Learnable Brain Connectivity Structures for Identifying Neurological Disorders

Zhengwang Xia, Tao Zhou, *Senior Member, IEEE*, Zhuqing Jiao, and Jianfeng Lu, *Member, IEEE*

***Abstract*— Brain networks/graphs have been widely recognized as powerful and efficient tools for identifying neurological disorders. In recent years, various graph neural network models have been developed to automatically extract features from brain networks. However, a key limitation of these models is that the inputs, namely brain networks/graphs, are constructed using predefined statistical metrics (e.g., Pearson correlation) and are not learnable. The lack of learnability restricts the flexibility of these approaches. While statistically-specific brain networks can be highly effective in recognizing certain diseases, their performance may not exhibit robustness when applied to other types of brain disorders. To address this issue, we propose a novel module called Brain Structure Inference (termed BSI), which can be seamlessly integrated with multiple downstream tasks within a unified framework, enabling end-to-end training. It is highly flexible to learn the most beneficial underlying graph structures directly for specific downstream tasks. The proposed method achieves classification accuracies of 74.83% and 79.18% on two publicly available datasets, respectively. This suggests an improvement of at least 3% over the best-performing existing methods for both tasks. In addition to its excellent performance, the proposed method is highly interpretable, and the results are generally consistent with previous findings.**

***Index Terms*— Deep learning, graph neural network, graph structure learning, brain disorder identification.**

I. Introduction

NEUROLOGICAL disorders encompass a range of diseases that involve dysfunction or damage to the nervous system, including Alzheimer's Disease (AD) [1], Autism Spectrum Disorder (ASD) [2], and Attention-Deficit/Hyperactivity Disorder (ADHD) [3]. The incidence rate of neurological disorders has been steadily rising, posing significant challenges to global public health [4]. These types of diseases present considerable difficulty in terms of finding a cure, and long-term treatment places a substantial financial burden on affected families.

Brain networks have been shown to be effective in identifying a range of neurological disorders such as AD, ASD, and ADHD. Over the past few decades, researchers have developed a comprehensive set of metrics to characterize specific topological properties of brain networks [5], [6], [7]. These metrics are widely recognized as potential biomarkers that can effectively differentiate between healthy individuals and those with brain disorders. For instance, Wadhera and Mahmud [5] introduced a novel metric called weighted hierarchical complexity to characterize the hierarchical organization of brain networks, achieving remarkable classification accuracies exceeding 95% in identifying individuals with ASD. In addition, Cai et al. [6] utilized the Fiedler value of the brain connectome to analyze the dynamic characteristics of the human brain over time. Their findings revealed a significant correlation between this metric and the severity of Parkinson's Disease (PD) in patients. Moreover, Avvaru and Parhi [7] proposed a novel causality measure called frequency-domain convergent cross-mapping (FDCCM) to analyze the underlying dynamics of the brain. Experimental results on multiple datasets demonstrated that FDCCM can successfully distinguish patients with PD from the control group. These studies highlight the potential of network metrics in characterizing brain disorders.

While the aforementioned methods have shown good performance in disease-specific recognition tasks, extracting these features typically requires the designer to have a deep understanding of the domain. Moreover, these predefined features often lack robustness and may exhibit limited generalization capabilities when applied to different scenarios [8]. Motivated by the impressive representational capabilities of graph neural networks (GNNs) in handling graph-structured data, many studies have adopted GNNs for automatic feature extraction from brain networks [9], [10], [11], thus replacing traditional feature design approaches. For example, Cui et al. [9] designed a novel graph convolution network to automatically extract spatio-temporal features from brain image data, the results derived from multiple independent datasets present compelling evidence regarding the robustness and effectiveness of this

Manuscript received 27 December 2023; revised 23 July 2024; accepted 16 August 2024. Date of publication 20 August 2024; date of current version 28 August 2024. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20221487, in part by the National Natural Science Foundation of China under Grant 62172228, in part by Jiangsu Provincial Key Research and Development Program under Grant BE2021636, and in part by Qing Lan Project of Jiangsu Province. *(Corresponding author: Jianfeng Lu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Board of ADNI and ABIDE.

Zhengwang Xia, Tao Zhou, and Jianfeng Lu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: xzhengwang@njust.edu.cn; taozhou.dreams@gmail.com; lujf@njust.edu.cn).

