Information Processing and Management 59 (2022) 102901

Contents lists available at ScienceDirect

Information Processing and Management

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

Dementia analysis from functional connectivity network with graph neural networks

Lujing Wang [a,1], Weifeng Yuan [b,1], Lu Zeng [c], Jie Xu [c], Yujie Mo [c], Xinxiang Zhao [a,*], Liang Peng [c,*]

[a] Department of Radiology, The Second Affiliated Hospital of Kunming Medical University, Kunming 650101, China
[b] Department of Radiology, Clinical Medical College and The First Affiliated Hospital of Chengdu Medical College, Chengdu 610500, China
[c] School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China

ARTICLE INFO

Keywords:
Brain Functional Connectivity
Graph neural networks
Structure learning
Self-attention
Feature selection

ABSTRACT

Deep learning methods have been widely applied for disease diagnosis on resting-state fMRI (rs-fMRI) data, but they are incapable of investigating global relationships between different brain regions as well as ignoring the interpretability. To address these issues, this paper presents a new graph neural network framework for brain disease diagnosis via jointly learning global relationships and selecting the most discriminative brain regions. Specifically, we first design a self-attention structure learning to capture the global interactions between brain regions for achieving diagnosis effectiveness, and theoretically integrate a feature selection method to reduce the noise influence as well as achieve interpretability. Experiment results on three neurological diseases datasets show the effectiveness of our method, compared to the comparison methods, in terms of diagnostic performance and interpretability.

1. Introduction

Dementia is a neurological disease syndrome in which patients experience deterioration of memory and daily living skills. There are many different types of dementia, such as Alzheimers disease (*i.e.,* AD) (Khachaturian, 1985), obsessive–compulsive disorder (*i.e.,* OCD) (Stein, 2002) and frontotemporal dementia (*i.e.,* FTD) (Neary, Snowden, & Mann, 2005). By 2021, there will be more than 50 million dementia patients worldwide with about 10 million new cases each year. Unfortunately, there are currently no effective treatments for dementia, but early diagnosis allows patients to slow the progress of disease, allowing them to function normally for a longer period of time (Zhu, Ma, Yuan, & Zhu, 2022). Measuring Blood Oxygen Level Dependent (*i.e.,* BOLD) indicators as a physiological indicator of brain activity, rs-fMRI has been extensively used to identify markers of functional pathology of brain regions for the diagnosis of MCI (Gan et al., 2021; Hu et al., 2021; Zhu et al., 2022). Defined as the relevance among brain regions, Functional Connectivity Networks (FCNs) have been extensively studied in the search for associations between underlying functional structures of brain regions and neurological diseases, and it has been extensively investigated in the field of medical dementia analysis (Zhu et al., 2021).

In the past decades, numerous shallow learning methods have been performed to analyze neurological diseases. As the most widely studied shallow learning methods, Support Vector Machine (SVM) is trained to perform boundary delineation on the row features and is used for medical image classification. Zhang, Zhang, Chen, Lee, and Shen (2017) point out that high order FCNs

* Corresponding authors.
E-mail addresses: zzhaoxinxiang06@126.com (X. Zhao), 202011081627@std.uestc.edu.cn (L. Peng).
[1] Lujing Wang and Weifeng Yuan contributed equally to this work.

https://doi.org/10.1016/j.ipm.2022.102901
Received 2 October 2021; Received in revised form 9 January 2022; Accepted 9 February 2022
Available online 8 April 2022

are essential for studying brain diseases, and proposed an SVM algorithm on low-order and high-order FCNs. Recently, Gan et al. (2021) jointly perform feature selection (Mishra & Singh, 2020) and feature extraction in a unified training framework and obtained remarkable results. Although shallow learning methods can achieve relatively consistent results on small and simple datasets, their ability to work on larger and complex datasets is limited (Yuan, Zhong, Lei, Zhu, & Hu, 2021). Recently, deep learning methods draw a lot of attention in disease diagnosis. In general, models trained with deep learning methods provide more accurate results than those trained with traditional methods. For Alzheimers disease, Farooq, Anwar, Awais, and Rehman (2017) argue that deep neural network-based approaches can provide better feature representation than traditional shallow learning methods. For multi-view data, deep multi-view learning methods executed in some unsupervised or semi-supervised scenarios also show remarkable results (Abdi, Shamsuddin, Hasan, & Piran, 2019; Peng, Kong, Liu, & Kuang, 2021; Xu et al., 2021). But, a typical Convolutional Neural Network (CNN) (Jin, McCann, Froustey, & Unser, 2017) cannot directly process data with non-Euclidean structures because the convolution kernel defined in the CNN is translationally invariant (Kauderer-Abrams, 2020). In other words, these methods cannot account for the structural information in the data. To fully utilize the structural and semantic information, graph convolutional neural networks (GCNs) (Isufi, Pocchiari, & Hanjalic, 2021; Kipf & Welling, 2017; Veličković et al., 2018) are more suitable for brain disease treatment by merging the features between brain regions based on the structural information provided during feature extraction. Compared to CNN-based methods, GCN-based methods can extract essential features from structural information in non-Euclidean data. With this background, in this research, we focus on the development of GCNs for brain disease diagnosis because FCNs represent relationships between brain regions.

Although significant advancement has been achieved in the diagnosis of brain diseases with the existing shallow learning methods and deep learning methods, the performances of these methods are limited due to the following drawbacks. Firstly, the significant challenge in brain disease diagnosis is to characterizes the relations among the brain regions which is a key factor in achieving the desired algorithmic goal. Traditional methods are implemented with shallow models (*e.g.,* kNN (Zhang, Li, Zong, Zhu, & Wang, 2018), SVM, random forests (Breiman, 2001), and decision trees (Quinlan, 1986)) leading to limited discriminative ability on real medical applications. On the contrary, deep models, such as CNNs and GNNs, try to extract high-level features from original semantic information or structure information in an iterative manner, leading highly discriminative representation for brain disease diagnosis. Secondly, the initial structure information obtained by original FCNs data containing some irrelevant features easily results in incorrect correlations among brain regions. Moreover, the initial structure information with incorrect correlations affects the process of feature learning and the result of the classification model. The original graph structure is constructed from raw data (*e.g.,* kNN) and cannot represent the relationship between brain regions well. Therefore, how to capture structural information between brain regions is crucial to boost the performance of neurological disease diagnosis. In this regard, the underlying relationship in the graph structure is critical for the GNN models as the wrong (*i.e.,* incorrect correlations) and insufficient (*i.e.,* lack correlations) relation information will be passed to the network construction to mislead and limit the model effectiveness. In addition, existing methods are difficult to provide useful experience for brain diseases. The main impediment is that most of researches have been focused on improving accuracy while ignoring the interpretability for brain disease diagnosis.

Considering the above challenges, we propose a deep graph convolutional network framework with self-attention structure learning and feature selection, as shown in Fig. 1. To handle the first challenge, in this study, we propose to adopt a deep graph neural network framework with a self-attention mechanism. More specifically, we first employed a GCN to capture and aggregate information according to the relationships between different brain regions. To exploit the latent relationships between brain regions, we develop a self-attention mechanism in the GCN layers. The graph structure in the proposed framework can be adaptively adjusted based on the representation of the feature learning result. In other words, the relations among the brain regions is iteratively updated during the learning process. With updated correlations among the brain regions, the latent relations can capture both global and local information, which encourages the network to understand the attributes of brain disease data. To handle the second challenge, inspired by the real diagnostic process, there are some brain areas that are important for brain disease diagnosis. We apply a representation selection module to capture the essential discriminating brain regions for the brain disease diagnosis. Comparing with previous approaches, selects significant brain regions in a shallow learning framework and thus has limited capability to discriminate. Our solution is to jointly optimize GCN with self-attention mechanism and to select significant features. Moreover, we discuss the interpretive diagnosis results of the potential brain area associations found in the self-attention mechanism and the important brain regions under feature selection in our method. To study the effect of our method for clinical implications, we show that the results of proposed framework which can provide a plausible explanation for brain disease diagnosis. In conclusion, we list the main contributions as follows.

(i) We introduce a new deep graph learning framework with self-attention structure learning to characterize the relationships between brain regions, and validate its effectiveness with extensive ablation studies.

(ii) We jointly learn the proposed deep graph learning model and select the discriminative brain regions. Feature selection reduces the impact of noise and promotes the diagnostic accuracy of brain diseases.

(iii) Our method takes medical interpretability into account, which has important clinical implications for diagnosing various brain diseases. On the contrary, a few deep learning methods consider the interpretability of the model for medical image.

