feature selection methods are the simple extension of the single-modality based method, which combine multi-modal data into a long vector at the beginning, and then apply the traditional feature selection methods [31, 32] to identify important features that contribute most to the disease diagnosis. However, this approach ignores the intrinsic correlations within the multi-modal data, which are shown to be useful for multi-modal data integration [30, 33, 34]. To address this problem, Jie et al [35] proposed a multi-modal feature selection method that used multi-task learning framework to preserve the intrinsic relatedness among multiple modalities of data. Zu et al [36] constructed a label-aligned multi-task learning framework that can uncover inherent structures among multi-modal data. In addition, Lei [37] proposed a canonical correlation analysis (CCA) based method to fuse multi-modal data for the diagnosis of AD/MCI patients.

Although much progress has been achieved through the incorporation of above frameworks into multi-modal data feature selection, one disadvantage of the existing methods is that most of them only consider the correlation within multi-modality data but ignore the data distribution information of each data modality. To the best of our knowledge, only Jie et al [35] took the pairwise relationship among different subjects into consideration. However, such pairwise relationship does not reflect the complex connections of different subjects in the real applications. Intuitively, the higher-order relationships among subjects can provide useful information to improve the machine learning performance of AD/MCI diagnosis tasks. Based on the above considerations, in this paper, we propose a hyper-graph-

based sparse canonical correlation analysis method that can identify important miRNA, gene and DNA methylation biomarkers for the diagnosis of AD (or MCI) from normal aging cohort. Specifically, our model extends our previously developed generalized sparse canonical correlation analysis (GSCCA) framework [28], which aims to find the best linear transformations for multi-modal data, so that the highest correlation within multi-modal data can be achieved. In addition, we constructed a hyper-graph in each modality and introduced a hyper-graph-based Laplacian regularization term to capture the high-order relationships among different subjects. To evaluate the effectiveness of the proposed method, we perform the classification on the dataset derived from Religious Orders Study and Memory and Aging Project (ROSMAP), and the superior performance over other existing methods demonstrate the effectiveness of our method.

2. MATERIALS AND METHODS

2.1 Datasets

ROSMAP is composed of two AD cohort studies ROS and MAP. The data were combined together, constituting a longitudinal clinical-pathologic cohort study of aging and AD project by Rush University [38, 39]. ROSMAP includes 2251 participants under age 90. In this paper, quality-controlled RNA-Seq gene expression data, DNA methylation data and miRNA expression data are downloaded from AMP-AD knowledge portal (Accelerating Medicines Partnership-Alzheimer's disease, https://www.82synapse.org/#!Synapse:syn9702085/). To perform the proposed multi-modal based Hyper-graph-based Sparse Canonical Correlation Analysis (HGSCCA) algorithm, only subjects with matched RNA-seq, miRNA and DNA methylation data are included. In total, we have 475 subjects (samples) including 168 Normal Control (NC), 125 Mild Cognitive Impairment (MCI) and 182 AD patients. Table 1 lists the demographic and clinical information of the cohort.

2.2 Overview of the HGSCCA Workflow

Fig 1 illustrates the workflow of the proposed HGSCCA method, which includes four steps: 1) for the multi-modal data, filter and pre-select the features that are highly correlated to AD status; 2) construct the hyper-graph for each data modality that can capture the high-order correlation among different subjects; 3) based on the derived hyper-graphs, a novel HGSCCA algorithm is used to further select the informative features from the pre-selected features in step 1; and 4) the support vector machine (SVM) classifier is adopted for the diagnosis of AD (or MCI) patients from the normal control subjects.

2.3 Feature Extraction and Feature Pre-selection

For the diagnosis of AD, different modalities of biomarkers (i.e., gene expression, miRNA expression, DNA methylation level) are used, which are expected to provide comprehensive and complementary information about the disease pathology and progression. Specifically, for the RNA-seq data, we follow the instruction in the AMP-AD knowledge portal for data preprocessing, which yields 55,890 gene expression features (probeset IDs) for each subject. The miRNA expression data were filtered to retain miRNAs with a call rate of 95% and with an absolute value of above 15 in at least 50% of the samples. This pre-processing step resulted in 309 miRNA features. Lastly, for the DNA methylation data, the normalized

beta-values were downloaded and used as the methylation level of each methylation site. This process generates 420,133 DNA methylation features (probeset IDs) for each subject.

For above multi-modal data, it may contain redundant features or noise that may negatively affect the final diagnosis performance. To address this problem, we pre-select the features by the following two steps: first, features with variance smaller than 0.3 in the training dataset were filtered out; second, for each classification task (i.e., AD vs NC, MCI vs NC), Student's t-tests was performed sequentially to evaluate whether the values of each remaining feature is significantly different across different classes in the training dataset. In this step, a multiple-testing compensation method described in [40] was applied to keep the features with False Discovery Rate (FDR) smaller than 0.05, 0.005, and 0.0005 for miRNA expression, RNA-seq, and DNA methylation data, respectively.

2.4 Generalized Sparse Canonical Correlation Analysis

Although the above t-test based method may capture the relevance of initial features with clinical phenotypes, the features were derived from different modalities independently, which does not take into account the intrinsic relationship across different data types (views). In the machine learning field, multi-view association has been widely applied as a key component of recently developed multi-view feature selection algorithms [33, 34]. Accordingly, we further implement the second-stage feature selection model with the Generalized Sparse Canonical Correlation Analysis (GSCCA) framework [28] that can capture the intrinsic correlation among the multi-modal data. Specifically, let $X_G \in R^{N \times p}, X_D \in R^{N \times q}, X_M \in R^{N \times r}$ denote the gene expression, DNA methylation and miRNA expression data, where N denotes the number of samples, p, q,r correspond to the dimensionalities for each modality of data. Then, the objective function of the GSCCA model can be formulated as follows:

$$\underset{w_i, Q}{\text{Min}} \Sigma_{i \in \{G, D, M\}} \|X_i w_i - Q\|_2^2 \quad (1)$$

$$s.t. \quad \|w_i\|_1 \leq t \quad i \in \{G, D, M\}$$

where the objective function is used to approximate the projection of each modality. *i.e.*, $X_i w_i$, to a shared space, so that the correlation among multi-modal could be captured. $\|w_i\|_1 \leq t, \quad i \in \{G, D, M\}$ are the regularization terms for the objective function, which can lead to the identification of a small subset of features from the pre-selected multi-modal genomic data.

2.5 Hyper-graph-based Sparse Canonical Correlation Analysis (HGSCCA)

In the GSCCA model, we only consider the inter-correlations among multi-modal data but ignore the mutual dependency among different subjects, which may result in large deviations even for very similar data after projecting to the common representation Q (shown in Eq.

(1)). To address this issue, several machine learning algorithms [35] [41] applied the graph structure to characterize the correlations among samples, where each sample is denoted as a node and the relationship between any pair of nodes is represented by an edge. The graph can be undirected or directed, depending on whether the pairwise relationships between subjects are symmetric or not. However, the graph can only characterize the pairwise correlation between different samples, not the complex relationships in real applications.

Different from the traditional graph, a hyper-graph can connect more than two vertices by a hyperedge. Hence, we can model complex relationships (i.e., high-order relationships) among samples of our interest using hyper-graph [42, 43]. We show an example of the hyper-graph on the left panel of Fig. 2, where the vertices involved in the same hyper-edge are circled by an ellipse, and different hyper-edges are represented by the ellipses with different colors. For instance, the hyper-edge e_1 represented by yellow ellipse is consisted of three vertices i.e., v_4, v_2,v_8 . Please note that the conventional graph is a special case of hyper-graph with each hyper-edge only containing two nodes.

