INTERPRETABLE GRAPH CONVOLUTIONAL NETWORK OF MULTI-MODALITY BRAIN IMAGING FOR ALZHEIMER'S DISEASE DIAGNOSIS

Houliang Zhou[1], Lifang He[1], Yu Zhang[2], Li Shen[3], Brian Chen[1]

[1] Department of Computer Science and Engineering, Lehigh University, PA, USA
[2] Department of Bioengineering, Lehigh University, PA, USA
[3] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, PA, USA

ABSTRACT

Identification of brain regions related to the specific neurological disorders are of great importance for biomarker and diagnostic studies. In this paper, we propose an interpretable Graph Convolutional Network (GCN) framework for the identification and classification of Alzheimer's Disease (AD) using multi-modality brain imaging data. Specifically, we extended the Gradient Class Activation Mapping (Grad-CAM) technique to quantify the most discriminative features identified by GCN from brain connectivity patterns. We then utilized them to find signature regions of interest (ROIs) by detecting the difference of features between regions in healthy control (HC), mild cognitive impairment (MCI), and AD groups. We conducted the experiments on the ADNI database with imaging data from three modalities, including VBM-MRI, FDG-PET, and AV45-PET, and showed that the ROI features learned by our method were effective for enhancing the performances of both clinical score prediction and disease status identification. It also successfully identified biomarkers associated with AD and MCI.

Index Terms— Interpretation, graph convolutional network, neuroimaging, multi-modality

1. INTRODUCTION

Neuroimaging pattern classification methods have demonstrated recent advances in predicting Alzheimer's disease (AD) and mild cognitive impairment (MCI) from magnetic resonance imaging (MRI) and positron emission tomography (PET) scans [1, 2, 3]. Since the brain is an extremely complex system, large improvements in understanding the brain's organization have been made by representing the brain as a connectivity graph [4]. In this graph, nodes are defined as brain regions of interest (ROIs) and edges are defined as the connectivity between those ROIs. This representation is highly compatible with Graph Convolutional Networks (GCN), a deep learning method with demonstrated capabilities for analyzing graph structure problems [5, 6, 7, 8].

In neuroimaging, GCN has been widely used to analyze the brain connectivity graph and discover neurological biomarkers [9, 10]. In the graph classification problem, the explainability of GCN predictions is crucial for helping to identify and localize biomarkers that contribute to AD or MCI. Several approaches have been proposed to explain the GCN model (e.g., [11, 12, 13]). Most of them did not target disease prediction. Generally, recent methods for interpreting brain networks focused only on single modality data [9, 10]. Because we assume that the subjects having the same disease may share the similar patterns in brain network, group-level explanations are significant for identifying salient ROIs and discovering biomarkers related to AD and MCI. Recent studies indicated that different imaging modalities can provide essential complementary information that can improve accuracy in disease diagnosis [14]. Thus, we consider here the gradient-weighted Class Activation Mapping method (Grad-CAM) for multi-modality imaging data, which has been demonstrated to produce explanations on graphs of moderate size with high fidelity, contrastivity, and sparsity [15].

In this paper, we propose an interpretable GCN framework for the identification and classification of Alzheimer's Disease (AD) using multi-modal brain imaging data. We extended the Grad-CAM technique to interpret the salient ROIs for each subject in HC, AD, and MCI categories. The major contributions of this work include the adaptation of Grad-CAM to the GCN model for interpreting salient ROIs, the integration of three modalities of imaging data to construct the brain connectivity graph, and the extension of the GCN model to discover biomarkers in AD and MCI. In our experimental results, our method exhibited high classification performance, and our interpretation method revealed that subjects in the same disease group share the similar patterns. We also found that the putamen and pallidum biomarkers were important for distinguishing HC, AD and MCI. The features learned by our method have a high correlation with the clinical test scores of Mini-Mental State Examination (MMSE) and Alzheimer's Disease Assessment Score 13 (ADAS13), suggesting that our learned features could potentially predict clinical test scores. These results point to applications for our method in the interpretation of ROIs from imaging data in multiple modalities and disease conditions.

2. METHODS

2.1. Data Acquisition and Preprocessing

In this work, we used the image data obtained from Alzheimer Disease Neuroimaging Initiative (ADNI) [16]. The neuroimaging data were from 755 non-Hispanic Caucasian participants, including 182 HC subjects, 476 MCI subjects, and 97 AD subjects, which consists of three modalities including structural Magnetic Resonance Imaging (VBM-MRI), fluorodeoxyglucose Positron Emission Tomography (FDG-PET), and 18-F florbetapir PET (AV45-PET).

