Multi-Modal Diagnosis of Alzheimer's Disease using Interpretable Graph Convolutional Networks

Houliang Zhou, Lifang He, Brian Y. Chen, Li Shen, Senior Member, IEEE, and Yu Zhang, Senior Member, IEEE

***Abstract*— The interconnection between brain regions in neurological disease encodes vital information for the advancement of biomarkers and diagnostics. Although graph convolutional networks are widely applied for discovering brain connection patterns that point to disease conditions, the potential of connection patterns that arise from multiple imaging modalities has yet to be fully realized. In this paper, we propose a multi-modal sparse interpretable GCN framework (SGCN) for the detection of Alzheimer's disease (AD) and its prodromal stage, known as mild cognitive impairment (MCI). In our experimentation, SGCN learned the sparse regional importance probability to find signature regions of interest (ROIs), and the connective importance probability to reveal disease-specific brain network connections. We evaluated SGCN on the Alzheimer's Disease Neuroimaging Initiative database with multi-modal brain images and demonstrated that the ROI features learned by SGCN were effective for enhancing AD status identification. The identified abnormalities were significantly correlated with AD-related clinical symptoms. We further interpreted the identified brain dysfunctions at the level of large-scale neural systems and sex-related connectivity abnormalities in AD/MCI. The salient ROIs and the prominent brain connectivity abnormalities interpreted by SGCN are considerably important for developing novel biomarkers. These findings contribute to a better understanding of the network-based disorder via multi-modal diagnosis and offer the potential for precision diagnostics. The source code is available at https://github.com/Houliang-Zhou/SGCN.**

***Index Terms*— Computer aided analysis, graph convolutional network, multi-modality, neuroimaging, sparse interpretation**

Manuscript received 3 March 2024. This work was supported in part by the National Institutes of Health under Grant U01AG068057, Grant U01AG-066833, Grant R01LM013463, Grant R01MH129694, Grant R21AG080425, and Grant R21MH130956, in part by Alzheimer's Association under Grant AARG-22-972541, and in part by Lehigh's grants under Accelerator S00010293, CORE 001250, and FIG FIGAWD35. Lifang He is partially supported by the NSF Grants (MRI-2215789, IIS-1909879, IIS-2319451), NIH Grant under R21EY034179, and Lehigh's Grants under Accelerator and CORE. (Corresponding author: Yu Zhang)

Houliang Zhou, Lifang He, and Brian Y. Chen are with the Department of Computer Science and Engineering, Lehigh University, PA 18015, USA (e-mail:hoz421@lehigh.edu, lih319@lehigh.edu, byc210@lehigh.edu).

Li Shen is with the Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, PA 19104, USA (e-mail:li.shen@pennmedicine.upenn.edu).

Yu Zhang is with the Department of Bioengineering and the Department of Electrical and Computer Engineering, Lehigh University, PA 18015, USA (e-mail: yuzi20@lehigh.edu).

I. Introduction

Neuroimaging-based diagnostics have demonstrated recent advances in predicting Alzheimer's disease (AD) and mild cognitive impairment (MCI) from multi-modal brain images, such as magnetic resonance imaging (MRI) and positron emission tomography (PET) scans [1]. In disease diagnosis, MRI images can detect structural changes in the brains of AD/MCI patients [2]. In contrast, the fluorodeoxyglucose PET, and florbetapir PET measure separately the metabolic abnormality or pathological process of their brains [3]. Thus, it is crucial to combine the contribution of all these modalities in a multi-modal analysis for the identification of AD/MCI. Recent neuroimaging studies have reached an agreement that the interactions between brain regions are the essential driving factor for neural development and neurological disorder analysis [4]. Large improvements in understanding the brain's organization have been made by representing the brain as a connectivity graph to describe the interactions between regions [4]. In this graph, nodes are defined as brain regions of interest (ROIs), and edges as the connections between ROIs. This representation is compatible with graph convolutional networks (GCNs) model with demonstrated capabilities for analyzing graph-structured data [5].

In brain imaging, GCNs have shown significant promise in finding abnormalities in brain connectivity and in discovering biomarkers for various mental disorders [6]–[9]. In recent years, the importance of explainable artificial intelligence (XAI) has been increasingly recognized in mental health to clarify the mechanism underlying the association between neural circuits and behavior/cognition [10], [11]. In medical diagnosis, the explainability of GCN predictions is crucial for helping identify biomarkers that contribute to the brain disorder. For example, Yang et al. applied an Edge-weighted Graph Attention Network with dense hierarchical pooling to understand the derivation of Bipolar disorder [12]. Cui et al. designed a globally explainable generator to highlight disorder-specific biomarkers related to the disorder [7]. Li et al. proposed BrainGNN with ROI-aware graph convolutional layers to analyze functional MRI for neurological biomarker prediction [4]. Although several approaches have been recently proposed to explain the GCN model [4], [6], [13], [14], most of them have focused only on data from a single modality.

The existing multi-modal GCN method, proposed by Zhang et al. [15], concatenates multi-modal features for disease prediction, limiting the multi-modal interpretation of salient ROIs and the most discriminative connections. Overall, recent methods for interpreting brain networks have applied limited consideration to multi-modal regional features and their connections in brain network-based disease analysis, even though recent studies have indicated that different imaging modalities provide essential complementary information that can improve accuracy in disease diagnosis [16]–[20]. We argue here that multi-modal interpretations create an improved opportunity for identifying salient ROIs and discovering prominent brain network connections related to AD and MCI. Given that ROIs can be partitioned into different neural systems based on their structural and functional roles [21], the neural system-level connectivity abnormalities via multi-modal analysis can facilitate the discovery of novel neurological biomarkers.

In this paper, we present a multi-modal sparse interpretable GCN framework (SGCN) for detecting AD and for explaining AD pathology as it relates to individual brain regions, connections, and neural systems. An overview of the multi-modal SGCN model for Alzheimer's diagnosis and biomarker interpretation is shown in Fig. 1. The innovation of SGCN is listed as follows: 1) SGCN is the first to introduce the importance probability to detect the salient ROIs and the most prominent subgraph structure to discriminate subjects between HC, AD, and MCI, which exhibited superior prediction performance. 2) SGCN provides interpretability of both brain regions and brain connectivity through the importance probability technique, which is confirmed by extensive statistical analyses of the learned topological patterns. We observed that these patterns correlate significantly with typical AD-related clinical measures including Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Score 13 (ADAS13), and Clinical Dementia Rating Scale Sum of Boxes (CDR-SOB). 3) SGCN further identifies biomarkers that are correlated to connectivity abnormalities in neural systems, to disease progression and to sex-related differences in AD/MCI. We observed that SGCN rediscovered multiple established findings relating to these applications, as well as several new ones. Altogether, these results point to potential applications of our SGCN method for identifying novel biomarkers and brain network connectivity abnormalities from multi-modal brain images.

We have demonstrated the prediction ability of the sparse interpretable GCN method originally presented at the MICCAI conference to distinguish AD from HC [22]. In this paper, we advance the original work through extensive experimental analyses: first, an identification of disease-related ROIs and of brain connectivity abnormalities. Second, a comparison and interpretation of neural system-level and sex-related abnormalities in brain connectivity, as observed in multiple modalities. Third, a statistical investigation of the predictability of circuit abnormalities for AD symptoms. Fourth, a prediction evaluation of generalizability on ADNI-2/GO and independent ADNI-1 test set. Finally, we describe a method for the multi-modal diagnosis of MCI and the progression from MCI into AD.

II. Methods

A. Notations

We parcellate the entire brain into N ROIs based on the automated anatomical labeling (AAL) atlas [23]. Multiple modalities are concatenated into the ROI's feature vector. We define a brain adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and node feature matrix $\boldsymbol{X} \in \mathbb{R}^{N\times D}$, where N denotes the number of ROIs and D denotes the dimension of multi-modal features. Given each ROI is considered as a node, we viewed the brain connectivity graph as an undirected weighted graph $G = (V, E)$. In this graph, the vertex set $V = \{v_1, \cdots, v_N\}$ is composed of ROIs in the brain. Meanwhile, the edge set E is composed of connections between ROIs, which are weighted by similar strength.