Mathematically, a hyper-graph can be represented by $G(V, E, a)$, where $V = \{v_1, v_2 \ldots v_N\}$, $E = \{e_1, e_2 \ldots, e_{N_E}\}$ denotes the set of vertices and hyper-edges respectively, and $a \in R^{N_E}$ is the weights for the hyper-edges. Each hyper-edge $e_i (i = 1, 2, \ldots N_E)$ is assigned a weight $a(e_i)$. In hyper-graphs, a hyper-edge can connect more than two vertices, through which high-order relationships can be modeled explicitly. We use the incidence matrix, i.e., $H \in R^{N \times N_E}$ to represent the relationships between hyper-edges and vertices, an entry $H(v_i, e_j)$ is set to 1 if the hyperedge e_j contains vertex v_i, and 0 otherwise. Based on this definition of H, the degree of each vertex and hyper-edge are denoted as:

$$d(v_i) = \Sigma_{e_j \in E} a(e_j) H(v_i, e_j) \tag{2}$$

$$\varepsilon(e_j) = \Sigma_{v_i \in V} H(v_i, e_j) \tag{3}$$

Suppose D_V and D_E represent the diagonal matrices containing the vertex degrees and the hyperedge degrees, respectively. Denote $A \in R^{N_E \times N_E}$ as the diagonal matrix of hyper-edge weights with the diagonal element $A_{i,i} = a(e_i)$. In order to capture the high-order correlation among different subjects, we follow the method in [42] to define the hyper-graph Laplacian matrix L^H as following:

$$L^H = I - \varnothing \tag{4}$$

where I is the identity matrix and $\varnothing = D_V^{-\frac{1}{2}} H A D_E^{-1} H^T D_V^{-\frac{1}{2}}$. As can be seen from Eq.(2) and Eq.(3), the most important step for the hyper-graph learning is to construct the hyper-graph

from the input data. Here, we apply the method introduced in [42] that used the K nearest neighbor (KNN) strategy to generate the hyper-graph. Specifically, for each modality of the input multi-modal data, we treat each sample as a vertex and calculate its Euclidean distance with other samples. Then we construct a hyper-edge by connecting each sample with its k nearest neighbors. Accordingly, we can derive n hyper-edges given n samples. Here, we set the weight of each hyper-edge as 1, i.e., $a(e_i) = 1$ and K is fixed as 6. More discussions for the parameter discussion on k are provided in Section 4.2.

Based on the hyper-graph Laplacian matrix (shown in Eq. (4)), we define the objective function of the hyper-graph based sparse canonical correlation analysis method as

$$min_{w_i, Q} \frac{1}{N} \sum_{i \in \{G, D, M\}} \|X_i w_i - Y\|_2^2 + \alpha \Sigma_{i \in \{G, D, M\}} \|X_i w_i - Q\|_2^2 + \beta \Sigma_{i \in \{G, D, M\}} w_i^T X_i^T L_i^H X_i w_i + \Sigma_{i \in \{G, D, M\}} r_i \|w_i\|_1 \tag{5}$$

where the second and the fourth terms in Eq. (5) are the same as stated in the GSCCA model. The first term corresponds to the linear discriminant function for each data modality , and $Y=[y_1, y_2, \ldots y_N]^T$ is the label vector for the N subjects, and y_i=1 or —1 indicates the disease case or normal control. L_i^H in the third term represents the hyper-graph Laplacian matrix for modality i. In comparison to the GSCCA model, our proposed HGSCCA model not only incorporates correlations among multi-modal data, but also preserves the high-order structure information in each modality by adding hyper-graph regularization terms, which can generate more discriminant features for the diagnosis of AD disease.

2.6 Optimization

The objective function in Eq.(5) is not jointly convex with respect to w_i and Q. However, it is convex with respect to w_i when Q is fixed, and also convex with respect to Q given a fixed w_i. In this study, we adopt an alternating optimization algorithm to solve the problem in (5). Specifically, we first fix $w_i, i \in \{G, D, M\}$, then the optimization problem for Q is:

$$min_Q \Sigma_{i \in \{G, D, M\}} \|X_i w_i - Q\|_2^2 \tag{6}$$

which has a closed form solution:

$$Q = \frac{1}{3} \Sigma_{i \in \{G, D, M\}} X_i w_i \tag{7}$$

On the other hand, if other variables are fixed, the optimization problem for each w_i can be formulated as:

$$min_{w_i} \frac{1}{N} \|X_i w_i - Y\|_2^2 + \alpha \|X_i w_i - Q\|_2^2 + \beta w_i^T X_i^T L_i^H X_i w_i + r_i \|w_i\|_1$$

(8)

Since the L1-norm is non-differentiable at zero, a smooth approximation is estimated for L1 term by including an extremely small value. Then, we can approximate $r_i\|w_i\|_1$ by

$$\|w_i\|_1 \approx \Sigma_{j=1}^{d_i} \sqrt[2]{(w_i^j)^2 + \epsilon}$$

(9)

Where d_i refers to the dimension for w_i. Then, the derivative of $\|w_i\|_1$ can be formulated by:

$$\frac{\partial \|w_i\|_1}{\partial w_i} = \left(\frac{(w_i^1)}{\sqrt[2]{(w_i^1)^2 + \epsilon}}, \frac{(w_i^2)}{\sqrt[2]{(w_i^2)^2 + \epsilon}}, \ldots, \frac{(w_i^{d_i})}{\sqrt[2]{(w_i^{d_i})^2 + \epsilon}} \right)^T = D_i w_i$$

(10)

Here, D_i is a diagonal matrix with the j-th element denoted as $\frac{1}{\sqrt[2]{(w_i^1)^2 + \epsilon}}$. Then, by taking the derivative of Eq.(8) with respect to w_i and let it be zero, we can obtain:

$$\frac{2}{N}(X_i)^T(X_i w_i - Y) + 2\alpha(X_i)^T(X_i w_i - Q) + 2\beta X_i^T L_i^H X_i w_i + r_i D_i w_i = 0$$

(11)

Then, we can derive

$$w_i = \left[\left(\alpha + \frac{1}{N} \right) X_i^T X_i + \beta X_i^T L_i^H X_i + \frac{r_i}{2} D_i \right]^{-1} \left(\alpha X_i^T Q + \frac{1}{N} X_i^T Y \right)$$

(12)

The above-mentioned two steps are performed iteratively until the maximum number of iteration is reached. In our experiments, the iterative number is set as 10. We show the optimization algorithm for the proposed HGSCCA method in Table. 4.

2.7 Gene Ontology Enrichment Analysis

For the derived multi-modal biomarkers, we also conduct the gene functional ontology enrichment analysis to identify enriched functions/pathways for the identified biomarkers using ToppGene Suit [44]. ToppGene Suit is a one-stop portal for gene list functional enrichment, candidate gene prioritization using either functional annotations or network analysis. (https://toppgene.cchmc.org/prioritization.jsp.)

3. Experimental Results

3.1 Cross-validation and performance evaluation

To validate the effectiveness of our proposed HGSCCA-based feature selection method, we test it on two different classification tasks, i.e. AD vs NC and MCI vs NC. For each task, we followed five-fold cross-validation procedure. Specifically, we randomly split the whole dataset into five folds with equal number of subjects (samples). Each time, samples within one fold are used as testing data, while all the others in the remaining four folds are used for model training. Since the pre-selected features in each fold are different, we list the number of pre-selected feature in each fold and the common features that are selected in most cases (i.e., at least three folds) of the cross-validation as follows:

From Tables 3 and 4 we observe that, most pre-selected features are co-appeared in different folds of the cross-validation, which shows the stability of our way for feature pre-selection.

In addition, in order to evaluate the performance of our methods, we chose six metrics including classification accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and the area under the receiver operating characteristic (ROC) curve (AUC). Denote True Positive, True Negative, False Positive, and False Negative as TP, TN, FP and FN, respectively. Those evaluation metrics can be defined as follows: $\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$, $\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, $\mathrm{SPE} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$, $\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, and $\mathrm{NPV} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FN})$. For these metrics, higher values indicate better classification performance. For parameter settings of our HGSCCA modal, α, β and $r_i (i \in \{G, D, M\})$ are tuned from {0.01,0.05,0.1}, {0.25,0.5,1,2} and {0.05,0.25,0.5,1} in the training dataset, respectively.