The multi-modality imaging data were aligned to each participant's same visit. The structural MRI scans were preprocessed with voxel-based morphometry (VBM) using the SPM software [17]. Generally, all scans were aligned to a T1-weighted template image, segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) maps, normalized to the standard Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2$ mm^3 voxels, and were smoothed with an 8 mm FWHM kernel. The FDG-PET and AV45-PET scans were also registered to the same MNI space by SPM. We subsampled the whole brain and obtained 90 ROIs (excluding the cerebellum and vermis) based on the AAL atlas [18]. ROI-level measures were calculated by averaging all the voxel-level measures within each ROI.

2.2. Brain Graph Construction

To construct the brain connectivity graph from the three modalities, we concatenate all of three modalities into the feature vector for each ROI. We use 90 ROIs for each subject. Each ROI in the image views as the node in the graph, which can be represented as an undirected weighted graph $G = (V, E)$. The vertex set $V = \{v_i, \cdots, v_n\}$ consists of ROIs in the brain and each edge in E is weighted by a connectivity strength, where n is the number of ROIs. In our work, we define $\widetilde{G} = (V, \widetilde{E})$ based on the multi-modal information of the region using K-Nearest Neighbor (KNN) graph [14]. The edges are weighted by the Gaussian similarity function of Euclidean distances, i.e., $e(v_i, v_j) = exp(-\frac{\|v_i - v_j\|^2}{2\sigma^2})$. We identify N_j as the set of K-Nearest Neighbors of vertex v_i, and connect v_i and v_j if $v_i \in N_j$ or $v_j \in N_i$. The weighted adjacency matrix $A \in \mathbb{R}^{N \times N}$ represents the similarity between each ROI and its nearest similar neighbor ROIs. The element of adjacency matrix A can be represented as follows:

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

2.3. GCN Model

The classification of graphs can be achieved by embedding node features into a low dimensional space, grouping nodes, and summarizing them [6]. The summarized vector for each graph is fed into a multilayer perceptron (MLP) classifier.

In this work, the whole architecture contains three types of layers, including graph convolutional layers, a node pooling layer, and a readout layer. The graph convolutional layer inductively learns a node representation by recursively transforming and aggregating the feature vectors of its neighboring nodes.

We define a brain adjacency matrix $A \in \mathbb{R}^{N \times N}$ and node feature matrix $X \in \mathbb{R}^{N \times d_{in}}$, where N is the number of ROIs and d_{in} is the dimension of multi-modality input feature. The propagation of GCN model or the forward-pass update of node representation is calculated as:

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

where $H^0 = X$, $H^l \in \mathbb{R}^{N \times d_l}$ is the output of the lth graph convolution layer, d_l is the number of output channels of layer l, $\widetilde{A} = A + I$ is the adjacency matrix of graph with self-loops, $I \in \mathbb{R}^{N \times N}$ is the indentity matrix, $W^l \in \mathbb{R}^{d_l \times d_{l+1}}$ are the learnable parameters, and $\widetilde{D}$ is the diagonal degree matrix with $\widetilde{D}_{i,i} = \sum_j \widetilde{A}_{i,j}$. We normalize the $\widetilde{A}$ by multiplying $\widetilde{D}^{-\frac{1}{2}}$ in order to keep a fixed feature scale after graph convolution. σ is the activation function.

The node pooling layer groups the nodes together to summarize the features of whole graph. After we get the final output H^L of graph convolution layer, we will summarize the whole feature matrix H^L into a single vector, which is fed into an MLP classifier with softmax activation function in the final layer. We use the negative log-likelihood as the loss function for graph classification.

2.4. Interpretability of GCN

We now present how we can use the above formulation for explaining GNN predictions and identifying important ROIs. Our idea draws inspiration from the recent work on GNN explainability using Gradient Class Activation Mapping (Grad-CAM) [15], which was originally proposed for producing visual explanations for CNN models. We first define the kth graph convolutional feature map at layer l as:

$$H^{l+1}_k = \sigma(\widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} H^l W^l_k) \quad (3)$$

where W^l_k is the kth column of learnable matrix W^l. According to this notation, we denote $H^l_{k,n}$ is the kth feature at the lth layer for each node n. We denote L as the final GCN layer. $H^L_{k,n}$ represents the feature map of the final convolutional layer.