B. Brain graph construction

In order to deal with the noisy edges, we define the K-Nearest Neighbor (KNN) graph $\widetilde{G} = (V, \widetilde{E})$ from the multi-modal regional information [16], where K is the number of the nearest neighbor. In this KNN graph, the edges are weighted using the Gaussian similarity function based on Euclidean distances. Mathematically, this function can be written as $e(v_i, v_j) = \exp(-\frac{\|v_i - v_j\|^2}{2\sigma^2})$, where σ is the standard deviation of the Gaussian function and influences the sensitivity of the similarity measure. Here, N_i denotes the set of K-nearest neighbors of vertex v_i and N_j denotes the set of K-nearest neighbors of vertex v_j. We build the similarity function between vertex v_i and vertex v_j if $v_i \in N_j$ or $v_j \in N_i$. Finally, the weighted adjacency matrix $\boldsymbol{A}$ reflects the similarity between ROIs and their nearest similar neighbors. The elements of defined adjacency matrix $\boldsymbol{A}$ can be denoted as follows:

$$\boldsymbol{a}_{i,j} = \begin{cases} e(v_i, v_j), & \text{if } v_i \in N_j \text{ or } v_j \in N_i \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

C. Graph convolutional network

In the graph classification problem, the Graph Convolutional Network (GCN) can embed node-level features into a low dimensional space, and summarize them into graph-level features [5]. The summarized graph-level features are flattened into a feature vector, which is fed into a multilayer perceptron (MLP) classifier. Our architecture composes of three types of layers: graph convolutional layers, a node pooling layer, and an MLP layer. In our architecture, the graph convolutional layer recursively learns a node representation by transforming and aggregating the neighboring feature vectors. Mathematically, the propagation update of node representation in our SGCN model can be calculated as:

$$\boldsymbol{H}^{l+1} = \sigma(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{\boldsymbol{l}}\boldsymbol{W}^{l}) \tag{2}$$

where $\boldsymbol{H}^0 = \boldsymbol{X}$, $\boldsymbol{H}^l \in \mathbb{R}^{N\times d_l}$ is the output of the l^{th} graph convolution layer, d_l is the number of output channels at layer l. We add self-loops into the adjacency matrix $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, where $\boldsymbol{I} \in \mathbb{R}^{N\times N}$ is the identity matrix. In this equation, we define that $\boldsymbol{W}^l \in \mathbb{R}^{d_l \times d_{l+1}}$ are the learnable parameters, $\widetilde{\boldsymbol{D}}$ is the diagonal degree matrix with $\widetilde{\boldsymbol{D}}_{i,i} = \sum_j \widetilde{A}_{i,j}$, and

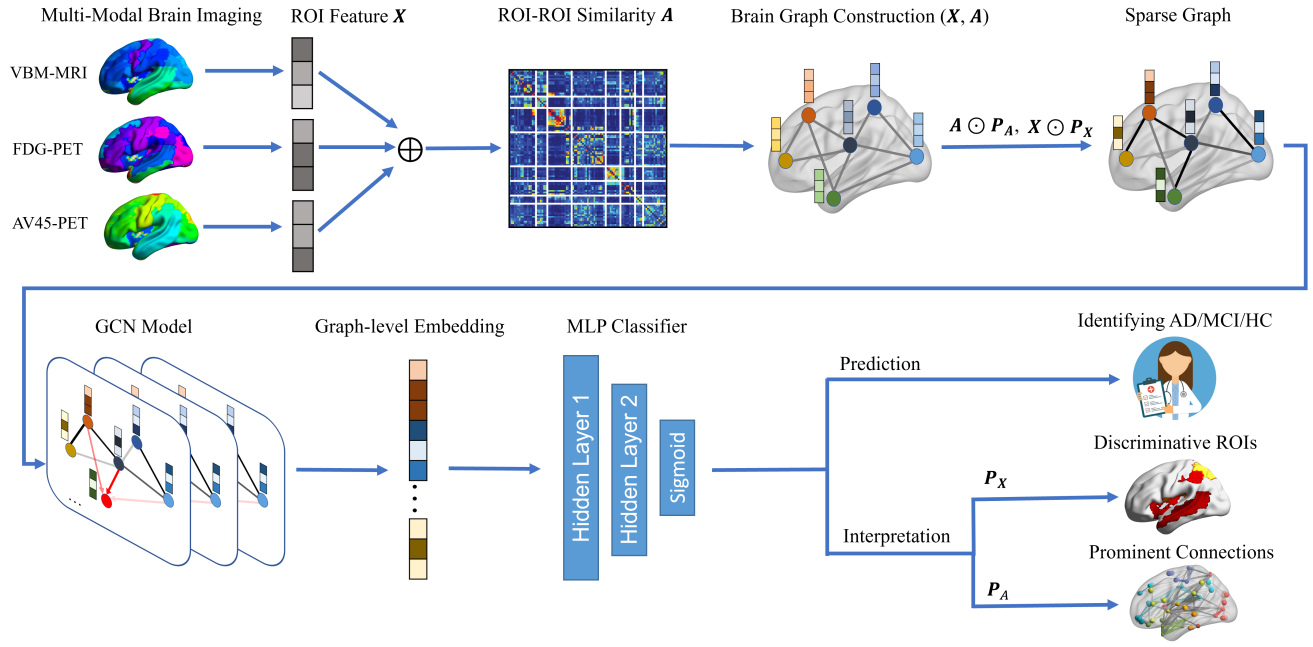


Fig. 1. An overview of the proposed SGCN model for Alzheimer's diagnosis and biomarker interpretation. The multi-modal brain images are converted to graphs by using the Gaussian similarity to construct the connections between ROIs. The graphs combined with the feature importance probability $\boldsymbol{P_X}$ and the edge importance probability $\boldsymbol{P_A}$ are sent to our proposed sparse GCN model to predict the disease. The importance probabilities $\boldsymbol{P_X}$ on nodes and $\boldsymbol{P_A}$ on edges provide the interpretation for the salient ROIs and the prominent disease-specific connections.

σ is the sigmoid function. Meanwhile, $\widetilde{\boldsymbol{A}}$ is normalized by multiplying $\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}$, which can keep a fixed feature scale after graph convolution layer.

After the graph convolution layer, the node pooling layer is applied to group the node-level features together to summarize the graph-level features. Next, the summarized output $\boldsymbol{H}^L$ of the graph convolution layer is flattened into a single feature vector, which is fed into an MLP classifier with a sigmoid function for the final classification. Finally, we apply the supervised cross-entropy loss function for disease prediction.

D. Sparse interpretability

1) Importance probabilities as the interpretation: Because the brain connectivity graph G and regional feature $\boldsymbol{X}$ may contain redundant or noisy information, the original graph G and feature matrix $\boldsymbol{X}$ into the GCN model $f(\cdot)$ are not highly beneficial for predicting disease. Hence, we hypothesize that an important subgraph $G_s \subseteq G$ and an important subset of multi-modal features $X_s = \{\boldsymbol{x_i} | v_i \in G_s\}$ contribute most to the disease prediction. In order to find such X_s and G_s, we propose to learn a shared multi-modal feature importance probability $\boldsymbol{P_X}$, and the individual edge importance probability $\boldsymbol{P_A}$ between nodes for each subject. Specifically, we define the important subgraph as $G_s = \boldsymbol{A} \odot \boldsymbol{P_A}$, and the important subset of multi-modal feature as $X_s = \boldsymbol{X} \odot \boldsymbol{P_X}$. Therefore, we mathematically expressed the final prediction output $\hat{y}$ of the GCN model $f(\cdot)$ as:

$$\hat{y} = f(\boldsymbol{A} \odot \boldsymbol{P_A}, \boldsymbol{X} \odot \boldsymbol{P_X}) \tag{3}$$

Generally, the problem of exploring the important subgraph and the important subset of node feature is translated into the inference of importance probability $\boldsymbol{P_A}$ on edges and $\boldsymbol{P_X}$ on nodes. The importance probabilities are applied to the individual brain network and multi-modal node feature across all subjects from HC, AD, and MCI.

Given that the different modalities of ROIs contribute differently to the disease prediction, we define the multi-modal feature importance probability $\boldsymbol{P_X} \in \mathbb{R}^{N \times D}$, where $\boldsymbol{P_X} = [\boldsymbol{p_1}, \boldsymbol{p_2}, ..., \boldsymbol{p_N}]$, and $\boldsymbol{p_i} \in \mathbb{R}^D$, $1 \leq i \leq N$, indicates the ROI's feature importance probability. Because the multi-modal node features are associated with the weight of their connections, we define the edge importance probability $\boldsymbol{P_A} \in \mathbb{R}^{N \times N}$ between node i and j by considering the joint connection between multi-modal node features $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$:

$$P_{A_{i,j}} = \sigma(\boldsymbol{v}^T[\boldsymbol{x_i} \odot \boldsymbol{p_i} || \boldsymbol{x_j} \odot \boldsymbol{p_j} || \boldsymbol{a}_{i,j}]) \tag{4}$$

where $\boldsymbol{p_i}$ is the feature importance probability from node i, $\boldsymbol{a}_{i,j}$ is the weight of edge between node i and j, $\boldsymbol{v} \in \mathbb{R}^{2D+1}$ denotes the learnable parameter, $\odot$ denotes the Hadamard element-wise product function, and $||$ denotes the concatenation function. The edge importance probability is calculated by incorporating multi-modal node features and their importance probabilities together. This mechanism is beneficial to discover the prominent connectivity abnormalities from the information of the multi-modal node features.