2. Related work

2.1. Graph convolutional networks

To aid deep learning on graph structured data, such as communication networks and citation networks, a wide number of GNN algorithms, *e.g.,* (Kipf & Welling, 2017; Li et al., 2021; Liao, Deng, Wan, & Liu, 2022; Veličković et al., 2018), have been

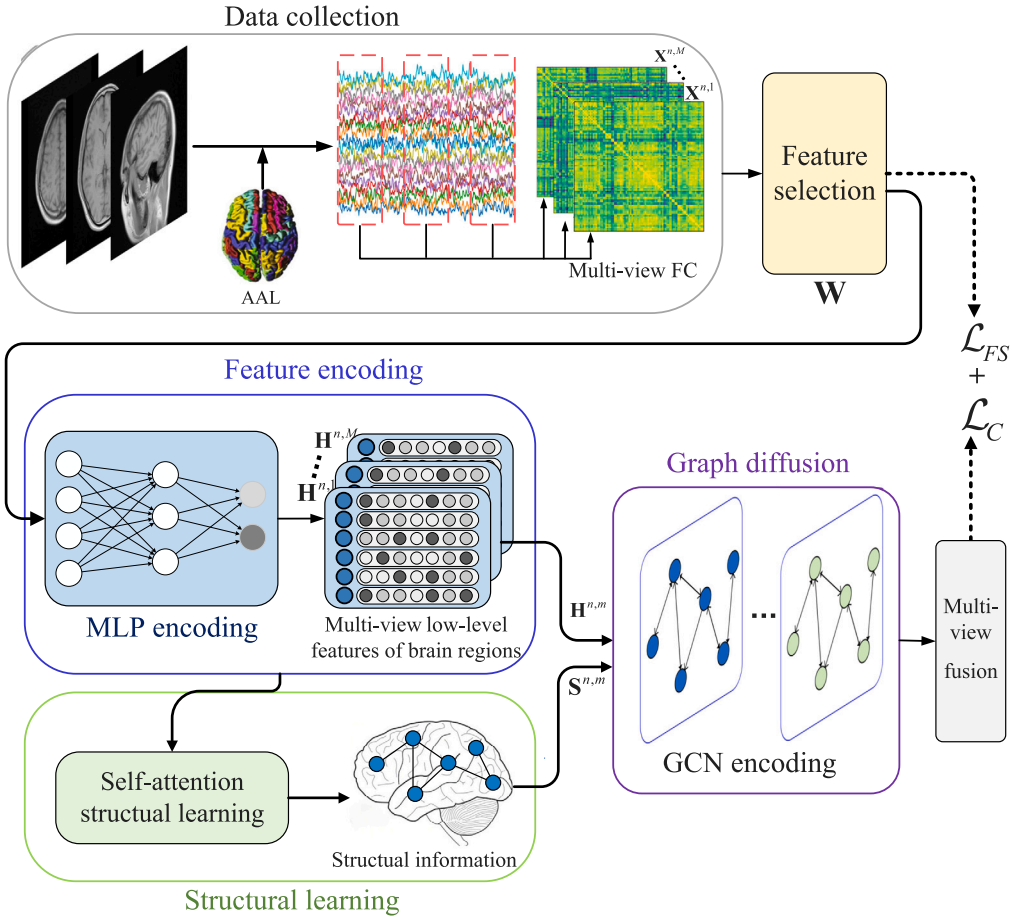


Fig. 1. The flowchart of our proposed framework for personalized neurological diseases, which first collects and pre-processes the datasets (*i.e.,* gray box), and then utilizes feature selection (*i.e.,* yellow box) to select discriminative brain regions and avoid the noise. Then, the feature encoding module (*i.e.,* blue box) is used for learning the local and low-level features $\mathbf{H}^{n,m}$ of each brain region. A self-attention structural learning module (*i.e.,* green box) is proposed to capture correlation $\mathbf{S}^{n,m}$ among brain regions. After that, the deep graph diffusion module (*i.e.,* purple box) is proposed to learn the high-level representation $\mathbf{H}_T^{n,m}$ for each view. Finally we fuse the features of multiple views to predict whether a subject is normal or not.

presented. Bruna, Zaremba, Szlam, and LeCun (2014) propose generalizing the convolution operation from Euclidean data to non-Euclidean data with the Fourier basis of a particular graph, with the goal of performing the classification task. To simplify spectral GNNs, Defferrard, Bresson, and Vandergheynst (2016) use Chebyshev polynomials as the convolution filter. To generate node representations, Graph Convolutional Networks (GCN) (Kipf & Welling, 2017) suggests a localized spectral convolution filter approximation averaging feature across first-order neighborhood nodes. Simple Graph Convolution (SGC) (Wu et al., 2019) reduces graph convolution to a linear model while maintaining competitive performance. However, the effectiveness of GCN highly relies on the quality of the graph. With a low-quality graph, *i.e.,* a graph with incorrect edges (Fan & Wang, 2022), the classification performance of the GCN model is affected. Graph Attention Networks (GAT) (Veličković et al., 2018) conducts graph learning by dynamically updating weights of edges through attention mechanism. Following in the footsteps of GAT, Graph Learning Convolutional Networks (GLCN) (Jiang, Zhang, Lin, Tang, & Luo, 2019) improves learning by combining graph learning and graph convolution in a single network architecture.

Compare with GCN and SGC, our method can dynamically update the graph structure. In contrast to GAT, our method is able to dynamically discover new relationships while selecting important brain regions. Unlike GLCN, our approach is automatically data-driven, *i.e.,* no loss function is specified for graph learning. Moreover, our method applies graph learning to brain disease diagnosis to capture structural information between brain regions which further provides an interpretation of the GCN in the diagnosis of brain diseases.

2.2. GCNs for disease prediction

In the medical area, there has recently been a greater emphasis on graph learning on unstructured data, such as functional connectivity network (Jiang, Cao, Xu, Yang, & Zaiane, 2020; Parisot et al., 2018; Zhu et al., 2022). For example, Parisot et al.

(2018) extend GCN model for semi-supervised brain disease prediction with neuroimaging data, where nodes represent patients and edges reflect the interaction and relationship between two patients. Hi-GCN (Jiang et al., 2020) presents a hierarchical GCN framework for learning graph feature embedding while also taking network structure and subject association into account for the classification of ASD and AD. Zhu et al. (2022) integrate dynamic graph learning and graph convolution to learn the optimal graph structure and provides feature interpretability for brain diseases diagnosis. However, two issues remain in the current GCN-based model for brain disease prediction. First, the graph structure is fixed and keep it invariant during the process of feature learning, which cannot represent the relationship between different brain regions well. Second, the input raw features are redundant and noisy, and thus reducing the performance.

In contrast to previous GCN-based disease prediction researches, *e.g.,* (Jiang et al., 2020; Parisot et al., 2018), we investigate an under-explored but realistic challenge of dynamic characterizing the connection between brain regions and address the unique challenge of medical imaging on the input feature with noise. Moreover, in contrast to Zhu et al. (2022), we theoretically prove the convergence of our method and apply self-attention structure learning to capture global information among brain regions.

3. Method

3.1. Notations

We abstract the rs-fMRI time series data, shown at the top of Fig. 1. Suppose there are N subjects involved in the training process, given the rs-fMRI data of each subject, Automated-Anatomical-Labeling (*i.e.,* AAL) template (Rolls, Huang, Lin, Feng, & Joliot, 2020) was applied to parcellate the collect rs-fMRI data into $B = 90$ brain regions. Following existing works on the rs-fMRI time series analysis (Hu et al., 2021), sliding window approach was used to cut the long time rs-fMRI into multiple short time series M. Furthermore, by calculating the correlation between all brain regions, we can construct a symmetric correlation matrix, which representing the relationship between brain regions. In particular, we denote the multiple FCNs of all subjects as $\mathbf{X} = \{\mathbf{X}^1 \cdots \mathbf{X}^N\}$, and each subject is assigned a label $y^n \in \{0, 1\}$ indicates whether the nth subject belongs to normal cases (*i.e.,* 0) or not (*i.e.,* 1). For each subject, we separate the time-series data into multiple small segments $\mathbf{X}^n = \{\mathbf{X}^{n,1} \cdots \mathbf{X}^{1,M}\}$ where $\mathbf{X}^{n,m} \in \mathbb{R}^{B\times B}$ represents the FCNs of the mth view in the nth subject. The overall architecture is illustrated in Fig. 1. Our proposed methods conceptually consists of three blocks, *i.e.,* (i) Brain features extraction (Section 3.2), (ii) Self-attention structure learning (Section 3.3), and (iii) Brain region selection (Section 3.4).