The kth average feature map after the final convolution layer and the class score for each class c can be calculated as:

$$e_k = \frac{1}{N} \sum_{n=1}^{N} H^L_{k,n}$$
$$y^c = \sum_k w^c_k e_k \quad (4)$$

where the learnable w_k^c encodes the kth feature importance for predicting the class c, and y^c is the class score. The class specific weights β_k^c of Grad-CAM for class c and feature k at final layer L can be calculated as:

$$\beta_k^c = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial y^c}{\partial H_{k,n}^L} \tag{5}$$

The Grad-CAM's heat map can be calculated at the final convolutional layer as:

$$L_n^c = ReLU(\sum_k \beta_k^c H_{k,n}^L) \tag{6}$$

where L_n^c is the heat map of node n for class c at the final GCN layer. We apply the equation (6) to get the node importance for each ROI in the brain connectivity graph.

3. RESULTS

3.1. Classification Performance

In our experiments, we separate the whole data into three groups including HC vs AD, HC vs MCI, and MCI vs AD to examine our classification results. We compare the classification performance of each group when using one, two or three modalities. This experimental design measures the performance of our method on different numbers of modalities, since it is a multimodal method, and on different groups.

For the brain connectivity graphs, we used $K = 10$ to build the KNN graph for each subject. After building the graph data, we used the GCN + MLP to calculate the classification result with the following configuration: three graph convolutional layers of size 10, 10, and 5, respectively, followed by three fully-connected layers, a dropout layer, and a softmax classifier. The models were trained for 100 epochs using the ADAM optimizer with learning rate 0.005, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The dropout rate is 0.5. We performed 5-fold cross validation and repeated the experiment for 10 times. The average classification accuracy, ROC-AUC, sensitivity, specificity, and their standard deviations are reported.

Tables 1, 2 and 3 show the classification result by using one, two and three modalities, respectively, where $\pm$ represents the standard deviation of evaluation scores in 5 folds. We compared our method with SVM (RBF kernel). The best classification performance was achieved by our method when using all three modalities. In testing, we concatenated the predicted possibilities and the labels of testing data in 5 folds to build the ROC curve by calculating the true positive rate and false positive rate. The ROC curve, using all three modalities, is plotted in Fig. 1. Based on these findings, all three modalities were used to evaluate the interpretability of our model.

3.2. Interpretation of Salient ROIs

After training models for the ADNI datasets, to summarize the salient ROIs, we apply the Grad-CAM method on all sub-

Table 1. **One Modality**. Classification results of different groups with VBM-MRI only.

Method	Group	Accuracy	ROC-AUC	Sensitivity	Specificity
	HC vs AD	.795 ±.049	.847 ±.024	.761 ±.039	.853 ±.078
SVM	HC vs MCI	.644 ±.035	.699 ±.046	.673 ±.080	.602 ±.084
	MCI vs AD	.732 ±.051	.715 ±.112	.134 ±.101	.947 ±.049
	HC vs AD	.849 ±.024	.851 ±.041	.803 ±.071	.895 ±.053
Ours	HC vs MCI	.703 ±.046	.701 ±.034	.757 ±.052	.635 ±.032
	MCI vs AD	.771 ±.081	.713 ±.090	.635 ±.163	.807 ±.064

Table 2. **Two Modalities**. Classification results of different groups with VBM-MRI and FDG-PET.

Method	Group	Accuracy	ROC-AUC	Sensitivity	Specificity
	HC vs AD	.824 ±.048	.838 ±.053	.782 ±.135	.846 ±.031
SVM	HC vs MCI	.632 ±.052	.689 ±.054	.710 ±.093	.531 ±.076
	MCI vs AD	.783 ±.035	.794 ±.057	.282 ±.127	.963 ±.039
	HC vs AD	.863 ±.051	.868 ±.042	.788 ±.103	.932 ±.017
Ours	HC vs MCI	.717 ±.021	.702 ±.034	.754 ±.068	.669 ±.115
	MCI vs AD	.848 ±.036	.804 ±.109	.620 ±.186	.926 ±.047

Table 3. **Three Modalities**. Classification results of different groups with three modalities.

| Method	Group	Accuracy	ROC-AUC	Sensitivity	Specificity
	HC vs AD	.831 ±.065	.858 ±.044	.792 ±.059	.845 ±.095
SVM	HC vs MCI	.638 ±.071	.689 ±.039	.705 ±.125	.522 ±.091
	MCI vs AD	.790 ±.023	.818 ±.042	.396 ±.080	.931 ±.054
	HC vs AD	.871 ±.025	.884 ±.048	.825 ±.092	.901 ±.027
Ours	HC vs MCI	.723 ±.046	.714 ±.047	.723 ±.106	.706 ±.110
	MCI vs AD	.861 ±.079	.849 ±.099	.773 ±.109	.887 ±.095
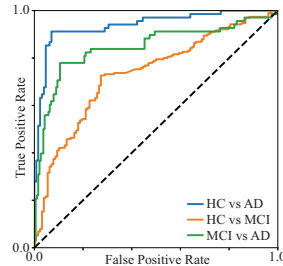
**Fig. 1**. ROC Curve of different groups with three modalities.

jects for each class and obtain a set of scalar scores for each ROI, which is the heatmap. For the heatmap of each subject, we rank the scores in descending order and only keep the top 20 salient ROIs. We used the BrainNet Viewer [19] to plot those salient ROIs on the brain surface.