Zhuqing Jiao is with the School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213000, China (e-mail: jzq@cczu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3446588

approach. Jiang et al. [10] developed a hierarchical graph network model that adeptly captures network topology information while simultaneously learning graph representations. The effectiveness of the proposed method is demonstrated by experimental results on two publicly available datasets. Zhang et al. [11] proposed a novel graph network model called the Local to Global Graph Neural Network (LG-GNN) for progressively integrating local and global features. This model achieved significant recognition results on two publicly available datasets.

The aforementioned research results fully demonstrate the benefits of using GNNs for feature extraction. However, the inputs to these GNNs (i.e., brain networks) are constructed using predefined metrics (e.g., Pearson correlation), which presents the same problem as before. It is widely recognized that the underlying pathogenic mechanisms, particularly the abnormal patterns among brain regions, exhibit significant variations across different neurological diseases [5], [6], [7]. While brain networks constructed using predefined computational metrics, such as Pearson correlation coefficients, can aid in identifying various brain diseases, they often fail to achieve optimal results. The primary issue lies in the fact that these brain network modeling approaches lack a direct feedback relationship with the subsequent step, leading to an independent execution of brain network modeling and feature extraction. The quality of subsequent recognition results is greatly influenced by the accuracy of brain network modeling. Consequently, such methods frequently yield suboptimal classification results.

To address these limitations, we propose a novel module termed Brain Structure Inference (BSI) to infer the causal relationship between different brain regions. This module takes time-series data extracted from imaging data as input and generates a brain network that represents causal effects between different brain regions as output. The brain network, represented as a graph, is a trainable parameter of the BSI module that can be seamlessly integrated with downstream tasks in an end-to-end manner. More details about the BSI module can be found in subsection III-B. The use of causality-based approaches to model relationships between brain regions is motivated by multiple studies that have demonstrated their superior robustness over correlation-based methods [12], [13], [14]. Moreover, constructing brain networks based on causality (asymmetric graphs) can provide more comprehensive information compared to correlation-based networks (symmetric graphs). In addition to inferring causality, we introduce two additional constraints for the learnable graph structure: promoting sparsity and low rank. It is worth noting that, the sparsity constraint is to prevent overfitting of the model [15], [16], and the low-rank constraint is used to encourage brain regions to form dense clusters with their neighboring nodes.

The main contributions of this paper can be summarized as follows:

- **Flexibility and Adaptability** (cf. **Section** III-B): We present a novel Brain Structure Inference Graph Neural Network (BSIGNN) that integrates graph structure learning and downstream tasks into a unified framework. Traditionally, brain network construction (i.e., graph structure learning) and neurological disease identification are conducted independently. In contrast, BSIGNN facilitates dynamic interaction between the two steps. By introducing a novel graph learning module, BSI, BSIGNN eliminates the need to design brain network modeling methodologies for each neurological disease identification task individually. The improvement in the diagnostic framework offers superior flexibility and adaptability to BSIGNN.
- **Nonlinear Fitting Capability** (cf. **Equation** 2): The BSI module is proposed to model the nonlinear interactions among brain regions, significantly surpassing conventional approaches that are limited to capturing linear interactions between brain regions.
- **Excellent Performance** (cf. **Table II**): The proposed BSIGNN achieves state-of-the-art performance on two publicly available datasets, highlighting the effectiveness of our approach.

II. Related Work

In this section, we will briefly review traditional and GNN-based methods for neurological disease recognition, with a particular emphasis on GNN-based methods.

A. Traditional Methods

Traditional methods typically involve a two-step process to identify neurological disorders. First, a brain network is constructed by estimating statistical dependencies (e.g., Pearson's correlation coefficient) between physiological signals in each brain region. Second, features are extracted from the constructed brain network to train classifiers for the recognition of neurological diseases. For example, Wee et al. [17] introduced the sliding window approach to construct dynamic brain functional networks and subsequently extracted network attributes (e.g., clustering coefficients) from these networks as features for mild cognitive impairment (MCI) classification. Zhang et al. [18] introduced a new method for constructing effective brain networks. Then, numerous node features were extracted as candidates to train a classifier for accurately identifying individuals with schizophrenia (SZ). However, these methods divide brain network modeling and feature selection into two separate steps. If either of the two steps is not handled well, it is likely to result in a suboptimal classification outcome.