3.2. Brain features extraction

Analysis of brain regions can help us determine neurological disorders at the level of brain regions. By representing the correlation between brain regions as a graph structure, Structural information can model the relationships between brain regions and provide high-level features. Thus, we have to extract useful information from graph structure among brain regions. Deep graph convolutional network have the ability to analyzes the graph-structured data via message passing between the nodes of graphs and thus is an intuitive choice regarding the learning of encoding brain regions structural data in our setting. Given the original brain functional connectivity data in M segments of N subjects, we first encode the low-level features for each brain region with local information. To obtain high-level features, we intend to take advantage of the underlying correlations among brain regions by adopting a deep graph neural networks. For the purpose of enhancing the information of representation, we fuse the output high-level features of M segments to predict results.

3.2.1. Feature encoding

To extract features of each brain region, in this paper, we first employ the Multi-Layer Perceptron (MLP) on each brain functional connectivity segment $\mathbf{X}^{n,m} \in \{\mathbf{X}^{n,1} \cdots \mathbf{X}^{n,M}\}$ to generate low-level features $\mathbf{H}^{n,m}$ for each brain regions with its local information, which is defined as

$$\mathbf{H}^{n,m} = \mathrm{MLP}(\mathbf{X}^{n,m}), \tag{1}$$

where $\mathrm{MLP}(\cdot)$ consists of multiple fully connected hidden layers with non-linear activation function after each hidden layer, and we obtain the d_h dimensional low-level representations of a set of brain regions as $\mathbf{H}^{n,m} \in \mathbb{R}^{B\times d_h}$.

3.2.2. Graph diffusion

In deep neural networks, GNN was introduced by Scarselli, Gori, Tsoi, Hagenbuchner, and Monfardini (2008) and is a generalization of recurrent neural network which can handle structural data. Recently, to conduct operation of convolution on structural data, Kipf and Welling (2017) propose deep graph convolutional neural network to learn the deep features while preserving the local structure stored in the fixed graph (Kipf & Welling, 2017). Specifically, given the graph structure $\mathbf{A}^{n,m} \in \mathbf{R}^{B\times B}$ storing the relationship among the brain regions for the mth segment in the nth subject (*i.e.,* $\mathbf{X}^{n,m}$), the GCN model includes several hidden layers and one perceptron layer while each hidden layer involves two operations, *i.e.,* propagating the local structure $\mathbf{A}$ and feature leaning. More specifically, the layer-wise propagation on the tth hidden GCN layer is

$$\mathbf{H}_t^{n,m} = \sigma((\mathbf{D}^{n,m})^{-\frac{1}{2}}\mathbf{A}^{n,m}(\mathbf{D}^{n,m})^{-\frac{1}{2}}\mathbf{H}_{t-1}^{n,m}\mathbf{\Theta}^{(t)}), \tag{2}$$

where $\mathbf{D}^{n,m}$ is a diagonal matrix of $\mathbf{A}^{n,m}$, and $\mathbf{W}^{(t-1)} \in \mathbf{R}^{d_{t-1}\times d_t}$ is a weight matrix which need to be trained in the $(t-1)$-th layer. $\mathbf{H}^{(t)} \in \mathbf{R}^{n\times d_{(t-1)}}$ denotes the representation in the tth layer and $\sigma(.)$ represent the function for activation operation. Here we use $\mathbf{H}_T^{n,m}$ to represents the output representation of the final diffusion layer.

3.2.3. Multi-view fusion

As mentioned above, for M segments of brain functional connectivity data of nth subject, we construct the MLP as well as the GCN networks to extract features for each segment and obtain the final M representations (*i.e.,* $\{\mathbf{H}_T^{n,1} \cdots \mathbf{H}_T^{n,M}\}$). Based on this, the final perceptron layer is defined as

$$\mathbf{z}^n = \mathrm{softmax}(\mathrm{Linear}(\mathrm{Concat}(\mathbf{H}_T^{n,1} \cdots \mathbf{H}_T^{n,M}))), \tag{3}$$

where $\mathrm{Concat}(\cdot)$ means channel-concatenate and $\mathrm{Linear}(\cdot)$ is used to regress the probability of each category. The final output $\mathbf{z}^n$ denotes the label prediction by Softmax operation for the nth sample. Finally, the cross-entropy loss function to minimize the parameters

$$\mathcal{L}_C = -\sum_{n=1}^{N} y^n \log(z^n), \tag{4}$$

where N indicates the set of labeled subjects, y^n is the ground truth and z^n is the corresponding predictions.

3.3. Self-attention structure learning

The structural information in the graph data is critical for the GNN models to discover discriminative representations as the noise (*e.g.,* incorrect edges) and insufficient (*e.g.,* lack edges) information will be passed to the network construction to mislead and limit the model effectiveness. We thus propose graph structure learning with self-attention mechanism to automatically capture the latent structural information that model the deep connection of regions located in the brain. Further, feature embeddings GNN method and self-attention graph structure learning method is trained in the end-to-end data-driven learning strategy simultaneously. Besides, it can provide interpretable result on the correlation between brain regions. Given a set of potentially related items, the self-attentive method establishes their relevance with each other based on their features (*e.g.,* a brain region is related to other brain regions).

3.3.1. Self-attention learning

Concretely, given the d_h dimensional representations of a set of brain regions $\mathbf{H}^{n,m} \in \mathbb{R}^{B\times d_h}$ in the mth segment of the nth sample, we propose a self-attention structure learning module $F(\mathbf{H}^{n,m}; \mathbf{W}_Q; \mathbf{W}_K)$, which is parameterized by $\mathbf{W}_Q \in \mathbb{R}^{d_h\times d_v}$ and $\mathbf{W}_K \in \mathbb{R}^{d_h\times d_v}$, to learn latent structure information among different brain regions that are denoted by $\mathbf{S}^{n,m} \in \mathbb{R}^{B\times B}$. Packed into the matrix $\mathbf{H}^{n,m}$, the self-attention structure learning module first transforms the input matrix $\mathbf{H}^{n,m}$ into queries $\mathbf{Q} \in \mathbb{R}^{B\times d_v}$ and keys $\mathbf{K} \in \mathbb{R}^{B\times d_v}$, to obtain sufficient expressive power as follows:

$$\begin{cases} \mathbf{Q} = \mathbf{H}^{n,m}\mathbf{W}_Q & \text{Embedding of the queries set,} \\ \mathbf{K} = \mathbf{H}^{n,m}\mathbf{W}_K & \text{Embedding of the keys set.} \end{cases} \tag{5}$$

It can be considered as the input representations go through two linear layer parameterized by $\mathbf{W}_Q$ and $\mathbf{W}_K$ respectively, to obtain higher-level features. Then, we capture the latent structural information $\mathbf{S} \in \mathbb{R}^{B\times B}$ where $s_{i,j}$ denote the weight to non-negative edge between brain region i and brain region j by computing the dot-product of the queries with all keys:

$$\mathbf{S}^{n,m} = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_v}}), \tag{6}$$

where $\mathrm{softmax}(\cdot)$ is used to normalize the attention values and d_v is dimension of $\mathbf{Q}$ and $\mathbf{K}$ which is used as the scaling factor. Based on but different from the GAT (Veličković et al., 2018) where the latent structural information $\mathbf{S}^{n,m}$ will be combine with the original structural information $\mathbf{A}^{n,m}$ (by taking a dot-product with $\mathbf{A}^{n,m}$), the $\mathbf{S}^{n,m}$ generated by proposed self-attention structure learning module does not need to be combined with $\mathbf{A}^{n,m}$ to avoid the effects of noise and lack of connectivity in the original structural information $\mathbf{A}^{n,m}$.

Next, we theoretically analyze how self-attention structure learning module can captures global relationships between brain regions which are critical for brain disease diagnosis. Given representation $\mathbf{H} \in \mathbb{R}^{B\times d_h}$ of a subject (*i.e.,* a node sample) including B brain regions feature, graph structural matrix $\mathbf{S} \in \mathbb{R}^{B\times B}$ learned by Eq. (6). In the case of deep GCN model (*i.e.,* see Eq. (2)), we can approximate the prediction (*i.e.,* p) of the GCN model for classification as $p \approx \theta^T\mathbf{S}\mathbf{H} + b$ by calculating the first-order Taylor expansion where θ is the weight vector and $\mathbf{b}$ is the bias vector. Let $I = \mathbf{S}\mathbf{H}$ for the simplicity, and $I_c \in I$ is one brain region in I. Therefore, the gradient of θ in binary classification at the point (brain region) I_c are defined by the class score derivative $\theta = \frac{\partial p}{\partial I}|_{I_c}$, where the magnitude of the derivative indicates which brain regions need to be changed the least to affect the class score the most.