3.2 Classification performance comparisons among five different methods

In this section, we compared the diagnosis power of the proposed HGSCCA with the following five algorithms on both AD vs NC and MCI vs NC tasks. 1) Baseline: pre-selected multi-modal data were directly concatenated for the biomarker identification. 2) SLASSO: LASSO was applied on each single modality data (i.e., gene expression, miRNA expression or DNA methylation data), where the data type which yields the best accuracy was reported. 3) LASSO: the multi-modal data were concatenated into a long vector, and LASSO algorithm was applied for feature selection. 4) GSCCA: a variant of HGSCCA, where the relationship among different patients were ignored (i.e., the third term in Eq.(5)). and 5) PGSCCA: only the pairwise correlation among different samples were considered, but not the high-order correlations among them. For the proposed and all the comparing methods, the SVM classifier with RBF kernel is adopted for the classification task on the selected features. The results i.e., mean accuracy (variance) are shown in Table 5 and Table 6, and the ROC curve for different methods are shown in Fig. 3

As can be seen from Table 5, Table 6 and Fig. 3, for both AD vs NC and MCI and NC tasks, The algorithms (i.e., LASSO, GSCCA, PGSCCA and HGSCCA) combining multi-modal genomic data can consistently achieve more accurate discrimination than the single-modality based algorithm (S-LASSO). These results clearly demonstrated the power

of the integrative analysis with multi-modal data for AD diagnosis. In addition, the GSCCA, PGSCCA and HGSCCA can achieve better classification performance than the baseline and LASSO methods, which shows the advantage of incorporating correlations among multi-modal data for the biomarker identification. Next, the HGSCCA and PGSCCA algorithm achieved higher classification accuracy than the other three methods. These results show that capturing the relationship among different samples can boost the diagnosis performance. Finally, the proposed HGSCCA classifies AD patients better than the PGSCCA method in most of the classification evaluation metrics including accuracy, sensitivity, negative predictive value (NPV) and AUC, which also demonstrates the advantage of taking the high-order relationship among samples account for feature selection. We specially emphasized that, the high sensitivity of HGSCCA of 0.837 for AD vs NC and 0.80 for MCI vs NC, respectively, is superior to all of the comparing methods. In the clinical diagnosis, high sensitivity is especially important for doctors and family to spot early cognition impairment and start intervention in the early stage of AD, which may significantly defer the disease onset and increase the patient's quality of life.

3.3 Biomarkers Detected by the Proposed HGSCCA Method

For biologists and clinicians, the identified biomarkers may harbor important biological information in terms of disease mechanism or progression. We mapped the identified features to the genes or nearest genes (DNA methylation features) and summarized them in Tables 7 and 8. As the selected features in each fold of the cross-validation are different, we only list common features that are selected in most cases (i.e., at least three times) of the five folds cross-validation for AD vs. NC and MCI vs. NC tasks in Table 7 and Table 8, respectively.

As seen from Table 7, for AD vs NC task, 31 gene level biomarkers are identified (from 9 DNA methylation feature, 17 RNA-seq features and 5 miRNA-seq features). Gene ontology enrichment analysis by the Toppgene (https://toppgene.cchmc.org/prioritization.jsp) shows that, of all these 31 genes, 9 are involved in functions of neurotransmission, synaptic signaling or their regulations. For example, RASGRF2 [45] helps regulate neuronal excitability, neuronal survival in response to ischemia, learning and memory formation. HSPB2 was associated with classic senile plaques in AD brains [46], and the expression of NEUROG2 is significantly downregulated from Amyloid-β precursor protein, which is associated with the progression of AD [47]. In addition, miR-132 was reported to play a key role in tau-metabolism, as it regulates exon splicing of tau, colocalizes with hyperphosphorylated tau and negatively correlates with neurofibrillary tangles-bearing neurons from cortical tissue [48].

On the other hand, for the MCI vs NC task (Table 8), we identify 18 genes close to the DNA methylation features. Among them, the receptor for GANAB was shown to play important roles in reducing amyloid-β production, which is a key step in AD pathogenesis [49]. And the expression level of DBI is elevated in patients with AD [50]. As for genes isolated from RNA-seq, it has been demonstrated that PLEKHB1 is involved with neuronal development [51]. In addition, low TFF3 in cerebrospinal fluid is a predictive factor for brain atrophy, and its potential role in the pathogenesis of AD has been suggested in [52]. Finally, for miRNA

features, it is reported that miR-129–5p regulates nerve injury and inflammatory response in AD by modulating SOX6 expression [53], while the overexpression of miR-203, a hub of a putative regulatory microRNA (miRNA) module, induces the neuronal cell death. These results again validate the advantage of the HGSCCA method that can identify biologically meaningful multidimensional genomic biomarkers for AD disease diagnosis.

4 Discussion

4.1 Results for Braak & Braak Stage Prediction Task.

Besides the classification of AD (MCI) from NC cohort, we also tried to identify the features and test our workflow on the prediction of AD staging task. Braak & Braak stage (BB score) is a universally adopted semiquantitative measure of severity of neurofibrillary tangle (NFT) pathology. The system divided the progression of AD into six different stages, with each stage being attributed to abnormal pathology in defined neurological structures, and the higher Braak stage indicates more severe form of AD disease. In ROSMAP dataset, we have a total of 470 subjects (40 stage I, 39 stage II, 142 stage III, 151 stage IV, 91 stage V and 7 stage VI samples) with matched RNA, DNA Methylation, miRNA data. Due to the limited sample size for stage VI patients, it is difficult to train machine learning models. We only aim at stratifying samples whose BB Score ranges from stage I to stage V. More specifically, we firstly follow the feature pre-selection method introduced in Section 2.3 to pre-select related multi-modal features in the training set. Then, HGSCCA is applied to further identify the Braak stage associated features, followed by one-vs-one SVM classifier for final classification. The experimental results are shown in Fig. 4.

In Fig. 4, we found that HGSCCA can achieve the mean classification accuracy of 0.493, which is better than LASSO method (0.474) and PGSCCA method (0.460) with p-values 0.043 and 0.024, respectively. The results in Fig. 4, Table 5, and Table 6 indicate that HGSCCA can consistently outperform the comparing methods for both diagnosis and Braak stage prediction task, which again confirms the advantages of capturing the high-order relationship among samples for the identification of essential biomarkers from multi-modal genomic data.

4.2 The Influence of Neighbor Size for Constructing Hyper-graph

As mentioned in Section 2, we used KNN to construct hyper-graph. Specifically, we chose K nearest vertices to generate a hyperedge. In this section, the influence of the neighbor size of KNN in HGSCCA workflow is discussed. The result is shown in Fig 5, where the neighbor size K varies in the range {3,5, 6,9,11,13,17}. For both classification tasks (AD vs. NC and MCI vs. NC), the classification accuracies increase if the number of vertices in the hyperedge is increased (but less than 6). We think it reflects the algorithm can characterize more complex relationship among different samples. However, the classification decreases rapidly if K is larger than 6. The underlying reason lies in that, the hyperedges with larger number of vertices may contain samples of different classes, and thus can no longer capture the genuine underlying structure of data.

4.3 The Influence of Different Loss Functions

As shown in Eq. (5), our HGSCCA model uses the least square loss to calculate the empirical risk in the training dataset. In this section, we will discuss the classification results if we replace the least square loss with the logistic regression loss, and the results are shown in Table 7. As can be seen from Table 7, we can clearly find that different loss functions have distinct advantage and disadvantage on different tasks (i.e., AD vs NC, MCI vs NC), and there is no significant difference between least square loss and logistic regression loss on various classification evaluation metrics. One possible reason is that rather than directly using the loss function to predict the outcome of different subjects, our HGSCCA only apply it to identify meaningful biomarkers. Then, the selected feature will be fed to the SVM classifier for final classification. It seems that the selected features by these two loss functions have similar ability of classification by the SVM classifier, and thus achieve similar results in the diagnosis of AD.