B. GNN-Based Methods

Recently, researchers have been increasingly inclined to utilize GNN-based methods for brain imaging analysis due to their remarkable representation capabilities on graph-structured data. Currently, GNN-based diagnostic methods for neurological diseases can be broadly classified into two categories based on how the subjects are represented: subject graph and brain graph.

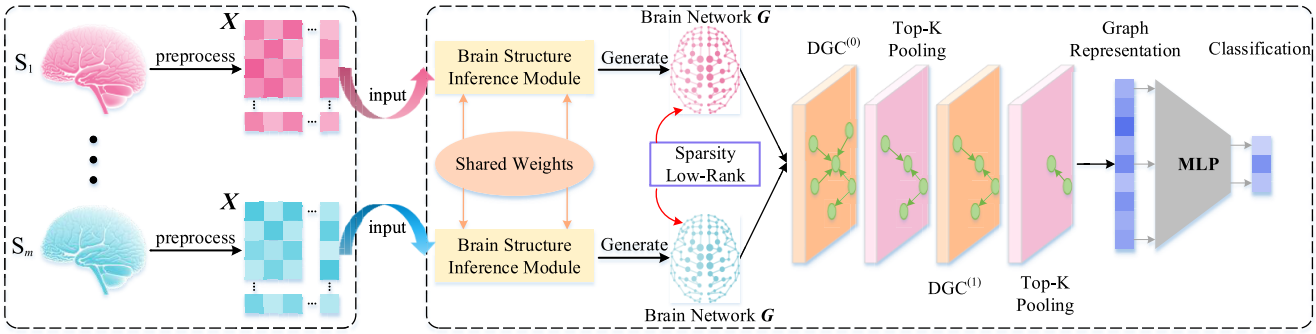


Fig. 1. The overall framework of the proposed BSIGNN for identifying neurological disorders. For each subject S_i, a matrix $\boldsymbol{X} \in \mathbb{R}^{T\times N}$ will be obtained after preprocessing. T represents the length of the time series, and N corresponds to the number of brain regions defined by the brain atlas. Then, the matrix $\boldsymbol{X}$ will be used as input for our newly designed module, which infers causal relationships between brain regions to generate brain networks $\boldsymbol{G} \in \mathbb{R}^{N\times N}$. Finally, we aggregate node information based on the inferred graph to obtain graph-level representations for subsequent classification.

1) Subject Graph: In this approach, each subject is treated as a node within the graph, and the relationships between subjects are evaluated using predefined similarity metrics to construct a subject relationship graph. Subsequently, a graph convolutional network (GCN) is employed to aggregate subject features and generate predictions. For example, Parisot et al. [19] comprehensively incorporated diverse types of information, including imaging and non-imaging information, to construct the subject graph. Subsequently, they utilized a GCN to classify subjects based on the constructed graph. Based on the aforementioned work, Song et al. [20] implemented a range of improvements that involved refining the construction methodology of the subject graph and optimizing the GCN algorithm. For this type of approach, the essence lies in constructing a robust subject graph. Unfortunately, the subject graph is generated using predefined metrics rather than a structure that can be learned. If the resulting subject graph proves to be unreliable, there is a high probability that subsequent downstream tasks will yield suboptimal outcomes.

2) Brain Graph: In this approach, each subject is represented by a graph, such as a brain functional network defined by the Pearson correlation coefficient. Following that, a GNN is adopted to classify the corresponding brain network of each subject in order to distinguish whether they have a disease or not. For example, Yang et al. [16] proposed an innovative method for constructing brain networks. Afterward, these networks were used as input for a Graph Attention Network (GAT) to extract discriminative features for recognizing ASD. Li et al. [21] designed a novel GNN for extracting more comprehensive information from constructed brain functional networks. The proposed method showed excellent performance in both brain disease diagnosis and brain function decoding tasks. However, similar to the first category of methods, these approaches also suffer from the same limitations. The graphs (brain networks) used for training GNNs are still generated using predefined metrics rather than being a learnable structure.

Generally, the prevailing diagnostic framework lacks the capability to dynamically adjust brain network modeling strategies tailored to each specific neurological disease. This limitation hinders diagnostic accuracy and necessitates the urgent development of innovative methods to enhance the diagnostic framework.