3.3.2. Multi-head self-attention

In general, for a single learning function for self-attention structure, it is difficult to uncover all potential correlations between brain regions in the single feature subspace. Inspired by priors self-attention approaches, we extend the single self-attention structure learning function to a multi-head self-attention structure learning function (*i.e.,* $MF(\mathbf{H}^{n,m}; \{\mathbf{W}_Q^1 \cdots \mathbf{W}_Q^C\}; \{\mathbf{W}_K^1 \cdots \mathbf{W}_K^C\})$ where C is the number of self-attention function) to effectively extract the potential correlations among brain regions in different feature subspace. In particular, $MF(\cdot)$ execute C times Eq. (5) in parallel and obtains multiple different latent structural information $\{\mathbf{S}_1^{n,m}, \mathbf{S}_2^{n,m}, \cdots, \mathbf{S}_C^{n,m}\}$, which is defined as

$$\hat{\mathbf{S}}^{n,m} = MF(\mathbf{H}^{n,m}; \{\mathbf{W}_Q^1 \cdots \mathbf{W}_Q^C\}; \{\mathbf{W}_K^1 \cdots \mathbf{W}_K^C\})$$

$$= \frac{1}{C}\sum_{c=1}^{C} F(\mathbf{H}^{n,m}; \mathbf{W}_Q^c; \mathbf{W}_K^c) \tag{7}$$
$$= \frac{1}{C}(\mathbf{S}_1^{n,m} + \mathbf{S}_2^{n,m} \cdots \mathbf{S}_C^{n,m}),$$

where $\hat{\mathbf{S}}^{n,m}$ is the average structure information from different attention subspaces which will be used to replace $\mathbf{A}^{n,m}$ in Eq. (2). Note that, $\mathbf{S}^{n,m}$ is dynamically updated during training proceeds under an end-to-end training strategy.

3.4. Brain region selection

Brain functional connectivity data on rs-fMRI is naturally a complex network and useful information to understand neurological disorders. On the one hand, original brain region correlation data contain large amount of unrelated feature and noisy which can disrupt the learning process of neural networks. On the other hand, the original feature contain the physical meanings which play an essential role for disease determination (Zhu, Zhang, Zhu, Zhu, & Gao, 2020). Hence, to keep some of discriminative features while decrease the risk of noisy and irrelevant features, we conduct a sparse learning-based feature selection on original data via the $\ell_{2,1}$-norm regularization. Specifically, For any matrix $\mathbf{X}^{n,m} \in \mathbb{R}^{p\times q}$, sparse learning-based feature selection contains a data reconstruction item and a sparse learning item in the regularization loss function:

$$\mathcal{L}_{FS} = \|\mathbf{X}^{n,m} - \mathbf{X}^{n,m}\mathbf{W}\|_F^2 + \alpha \|\mathbf{W}\|_{2,1}, \tag{8}$$

where $\|\cdot\|_F^2$ is the Frobenius norm and $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{p}\sqrt{\sum_{j=1}^{q}\mathbf{W}(i,j)^2}$ is $\ell_{2,1}$-norm. $\mathbf{W} \in \mathbb{R}^{q\times q}$ is the learnable parameters for channels in brain region $\mathbf{X}$ and α is a tuning parameter to balance two items. By optimizing above equation, $\mathbf{W}$ is optimized as a sparse matrix in which some rows are equal to a small values. Then the corresponding features multiply these rows and the value in the original corresponding features will be eliminated in the corresponding new features $\mathbf{X}^{n,m}\mathbf{W}$. Whats more, we will process the new features $\mathbf{X}^{n,m}\mathbf{W}$ instead of the original features $\mathbf{X}^{n,m}$ in Eq. (1) for feature extraction. To this end, coupling with the classification objective $\mathcal{L}_C$ in Eq. (4), the proposed method is trained with

$$\mathcal{L} = \mathcal{L}_C + \beta\mathcal{L}_{FS}, \tag{9}$$

to learn useful representations and discriminative features where β is a tuning parameter to balance two terms. Note that our proposed self-attention structure learning technique do not need loss function to optimize, which means it is an automated data-driven technique. The full training process of the presented frameworks can be concluded in the Algorithm 1.

Algorithm 1: The pseudo of our proposed method.

Require: MLP encoder $f(\cdot)$; GCN encoder $g(\cdot,\cdot)$; self-attention structure learning $(\{\mathbf{W}_Q^1 \cdots \mathbf{W}_Q^C\}; \{\mathbf{W}_K^1 \cdots \mathbf{W}_K^C\})$; feature selection $\mathbf{W}$.
Input: Multiple FCNs matrix $\mathbf{X}$; The number of multi-head attention C.
Output: Class prediction $\mathbf{Z}$.
Initialize parameters.
for epoch $\leftarrow 1, 2, \cdots$ **do**
  Generate selected brain region features $\mathbf{X}^{n,m}\mathbf{W}$ by the sparse matrix $\mathbf{W}$.
  Obtain low-level features of each brain region $\mathbf{H}^{n,m}$ by MLP encoder $f(\mathbf{X}^{n,m}\mathbf{W})$ (*i.e.,* Eq. (1)).
  Capture the latent graph structural $\hat{\mathbf{S}}^{n,m}$ by Eqs. (5), (6) and (7).
  Compute high-level feature $\mathbf{H}_T^{n,m}$ by GCN encoder $g(\mathbf{H}^{n,m}, \hat{\mathbf{S}}^{n,m})$ (*i.e.,* Eq. (2)).
  Fuse multi-view representations and predict label $\mathbf{z}^n$ by Eq. (3).
  Compute objective function $\mathcal{L}$ by Eq. (9).
  Back-propagate gradients to update weights in MLP $f(\cdot)$, self-attention structure learning $\{\mathbf{W}_Q^1 \cdots \mathbf{W}_Q^C\}; \{\mathbf{W}_K^1 \cdots \mathbf{W}_K^C\})$, GCN $g(\cdot,\cdot)$ and feature selection $\mathbf{W}$.
end for

3.5. Convergence analysis

According to Wen and Yin (2013), the $\ell_{2,0}$-norm constraint is able to output sparse results directly, however the $\ell_{2,0}$-norm is non-convexity, so it is hard to utilize the $\ell_{2,0}$-norm constraint directly. Because, the $\ell_{2,1}$-norm is a approximate solution of $\ell_{2,0}$-norm, we employ the $\ell_{2,1}$-norm regularization term in our method. Consequently, we discuss the convergence of the proposed method and have the following Theorem 1 to prove the convergence of Algorithm 1 with object function Eq. (9).

Theorem 1. *The objective function value of Eq. (9) monotonically decreases until Algorithm 1 converges.*

Proof. There are only two item in Eq. (9), *i.e.,* $\mathcal{L}_C$ and $\mathcal{L}_{FS}$. After the t-th iteration, we denote the objective function value of $\mathcal{L}_C$ as $J(\mathbf{W}^{(t)}, \Theta^{(t)}, \mathbf{X}, \mathbf{Y})$, where $\mathbf{W}^{(t)}$ and $\Theta^{(t)}$ denote the learnable matrix in feature selection and learnable parameters in remaining

network (*i.e.,* feature encoding, structural learning and graph diffusion), respectively. Our goal is to prove the following inequality

$$\begin{aligned} &J(\mathbf{W}^{(t+1)}, \Theta^{(t+1)}, \mathbf{X}, \mathbf{Y}) + \beta \left(\left\| \mathbf{X} - \mathbf{X}\mathbf{W}^{(t+1)} \right\|_F^2 + \alpha \|\mathbf{W}^{(t+1)}\|_{2,1} \right) \\ &\leq J(\mathbf{W}^{(t)}, \Theta^{(t)}, \mathbf{X}, \mathbf{Y}) + \beta \left(\left\| \mathbf{X} - \mathbf{X}\mathbf{W}^{(t)} \right\|_F^2 + \alpha \|\mathbf{W}^{(t)}\|_{2,1} \right), \end{aligned} \tag{10}$$

where α and β are fixed variables in Eqs. (8) and (9).