4.4 Convergence Analysis

In this section, we show the convergence property of the proposed HGSCCA method. Actually, it is obvious that the objective function (Eq. (5)) of the HGSCCA is bi-convex with respect to w_i and Q. Hence, we use an alternative optimization algorithm to solve the proposed problem, which can theoretically guarantee its convergence to the optimal solution [54]. In addition, we also plot the change of the objective function values versus iteration number on both the NC vs MCI and NC vs AD classification tasks in Fig. 6. From Fig. 6, we can see that the objective function value decreases rapidly within several initial iterations, illustrating the fast convergence of the proposed optimization algorithm.

5 Conclusion

In this paper, we proposed a novel hyper-graph-based sparse canonical correlation analysis (HGSCCA) framework for the diagnosis of Alzheimer's disease using multi-dimensional biological data. The proposed HGSCCA algorithm is within generalized sparse canonical correlation analysis framework, but we extended it by introducing a hyper-graph-based regularization term to capture the high-order correlation among different subjects. Different from the existing multi-modal learning algorithms, HGSCCA can incorporate the distribution knowledge of multi-modal data to find more discriminant features with biological meanings. The classification results on ROSMAP dataset demonstrated that HGSCCA can achieve significantly better classification performance for the diagnosis of AD and MCI than the currently available methods that also handle multi-modal genomic data. In the future, we plan to incorporate more data types (e.g., imaging data, metabolomic data or proteomic data) to the HGSCCA algorithm to further refine the search for biomarkers in the diagnosis of AD and predict the disease stage. In addition, in this study, we used the least square loss to calculate the empirical risk on the training set, which may not fully reflect the complicated statistical structures of the multi-modal genomic data. In the future work, we will explore potentially more accurate loss function to further improve the AD classification performance.

Acknowledgment

This work is supported by the IU Precision Health Initiative (Kun Huang, Jie Zhang) and NIH U54 AG065181 grant (Kun Huang, Jie Zhang). We also acknowledge the computing support from the IU Supercomputers.

Reference

[1]. Yang L-B, Lindholm K, Yan R, Citron M, Xia W, Yang X-L, et al. , "Elevated β-secretase expression and enzymatic activity detected in sporadic Alzheimer disease," Nature Medicine, vol. 9, pp. 3–11, 2003.

[2]. Brookmeyer R, Johnson E, Ziegler-Graham K, and Arrighi HM, "Forecasting the global burden of Alzheimer's disease," Alzheimer's & Dementia, vol. 3, pp. 186–191, 2007.

[3]. Colangelo V, Schurr J, Ball MJ, Pelaez RP, Bazan NG, and Lukiw WJ, "Gene expression profiling of 12633 genes in Alzheimer hippocampal CA1: transcription and neurotrophic factor down - regulation and up - regulation of apoptotic and pro inflammatory signaling," Journal of Neuroscience Research, vol. 70, pp. 462–473, 2002. [PubMed: 12391607]

[4]. Lanke V, Moolamalla S, Roy D, and Vinod P, "Integrative analysis of hippocampus gene expression profiles identifies network alterations in ageing and Alzheimer's disease," Frontiers in Aging Neuroscience, vol. 10, pp. 153–163, 2018. [PubMed: 29875655]

[5]. Xu H, Yue C, and Chen L, "Post-Transcriptional Regulation of Soluble Guanylate Cyclase that Governs Neuropathic Pain in Alzheimer's Disease," Journal of Alzheimer's Disease, pp. 1–8, 2019.

[6]. Hansson O, Zetterberg H, Buchhave P, Londos E, Blennow K, and Minthon L, "Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study," Lancet Neurology, vol. 5, pp. 22834, 2006.

[7]. Cooper-Knock J, Kirby J, Ferraiuolo L, Heath PR, Rattray M, and Shaw PJ, "Gene expression profiling in human neurodegenerative disease," Nature Reviews Neurology, vol. 8, p. 518, 2012. [PubMed: 22890216]

[8]. Saugstad J, Wiedrick J, Lapidus J, Sandau U, Lusardi T, Harrington C, et al. , "Validation of human cerebrospinal fluid microRNAs as biomarkers for Alzheimer's disease," Journal of Extracellular Vesicles, vol. 7, pp. 199–199, 2018.

[9]. Silvestro S, Bramanti P, and Mazzon E, "Role of miRNAs in Alzheimer's disease and possible fields of application," International Journal of Molecular Sciences, vol. 20, p. 3979, 2019. [PubMed: 31443326]

[10]. Qazi TJ, Quan Z, Mir A, and Qing H, "Epigenetics in Alzheimer's disease: perspective of DNA methylation," Molecular Neurobiology, vol. 55, pp. 1026–1044, 2018. [PubMed: 28092081]

[11]. Smith AR, Smith RG, Pishva E, Hannon E, Roubroeks JA, Burrage J, et al. , "Parallel profiling of DNA methylation and hydroxymethylation highlights neuropathology-associated epigenetic variation in Alzheimer's disease," Clinical Epigenetics, vol. 11, pp 52–62, 2019. [PubMed: 30898171]

[12]. Agarwal S, Ghanty P, and Pal NR, "Identification of a small set of plasma signalling proteins using neural network for prediction of Alzheimer's disease," Bioinformatics, vol. 31, pp. 2505–2513, 2015. [PubMed: 25819077]

[13]. Li X, Wang H, Long J, Pan G, He T, Anichtchik O, et al. , "Systematic analysis and biomarker study for Alzheimer's disease," Scientific Reports, vol. 8, pp. 173–194, 2018. [PubMed: 29317698]

[14]. Mukherjee S, Perumal T, Daily K, Sieberts S, Omberg L, Preuss C, et al. , "Identifying and ranking potential driver genes of Alzheimer's Disease using multi-view evidence aggregation," BioRxiv, pp. 534–545, 2019.

[15]. Zhang L and Huang X, "Multiple SVM-RFE for multi-class gene selection on DNA Microarray data," In International Joint Conference on Neural Networks, pp. 1–6, 2005.

[16]. Zhang J, Xiang Y, Ding L, Borlawsky TB, Ozer HG, Jin R, et al. , "Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia," BMC Bioinformatics, pp. 5–11, 2010, [PubMed: 20047655]

[17]. Han Y, Ye X, Wang C, Liu Y, Zhang S, Feng W, et al. , "Integration of molecular features with clinical information for predicting outcomes for neuroblastoma patients," Biology Direct, vol. 14, pp. 1–16, 2019. [PubMed: 30646935]

[18]. Han Y, Ye X, Cheng J, Zhang S, Feng W, Han Z, et al. , "Integrative analysis based on survival associated co-expression gene modules for predicting Neuroblastoma patients' survival time," Biology Direct, vol. 14, pp. 4–11, 2019. [PubMed: 30760313]

[19]. Wang L and Liu Z-P, "Detecting Diagnostic Biomarkers of Alzheimer's Disease by Integrating Gene Expression Data in Six Brain Regions," Frontiers in Genetics, vol. 10, pp. 157-157, 2019. [PubMed: 30915100]

[20]. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. , "A blood based 12-miRNA signature of Alzheimer disease patients," Genome biology, vol. 14, pp. 78–84, 2013.

[21]. Fransquet PD and Ryan J, "Micro RNA as a potential blood-based epigenetic biomarker for Alzheimer's disease," Clinical biochemistry, vol. 58, pp. 5–14, 2018. [PubMed: 29885309]

[22]. Dong H, Li J, Huang L, Chen X, Li D, Wang T, et al. , "Serum microRNA profiles serve as novel biomarkers for the diagnosis of Alzheimer's disease," Disease Markers, vol. 2015, 2015.

[23]. Li P, Marshall L, Oh G, Jakubowski JL, Groot D, He Y, et al. , "Epigenetic dysregulation of enhancers in neurons is associated with Alzheimer's disease pathology and cognitive symptoms," Nature communications, vol. 10, pp. 2246–2261, 2019.

[24]. Vasanthakumar A, Davis JW, Idler K, Kim S, Nho K, Nudelman KN, et al. , "DNA methylation dynamics in Alzheimer's disease diagnosis and progression," Alzheimer's & Dementia: The Journal of the Alzheimer's Association, vol. 13, pp.297–287, 2017.