2) Loss function: In this section, we define the conditional entropy loss to determine the importance probabilities $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$, as well as the ℓ_1 loss and entropy regularization loss to promote sparsity on them. The interpretability of the GCN model is achieved by exploring the important subgraph G_s and the important subset of node feature X_s, which exhibit

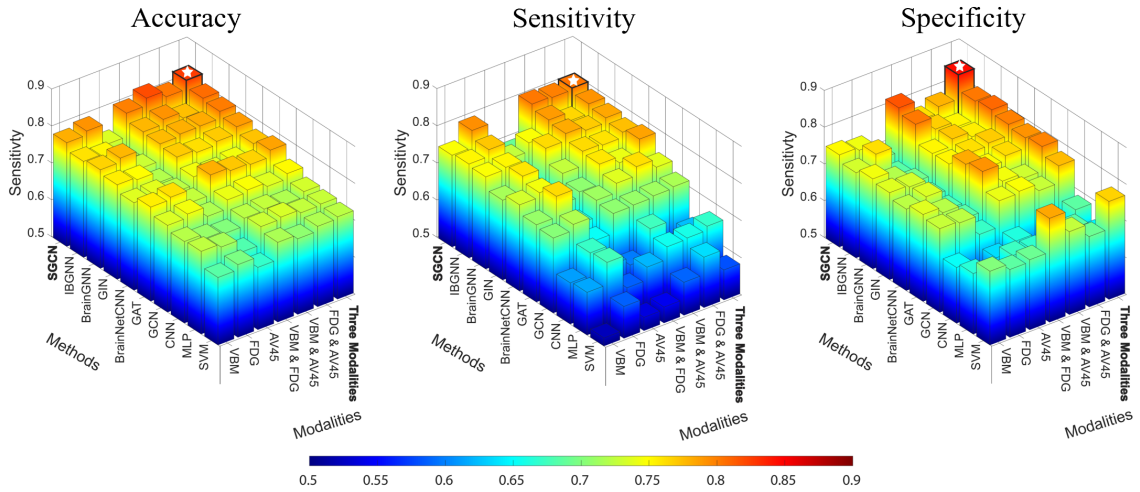


Fig. 2. Multiclass classification comparison between the state-of-the-art machine learning models and our proposed SGCN by using different modalities. The highest a) accuracy, b) sensitivity, and c) specificity labeled with a white star are 0.826, 0.804, and 0.845 respectively, which are achieved by our SCGN when using all three modalities.

the maximum mutual information with the distribution of truth labels. Specifically, we train the SGCN model and find the $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$ by maximizing the mutual information between the true label y from the original graph and the predictive output $\hat{y}$ learned from the G_s and X_s. Maximizing this mutual information is equivalent to minimizing conditional entropy [24]. Assuming there are C disease classes, the conditional entropy loss $\mathcal{L}_m$ is expressed as:

$$\mathcal{L}_m = -\sum_{c=1}^{C} \mathbb{1}[y=c] \log P_f(\hat{y}=y \mid G_s = \boldsymbol{A} \odot \boldsymbol{P_A}, X_s = \boldsymbol{X} \odot \boldsymbol{P_X}) \quad (5)$$

Our method minimized the conditional entropy loss to determine the importance probability $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$ for disease prediction. The ℓ_1 and entropy regularization were further applied to promote the sparsity on $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$. We define the ℓ_1 regularization on $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$ as:

$$\mathcal{L}_s = \|\boldsymbol{P_X}\|_1 + \|\boldsymbol{P_A}\|_1 \quad (6)$$

Meanwhile, element-wise entropy regularization is applied to encourage discreteness in the probability distribution:

$$\begin{aligned} \mathcal{L}_{P_A,e} &= -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{A_{i,j}}\log(P_{A_{i,j}}) + (1-P_{A_{i,j}})\log(1-P_{A_{i,j}}) \\ \mathcal{L}_{P_X,e} &= -\frac{1}{ND}\sum_{i=1}^{N}\sum_{m=1}^{D} P_{X_{i,m}}\log(P_{X_{i,m}}) + (1-P_{X_{i,m}})\log(1-P_{X_{i,m}}) \end{aligned} \quad (7)$$

where $\mathcal{L}_{P_A,e}$ denotes the entropy regularization on $\boldsymbol{P_A}$, and $\mathcal{L}_{P_X,e}$ denotes the entropy regularization on $\boldsymbol{P_X}$. The total entropy regularization is summarized as $\mathcal{L}_e = \mathcal{L}_{P_X,e} + \mathcal{L}_{P_A,e}$. Both ℓ_1 and entropy regularization serve to induce the sparsity on $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$, which encourage the probabilities of unimportant or noisy entries to approach zero. Simultaneously, the mechanism of entropy regularization ensures that important features and connections have higher probabilities closer to one, facilitating disease prediction.

After combing all of the loss functions, the final training objective of our SGCN can be expressed as:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1\mathcal{L}_m + \lambda_2\mathcal{L}_s + \lambda_3\mathcal{L}_e \quad (8)$$

where $\mathcal{L}_c$ is the supervised cross-entropy loss used for disease prediction, and λ's denote the tunable hyper-parameters serving as penalty coefficients for the various loss terms. The optimized solution on our objective function $\mathcal{L}$ is similar to the regular GCN model, with the exception of the learnable parameters $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$. In our result, the importance probability $\boldsymbol{P_X}$ and $\boldsymbol{P_A}$ learned from SGCN provide the interpretation regarding the salient ROIs and the prominent disease-related brain connectivity abnormalities.

III. Results

A. Dataset and preprocessing

In this work, we evaluated the SGCN framework on a multimodal brain imaging dataset from the public Alzheimer's Disease Neuroimaging Initiative (ADNI) [25], which consisted of three modal brain images including structural Magnetic Resonance Imaging (VBM-MRI), 18F-fluorodeoxyglucose Positron Emission Tomography (FDG-PET), and 18F-florbetapir PET (AV45-PET). These brain imaging data were gathered from 739 non-Hispanic Caucasian participants, including 172 HC subjects, 471 MCI subjects, and 96 AD subjects. All scans in this study meet all quality-controlled criteria described in [26]. In the MCI group, 142 MCI subjects progress into AD (pMCI) after more than one year while the rest 329 MCI subjects remain stable (sMCI).

In the preprocessing step, the multi-modal brain imaging scans were aligned to the corresponding visit of each participant. Specifically, all brain imaging scans were aligned to a T1-weighted template image. Subsequently, these scans were segmented into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) maps. They were then normalized to the standard Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2$ mm^3 voxels and further smoothed using an 8 mm full-width at half-maximum (FWHM) kernel. The structural MRI was preprocessed with voxel-based morphometry (VBM) and the FDG-PET and AV45-PET were registered to the MNI space by applying the SPM software [27]. The entire brain was subsampled to 90 ROIs (excluding the cerebellum and vermis) based on the automated anatomical labeling (AAL) atlas [23]. Finally, we summarized ROI-level measures by averaging all of the voxel-level measures within each ROI.

TABLE I
BINARY CLASSIFICATION COMPARISON BETWEEN THE STATE OF THE ART MACHINE LEARNING MODELS AND SGCN USING ALL MODALITIES UNDER SMCI VS. PMCI CONTRAST.

Methods	Accuracy	ROC-AUC	Sensitivity	Specificity
SVM	.562 ±.131	.625 ±.078	.531 ±.089	.642 ±.083
MLP	.607 ±.084	.645 ±.069	.582 ±.091	.667 ±.075
CNN	.589 ±.109	.638 ±.052	.579 ±.087	.668 ±.068
GCN	.635 ±.063	.669 ±.068	.594 ±.094	.671 ±.055
GAT	.657 ±.057	.682 ±.045	.605 ±.076	.676 ±.061
BrainNetCNN	.642 ±.071	.691 ±.073	.597 ±.069	.679 ±.064
GIN	.654 ±.048	.698 ±.065	.611 ±.071	.672 ±.078
BrainGNN	.673 ±.062	.714 ±.069	.607 ±.065	.691 ±.055
IBGNN	.669 ±.055	.708 ±.072	.614 ±.087	.683 ±.068
SGCN	.702 ±.041	.736 ±.065	.637 ±.072	.714 ±.059

B. Experimental setup

We trained and tested our proposed method on Pytorch framework by using a Nvidia RTX A5000 with 24GB GPU memory. The model architecture included the brain graph construction and GCN model learning as shown in Fig. 1. Because we used K-Nearest Neighbor (KNN) graphs to construct the brain connectivity graphs, we tested $K \in \{3, 5, 10, 15, 20, 25, 30, 35, 40\}$ to compare the classification performance. We noted that the smaller K is not enough to exploit the intrinsic neighborhood structure to identify the AD, and the larger K brings noisy information to affect the performance, which can be supported by both previous Graph Laplacian study [28] and a brain connectivity study [29] to apply $K = 10$ in KNN Graph to construct multiple sparse brain functional connectivity networks for various neuro-disease analyses. Therefore, we used $K = 10$ and $\sigma = 1$ to build the KNN graph in our experiment. After building the graph data, we used a GCN model with three graph convolutional layers followed by three fully-connected layers, a dropout layer, and a softmax classifier with parameters $N = 90$, $D = d_0 = 3$, $d_1 = 3$, $d_2 = 3$, and $C = 2$. The hidden dimensions of the three fully-connected layers are 16, 16, and 1 respectively. The dropout rate is 0.5. In the experiment, we use grid search to find the best hyperparameters and set λ_1 to 1.0, λ_2 to 0.1, and λ_3 to 0.1 in the loss terms of our method. We trained the SGCN for 100 epochs with a learning rate of 0.001. Adam was used as the learning optimizer. Each batch contained 32 graphs during training. Meanwhile, we performed the 5-fold cross validation to examine the performance.

C. Classification performance

In our experiments, three contrasts including HC vs. AD, HC vs. MCI, and MCI vs. AD were examined to evaluate classification results. We use the one-against-one strategy to classify three contrasts and compare the results via one, two, or three modalities. This experimental design measures our method's classification performance on different modalities since it is a multi-modal method, and on different contrasts. We performed the 5-fold cross validation to examine the classification performance. The average classification accuracy, the area under the receiver operating characteristic curve (ROC-AUC), sensitivity, specificity, and standard deviations are reported.