According to Amaran, Sahinidis, Sharda, and Bury (2016) and De Boer, Kroese, Mannor, and Rubinstein (2005), the cross-entropy loss encourage convex behavior and the method is guaranteed to converge to a local optimum with gradient descent optimization algorithms, so we have:

$$J(\mathbf{W}^{(t+1)}, \Theta^{(t+1)}, \mathbf{X}, \mathbf{Y}) \leq J(\mathbf{W}^{(t)}, \Theta^{(t)}, \mathbf{X}, \mathbf{Y}), \tag{11}$$

For feature selection loss $\mathcal{L}_{FS}$ in Eq. (9), with the convex behavior of the Frobenius norm $\|\cdot\|_F^2$ (Zhu et al., 2021), inequality Eq. (10) can be changed to:

$$\begin{aligned} &J(\mathbf{W}^{(t+1)}, \Theta^{(t+1)}, \mathbf{X}, \mathbf{Y}) + \beta \left(\left\| \mathbf{X} - \mathbf{X}\mathbf{W}^{(t+1)} \right\|_F^2 + \alpha \sum_{i=1}^{q} \frac{(\|\mathbf{w}_{i,\cdot}^{(t+1)}\|_2)^2}{2(\|\mathbf{w}_{i,\cdot}^{(t)}\|_2)} \right) \\ &\leq J(\mathbf{W}^{(t)}, \Theta^{(t)}, \mathbf{X}, \mathbf{Y}) + \beta \left(\left\| \mathbf{X} - \mathbf{X}\mathbf{W}^{(t)} \right\|_F^2 + \alpha \sum_{i=1}^{q} \frac{(\|\mathbf{w}_{i,\cdot}^{(t)}\|_2)^2}{2(\|\mathbf{w}_{i,\cdot}^{(t)}\|_2)} \right). \end{aligned} \tag{12}$$

According to Zhu et al. (2020), we obtain the following inequality

$$(\|\mathbf{w}_{i,\cdot}^{(t+1)}\|_2) - \frac{(\|\mathbf{w}_{i,\cdot}^{(t+1)}\|_2)^2}{2(\|\mathbf{w}_{i,\cdot}^{(t)}\|_2)} \leq (\|\mathbf{w}_{i,\cdot}^{(t)}\|_2) - \frac{(\|\mathbf{w}_{i,\cdot}^{(t)}\|_2)^2}{2(\|\mathbf{w}_{i,\cdot}^{(t)}\|_2)}. \tag{13}$$

By plugging Eq. (13) into Eq. (12) and considering qth $\mathbf{w}_{i,\cdot}$ together, we obtain Eq. (10). After this, we prove that the value of objective function Eq. (9) gradually decrease until Algorithm 1 converges. Hence, Theorem 1 is proved. □

4. Experiments

4.1. Setting

4.1.1. Datasets

Our designed brain functional connectivity analysis framework is conducted on three real brain disease datasets, including frontotemporal dementia (*i.e.,* FTD for short),[2] obsessive–compulsive disorder (*i.e.,* OCD for short) and Alzheimers Disease Neuroimaging Initiative (*i.e.,* ADNI for short).[3] In particular, the dataset FTD includes a total of 181 subjects, which consists of 86 normal cases (HC), 95 characterized as FTD. The dataset OCD contains 20 normal cases (HC) subjects and 62 OCD subjects. The dataset ADNI has 48 normal cases (HC) subjects and 59 AD subjects. The datasets used in our experiment are provided by different hospitals. To collect rs-fMRI data, all images were acquired on the 3.0T scanner at different centers with a gradient field strength of 80mT/m and gradient switching rate of 200mT/m/ms, using an eight-channel phased-array receiver coil.

We followed the literature (Gan et al., 2021) to process these rs-fMRI data. Briefly, the generated long-term rs-fMRI series were preprocessed using the DPARSF toolbox (Yan & Zang, 2010). And the first 10 time points of each sample were deleted to ensure the balance of magnetization. Then the rs-fMRI data are preprocessed with the data processing pipeline to reduce the impact of noise on FCNs construction: (i) slice timing correction; (ii) head motion correction; (iii) spatial normalization to the Montreal Neurological Institute (MNI) template; (iv) spatial smoothing using a full half-width Gaussian smoothing kernel; and (v) linear detrending and temporal bandpass filtering (0.01–0.10 Hz) for BOLD signals. Finally, the registered fMRI volumes were parcellated into 90 ROIs according to the AAL template. For multiview FCNs estimation, the number of time sub-series was set to $M = 5$ according to previous studies. The network structure can be denoted as $\mathbf{X}$ - $\mathbf{FC}_{200}$ - $\mathbf{FC}_{64}$ - $\mathbf{GCN}_{64}$ - $\mathbf{GCN}_{16}$, where $\mathbf{FC}_{200}$ represents the fully connected neural network with 200 neurons and $\mathbf{GCN}_{64}$ represents the GCN layer with 64 neurons.

4.1.2. Comparison methods

The comparative methods include five traditional algorithm (*i.e.,* SVM (Fan, Chang, Hsieh, Wang, & Lin, 2008), High-order Functional Connectivity (*i.e.,* HFC) (Zhang et al., 2017), Connectivity Network analysis with Hub Detection (*i.e.,* CNHD) (Wang, Huang, Liu, & Zhang, 2019), Sparse Connectivity Pattern (*i.e.,* SCP) (Eavani et al., 2015), and Brain FCNs analysis on Multiple Graph Fusion (*i.e.,* BMGF) (Gan et al., 2021)) and 3 deep learning algorithms (*i.e.,* Graph Convolutional Networks (*i.e.,* GCN) (Kipf & Welling, 2017), Deep Iterative and Adaptive Learning (*i.e.,* DIAL) (Chen, Wu, & Zaki, 2020) and Simplify Graph Convolutional networks (*i.e.,* SGC) (Wu et al., 2019)). We provide details of each comparison method below.

- SVM (Fan et al., 2008) is widely used for classification. SVM separates the points in the input raw space according to their class by selecting the hyper-plane.

[2] https://cind.ucsf.edu/research/grants/frontotemporal-lobar-degeneration-neuroimaging-initiative-0.

[3] http://adni.loni.usc.edu/.

Table 1
Comparison of performance (*i.e.,* ACC (%), SEN (%), SPE (%) and AUC (%)) of all methods on the dataset FTD.

Methods	ACC	SEN	SPE	AUC
SVM (Fan et al., 2008)	64.72 ± 3.28	62.85 ± 1.51	67.43 ± 2.84	65.80 ± 1.59
SCP (Eavani et al., 2015)	84.07 ± 4.23	82.59 ± 3.78	85.25 ± 3.82	83.95 ± 5.46
HFC (Zhang et al., 2017)	79.34 ± 4.42	78.57 ± 5.37	82.31 ± 4.42	79.57 ± 3.58
GCN (Kipf & Welling, 2017)	83.06 ± 3.71	83.67 ± 4.38	85.31 ± 3.36	85.31 ± 3.36
CNHD (Wang et al., 2019)	83.55 ± 3.69	81.20 ± 2.85	85.36 ± 3.13	82.64 ± 3.78
SGC (Wu et al., 2019)	84.56 ± 3.90	84.18 ± 3.13	84.35 ± 4.83	84.64 ± 3.53
DIAL (Chen et al., 2020)	85.87 ± 2.59	86.43 ± 3.15	85.59 ± 2.74	84.23 ± 2.17
BMGF (Gan et al., 2021)	86.96 ± 3.37	87.57 ± 3.08	84.10 ± 2.53	**87.19** ± 3.57
Proposed	**87.22**± 2.83	**87.90** ± 3.51	**86.14** ± 2.27	87.15 ± 2.21

Table 2
Comparison of performance (*i.e.,* ACC (%), SEN (%), SPE (%) and AUC (%)) of all methods on the dataset OCD.