Fig. 2(a-c) illustrates the lateral, medial, and ventral view of brain surface for two subjects selected from the HC, AD, and MCI groups. We named the selected ROIs as the biomarkers for identifying each group: node importance values were encoded as the color of regions on brain surface. In Fig. 2(a), putamen, pallidum, Superior parietal gyrus, and gyrus rectus were selected for HC; In Fig. 2(b), putamen, pallidum, Superior occipital and parietal gyrus were selected for MCI; In Fig. 2(c), putamen, pallidum, thalamus, Caudate nucleus, Superior occipital and Lingual gyrus were selected for AD. Putamen and pallidum were important for each group.

Given that each ROI represents a node in the graph, after we computed the node importance of each subject, we average those node importance in HC, AD, and MCI group separately. Then we calculate the Euclidean distance between ROIs based on the average node importance for three groups.

(a) HC Subject

(b) MCI Subject

(c) AD Subject
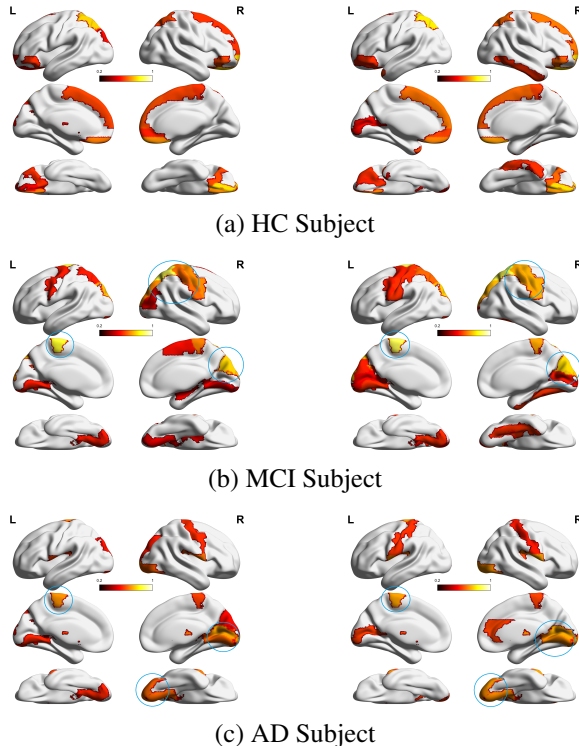
Fig. 2. Interpreting top 20 selected salient ROIs of two different individuals in HC, MCI, and AD group respectively. The color bar ranges from 0.2 to 1.0. The bright-yellow color indicates a high score, while dark-red color indicates a low score. The common detected salient ROIs across different subjects inside MCI and AD are circled in blue.
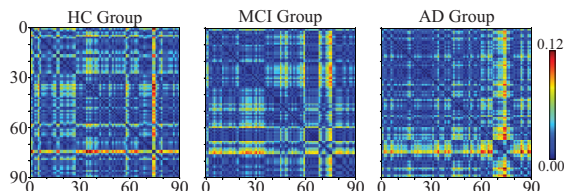

Fig. 3. The Euclidean distance between ROIs by averaging node importance of subjects inside each group.

In Fig. 3, we plot the distance matrices between ROIs. The average node importance of HC, AD, and MCI groups were substantially different from each other, suggesting that the salient ROIs are different in each group.

3.3. Prediction Score

In this subsection, we collected the cognitive test scores including the MMSE and ADAS13 from ADNI dataset. We used the learned features of ROIs in the last GCN layer to predict the cognitive tests scores of HC, AD, and MCI subjects. We normalized the learned features as the mean of 0 and the standard deviation of 1. The linear regression model
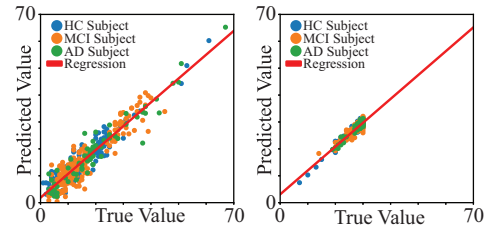
Fig. 4. Regression between the predicted and true values for ADAS13 and MMSE test scores. The HC, AD, and MCI subjects are plotted as the blue, orange, and green color respectively. The red line is the fitted regression line.