[25]. Bihlmeyer NA, Merrill E, Lambert Y, Srivastava GP, Clark TW, Hyman BT, et al. , "Novel methods for integration and visualization of genomics and genetics data in Alzheimer's disease," Alzheimer's & Dementia, 2019.

[26]. Humphries CE, Kohli MA, Nathanson L, Whitehead P, Beecham G, Martin E, et al. , "Integrated whole transcriptome and DNA methylation analysis identifies gene networks specific to late-onset Alzheimer's disease," Journal of Alzheimer's Disease, vol. 44, pp. 977–987, 2015.

[27]. Jiang W, Zhang Y, Meng F, Lian B, Chen X, Yu X, et al. , "Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease," Bioinformatics, vol. 29, pp. 2596–2602, 2013. [PubMed: 23990414]

[28]. Shao W, Han Z, Cheng J, Cheng L, Wang T, Sun L, et al. , "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," IEEE Transactions on Medical Imaging, vol. 39, 2019.

[29]. Madabhushi A, Agner S, Basavanhally A, Doyle S, and Lee G, "Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data," Computerized Medical Imaging and Graphics, vol. 35, pp. 506–514, 2011. [PubMed: 21333490]

[30]. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. , "SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer," Frontiers in Genetics, vol. 10, pp. 166–176, 2019. [PubMed: 30906311]

[31]. Ye J, Farnum M, Yang E, Verbeeck R, Lobanov V, Raghavan N, et al. , "Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data," BMC Neurology, vol. 12, p. 46–56, 2012. [PubMed: 22731740]

[32]. Varatharajah Y, Ramanan VK, Iyer R, and Vemuri P, "Predicting Short-term MCI-to-AD Progression Using Imaging, CSF, Genetic Factors, Cognitive Resilience, and Demographics," Scientific Reports, vol. 9, pp. 2235–2245, 2019. [PubMed: 30783207]

[33]. Xu C, Tao D, and Xu C, "Multi-view intact space learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, pp. 2531–2544, 2015. [PubMed: 26539856]

[34]. Liu M, Zhang D, and Shen D, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," IEEE Transactions on Medical Imaging, vol. 35, pp. 1463–1474, 2016. [PubMed: 26742127]

[35]. Jie B, Zhang D, Cheng B, and Shen D, "Manifold regularized multitask feature learning for multimodality disease classification," Human Brain Mapping, vol. 36, pp. 489–507, 2015. [PubMed: 25277605]

[36]. Zu C, Jie B, Liu M, Chen S, Shen D, Zhang D, et al. , "Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment," Brain Imaging and Behavior, vol. 10, pp. 1148–1159, 2016. [PubMed: 26572145]

[37]. Lei B, Chen S, Ni D, and Wang T, "Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion," Frontiers in Aging Neuroscience, vol. 8, pp. 77–84, 2016. [PubMed: 27242506]

[38]. A Bennett D, A Schneider J, Arvanitakis Z, and S Wilson R, "Overview and findings from the religious orders study," Current Alzheimer Research, vol. 9, pp. 628–645, 2012. [PubMed: 22471860]

[39]. De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, et al. , "A multiomic atlas of the human frontal cortex for aging and Alzheimer's disease research," Scientific Data, vol. 5, pp. 121–142, 2018.

[40]. Storey JD, Taylor JE, and Siegmund D, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 66, pp. 187–205, 2004.

[41]. Shao W, Sun L, and Zhang D, "Deep active learning for nucleus classification in pathology images," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 199–202, 2018

[42]. Zhou D, Huang J, and Schölkopf B, "Learning with hypergraphs: Clustering, classification, and embedding," in Advances in neural information processing systems, pp. 1601–1608, 2007.

[43]. Gao Y, Wang M, Tao D, Ji R, and Dai Q, "3-D object retrieval and recognition with hypergraph analysis," IEEE Transactions on Image Processing, vol. 21, pp. 4290–4303, 2012. [PubMed: 22614650]

[44]. Chen J, Bardes EE, Aronow BJ, and Jegga AG, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," Nucleic Acids Research, vol. 37, pp. 305–311, 2009.

[45]. Roussotte FF, Gutman BA, Hibar DP, Jahanshad N, Madsen SK, Jack CR Jr, et al. , "A single nucleotide polymorphism associated with reduced alcohol intake in the RASGRF2 gene predicts larger cortical volumes but faster longitudinal ventricular expansion in the elderly," Frontiers in Aging Neuroscience, vol. 5, p. 93–99, 2013. [PubMed: 24409144]

[46]. Wilhelmus M, Otte-Höller I, Wesseling P, De Waal R, Boelens W, and Verbeek M, "Specific association of small heat shock proteins with the pathological hallmarks of Alzheimer's disease brains," Neuropathology and Applied Neurobiology, vol. 32, pp. 119–130, 2006. [PubMed: 16599941]

[47]. Bolós M, Hu Y, Young KM, Foa L, and Small DH, "Neurogenin 2 mediates amyloid-β precursor protein-stimulated neurogenesis," Journal of Biological Chemistry, vol. 289, pp. 31253–31261, 2014. [PubMed: 25217641]

[48]. Pichler S, Gu W, Hartl D, Gasparoni G, Leidinger P, Keller A, et al. , "The miRNome of Alzheimer's disease: consistent downregulation of the miR-132/212 cluster," Neurobiology of Aging, vol. 50, pp. 167–174 2017.

[49]. Masliah E, Xie F, Dayan S, Rockenstein E, Mante M, Adame A, et al. , "Genetic deletion of Nogo/Rtn4 ameliorates behavioral and neuropathological outcomes in amyloid precursor protein transgenic mice," Neuroscience, vol. 169, pp. 488–494, 2010. [PubMed: 20433905]

[50]. Ferrarese C, Appollonio I, Frigo M, Meregalli S, Piolti R, Tamma F, et al. , "Cerebrospinal fluid levels of diazepam - binding inhibitor in neurodegenerative disorders with dementia," Neurology, vol. 40, pp. 632-632, 1990. [PubMed: 2138718]

[51]. Fernando F, Keijser R, Henneman P, van der Kevie A-MF, Mannens MM, van der Post JA, et al. , "The idiopathic preterm delivery methylation profile in umbilical cord blood DNA," BMC Genomics, vol. 16, pp. 736–745, 2015. [PubMed: 26419829]

[52]. Paterson R, Bartlett J, Blennow K, Fox N, Shaw L, Trojanowski J, et al. , "Cerebrospinal fluid markers including trefoil factor 3 are associated with neurodegeneration in amyloid-positive individuals," Translational Psychiatry, vol. 4, pp. e419–e425, 2014. [PubMed: 25072324]

[53]. Zeng Z, Liu Y, Zheng W, Liu L, Yin H, Zhang S, et al. , "MicroRNA-129–5p alleviates nerve injury and inflammatory response of Alzheimer's disease via downregulating SOX6," Cell Cycle, vol. 18, pp. 3095–3110, 2019. [PubMed: 31564203]

[54]. Gorski J, Pfeuffer F, and Klamroth K, “Biconvex sets and optimization with biconvex functions: a survey and extensions,” Mathematical Methods of Operations Research, vol. 66, pp. 373–407, 2007.

1. Proposed a novel integration method to combine multi-modal genomic data for the diagnosis of AD disease.
2. Implement the multi-modal feature selection method under the generalized canonical correlation analysis framework, which can take the correlation among multi-modal data into consideration.
3. Consider the high-order correlations among different subjects and introduce a hypergraph-based regularization term that will lead to the selection of more discriminative biomarkers.
4. Explain the identified biomarkers, which validate the advantage of the proposed HGSCCA method that can find biologically meaningful multi-dimensional genomic biomarkers for AD disease diagnosis.