Methods	ACC	SEN	SPE	AUC
SVM (Fan et al., 2008)	76.74 ± 3.38	73.39 ± 5.54	77.28 ± 2.72	78.61 ± 1.94
SCP (Eavani et al., 2015)	85.63 ± 3.32	85.64 ± 4.20	86.78 ± 3.93	86.85 ± 3.42
HFC (Zhang et al., 2017)	83.81 ± 2.47	83.31 ± 5.11	84.01 ± 2.62	83.26 ± 2.53
GCN (Kipf & Welling, 2017)	86.80 ± 3.15	86.67 ± 4.22	86.90 ± 3.97	87.21 ± 3.06
CNHD (Wang et al., 2019)	86.93 ± 4.11	87.01 ± 3.46	86.71 ± 4.46	84.67 ± 4.14
SGC (Wu et al., 2019)	86.25 ± 3.03	85.64 ± 4.24	87.65 ± 4.67	86.28 ± 4.49
DIAL (Chen et al., 2020)	87.37 ± 3.61	85.40 ± 3.68	86.43 ± 2.86	86.33 ± 2.55
BMGF (Gan et al., 2021)	88.16 ± 4.37	**87.65** ± 4.75	89.5 ± 3.60	88.59 ± 3.28
Proposed	**89.24** ± 3.44	87.55 ± 4.11	**90.31** ± 2.83	**88.89** ± 2.06

- SCP (Eavani et al., 2015) include positive as well as and negative correlations with low rank decomposition for the analysis of functional connectivity.
- HFC (Zhang et al., 2017) takes into account low-order, traditional high-order, and new correlated high-order FCNs of the subjects, and uses SVM classifiers based on multiple linear kernels to diagnosis.
- GCN (Kipf & Welling, 2017) aggregated the features of neighborhoods as well as updates the node features at each layer to obtain a new representation. This is the baseline method for deep graph neural networks.
- SGC (Wu et al., 2019) simplified the online neighborhood aggregation step to a pre-processing step. The network employed a neighborhood aggregation pre-processing step and multi-class logistic regression to simple vanilla GCN.
- CNHD (Wang et al., 2019) incorporate feature extraction and network-based classification of brain networks into a unified model based on the rs-fMRI data, while discriminative centers can be automatically identified from the data by $\ell_{2,1}$-norm regularizes.
- DIAL (Chen et al., 2020) aimed to iteratively update underlying graph at each graph learning layer and proposed a deep iterative adaptive graph learning methods which is able to iteratively update the new graph.
- BMGF (Gan et al., 2021) proposed a multi-graph fusion method based on shallow learning. It simultaneously fuses multiple FCNs and learning representation, and it automatically learns associations between brain regions and obtains significant results.

4.1.3. Implementation details

All experiments were implemented in PyTorch and conducted on a server with 8 NVIDIA GeForce 3090 (24 GB memory for each). In all experiments, we repeated the experiments 5 times for each method with random seeds to report their average performances and corresponding standard deviation (std). We obtained the author-verified codes for all comparative methods. For proposed method, we set the maximal number of epochs as 500 for the training process with the Adam optimizer (Kingma & Ba, 2015) where the initial Learning-rate (*i.e.,* Lr) and the Weight-decay (*i.e.,* Wd) are set to be 0.01 and 0.0005, respectively. Additionally, all parameters were initialized by the Glorot initialization, the activation function is ReLU (Nair & Hinton, 2010) and the last layer is Softmax which is used for classification. (Liu, Wen, Yu, & Yang, 2016). The hyper-parameter α in Eq. (8) is set to 0.1 and β in Eq. (9) is set to 1, for balancing the magnitudes of two terms, such as $\mathcal{L}_C$ and $\mathcal{L}_{FS}$. In Eq. (7), the number of self-attention heads C is set to 5 to balance performance and computational complexity. Moreover, we use an end-to-end training manner, which helps make our method easier to apply to real-world applications. The performance is evaluated by four widely used metrics (*i.e.,* ACC, SEN, SPE and AUC), for all of these metrics, a higher value denotes better performance.

4.2. Results and analysis

We evaluated the effectiveness of our proposed method on personalized diagnosis under supervised learning setting (*i.e.,* 80% labeled training subjects). Table 1, 2 and 3 illustrate the personalized diagnosis performance of our method and comparison methods on three real neurological disease datasets, respectively. From the tables, we have the conclusion as follows. First, the proposed framework outperforms or close to state-of-the-art methods (*i.e.,* SVM, HFC, SCP, CNHD, BMGF, GCN, SGC and DIAL) in terms of ACC (%), SEN (%), SPE (%) and AUC (%). For example, our method improves by 1.90%, 1.95%, 2.22% and 2.70% in terms of

Table 3
Comparison of performance (*i.e.,* ACC (%), SEN (%), SPE (%) and AUC (%)) of all methods on the dataset ADNI.

Methods	ACC	SEN	SPE	AUC
SVM (Fan et al., 2008)	76.90 ± 4.34	77.9 ± 3.62	76.21 ± 2.59	76.10 ± 2.47
SCP (Eavani et al., 2015)	84.97 ± 4.01	85.2 ± 3.42	84.89 ± 2.36	84.98 ± 3.43
HFC (Zhang et al., 2017)	80.31 ± 1.78	78.9 ± 2.17	81.44 ± 2.47	81.32 ± 3.95
GCN (Kipf & Welling, 2017)	87.48 ± 3.43	88.16 ± 3.73	88.27 ± 3.25	88.39 ± 4.03
CNHD (Wang et al., 2019)	86.03 ± 4.47	87.3 ± 5.72	84.60 ± 4.09	84.72 ± 4.58
SGC (Wu et al., 2019)	87.07 ± 2.83	88.01 ± 2.39	87.66 ± 3.72	87.57 ± 4.64
DIAL (Chen et al., 2020)	88.03 ± 4.88	88.20 ± 2.81	88.54 ± 1.3	88.72 ± 2.81
BMGF (Gan et al., 2021)	88.84 ± 3.22	89.55 ± 1.85	88.25 ± 2.49	90.22 ± 3.45
Proposed	**89.78** ± 3.85	**90.52** ±2.77	**89.90** ±3.45	**90.39** ±3.12

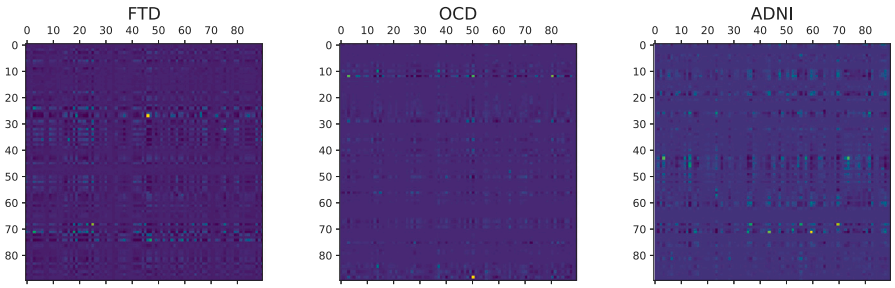
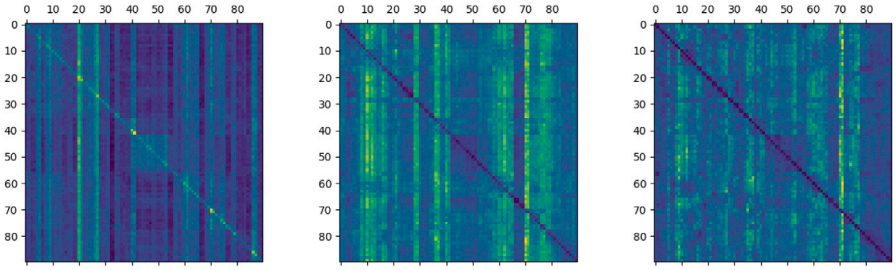


Fig. 2. Visualization of matrix **W** in the feature selection stage on all datasets.



(a) FTD (b) OCD (c) ADNI

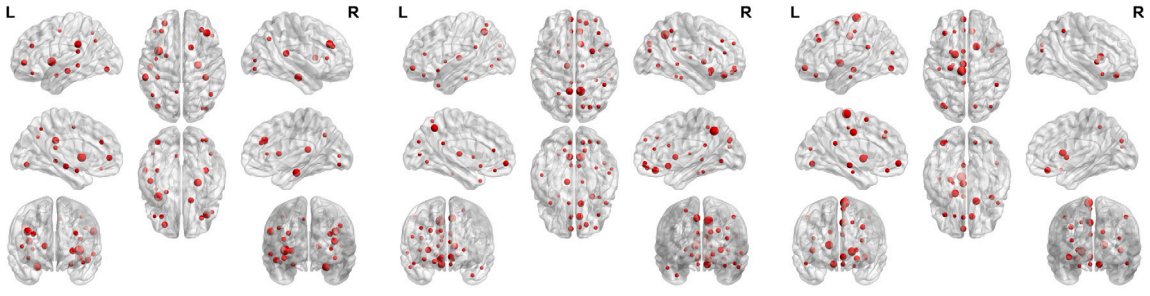
Fig. 3. Visualization of self-attention on all datasets.