Fig. 2 shows the multiclass classification comparison between the state-of-the-art machine learning models and our proposed SGCN method via different modalities. Our method was compared with Support Vector Machine (SVM) using a Radical Basis Function (RBF) kernel, Convolutional Neural Network (CNN) models, and BrainNetCNN [30]. The SVM and CNN models utilized vectorized adjacency matrices as inputs, and BrainNetCNN utilized brain network correlation as inputs. In addition to the traditional machine learning methods, our method was also compared with other GCN-based methods including GCN [5], GAT [31], GIN [32], BrainGNN [4], and IBGNN [33]. In our result, after combining the VBM-MRI and FDG-PET modalities, the SGCN model achieves an accuracy, sensitivity, and specificity increase of 3%, 4%, and 6% respectively compared to using only the VBM-MRI modality. When using all three modalities, the accuracy, sensitivity, and specificity of our SGCN model increase by between 3% and 9%. The best accuracy is achieved by combining all different modalities, indicating that different imaging modalities can provide essential complementary information that can improve accuracy in disease diagnosis [16], [17]. Based on these classification results, multi-modal brain images were used to evaluate the interpretability.

To predict the conversion of MCI into AD, a binary classification experiment for sMCI vs. pMCI under the 5-fold cross validation was conducted. Table. I shows the binary classification comparison between the state-of-the-art machine learning models and SGCN using all modalities for sMCI vs. pMCI. The best classifying result was achieved by using our method. The accuracy and sensitivity of SGCN increase around 3% compared to SOTA brain network-based methods including BrainGNN and IBGNN. The standard deviation of classification scores in SGCN is also small.

D. Interpretation Analysis

To interpret the salient ROIs for distinguishing HC and AD, we average the feature importance probability $\boldsymbol{P_X}$ learned by our method across different modalities and obtain a scalar

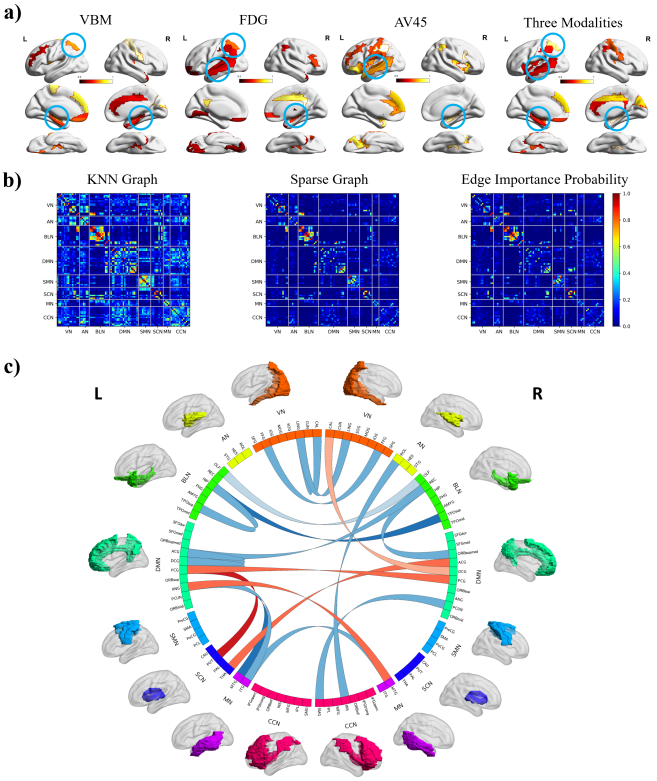


Fig. 3. The interpretation of salient ROIs and the most discriminative brain connections in **distinguishing AD from HC.** a) Interpreting top 20 salient ROIs based on the importance probability P_X between different modalities. The commonly detected salient ROIs across different modalities are circled in blue. b) Comparison between the KNN graph and the sparse interpretation of prominent brain network connections in AD group. c) The significant difference of the interpreted most discriminative connections for distinguishing HC and AD was evaluated by two-sample t-tests with false discovery rate (FDR) corrected p-value < 0.05. Here, the top 20 most discriminative ROI connections are visualized for interpretation by using multi-modalities. The dark-red and dark-blue color indicates the high positive and low negative t values.

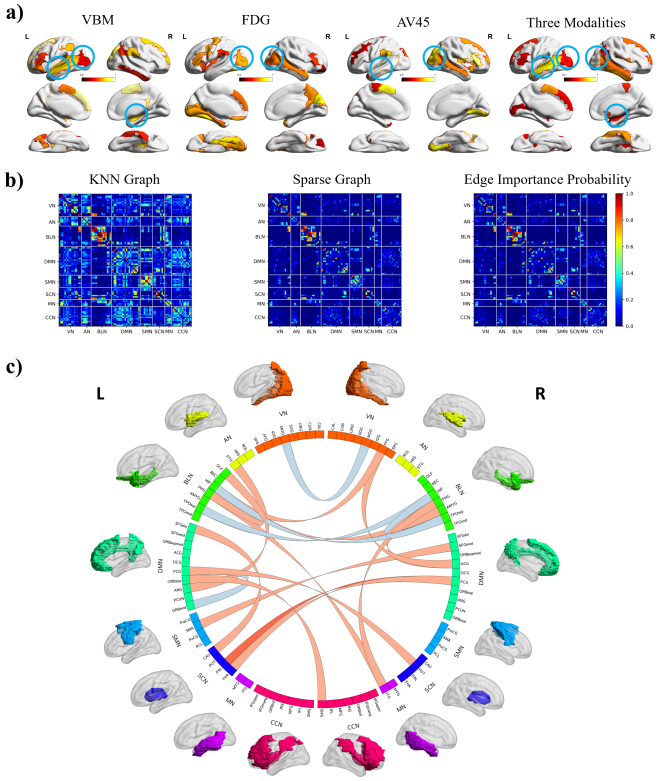
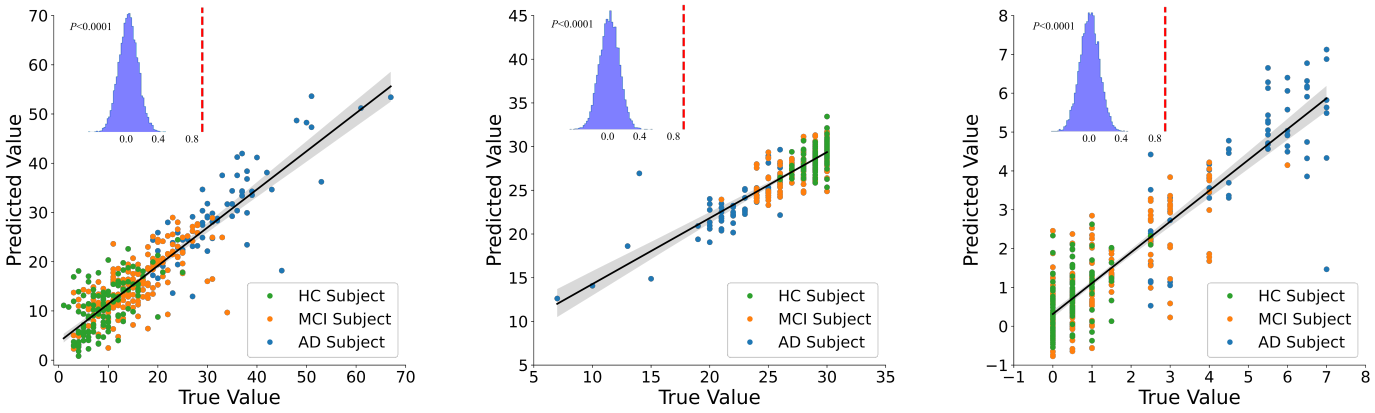


Fig. 4. The interpretation of salient ROIs and the most discriminative brain connections in **distinguishing MCI from HC.** This interpretation in MCI was reported by using the same strategy from AD analysis.

important score for each ROI. After ranking these scores in descending order, we visualized the top 20 most salient brain regions between HC and AD in the Fig. 3(a) identified by different modalities as well as the multi-modal analysis. The BrainNet Viewer [34] was used to plot the top 20 most salient ROIs in lateral views, medial views, and ventral views of the brain surface via the different modalities. We found the salient ROIs including the hippocampus, the parahippocampus, the parietal lobe, the temporal lobe, and the cingulate gyri regions were important for identifying AD, which was highly consistent to AD pathology based on previous studies [35], [36]. In Fig. 3(a), these commonly detected ROIs across different modalities are circled in blue. The right hippocampus, right cuneus, left superior parietal gyrus, and left angular gyrus achieved higher important probabilities by using multiple modalities than a single modality.

We further visualized the top 20 most salient brain regions between HC and MCI in the Fig. 4(a) identified by different modalities as well as the multi-modal analysis. It illustrates the interpreted top 20 most salient ROIs for identifying MCI from different modalities. We found that the middle occipital gyrus and middle temporal gyrus were important for identifying MCI via all different modalities. The right middle occipital gyrus, and left inferior temporal gyrus achieved higher important probabilities by using multiple modalities than a single modality. The olfactory cortex was only identifiable under the multi-modality analysis, which was highly related to MCI pathology based on previous studies [37]–[39]. The top 20 most salient brain regions in MCI vs. AD from Supplementary Fig. S2(a) suggest that Parahippocampal and Posterior cingulate gyrus were important by using multiple modalities. We can also discover similar salient patterns in sMCI vs. pMCI contrast from Supplementary Fig. S3(a) under multiple modalities. These salient brain regions around the limbic structure are highly associated with the studies on progressive MCI [40]. The interpretation of salient ROIs in AD and MCI contrasts suggests that incorporating all modalities can provide enhanced support for the interpretation of these salient ROIs in biomarker detection.