III. Proposed Method

The proposed framework BSIGNN for identifying neurological disorders using resting-state functional magnetic resonance imaging (rs-fMRI) is illustrated in Fig. 1. Different from previous methods, our approach (illustrated in the right box) directly utilizes the time series data $\boldsymbol{X}$ extracted from rs-fMRI images as input, without relying on predefined graphs. To capture causal effects among different brain regions, we propose a BSI module that incorporates these causal relationships as learnable parameters. This module can be seamlessly integrated with subsequent tasks, such as classification, within a unified framework, facilitating end-to-end optimization. Additionally, we impose two additional constraints on the latent graph structure, and leverage directed graph convolution (DGC) to automatically extract features that are effective in identifying neurological disorders. A detailed description of the network model, BSIGNN, will be provided in the subsequent subsections.

A. Preliminaries

The methods for constructing brain networks can be divided into two groups: correlation-based and causality-based methods. Correlation-based methods are limited to capturing temporal correlations between blood-oxygen signals from different brain regions. In contrast, causality-based methods offer two distinct advantages over correlation-based methods: (1) numerous studies have demonstrated that causality-based methods are more robust and reliable [12], [22]; (2) brain networks estimated by causality-based methods provide valuable insights into the direction of information flow, making them more advantageous in recognizing brain disorders [14], [23].

Granger causality (GC) [24] analysis is widely employed to estimate causal effects between brain regions, and it distinguishes between cause and effect based on the underlying principle that cause always precedes effect. For the two time

series $\boldsymbol{X}_1 \in \mathbb{R}^T$ and $\boldsymbol{X}_2 \in \mathbb{R}^T$ with T time steps, if the ability of $\boldsymbol{X}_2$ to predict its own future is enhanced by incorporating the past information of variable $\boldsymbol{X}_1$, then $\boldsymbol{X}_1$ is considered to be the cause of $\boldsymbol{X}_2$. Thus, the general form of causal effect between two variables can be defined by

$$\begin{aligned} \boldsymbol{X}_1^t &= \sum_{i=1}^{L} \alpha_i \boldsymbol{X}_2^{t-i} + \sum_{i=1}^{L} \beta_i \boldsymbol{X}_1^{t-i} + \mu_{1,t}, \\ \boldsymbol{X}_2^t &= \sum_{i=1}^{L} \lambda_i \boldsymbol{X}_1^{t-i} + \sum_{i=1}^{L} \theta_i \boldsymbol{X}_2^{t-i} + \mu_{2,t}, \end{aligned} \tag{1}$$

where $\boldsymbol{X}_1^t$ represents the t-th element of the vector $\boldsymbol{X}_1$, and L denotes the time lag length. α_i, β_i, λ_i and θ_i are regression coefficients, $\mu_{1,t}$ and $\mu_{2,t}$ represent noise terms. In fact, there are four possible relationships between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$: (1) If $\boldsymbol{X}_1$ is the cause of $\boldsymbol{X}_2$, then α should be zero overall and λ should be non-zero overall; (2) If $\boldsymbol{X}_2$ is the cause of $\boldsymbol{X}_1$, then α should be non-zero overall and λ should be zero overall; (3) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent of each other, then both α and λ should be zero overall; (4) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have an effect on each other, then both α and λ should be non-zero overall. In summary, causal relationships between variables can be inferred by examining the correlation coefficients between them.

B. Brain Structure Inference Module

To extend Eq. (1) to a multivariate scenario, the past information of all other variables is integrated into the predictive model for the i-th variable. Variables that play a more substantial role in enhancing the predictive accuracy of the i-th variable are considered as causes of the i-th variable. The general expression for the multivariate case can be defined as follows:

$$\boldsymbol{X}_i^t = f_i\left(\boldsymbol{X}_1^{<t}, \ldots \boldsymbol{X}_N^{<t}\right) + \mu_{i,t}, \tag{2}$$

where $\boldsymbol{X}_i^{<t} = [\boldsymbol{X}_i^{t-L}, \boldsymbol{X}_i^{t-L+1}, \ldots, \boldsymbol{X}_i^{t-1}]$ represents the sequence of L preceding values of variable $\boldsymbol{X}_i$ up to moment t, N equals the number of variables, and $\mu_{i,t}$ is the noise term. Notably, f_i can be either a linear or a nonlinear activation function. However, linear models have limited fitting ability. Therefore, in this paper, we choose to set f_i as a nonlinear activation function.