ACC, SEN, SPE and AUC, respectively, compared to the best deep graph learning comparison method DIAL on average. Moreover, the proposed method improves by 22.48% and 0.79% respectively, compared to the lowest effective comparison shallow learning method (*i.e.,* SVM) and the best comparison shallow learning method BMGF on average. This contributes to that our proposed method (i) adapts a deep graph learning framework to extract low-level and high-level features of brain regions; (ii) designs a self-attention structure learning funtion to automatically capture the latent structural information; (iii) jointly selects the critical brain regions and trains the GNN model.

Second, it can be seen that the performance obtained with deep graph neural network (*i.e.,* GCN and SGC) generally outperforms that of the traditional shallow learning algorithm (*i.e.,* SVM and HFC), which shows that the deep graph learning methods can utilize the structural information of the graph more effectively. However, small sample datasets still limit the effectiveness of deep learning methods (*i.e.,* outperform shallow learning methods by a large margin), which is due to the difficulty of collecting real datasets for neurological diseases. For small sample datasets (such as the real brain disease datasets we used), deep learning approaches struggle to be more powerful, while traditional methods tend to be more stable and interpretable.

4.3. Structural learning interpretability analysis

We visualize the results of the proposed self-attention structural learning. We implemented it by visualizing the $\mathbf{S}^{n,m}$ with random selection on all datasets. From Fig. 3, we can see that the attention matrix can be interpreted as a directed graph since the keys and queries contain different learnable parameters. For the self-attention to be symmetric, we would need to use the same projection matrix for the queries and the keys. This would make the graph shown indirect. The visualization is consistent with our assumptions



(a) FTD (b) OCD (c) ADNI

Fig. 4. Visualization of selected brain regions on all datasets.

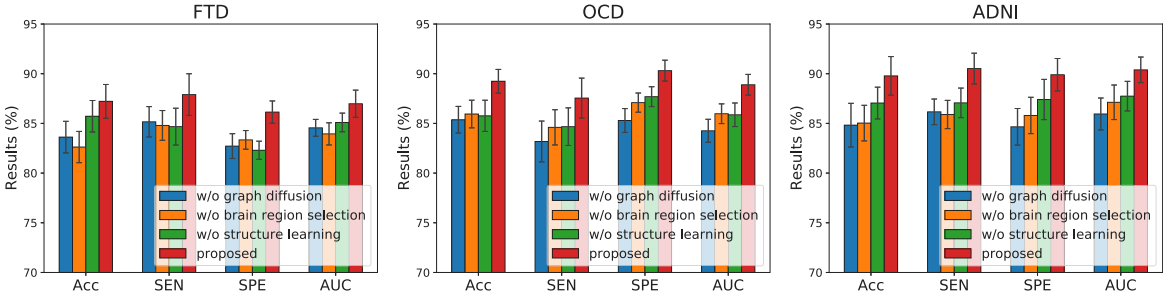


Fig. 5. Ablation experiments of our method on all datasets.

and analysis. Specifically, for the ADNI dataset, the large weights in the $\mathbf{S}^{n,m}$ are related to brain diseases, *i.e.,* amygdala, precentral gyrus and parahippocampal gyrus.

4.4. Feature selection interpretability analysis

Fig. 2 illustrates some of the visual results of proposed feature selection. We can see that the rows of the matrix are sparse, which is due to the $\ell_{2,1}$-norm. In addition, rows with different colors represent the corresponding brain regions that are important for its corresponding disease diagnosis. We have also marked the important brain regions in Fig. 4. From Fig. 4, we can see that our results have clinical implications for the brain regions corresponding to the onset of the disease. For example, the important brain regions selected for fronto-temporal dementia mainly focused on the frontal and temporal areas. Furthermore, the results were confirmed by two radiologists with 10 years of experience and found to be approximately accurate.

To this end, experimental results indicate that our method outperforms alternative shallow learning methods and GCN-based methods with significant margins on three neuroimaging datasets and can provide interpretability result for clinical research. Moreover, the interpretability analysis of proposed method opens up the possibility of bringing GCN-based disease prediction methods from the experiments to the clinic by allowing researchers to gain insight from structure learning and feature selection, as well as develop more accurate and robust machine learning models.

4.5. Ablation study

4.5.1. The effectiveness of proposed module

A variety of ablation studies were conducted to investigate the impacts of each proposed module on each dataset and is reported in Fig. 5. The comparative results and detailed analysis of the important modules proposed in our framework are as follows:

(1) *Without graph diffusion*: we remove the GCN encoding from our framework, and it can be seen that the accuracy of brain diseases is particularly poor. The reason can be that the graph diffusion module makes full use of structure information among brain regions, and it generates the high-level features.

(2) *Without brain region selection*: we then remove feature selection module from our framework in order to demonstrate the validity of brain region selection and prove its effectiveness. It is easy to see that the feature selection module can greatly improve the performance of personalized diagnosis.

(3) *Without structure learning*: to further analyze the effectiveness of learned graph structure information on brain disease diagnosis, we remove self-attention structure learning module from out framework. That is the original graph structure (*i.e.,* $\mathbf{A}^{n,m}$) is used for graph diffusion module. In Fig. 5, we can see that the performance is limited without self-attention structure learning.

In summary, the complete module (*i.e.,* graph diffusion, brain region selection and structure learning) achieves the best performance for personalized diagnosis, thus the effectiveness of each part of our proposed framework is verified.

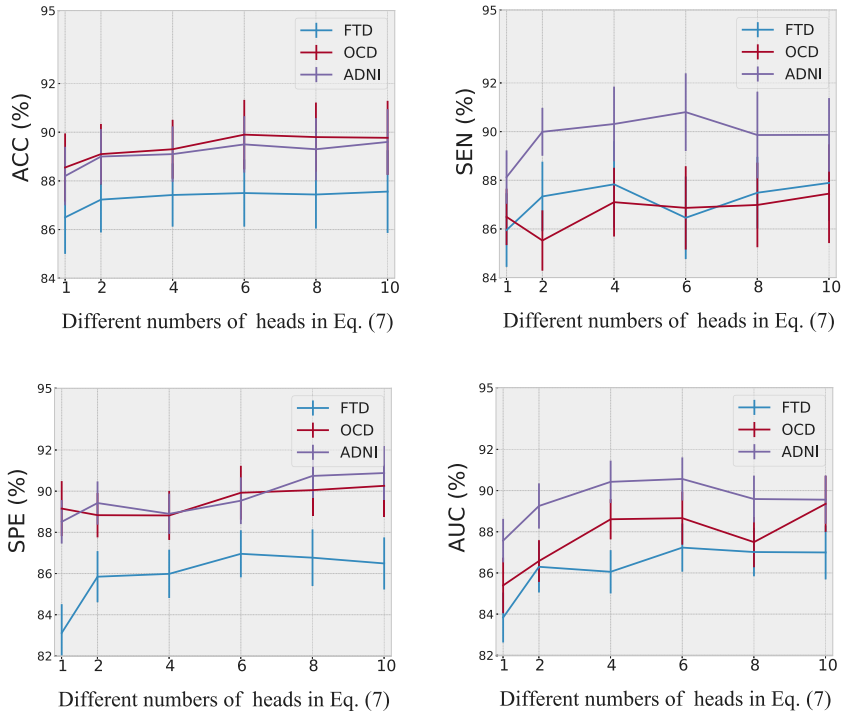


Fig. 6. Experimental results of different number of heads in the multiple self-attention on all datasets.

4.5.2. The number of the head in the multiple self-attention

In this subsection, we analyze the impacts of different numbers of the head in the multiple self-attention on our methods. For this purpose, we set the range of the number of self-attention head C as $\{1, 2, 4, 6, 8, 10\}$. As shown in Fig. 6, the classification performance of our method increased with the increasing values of C, *i.e.,* from $C = 1$ to $C = 6$. The performance peaks around $C = 6$ and then increases slowly. The reason is that small C value cannot fully exert the ability of multi-head self-attention.

5. Discussion

In this section, we discuss interpretability analysis of our results and practical implications of our research. Interpretation is an open problem in deep learning, but interpretations of medical images are critical for clinical research. In medical image analysis, shallow learning methods have been interpreted in a variety of ways, *e.g.,* (Zhang et al., 2017, 2019; Zhu et al., 2021). Deep learning methods are more effective than shallow learning methods in medical image analysis, *e.g.,* (Gan et al., 2021; Jiang et al., 2020; Parisot et al., 2018), but previous methods have concentrated on boosting accuracy while disregarding interpretability for brain diseases diagnosis. To tackle the realistic but ignored issue of interpretable learning for medical image analysis, we perform interpretability analysis from two aspects, *i.e.,* structural learning interpretability analysis (Section 4.3) and feature selection interpretability analysis (Section 4.4). Consequently, our method can helps clinicians to obtain interpretable results and can be used on other disease data as well.