In brain connectivity, we categorized the ROIs into different neural systems based on their structural and functional roles using a specific atlas, which offers valuable benefits for verifying our interpretation result from a neuroscience perspective [41]. Those ROIs on the AAL-90 atlas were mapped into eight commonly used neural systems [42], including Visual Network (VN), Auditory Network (AN), Bilateral Limbic Network (BLN), Default Mode Network (DMN), Somato-Motor Network (SMN), Subcortical Network (SCN), Memory Network (MN), and Cognitive Control Network (CCN).



(a) ADAS13 (b) MMSE (c) CDR-SOB

Fig. 5. Prediction of ADAS13, MMSE, and CDR-SOB test scores using multiple linear regression based on the ROI features learned by the last GCN layer. The prediction performance was evaluated using 5-fold cross-validation. The significance of the prediction was confirmed by random permutation tests of 10000 times. The actual correlation coefficients between the predicted scores and true scores are indicated by red dashed lines.

In Fig. 3(b), we evaluated the efficacy of edge importance probabilities $\boldsymbol{P_A}$ learned by our method in identifying AD. It shows the average KNN graph, average sparse graph, and edge importance probability on AD contrast perspectively. The average sparse graph was visualized by element-wise multiplying the original KNN graph with edge importance probability $\boldsymbol{P_A}$. Moreover, the figure from edge importance probability illustrates the sparsity of average $\boldsymbol{P_A}$ on AD contrast. These results revealed that our method effectively assigned the important connections to have higher probabilities towards one and those unimportant connections towards zero, which demonstrates the sparse effectiveness of $\boldsymbol{P_A}$ learned by our method on the KNN graph. The same interpreted measures were used in MCI contrast. The result of the average KNN graph, sparse graph, and the edge importance probability in MCI contrast were visualized in Fig. 4(b). The similar results in MCI vs. AD and sMCI vs. pMCI contrasts were visualized in Supplementary Fig. S2(b) and S3(b) respectively.

We further interpreted the most discriminative connections between ROIs based on the edge importance probability $\boldsymbol{P_A}$. We applied two-sample t-tests on sparse brain graphs to detect the most discriminative connections between HC and AD/MCI. Fig. 3(c) listed the top 20 most discriminative connections with false discovery rate (FDR) correlated p-value < 0.05 between brain ROIs in AD contrast. The lines between ROIs showed the t value of the discriminative connections in multi-modalities. In addition, Fig. 4(c) listed the top 20 most discriminative connections based on the same standard in MCI contrast. The Supplementary Fig. S2(c) and S3(c) listed the top 20 most discriminative connections based on the same standard in MCI vs. AD and sMCI vs. pMCI contrasts.

E. Prediction ability of circuit abnormalities for AD symptoms

In this subsection, we further investigated the relationship between the identified circuit abnormalities and AD-related clinical symptoms including ADAS13, MMSE, and CDR-SOB. We used the learned topological patterns in the last GCN layer of SGCN to predict each of these clinical measures across HC, AD, and MCI subjects. These patterns were z-score normalized and then used to train standard linear regression models for the prediction via the 5-fold cross validation.

Fig. 5 depicted the results for predicting ADAS13, MMSE, and CDR-SOB scores. It provided a visual perception of how accurate the prediction result is for the given test. Meanwhile, the fit of the regression line indicated that there is a substantial correlation between the prediction scores and ground truth. The significance of the prediction was further confirmed by random permutation tests of 10000 times. Table II summarized the numeric performance comparison between brain network-based methods and our SGCN for the regression results across all the HC, AD, and MCI subjects. It contained Pearson's correlation coefficient, mean absolute error (MAE), root mean squared error (RMSE), and R-squared measure. For all three clinical measures, the identified circuit biomarkers showed significant prediction performance, indicating their underlying associations with AD/MCI symptoms.

IV. Discussion

A. Novel biomarkers identified by multi-modal analysis

In both AD and MCI, we observed that the hippocampus, angular gyrus, and temporal gyrus are the commonly detected salient ROIs in AD and MCI patients. Most salient ROIs are identified by using multiple modalities, indicating that multi-modal prediction is superior to that of a single modality. The modality-specific salient ROIs suggest that three modalities contribute differently to discriminating AD/MCI from HC. Furthermore, the detected salient ROIs are associated with previous evidence that the hippocampus is important in memory and recognition [43]. Specifically, we found the left hippocampus (HIP.L), and right hippocampus (HIP.R) were the commonly detected salient ROIs between the multiple

TABLE II
REGRESSION COMPARISON BETWEEN BRAIN NETWORK-BASED METHODS AND SGCN TO PREDICT AD SYMPTOMS. THE EVALUATION METRICS BETWEEN THE PREDICTED AND TRUE SCORES OF ADAS13, MMSE, AND CDR-SOB WERE REPORTED.

Metrics	Methods	ADAS13	MMSE	CDR-SOB
P Value	SGCN	1.58E-12	3.23E-12	1.42E-13
	IBGNN	2.98E-10	4.37E-10	1.98E-11
	BrainGNN	7.21E-9	4.69E-10	5.43E-10
	GIN	1.92E-8	2.48E-8	1.26E-7
	BrainNetCNN	2.48E-6	1.65E-5	4.71E-6
Correlation	SGCN	0.864	0.857	0.872
	IBGNN	0.789	0.763	0.791
	BrainGNN	0.753	0.761	0.759
	GIN	0.692	0.687	0.645
	BrainNetCNN	0.654	0.602	0.637
RMSE	SGCN	4.954	1.597	0.854
	IBGNN	5.237	1.873	0.935
	BrainGNN	5.468	1.949	1.181
	GIN	6.137	2.536	1.543
	BrainNetCNN	7.583	2.778	1.672
R Squared	SGCN	0.775	0.754	0.793
	IBGNN	0.692	0.663	0.681
	BrainGNN	0.648	0.605	0.619
	GIN	0.564	0.539	0.487
	BrainNetCNN	0.432	0.397	0.401

modalities. Numerous studies have reported a high volume reduction in the left and right hippocampus in AD patients than in HC [44], implying that the structural change of the left and right hippocampus detected by our model are highly associated with the derivation of AD. Meanwhile, the atrophy in the hippocampus has been found as the early biomarker for the identification of AD/MCI in some studies [45], again supporting the outputs of our model. Additionally, we also found right olfactory cortex (OLF.R) kept a high important probability via multiple analyses, which was associated with the findings of some studies that AD preferentially attacked the patient's central olfactory structures and led to earlier symptoms of olfactory deficits than clinical cognitive and memory deficits [37]. Thus, the salient olfactory cortex (OLF) can be regarded as an earlier biomarker for the identification of HC and MCI. The high importance probabilities in the middle occipital gyrus and olfactory cortex for distinguishing HC and MCI imply that these regions are the most salient brain regions in the early identification of AD/MCI. In addition, parahippocampal (PHG) and posterior cingulate gyrus (PCG) were important in predicting the pMCI, suggesting that the alteration of limbic structures and hypometabolism in the posterior cingulate context were related to the progression to dementia [46]. Several other biomarkers including cuneus, superior parietal gyrus, and median cingulate gyrus may also help to identify patients.

B. The most discriminative brain connections identified by multi-modal analysis

Our interpretation of brain connectivity abnormalities indicated that the most discriminative connections between HC and AD were ANG.L-MTG.L, HIP.L-TPOsup.R, and MTG.R-ITG.L. The regions within these connections are associated with a recent study that showed the functional connectivity alterations of the temporal lobe (e.g., MTG.L, TPOsup.R, ITG.L) and hippocampus (e.g., HIP.L) in AD [47]. Moreover, we found that OLF.L-OLF.R and ACG.L-PCG.L were the top discriminative connections, which were associated with findings on abnormal connections around the cingulate gyrus within the DMN system in AD patients [48], [49]. Accordingly, the identified abnormal connections around the cingulate gyrus support our previous regional finding and imply that the structural atrophy and connectivity dysfunction around the cingulate gyrus were related to severe cognitive impairment.

In addition, the interpreted discriminative connections between HC and MCI included HIP.L-TPOsup.L, OLF.L-PHG.L, ACG.R-HIP.R, and MOG.R-MOG.L. The HIP, OLF in the BLN system, and ACG, PCG in the DMN system corroborate previous studies, which have indicated that patients of AD and MCI have the similar brain regional connectivity abnormalities in BLN and DMN compared with HC [50]. The strong connections on OLF and MOG also support our previous finding that the middle occipital gyrus and olfactory cortex can be the earlier biomarkers to identify MCI patients. Although the occipital gyrus and olfactory cortex can support our previous salient regional finding, the most discriminative connections provide new insights into brain connectivity dysfunctions around BLN and DMN systems via multi-modal analysis. The interpretation of the most discriminative connections suggests that ROIs associated with these connections could serve as potential biomarkers for the identification of AD/MCI.