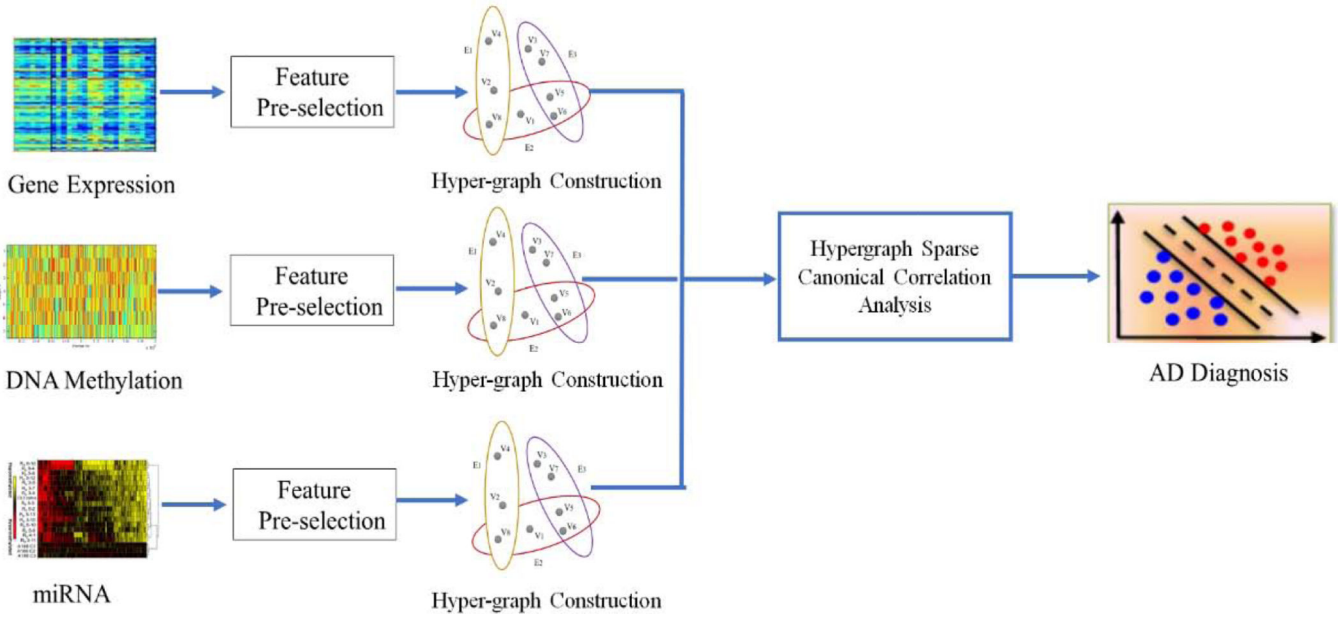
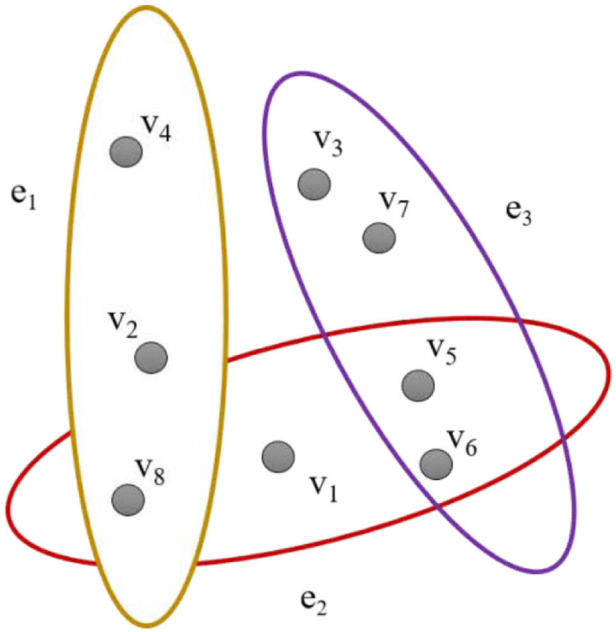


Fig. 1.
The flowchart of HGSCCA method.



	e1	e2	e3
v1	0	1	0
v2	1	0	0
v3	0	0	1
v4	1	0	0
v5	0	1	1
v6	0	1	1
v7	0	0	1
v8	1	1	0

Fig. 2.

An illustration of hyper-graph representation. Left: an example of a hyper-graph, where the vertices involved in the same hyper-edge are encircled by an ellipse, and different hyperedges are represented by the ellipses with different colors. Right: incidence matrix of the hypergraph in the left part (i.e.,H). The entry $H(v_i, e_j)$ is set to 1 if the hyper-edge e_j contains the vertex v_i, otherwise $H(v_i, e_j)$ equals 0.

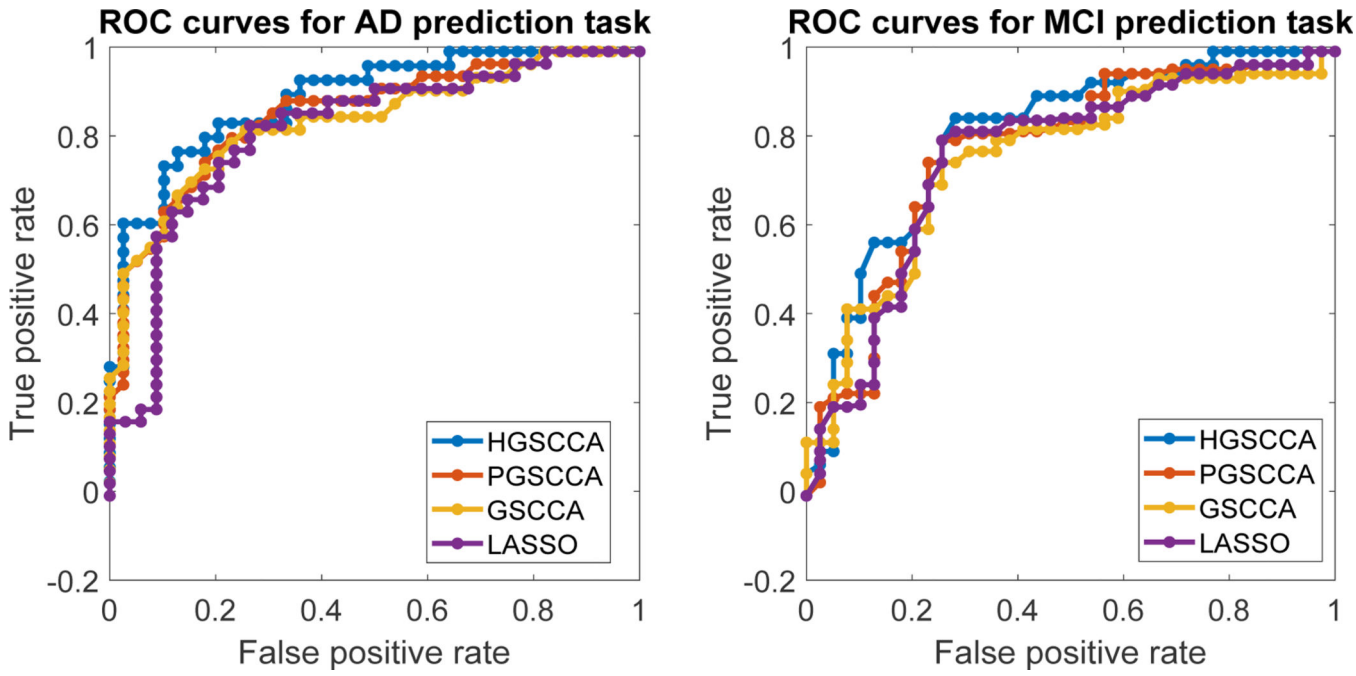


Fig. 3.
The ROC curves of AD vs NC and MCI vs NC tasks with four different feature selection methods.