In addition, in terms of results, some methods reported different accuracy on the ADNI dataset. For example, Parisot et al. (2018) obtain a lower performance (*i.e.,* 78% ACC with 86% AUC), and Liu, Cheng, Wang, and Wang (2018) report a slightly higher performance (*i.e.,* 93% ACC with 95% AUC), compare to ours (*i.e.,* 89% ACC with 90% AUC). However, the results of these methods are not comparable, because the data are pre-processed in different ways in the field of medical image. To summarize, our results demonstrate the superiority of proposed method, and the interpretability of our method takes GCN-based disease prediction methods from research to clinic by allowing clinicians gain insight from the model.

6. Conclusion

This study proposed a new deep learning based method by jointly conducting feature selection, self-attention graph structure learning and graph diffusion for personalize neurological diseases diagnosis. Particularly, feature selection module is used to select important brain regions for the specific brain disease diagnosis. Self-attention graph structure learning module is used to learn and capture the relationship between brain regions automatically. With learned structure information among brain regions, we use GCN to obtain high-level features in graph diffusion process. Finally, we combine the features of each view to output a more

accurately prediction. Extensive experiments on three public brain disease datasets confirmed that our method obtains state-of-the-art performance. Moreover, the interpretability of our model can provide more insight for clinical deployment. Our future work will address the noise label issue and few-shot learning task to further improve performance and robust.

CRediT authorship contribution statement

Lujing Wang: Conceptualization, Methodology, Formal analysis, Data curation, Supervision, Project administration, Funding acquisition, Writing – original draft & review. **Weifeng Yuan:** Data curation, Methodology, Resources, Supervision, Investigation, Writing – original draft & review. **Lu Zeng:** Software, Formal analysis, Writing – review. **Jie Xu:** Software, Validation, Investigation. **Yujie Mo:** Validation, Resources. **Xinxiang Zhao:** Resources, Data curation, Formal analysis, Supervision. **Liang Peng:** Formal analysis, Writing – review, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Innovation Fund of Kunming Medical University, China (2021S062).

References

Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, *56*(4), 1245–1259.
Amaran, S., Sahinidis, N. V., Sharda, B., & Bury, S. J. (2016). Simulation optimization: a review of algorithms and applications. *Annals of Operations Research*, *240*(1), 351–380.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2014). Spectral networks and locally connected networks on graphs. In *ICLR*.
Chen, Y., Wu, L., & Zaki, M. J. (2020). Deep iterative and adaptive learning for graph neural networks. In *AAAI*.
De Boer, P. T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, *134*(1), 19–67.
Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS* (pp. 3844–3852).
Eavani, H., Satterthwaite, T. D., Filipovych, R., Gur, R. E., Gur, R. C., & Davatzikos, C. (2015). Identifying sparse connectivity patterns in the brain using resting-state fMRI. *Neuroimage*, *105*, 286–299.
Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874.
Fan, T., & Wang, H. (2022). Research of Chinese intangible cultural heritage knowledge graph construction and attribute value extraction with graph attention network. *Information Processing & Management*, *59*(1), Article 102753.
Farooq, A., Anwar, S., Awais, M., & Rehman, S. (2017). A deep CNN based multi-class classification of alzheimers disease using MRI. In *IST* (pp. 1–6).
Gan, J., Peng, Z., Zhu, X., Hu, R., Ma, J., & Wu, G. (2021). Brain functional connectivity analysis based on multi-graph fusion. *Medical Image Analysis*, *71*, Article 102057.
Hu, R., Peng, Z., Zhu, X., Gan, J., Zhu, Y., Ma, J., et al. (2021). Multi-band brain network analysis for functional neuroimaging biomarker identification. *IEEE Transactions on Medical Imaging*, *40*, 3843–3855.
Isufi, E., Pocchiari, M., & Hanjalic, A. (2021). Accuracy-diversity trade-off in recommender systems via graph convolutions. *Information Processing & Management*, *58*(2), Article 102459.
Jiang, H., Cao, P., Xu, M., Yang, J., & Zaiane, O. (2020). Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction. *Computers in Biology and Medicine*, *127*, Article 104096.
Jiang, B., Zhang, Z., Lin, D., Tang, J., & Luo, B. (2019). Semi-supervised learning with graph learning-convolutional networks. In *CVPR* (pp. 11313–11320).
Jin, K. H., McCann, M. T., Froustey, E., & Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, *26*(9), 4509–4522.
Kauderer-Abrams, E. (2020). Quantifying translation-invariance in convolutional neural networks. In *CVPR*.
Khachaturian, Z. S. (1985). Diagnosis of alzheimers disease. *Archives of Neurology*, *42*(11), 1097–1105.
Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.
Kipf, N. T., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR*.
Li, L., Zhao, K., Gan, J., Cai, S., Liu, T., Mu, H., et al. (2021). Robust adaptive semi-supervised classification method based on dynamic graph and self-paced learning. *Information Processing & Management*, *58*(1), Article 102433.
Liao, G., Deng, X., Wan, C., & Liu, X. (2022). Group event recommendation based on graph multi-head attention network combining explicit and implicit information. *Information Processing & Management*, *59*(2), Article 102797.
Liu, M., Cheng, D., Wang, K., & Wang, Y. (2018). Multi-modality cascaded convolutional neural networks for alzheimers disease diagnosis. *Neuroinformatics*, *16*(3), 295–308.
Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016). Large-margin softmax loss for convolutional neural networks.. In *ICML, vol. 2, no. 3* (p. 7).
Mishra, N. K., & Singh, P. K. (2020). FS-MLC: Feature selection for multi-label classification using clustering in feature space. *Information Processing & Management*, *57*(4), Article 102240.
Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML*.
Neary, D., Snowden, J., & Mann, D. (2005). Frontotemporal dementia. *The Lancet Neurology*, *4*(11), 771–780.
Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., et al. (2018). Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimers disease. *Medical Image Analysis*, *48*, 117–130.
Peng, L., Kong, F., Liu, C., & Kuang, P. (2021). Robust and dynamic graph convolutional network for multi-view data classification. *The Computer Journal*, *64*, 1093–1103.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Rolls, E. T., Huang, C. C., Lin, C. P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *Neuroimage*, *206*, Article 116189.
Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.
Stein, D. J. (2002). Obsessive-compulsive disorder. *The Lancet*, *360*(9330), 397–405.
Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. In *ICLR*.
Wang, M., Huang, J., Liu, M., & Zhang, D. (2019). Functional connectivity network analysis with discriminative hub detection for brain disease identification. In *AAAI, vol. 33, no. 01* (pp. 1198–1205).
Wen, Z., & Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, *142*(1), 397–434.
Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., & Weinberger, K. (2019). Simplifying graph convolutional networks. In *ICML* (pp. 6861–6871).
Xu, J., Ren, Y., Li, G., Pan, L., Zhu, C., & Xu, Z. (2021). Deep embedded multi-view clustering with collaborative training. *Information Sciences*, *573*, 279–290.
Yan, C., & Zang, Y. (2010). DPARSF: a MATLAB toolbox for" pipeline" data analysis of resting-state fMRI. *Frontiers in Systems Neuroscience*, *4*, 13.
Yuan, C., Zhong, Z., Lei, C., Zhu, X., & Hu, R. (2021). Adaptive reverse graph learning for robust subspace learning. *Information Processing & Management*, *58*(6), Article 102733.
Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2018). Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1774–1785.
Zhang, Y., Zhang, H., Chen, X., Lee, S. W., & Shen, D. (2017). Hybrid high-order functional connectivity networks using resting-state functional MRI for mild cognitive impairment diagnosis. *Scientific Reports*, *7*(1), 1–15.
Zhang, Y., Zhang, H., Chen, X., Liu, M., Zhu, X., Lee, S. W., et al. (2019). Strength and similarity guided group-level brain functional network construction for MCI diagnosis. *Pattern Recognition*, *88*, 421–430.
Zhu, Y., Ma, J., Yuan, C., & Zhu, X. (2022). Interpretable learning based dynamic graph convolutional networks for alzheimers disease analysis. *Information Fusion*, *77*, 53–61.
Zhu, X., Song, B., Shi, F., Chen, Y., Hu, R., Gan, J., et al. (2021). Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. *Medical Image Analysis*, *67*, Article 101824.
Zhu, X., Zhang, S., Zhu, Y., Zhu, P., & Gao, Y. (2020). Unsupervised spectral feature selection with dynamic hyper-graph learning. *IEEE Transactions on Knowledge and Data Engineering*, *PP*, 1.