In this study, we propose a BSI module to model the nonlinear function f_i in Eq. (2). The detailed structure of BSI is shown in Fig. 2. For N variables, the previous L values of each variable before time t are separately fed into a sub-module f_i. The sub-module f_i is defined as follows:

$$[\boldsymbol{C}_{i\to 1}^t, \boldsymbol{C}_{i\to 2}^t, \ldots, \boldsymbol{C}_{i\to N}^t] = FC(LSTM(\boldsymbol{X}_i^{<t})), \tag{3}$$

where $LSTM$ refers to the Long Short-Term Memory (LSTM) network and FC represents the fully connected layer. The element $\boldsymbol{C}_{i\to j}^t$ represents the contribution of past information from the i-th variable to accurately predict the next time step of the j-th variable. The function of LSTM is to capture the temporal dependence between the value at time t and the previous L steps. The role of the fully connected layer is to reduce the LSTM output to N dimensions, matching the number of variables.

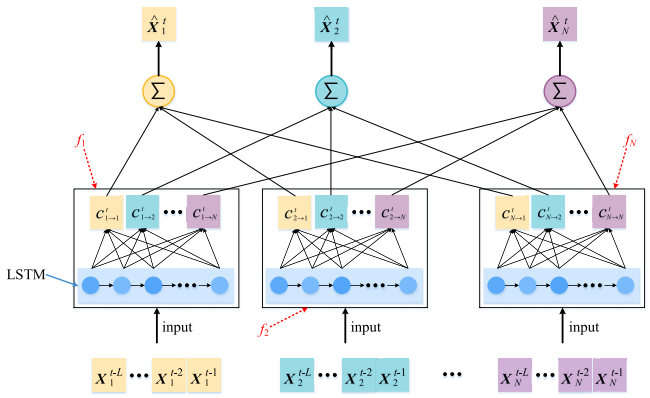


Fig. 2. The architecture of BSI module.

After obtaining the outputs from these N sub-modules, we aggregate the predictions of each variable by performing a summation operation, which can be defined as follows:

$$\hat{\boldsymbol{X}}_i^t = \sum_{v=1}^{N} \boldsymbol{C}_{v\to i}^t + \delta_i, \tag{4}$$

where δ_i denotes the error term, $\hat{\boldsymbol{X}}_i^t$ represents the value of the estimated i-th variable at time t. The purpose of aggregating the results is to consider both the influence of the previous information of the corresponding variable on predicting its future value and the influence of other variables on it. This is in accordance with the guidelines for determining causality in Granger causality analysis.

The remaining question is how to obtain the causal graph $\boldsymbol{G} \in \mathbb{R}^{N\times N}$, i.e., the brain network, which involves determining the causal relationships between the variables. According to the definition of GC, the first variable can be regarded as the cause of the second variable if the past information of the first variable is helpful in improving the prediction of the subsequent moments of the second variable. Thus, for any two variables $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, $\boldsymbol{X}_i$ is considered to be useful in predicting the future of $\boldsymbol{X}_j$ if $\boldsymbol{X}_j^t - (\boldsymbol{C}_{i\to j}^t + \boldsymbol{C}_{j\to j}^t) < \boldsymbol{X}_j^t - \boldsymbol{C}_{j\to j}^t$. In fact, the inequality holds as long as $\boldsymbol{C}_{i\to j}^t$ and $\boldsymbol{X}_j^t$ have the same sign, for example, both positive or both negative. Therefore, the causal effect of the i-th brain region on the j-th brain region, denoted as $\boldsymbol{G}_{i,j}$, can be defined as follows:

$$\boldsymbol{G}_{i,j} = \frac{1}{T-L} \sum_{t=L+1}^{T} \frac{\boldsymbol{C}_{i\to j}^t}{\boldsymbol{X}_j^t}, \tag{5}$$

where T represents the sampling frequency of each subject, which is equivalent to the length of the time series. The objective of $\frac{\boldsymbol{C}_i^t}{\boldsymbol{X}_j^t}$ is to quantitatively evaluate the influence of variable i on variable j.

After preprocessing, each subject yields a matrix $\boldsymbol{X} \in \mathbb{R}^{T\times N}$ that captures the dynamics of blood-oxygen signals across N brain regions over time. For BSI, the input consists of N time series, each with a length of L. The objective is to predict the next value of these N variables. Consequently, the contributions among all variables are evaluated a total of $T-L$ times, denoted as $\boldsymbol{C}^t$, where t ranges from $L+1$ to T.