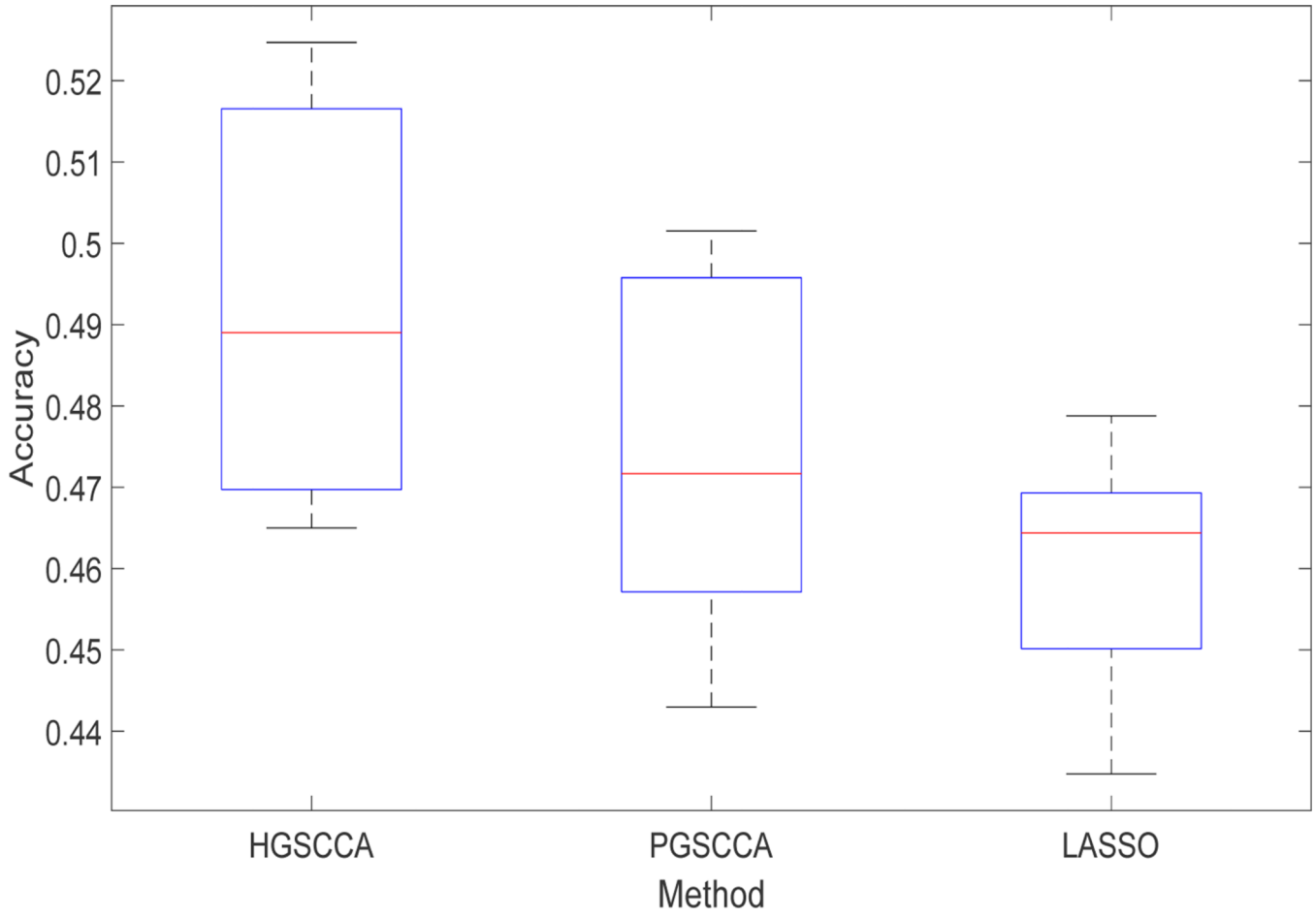


Fig. 4.
The classification accuracy for the Braak Stage Prediction task.

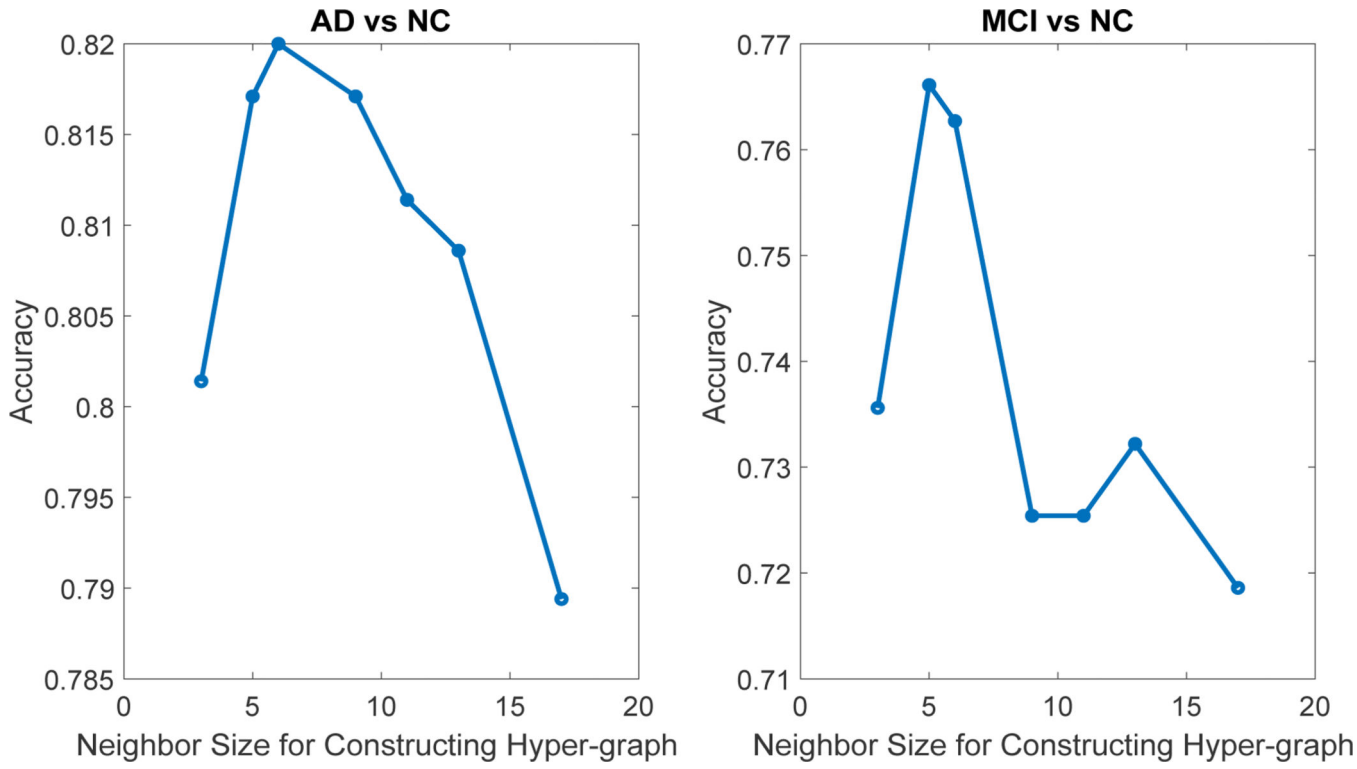


Fig. 5.
The effects of the neighbor size for the construction of the hyperedge.