Furthermore, the important connections to discriminate AD from MCI included OLF.L-PHG.L, PCG.R-PAL.L, and AMYG.R-HIP.L. Beside, the important connections to discriminate pMCI from sMCI included OLF.R-PHG.R, PCG.R-PAL.L, and AMYG.L-HIP.L. The brain regions within these discriminative connections are associated with the observations reported in the previous neuroimaging studies [50]–[52]. Specifically, we found that the hippocampus (HIP) and amygdala (AMYG) were identified by both contrasts, which were consistent with the previous finding that the hippocampus and amygdala were the effective biomarkers to detect the cognitive impairment of patients by identifying the volumes of regional gray matter in MRI modality [52]. Accordingly, another study showed that the posterior cingulate gyrus (PCG) is significantly correlated with the progression of MCI based on their thinning rate from MCI to AD [51]. The interpretation of the discriminative connections on the progression of MCI suggests that these ROIs could serve as potential biomarkers to discover the conversion into dementia.

C. Neural system-level connectivity abnormality

The neural system-level interpretation via multiple modalities in Fig. 6 indicated that the DMN, BLN, and MN sys-

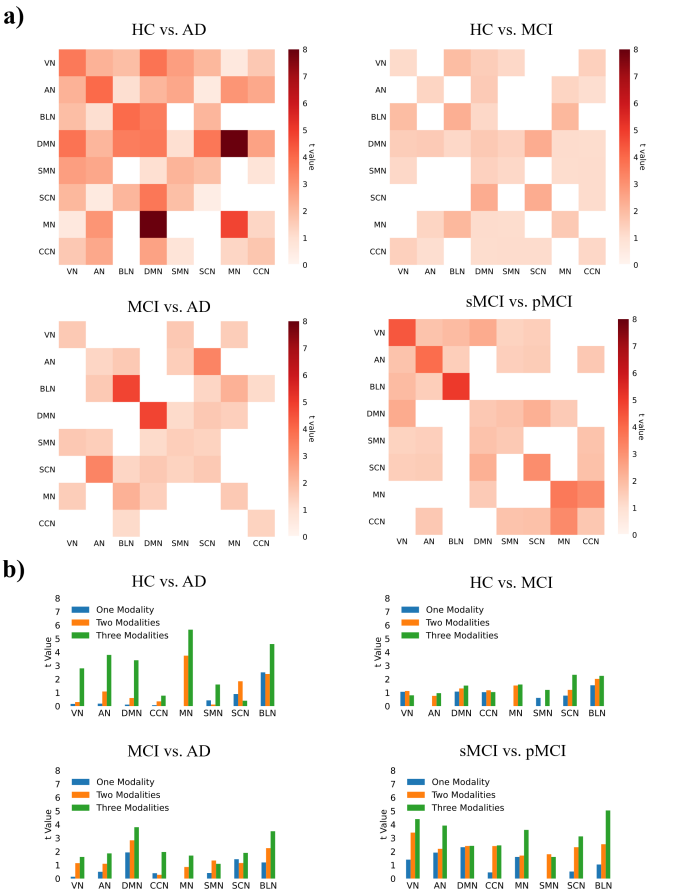


Fig. 6. Neural system-level interpretation of the most discriminative connections. a) The absolute t value of most discriminative ROI connections with FDR correlated p-value < 0.05 were reported between neural systems by using multi-modalities. The dark-red color indicates a high score. The non-significant connections are marked as white. b) Such t values were reported by using different modalities in each neural system. Here, the reported t values of one modality were the average results over all single modalities. Similarly, the t values of two modalities were the average results over all three pairs of modalities.

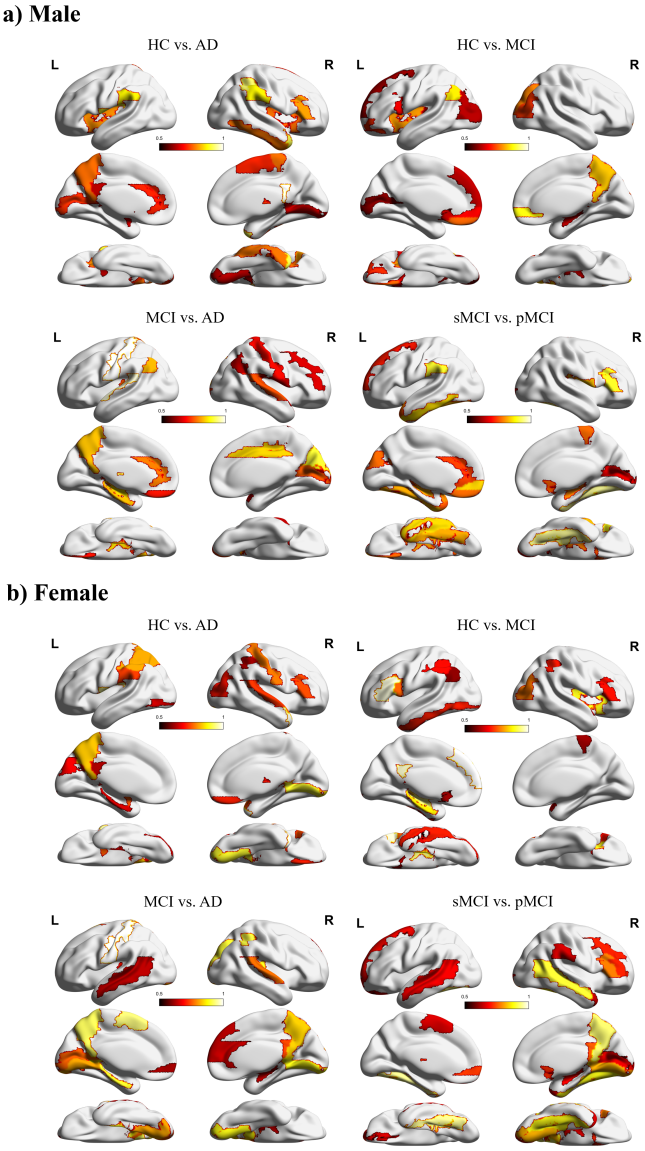


Fig. 7. Interpreting top 20 salient ROIs between males and females under multi-modalities.

tems contained stronger discriminative connections than single modality and double modalities. In fact, the connections within MN disappeared when only using a single modality, while they became the strongest after fusing all three modalities. Most discriminative connections within MN and DMN systems were found when using all three modalities instead of single or double modality. These discriminative connection patterns showed high correspondence with some previous neuroimaging evidence for the derivation of AD and MCI [47], [51], implying that combining multiple modalities can enhance the identification of neural system-level connectivity abnormalities in cognitively impaired patients. The neural system-level connections via multiple modalities including BLN-DMN, DMN-DMN, and DMN-MN in AD were more discriminative than using any single or double modalities, suggesting that the connectivity dysfunctions around DMN were strong in the multi-modal joint analysis. Moreover, the connections of BLN-MN and DMN-SCN in MCI also became stronger via multiple modalities. Besides, the connections around BLN and DMN were strong to discriminate AD from MCI, and the connections around BLN, VN, AN, and MN were the key to identify the progressive MCI under multiple modalities. Therefore, these observations suggest that the discriminative connectivity patterns interpreted by our method within the neural systems are enhanced with additional modalities of imaging data. Our finding further provides evidence that the multi-modal joint analysis can capture the structural and pathological changes in the brain via multiple modalities and identify the neural system-level connectivity dysfunctions in dementia.

D. Sex-related differences in the biomarkers

Our results investigated sex-related abnormalities based on the regional importance probabilities between males and females, given that the female sex is a major risk factor in AD with a higher incidence of the disease [53]. Especially, for identifying AD, our result indicated that the temporal

TABLE III
ABLATION STUDY OF LOSS TERMS IN SGCN.

Model	Accuracy	ROC-AUC	Sensitivity	Specificity
SGCN(all)	.826 ±.039	.841 ±.043	.804 ±.054	.845 ±.051
SGCN(w/o $\mathcal{L}_s$)	.810 ±.047	.808 ±.048	.775 ±.050	.821 ±.074
SGCN(w/o $\mathcal{L}_e$)	.814 ±.041	.811 ±.044	.772 ±.047	.828 ±.055
SGCN(w/o $\mathcal{L}_s$ & $\mathcal{L}_e$)	.805 ±.060	.799 ±.036	.767 ±.051	.824 ±.048
SGCN(w/o $\mathcal{L}_m$)	.791 ±.049	.784 ±.068	.743 ±.062	.807 ±.070

gyrus, precuneus, and parahippocampus were identified as the most salient ROIs in females than males in Fig. 7. A similar finding was reported that females with AD dementia could have sharper declines in cortical thickness around temporal regions, and left precuneus [53]. For identifying MCI, we found the frontal gyrus and posterior cingulate gyrus were more important in females than males. The posterior cingulate gyrus has been identified as a cortical region affected during the prodromal stage of AD by neuroimaging studies [54]. Meanwhile, we also found the posterior cingulate gyrus in females was also highly related to the progression of MCI. Our result suggests the atrophy of the posterior cingulate gyrus in females plays a key role in the derivation of early dementia and the progression into dementia. Although these findings have been associated with cognitive impairment in multiple studies [47], [49], but never in a multi-modal sex-specific analysis. Therefore, our work provides novel evidence on sex-related differences in biomarkers and their connectivity dysfunctions related to cognitive impairment.