Table 4. The reported value of evaluation metrics between the predicted and true scores of ADAS13 and MMSE.

Evaluation Metrics	ADAS13	MMSE
Pearson Correlation Coefficient	0.941	0.942
Mean Absolute Error	2.682	0.848
Root Mean Squared Error	3.527	1.084
R Squared	0.886	0.887

can be calculated to fit the relationship between the learned features and the cognitive test scores.

In Fig. 4, we show the regression measures between the predicted and true values for ADAS13 and MMSE test scores. It provides a visual perception of how accurate the prediction is for the given test. Based on the fit of the regression line, we can infer that there is substantial correlation between the prediction and ground truth.

In Table 4, we show the numeric performance of the regression results for all of HC, AD, and MCI subjects. It contains the value of Pearson correlation coefficient, mean absolute error, root mean squared error (RMSE), and R-squared measure for the ADAS13 and MMSE test scores. From the results of table, the correlation measures of these two cognitive test are higher than 0.9 and the R-squared measures are also very high.

4. CONCLUSIONS

In this paper, we proposed an interpretable Graph Convolutional Network (GCN) framework for the identification and classification of Alzheimer's Disease (AD) using multi-modality brain imaging data. We applied the interpretable Gradient Class Activation Mapping (Grad-CAM) technique to interpret the salient ROIs and found that the putamen and pallidum biomakers were very important to identify the HC, AD and MCI. We extended the current explanation method of GCN model to discover the neurological biomakers for multi-modal brain imaging analysis. Besides the promising classification performance, our interpretation of salient ROIs demonstrated the individual- and group-level patterns in HC, AD, and MCI groups respectively. This suggests that our method can be applied to interpret the salient ROIs for much more modalities imaging data.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by ADNI [16]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] Hua Wang, Feiping Nie, Heng Huang, Shannon Risacher, Chris Ding, Andrew J Saykin, and Li Shen, "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance," in *ICCV*. IEEE, 2011, pp. 557–562.

[2] Lei Du, Kefei Liu, Xiaohui Yao, Shannon Risacher, Junwei Han, Andrew Saykin, Lei Guo, and Li Shen, "Multitask sparse canonical correlation analysis with application to multi-modal brain imaging genetics," *IEEE/ACM Transactions On Computational Biology and Bioinformatics*, 2019.

[3] Warold Izquierdo, Harold Martin, Mercedes Cabrerizo, Armando Barreto, Jean Andrian, Naphtali Rishe, Sergio Gonzalez-Arias, David Loewenstein, Ranjan Duara, and Malek Adjouadi, "Robust prediction of cognitive test scores in alzheimer's patients," in *SPMB*. IEEE, 2017, pp. 1–7.

[4] Elena Rykhlevskaia, Gabriele Gratton, and Monica Fabiani, "Combining structural and functional neuroimaging data for studying brain connectivity: a review," *Psychophysiology*, vol. 45, no. 2, pp. 173–187, 2008.

[5] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen, "An end-to-end deep learning architecture for graph classification," in *AAAI*, 2018.

[6] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[7] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *NeurIPS*, 2015, pp. 2224–2232.

[8] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl, "Neural message passing for quantum chemistry," in *ICML*. PMLR, 2017, pp. 1263–1272.

[9] Xiaoxiao Li, Zhou Yuan, Da Nicha, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan, "BrainGNN: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, p. 102233, 2021.

[10] Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang, "BrainNNExplainer: An interpretable graph neural network framework for brain network based disease analysis," in *ICML Workshop on Interpretable Machine Learning in Healthcare*, 2021.

[11] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *NeurIPS*. 2019, vol. 32, p. 9240, NIH Public Access.

[12] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang, "Parameterized explainer for graph neural network," *arXiv preprint arXiv:2011.04573*, 2020.

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

[14] Xi Zhang, Lifang He, Kun Chen, Yuan Luo, Jiayu Zhou, and Fei Wang, "Multi-view graph convolutional network and its applications on neuroimage analysis for parkinson's disease," in *AMIA Annual Symposium Proceedings*. AMIA, 2018, vol. 2018, pp. 1147–1156.

[15] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann, "Explainability methods for graph convolutional neural networks," in *CVPR*, 2019, pp. 10772–10781.

[16] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett, "The alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, pp. 869, 2005.

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

[18] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc 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.

[19] Mingrui Xia, Jinhui Wang, and Yong He, "BrainNet viewer: a network visualization tool for human brain connectomics," *PloS one*, vol. 8, no. 7, pp. e68910, 2013.