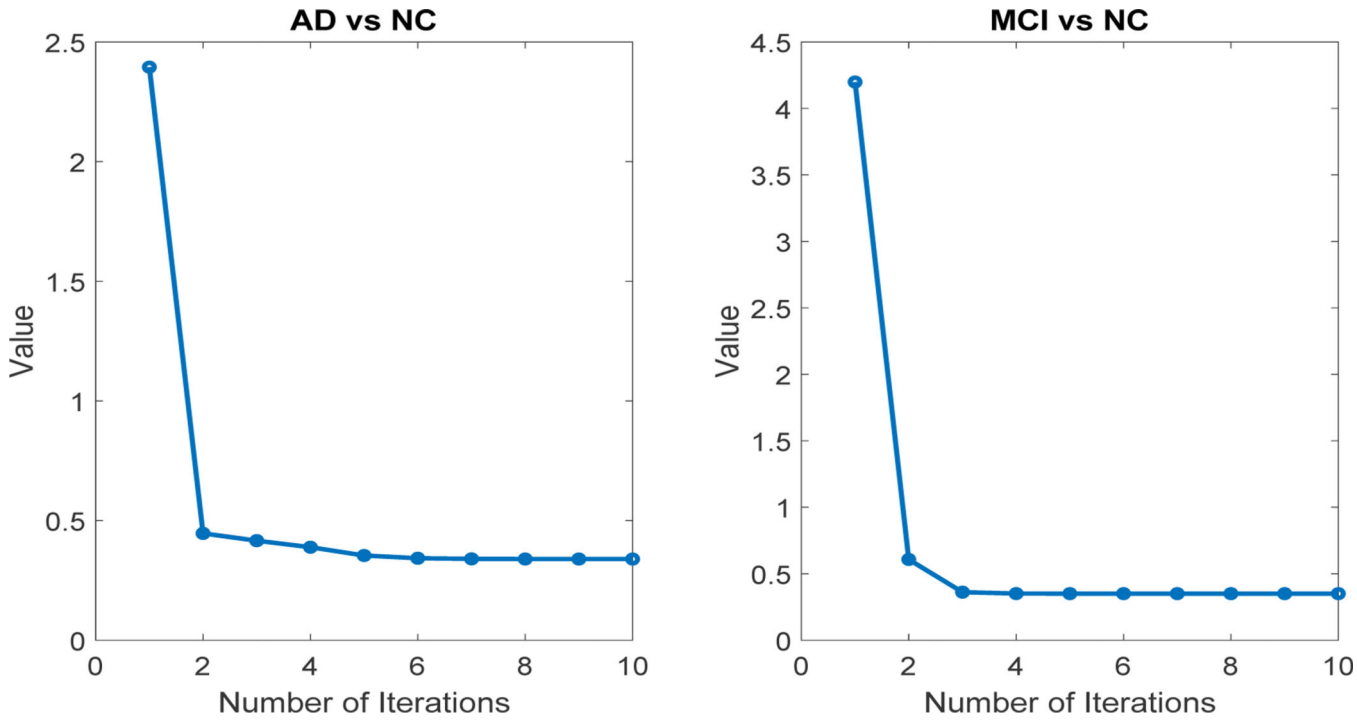


Fig. 6:

Change of the proposed objective function values achieved by our method on both AD vs NC and MCI vs NC tasks.

Table 1.

Demographic and clinical information for the participants in this study.

Diagnosis	**NC**	**MCI**	**AD**
Subject Number	168	125	182
Male/Female	72/96	57/68	81/101
Education (Mean ± std)	15.9 ± 4.2	16.4 ± 3.7	16.3 ± 3.8
Age (Mean ± std)	82.7 ± 4.7	84.6 ± 3.9	85.8 ± 3.6

Table 2.

The optimization algorithm for the proposed HGSCCA method.

The optimization procedure for HGSCCA metdod
Input:
$X_i(i \in G, D, M)$ % The data matrix for the multi-modal data.
Initialize:
$w_i = 0 \quad (i \in G, D, M), Q = 0$ % %Initialized as zero vector
$k = 0$ % Iteration step is initialized as zero.
While no convergence and $k \leq 10$:
Update $w_i(i \in G, D, M)$ according to Eq. (12)
Update Q according to Eq. (7)
End While
Output:
$w_i(i \in G, D, M)$ % The non-zero elements in w_i correspond to the selected features in each modality

Table 3.

Number of the pre-selected feature in each fold and the common features that are selected in most cases for MCI vs NC classification task.

	Methylation	RNA	miRNA
Fold 1	121	106	20
Fold 2	85	68	20
Fold 3	204	81	18
Fold 4	100	72	20
Fold 5	74	79	18
Common Feature	77	68	18

Table 4.

Number of the pre-elected feature in each fold and the common features that are selected in most cases for AD vs NC classification task.

	Methylation	RNA	miRNA
Fold 1	144	126	44
Fold 2	140	153	49
Fold 3	155	138	46
Fold 4	197	183	45
Fold 5	85	94	44
Common Feature	112	126	28

Table 5.

The average classification performance comparison (%) for AD vs. NC.

	ACC	SEN	SPE	PPV	NPV	AUC
Baseline	73.7(6.13)	72.2(7.34)	77.3(5.74)	74.5(7.34)	70.2(5.93)	77.7(6.34)
SLASSO	69.4(4.62)	73.4(4.89)	64.2(8.98)	68.7(5.47)	72.9(4.78)	74.3(6.78)
LASSO	77.6(6.24)	79.5(4.78)	73.7(7.75)	74.3(5.43)	78.9(4.92)	78.9(5.47)
GSCCA	78.4(5.79)	76.0(4.72)	**81.7(6.36)**	79.1(4.43)	74.7(6.77)	80.9(6.53)
PGSCCA	80.2(4.57)	81.1(2.57)	77.5(1.67)	79.5(3.01)	81.4(3.34)	83.7(3.03)
HGSCCA	**82.0(3.17)**	**83.7(3.06)**	80.3(3.31)	**80.4(3.36)**	**83.6(3.29)**	**86.2(2.94)**

Table 6.

The average classification performance comparison (%) for MCI vs. NC.

	ACC	SEN	SPE	PPV	NPV	AUC
Baseline	69.8(6.52)	71.5(7.13)	65.2(5.36)	66.4(6.83)	74.7(6.74)	70.4(5.63)
SLASSO	65.4(4.45)	69.7(5.78)	63.4(4.36)	62.7(3.68)	67.4(3.99)	65.7(4.12)
LASSO	71.7(4.36)	73.1(4.32)	67.4(3.57)	68.7(5.63)	80.3(5.42)	72.7(2.72)
GSCCA	70.4(5.74)	73.6(5.77)	66.8(4.65)	68.3(6.34)	81.3(4.72)	72.5(4.96)
PGSCCA	73.1(4.32)	77.1(4.21)	72.1(7.75)	**69.3(5.60)**	83.3(4.11)	73.5(6.45)
HGSCCA	**76.3(3.69)**	**80.0(3.54)**	**74.3(4.81)**	64.1(3.98)	**87.9(4.91)**	**76.5(3.01)**

Table 7.

Identified biomarkers for AD vs.NC task

DNA Methylation	CCND1, ESM1, RASGRF2, EFEMP2, NEUROG2, TP6AP1L, PCDH12, NGFR, SUSD4
RNA	PPDPF, CCDC69, PRELP, ARRDC2, CARHSP1, CSRP1, CCKBR, TMEM102, HSPB2, SLC25A48, CARHSP1, ITPKB, MRGPRF, HSPA2, DYNC1LI2, CEP83, TMEM107
miRNA	miR-129-5p, miR-132, miR-99a, miR-744, let-7i

Table 8.

Identified biomarkers for MCI vs. NC task

DNA Methylation	GANAB, FAM176A, C19orf25, ZNF300, PLEKHB1, MPRIP, COL5A3, DBI, RERE, ZNF432, OR5D18, OR5D18, PLVAP, LINC02405, TMEM41A, CDMR, CCDC6, RDMR.
RNA	PLEKHB1, GUSBP4, TCEA3, PRRG3, HMGB1P5, PDK1, CAMP, SPACA6, CRNDE, PINLYP, GAPDHP1, SWSAP1, CLC5A11, MORC1, LRRC1, QDPR, MT1F, HMBS, TFF3.
miRNA	miR-34b, miR-95, miR-129-5p, miR-362-3p, miR-580, miR-129-3p, miR-146b-5p, miR-K12-2, miR-615-5p, miR-203

Table 9.

The classification results for AD vs NC and MCI and NC tasks with different loss function. (ACC=Accuracy, SEN=Sensitivity, SPE= Specificity, AUC=Area under the curve, LSL=Least square loss. LGL=Logistic regression loss)

	AD vs NC				MCI vs NC			
	ACC	**SEN**	**SPE**	**AUC**	**ACC**	**SEN**	**SPE**	**AUC**
LSL	82.0	83.7	80.3	86.2	76.3	80.0	74.3	76.5
LGL	83.1	82.1	82.9	87.1	75.2	82.3	73.4	76.1