E. Ablation study

An ablation study was conducted to validate the effectiveness of loss term in SGCN for classifying three contrasts including HC vs. AD, HC vs. MCI, and MCI vs. AD. The one-vs-one strategy was used to report the classification result. Specially, we quantitatively measured the impact of different loss terms for identifying diseases via multi-modalities in Table III. Without the conditional entropy loss $\mathcal{L}_m$, the sensitivity score would drop a lot, suggesting that combing $\mathcal{L}_m$ is crucial to discover the important subset of node feature X_s and the important subgraph G_s related to the disease. We also ablated the sparse regularization $\mathcal{L}_s$ and $\mathcal{L}_e$, which resulted in the important features and edges having higher important probabilities towards one, and unimportant ones towards zero. Without them, the accuracy and sensitivity of SGCN would drop a little bit, and the salient ROIs and the prominent connections identified by the model would become less interpretable. The best performance was achieved after fusing all loss terms together. Our work provides evidence that combining conditional entropy loss and sparse regularization loss can improve the interpretable GCN model to identify the important disease-related subgraphs and subsets of features.

F. Generalizability of SGCN model on ADNI-2/GO and independent ADNI-1 test set

In order to analyze the generalizability of SGCN model, we separated the ADNI dataset into different phases including ADNI-1, ADNI-GO, and ADNI-2. The ADNI-1 included 20 HC, and 27 MCI subjects, given its goal is to identify biomarkers and genetic characteristics that would support the early detection of AD. After combining ADNI-GO and ADNI-2, the ADNI-2/GO dataset included 152 HC, 444 MCI, and 96 AD subjects.

TABLE IV
CLASSIFICATION PERFORMANCE OF OUR SGCN METHOD BY USING ADNI-2/GO AND INDEPENDENT ADNI-1 TEST SET.

Modalities	Measures	ADNI-1	ADNI-2/GO	
		HC vs. MCI	HC vs. MCI	Multi-class
VBM	Accuracy	.668 ±.059	.645 ±.073	.735 ±.048
	ROC-AUC	.656 ±.047	.639 ±.058	.751 ±.046
	Sensitivity	.640 ±.048	.553 ±.084	.693 ±.052
	Specificity	.651 ±.045	.678 ±.061	.724 ±.049
VBM FDG	Accuracy	.688 ±.044	.667 ±.058	.759 ±.043
	ROC-AUC	.691 ±.037	.672 ±.045	.793 ±.039
	Sensitivity	.663 ±.058	.583 ±.068	.725 ±.067
	Specificity	.709 ±.051	.685 ±.073	.787 ±.055
VBM FDG AV45	Accuracy	.706 ±.035	.714 ±.045	.806 ±.044
	ROC-AUC	.711 ±.046	.721 ±.043	.812 ±.045
	Sensitivity	.688 ±.045	.626 ±.056	.763 ±.056
	Specificity	.717 ±.039	.741 ±.047	.797 ±.048

In our experiments, we performed the 5-fold cross validation to train and test the performance on ADNI-2/GO dataset. We used the one-against-one strategy to conduct the multiclass classification. We regarded the ADNI-1 as the unseen test set. Given there are only HC and MCI subjects in ADNI-1, the SGCN model trained on ADNI-2/GO between HC vs. MCI was further used to test the performance on the unseen ADNI-1 dataset. Table IV shows the classification performance in ADNI-2/GO and ADNI-1 using different modalities. Under the HC vs. MCI contrast, the difference in accuracy between ADNI-1 and ADNI-2/GO was within 2%. The classification scores were comparable between ADNI-1 and ADNI-2/GO. Under the multiclass classification performance, the result was still comparable between ADNI-2/GO and the whole ADNI dataset. This result suggests that the SGCN model achieved great generalizability to new, previously unseen data.

V. CONCLUSIONS

In summary, we presented a multi-modal sparse interpretable GCN framework for identifying AD via multi-modal brain images. Our method applied sparse importance probabilities to discover novel neurological biomarkers under multi-modal analysis in AD and MCI. Besides the promising prediction performance, the disease-related network-based patterns identified by our method show significant predictability for typical AD-related clinical measures. Our results revealed that the hippocampus, olfactory cortex, angular, and temporal gyrus were potential regional biomarkers for detecting AD/MCI, and that prominent brain connectivity abnormalities within the memory, bilateral limbic, and default mode networks were most important for distinguishing AD/MCI from HC. These

findings show a high correspondence with established neuroimaging evidence associated with AD and MCI [37], [39], [55]. This observation suggests that our method is suitable for interpreting the most salient ROIs, the most discriminative brain network connections, and neural systems with additional imaging modalities.

The possible limitations were the robustness of our method and the generalization to other neurodegenerative disease datasets. In the data preprocessing, we applied the standard AAL atlas to subsample the whole brain and obtain 90 ROIs. However, it has been studied that the different atlases showed a considerable influence in the identification of mental disorders including AD and MCI for ROI-based analysis [56], [57]. We will investigate how the biomarker findings are robust to the selection of brain atlases. It is also important to further test the generalization ability of our model on many more datasets. In the future, we plan to apply our SGCN model to the Open Access Series of Imaging Studies (OASIS) [58] and the Parkinson's Progression Markers Initiative (PPMI) [59] cohorts to test the performance. For addressing the real clinical needs regarding the derivation of AD, it is also worth further exploring to apply our SGCN model to the longitudinal data to predict how and when the MCI will be converted into AD. Because our interpretable approach is model-agnostic, it is highly generalizable to other brain diseases for developing novel multi-modal biomarkers.

REFERENCES

[1] L. Du, K. Liu, X. Yao, S. Risacher, J. Han, A. Saykin, L. Guo, and L. Shen, "Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics," *IEEE/ACM Transactions On Computational Biology and Bioinformatics*, 2019.

[2] A. L. Jefferson, K. A. Gifford, S. Damon, G. W. Chapman, D. Liu, J. Sparling, V. Dobromyslin, and D. Salat, "Gray & white matter tissue contrast differentiates mild cognitive impairment converters from non-converters," *Brain imaging and behavior*, vol. 9, no. 2, pp. 141–148, 2015.

[3] V. Camus, P. Payoux, L. Barré, B. Desgranges, T. Voisin, C. Tauber, R. La Joie, M. Tafani, C. Hommet, G. Chételat *et al.*, "Using pet with 18f-av-45 (florbetapir) to quantify brain amyloid load in a clinical environment," *European journal of nuclear medicine and molecular imaging*, vol. 39, no. 4, pp. 621–631, 2012.

[4] X. Li, Z. Yuan, D. Nicha, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan, "BrainGNN: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, p. 102233, 2021.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.

[6] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease," *Medical image analysis*, vol. 48, pp. 117–130, 2018.

[7] H. Cui, W. Dai, Y. Zhu, X. Li, L. He, and C. Yang, "BrainNNExplainer: An interpretable graph neural network framework for brain network based disease analysis," in *ICML Workshop on Interpretable Machine Learning in Healthcare*, 2021.

[8] H. Zhou, L. He, Y. Zhang, L. Shen, and B. Chen, "Interpretable graph convolutional network of multi-modality brain imaging for alzheimer's disease diagnosis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[9] H. Zhou, Y. Zhang, L. He, L. Shen, and B. Y. Chen, "Interpretable graph convolutional network for alzheimer's disease diagnosis using multi-modal imaging genetics," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 1004–1007.

[10] Z. S. Chen, I. R. Galatzer-Levy, B. Bigio, C. Nasca, Y. Zhang *et al.*, "Modern views of machine learning for precision psychiatry," *Patterns*, vol. 3, no. 11, 2022.

[11] Y.-h. Sheu, "Illuminating the black box: Interpreting deep neural network models for psychiatric research," *Frontiers in Psychiatry*, p. 1091, 2020.

[12] H. Yang, X. Li, Y. Wu, S. Li, S. Lu, J. S. Duncan, J. C. Gee, and S. Gu, "Interpretable multimodality embedding of cerebral cortex using attention graph network for identifying bipolar disorder," in *MICCAI*. Springer, 2019, pp. 799–807.

[13] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *NeurIPS*, 2020.

[14] M. N. Vu and M. T. Thai, "PGM-Explainer: Probabilistic graphical model explanations for graph neural networks," in *NeurIPS*, 2020.

[15] Y. Zhang, L. Zhan, W. Cai, P. Thompson, and H. Huang, "Integrating heterogeneous brain networks for predicting brain disease conditions," in *MICCAI*. Springer, 2019, pp. 214–222.

[16] X. Zhang, L. He, K. Chen, Y. Luo, J. Zhou, and F. Wang, "Multi-view graph convolutional network and its applications on neuroimage analysis for parkinson's disease," *AMIA Annual Symposium Proceedings*, vol. 2018, p. 1147, 2018.

[17] Y. Li, J. Liu, X. Gao, B. Jie, M. Kim, P.-T. Yap, C.-Y. Wee, and D. Shen, "Multimodal hyper-connectivity of functional networks using functionally-weighted lasso for mci classification," *Medical image analysis*, vol. 52, pp. 80–96, 2019.

[18] S. Qiu, M. I. Miller, P. S. Joshi, J. C. Lee, C. Xue, Y. Ni, Y. Wang, D. Anda-Duran, P. H. Hwang, J. A. Cramer *et al.*, "Multimodal deep learning for alzheimer's disease dementia assessment," *Nature communications*, vol. 13, no. 1, pp. 1–17, 2022.

[19] R. Zhou, H. Zhou, B. Y. Chen, L. Shen, Y. Zhang, and L. He, "Attentive deep canonical correlation analysis for diagnosing alzheimer's disease using multimodal imaging genetics," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 681–691.

[20] R. Zhou, H. Zhou, L. Shen, B. Y. Chen, Y. Zhang, and L. He, "Integrating multimodal contrastive learning and cross-modal attention for alzheimer's disease prediction in brain imaging genetics," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 1806–1811.

[21] T. D. Figley, B. Mortazavi Moghadam, N. Bhullar, J. Kornelsen, S. M. Courtney, and C. R. Figley, "Probabilistic white matter atlases of human auditory, basal ganglia, language, precuneus, sensorimotor, visual and visuospatial networks," *Frontiers in human neuroscience*, vol. 11, p. 306, 2017.

[22] H. Zhou, Y. Zhang, B. Y. Chen, L. Shen, and L. He, "Sparse interpretation of graph convolutional networks for multi-modal diagnosis of alzheimer's disease," in *MICCAI*. Springer, 2022, pp. 469–478.

[23] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[24] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *NeurIPS*, vol. 32. NIH Public Access, 2019, p. 9240.

[25] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics*, vol. 15, no. 4, pp. 869–877, 2005.

[26] S. Cong, S. L. Risacher, J. D. West, Y.-C. Wu, L. G. Apostolova, E. Tallman, M. Rizkalla, P. Salama, A. J. Saykin, and L. Shen, "Volumetric comparison of hippocampal subfields extracted from 4-minute accelerated vs. 8-minute high-resolution t2-weighted 3t mri scans," *Brain imaging and behavior*, vol. 12, pp. 1583–1595, 2018.

[27] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[28] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[29] J. Gan, Z. Peng, X. Zhu, R. Hu, J. Ma, and G. Wu, "Brain functional connectivity analysis based on multi-graph fusion," *Medical image analysis*, vol. 71, p. 102057, 2021.

[30] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh, "Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.

[31] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *ICLR*, 2018.

[32] B.-H. Kim and J. C. Ye, "Understanding graph isomorphism network for rs-fmri functional connectivity analysis," *Frontiers in neuroscience*, p. 630, 2020.

[33] H. Cui, W. Dai, Y. Zhu, X. Li, L. He, and C. Yang, "Interpretable graph neural networks for connectome-based brain disorder analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 375–385.

[34] M. Xia, J. Wang, and Y. He, "BrainNet viewer: a network visualization tool for human brain connectomics," *PloS one*, vol. 8, no. 7, p. e68910, 2013.

[35] Y. Mu and F. H. Gage, "Adult hippocampal neurogenesis and its role in alzheimer's disease," *Molecular neurodegeneration*, vol. 6, no. 1, pp. 1–9, 2011.

[36] S. F. Eskildsen, P. Coupé, V. S. Fonov, J. C. Pruessner, D. L. Collins, A. D. N. Initiative *et al.*, "Structural imaging biomarkers of alzheimer's disease: predicting disease progression," *Neurobiology of aging*, vol. 36, pp. S23–S31, 2015.

[37] M. M. Vasavada, J. Wang, P. J. Eslinger, D. J. Gill, X. Sun, P. Karunanayaka, and Q. X. Yang, "Olfactory cortex degeneration in alzheimer's disease and mild cognitive impairment," *Journal of Alzheimer's disease*, vol. 45, no. 3, pp. 947–958, 2015.

[38] Y. Li, J. Liu, Z. Peng, C. Sheng, M. Kim, P.-T. Yap, C.-Y. Wee, and D. Shen, "Fusion of uls group constrained high-and low-order sparse functional connectivity networks for mci classification," *Neuroinformatics*, vol. 18, no. 1, pp. 1–24, 2020.

[39] X. Song, F. Zhou, A. F. Frangi, J. Cao, X. Xiao, Y. Lei, T. Wang, and B. Lei, "Graph convolution network with similarity awareness and adaptive calibration for disease-induced deterioration prediction," *Medical Image Analysis*, vol. 69, p. 101947, 2021.

[40] N. Garg, M. S. Choudhry, and R. M. Bodade, "A review on alzheimer's disease classification from normal controls and mild cognitive impairment using structural mr images," *Journal of neuroscience methods*, vol. 384, p. 109745, 2023.

[41] M. Xu, Z. Wang, H. Zhang, D. Pantazis, H. Wang, and Q. Li, "A new graph gaussian embedding method for analyzing the effects of cognitive training," *PLoS computational biology*, vol. 16, no. 9, p. e1008186, 2020.

[42] N. Chen, J. Shi, Y. Li, S. Ji, Y. Zou, L. Yang, Z. Yao, and B. Hu, "Decreased dynamism of overlapping brain sub-networks in major depressive disorder," *Journal of psychiatric research*, vol. 133, pp. 197–204, 2021.

[43] W. Jaroudi, J. Garami, S. Garrido, M. Hornberger, S. Keri, and A. A. Moustafa, "Factors underlying cognitive decline in old age and alzheimer's disease: the role of the hippocampus," *Reviews in the Neurosciences*, vol. 28, no. 7, pp. 705–714, 2017.

[44] F. Adriano, C. Caltagirone, and G. Spalletta, "Hippocampal volume reduction in first-episode and chronic schizophrenia: a review and meta-analysis," *The Neuroscientist*, vol. 18, no. 2, pp. 180–200, 2012.

[45] E. McLachlan, J. Bousfield, R. Howard, and S. Reeves, "Reduced parahippocampal volume and psychosis symptoms in alzheimer's disease," *International journal of geriatric psychiatry*, vol. 33, no. 2, pp. 389–395, 2018.

[46] C. Cerami, P. A. Della Rosa, G. Magnani, R. Santangelo, A. Marcone, S. F. Cappa, and D. Perani, "Brain metabolic maps in mild cognitive impairment predict heterogeneity of progression to dementia," *NeuroImage: Clinical*, vol. 7, pp. 187–194, 2015.

[47] S. Schwab, S. Afyouni, Y. Chen, Z. Han, Q. Guo, T. Dierks, L.-O. Wahlund, and M. Grieder, "Functional connectivity alterations of the temporal lobe and hippocampus in semantic dementia and alzheimer's disease," *Journal of Alzheimer's disease*, vol. 76, no. 4, pp. 1461–1475, 2020.

[48] C. Salvatore, A. Cerasa, P. Battista, M. C. Gilardi, A. Quattrone, and I. Castiglioni, "Magnetic resonance imaging biomarkers for the early diagnosis of alzheimer's disease: a machine learning approach," *Frontiers in neuroscience*, vol. 9, p. 307, 2015.

[49] E. Yu, Z. Liao, D. Mao, Q. Zhang, G. Ji, Y. Li, and Z. Ding, "Directed functional connectivity of posterior cingulate cortex and whole brain in alzheimer's disease and mild cognitive impairment," *Current Alzheimer Research*, vol. 14, no. 6, pp. 628–635, 2017.

[50] R. Yu, L. Qiao, M. Chen, S.-W. Lee, X. Fei, and D. Shen, "Weighted graph regularized sparse brain network construction for mci identification," *Pattern recognition*, vol. 90, pp. 220–231, 2019.

[51] J. P. Lerch, J. C. Pruessner, A. Zijdenbos, H. Hampel, S. J. Teipel, and A. C. Evans, "Focal decline of cortical thickness in alzheimer's disease identified by computational neuroanatomy," *Cerebral cortex*, vol. 15, no. 7, pp. 995–1001, 2005.

[52] Y. Zhang, N. Schuff, M. Camacho, L. L. Chao, T. P. Fletcher, K. Yaffe, S. C. Woolley, C. Madison, H. J. Rosen, B. L. Miller *et al.*, "Mri markers for mild cognitive impairment: comparisons between white matter integrity and gray matter volume measurements," *PloS one*, vol. 8, no. 6, p. e66367, 2013.

[53] F. Cieri, X. Zhuang, D. Cordes, N. Kaplan, J. Cummings, and J. Caldwell, "Relationship of sex differences in cortical thickness and memory among cognitively healthy subjects and individuals with mild cognitive impairment and alzheimer disease," *Alzheimer's research & therapy*, vol. 14, no. 1, pp. 1–12, 2022.

[54] S. W. Scheff, D. A. Price, M. A. Ansari, K. N. Roberts, F. A. Schmitt, M. D. Ikonomovic, and E. J. Mufson, "Synaptic change in the posterior cingulate gyrus in the progression of alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 43, no. 3, pp. 1073–1090, 2015.

[55] Y. Li, J. Liu, Z. Tang, and B. Lei, "Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for mci identification," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2818–2830, 2020.

[56] Z. Wu, D. Xu, T. Potter, Y. Zhang, A. D. N. Initiative *et al.*, "Effects of brain parcellation on the characterization of topological deterioration in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 11, p. 113, 2019.

[57] K. Dadi, M. Rahim, A. Abraham, D. Chyzhyk, M. Milham, B. Thirion, G. Varoquaux, A. D. N. Initiative *et al.*, "Benchmarking functional connectome-based predictive models for resting-state fmri," *NeuroImage*, vol. 192, pp. 115–134, 2019.

[58] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults," *Journal of cognitive neuroscience*, vol. 22, no. 12, pp. 2677–2684, 2010.

[59] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury *et al.*, "The parkinson progression marker initiative (ppmi)," *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.