Furthermore, as this is a prediction problem, the module is trained using the Mean Squared Error (MSE) loss. The formulation for MSE loss is as follows:

$$\mathcal{L}_G = MSE(\boldsymbol{X}_{:}^{t}, \hat{\boldsymbol{X}}_{:}^{t}), \quad \text{with} \quad t = [L+1, L+2, \ldots, T], \tag{6}$$

where $\boldsymbol{X}_{:}^{t}$ represents the true values of all variables at time t, and $\hat{\boldsymbol{X}}_{:}^{t}$ refers to the values of all variables at time t predicted by the module.

In clinical practice, standardized and universally accepted assessment criteria are often essential for disease diagnosis. Similarly, deep learning-based methods aim to learn a uniform criterion from training data that aligns with downstream tasks, such as identifying patients with neurological disorders. The BSI module designed in this paper employs a weight-sharing strategy to standardize the evaluation process, facilitating the extraction of general features from data [25]. This strategy enhances performance in downstream tasks by ensuring that the parameters of the BSI module are shared across all subjects, rather than tailored individually. This consistency in parameter sharing maintains a standardized diagnostic process for each disease, enabling the model to learn generalized features across individuals and mitigating overfitting issues associated with individual differences.

C. Constraint Rules

In addition to the MSE loss, we incorporate two additional constraints, namely the sparsity constraint and the low-rank constraint, to impose restrictions on the underlying graph structure. The inclusion of the sparsity constraint is motivated by two primary factors. Firstly, a substantial body of research has demonstrated that sparse brain networks exhibit greater robustness in recognizing neurological disorders [16], [26], [27]. Secondly, several observational studies have indicated that only a limited number of neurons are activated in the brain at any given moment [28], [29]. The sparsity constraint is defined as:

$$\mathcal{L}_{sparse} = ||\boldsymbol{G}||_2, \tag{7}$$

where the notation $||\cdot||$ represents the L_2 norm.

The low-rank constraint is introduced based on the fact that numerous studies have demonstrated the presence of small-world properties in brain networks [30], [31]. This suggests that nodes in a network tend to be directly connected, leading to the formation of multiple aggregated structures with strong internal connections. The low-rank constraint is defined as:

$$\mathcal{L}_{rank} = Rank(\boldsymbol{G}), \tag{8}$$

where the function $Rank(\cdot)$ returns the rank of the matrix $\boldsymbol{G} \in \mathbb{R}^{N \times N}$.

However, determining the rank of a matrix is a non-convex task. Therefore, we relax the low-rank constraint by using the nuclear norm. The low-rank constraint can be redefined as follows:

$$\mathcal{L}_{rank} = ||\boldsymbol{G}||_*, \tag{9}$$

where $||\cdot||_*$ refers to the nuclear norm.

D. Directed Graph Convolution and Pooling Operations

Once the BSI module has generated the brain network $\boldsymbol{G} \in \mathbb{R}^{N \times N}$ for each subject using the time-series data $\boldsymbol{X} \in \mathbb{R}^{T \times N}$, the next step is to extract informative features that facilitate graph classification. It is worth noting that the brain network obtained is an asymmetric directed graph. Therefore, we employ the directed graph convolution (DGC) as described in [32] to aggregate the information between neighboring nodes. The layer-wise propagation rule is defined as follows:

$$\begin{aligned}
\mathbf{H}^{(l+1)} &= Con\left(\sigma(\boldsymbol{G}^{F}\boldsymbol{N}_e^{(l)}), \sigma(\boldsymbol{G}^{S_{in}}\boldsymbol{N}_e^{(l)}), \sigma(\boldsymbol{G}^{S_o}\boldsymbol{N}_e^{(l)})\right), \\
\boldsymbol{N}_e^{(l)} &= \mathbf{H}^{(l)}\boldsymbol{W}^{(l)}, \\
\boldsymbol{G}_{i,j}^{F} &= \left(\boldsymbol{G}_{i,j}^{(l)} + \boldsymbol{G}_{j,i}^{(l)}\right)/2, \\
\boldsymbol{G}_{i,j}^{S_{in}} &= \sum_k \frac{\boldsymbol{G}_{k,i}^{(l)}\boldsymbol{G}_{k,j}^{(l)}}{\sum_v \boldsymbol{G}_{k,v}^{(l)}}, \\
\boldsymbol{G}_{i,j}^{S_o} &= \sum_k \frac{\boldsymbol{G}_{i,k}^{(l)}\boldsymbol{G}_{j,k}^{(l)}}{\sum_v \boldsymbol{G}_{v,k}^{(l)}},
\end{aligned} \tag{10}$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{n^{(l)} \times c^{(l)}}$ refers to the node feature matrix of the l-th DGC layer. $n^{(l)}$ equals the number of nodes in the l-th DGC layer, and $c^{(l)}$ denotes the number of features in the l-th DGC layer. $\mathbf{H}^{(0)}$ is equal to the transpose of the time series data $\boldsymbol{X}$, denoted as $\boldsymbol{H}^{(0)} = \boldsymbol{X}^T$. $\boldsymbol{G}^F$, $\boldsymbol{G}^{S_{in}}$ and $\boldsymbol{G}^{S_o}$ represent the normalized first-order proximity matrix, the normalized second-order in-degree proximity matrix, and the normalized second-order out-degree proximity matrix, respectively. $\boldsymbol{G}^{(l)} \in \mathbb{R}^{n^{(l)} \times n^{(l)}}$ refers to the adjacency matrix of the l-th DGC layer. $\boldsymbol{G}^{(0)}$ is equal to the generated brain network $\boldsymbol{G}$, denoted as $\boldsymbol{G}^{(0)} = \boldsymbol{G}$. $\boldsymbol{G}_{i,j}$ denotes the element in row i and column j. $\boldsymbol{N}_e^{(l)}$ is the matrix obtained after performing the transformation operation on the input node feature matrix $\mathbf{H}^{(l)}$ using a shared trainable weight matrix $\boldsymbol{W}^{(l)}$. $\sigma(\cdot)$ is an activation function. $Con(\cdot)$ represents the concatenation operation.

In addition to the DGC layer, which performs transformation operations on the original node features, the inclusion of a pooling layer is crucial. The pooling layer offers two key advantages: firstly, it reduces the number of parameters to be trained; secondly, it enhances the interpretability of the model by identifying which subgraph structures are more informative in recognizing neurological disorders. Inspired by the work of [33], we employ the top-k pooling strategy to preserve crucial sub-graph structures, where "k" denotes the number of nodes to be retained. The pooling strategy can be formulated as follows:

$$\begin{aligned}
\mathbf{s}^{(l)} &= \mathbf{H}^{(l+1)}\mathbf{w}^{(l)}/||\mathbf{w}^{(l)}||_2, \\
index &= top-k\left(\mathbf{s}^{(l)}, k\right), \\
\mathbf{H}_{retain}^{(l+1)} &= \left(\mathbf{H}^{(l+1)} \odot \text{sigmoid}(\mathbf{s}^{(l)})\right)_{index,:}, \\
\boldsymbol{G}_{retain}^{(l+1)} &= \boldsymbol{G}_{index,index}^{(l+1)},
\end{aligned} \tag{11}$$

where $\mathbf{w}^{(l)} \in \mathbb{R}^{c^{(l+1)}}$ is a trainable vector used to assign a score to each node. $\mathbf{s}^{(l)} \in \mathbb{R}^{n^{(l+1)}}$ is a vector that contains the scores of all nodes. $||\cdot||_2$ is the L_2 norm. The $top-k(\cdot)$ function returns the indices corresponding to the k largest elements in the

score vector $\mathbf{s}^{(l)}$. $\odot$ represents the (broadcasted) element-wise multiplication operation, $(\cdot)_{i,j}$ denotes an indexing operation that retrieves elements with row indices specified by i and column indices specified by j (colons indicate all indices).

After applying l graph convolution and graph pooling operations on each subject's brain network, we retain the subgraph structures that are closely related to the downstream task, along with their node feature matrices (denoted as $\mathbf{H}^{(l)}_{retain}$). To facilitate graph classification, we further convert the node representations of these retained subgraph structures into graph-level representations using the following approach:

$$\boldsymbol{G}_r = vec\left(\mathbf{H}^{(l)}_{retain}\right), \tag{12}$$

where the symbol $vec(\cdot)$ indicates that the input matrix is converted to a vector form.

E. Total Loss

For each subject, when the time series data $\boldsymbol{X}$ is fed into the network model, the corresponding graph-level representation $\boldsymbol{G}_r$ of the subject will be obtained. The primary objective of this study is to predict whether a subject has a specific neurological disorder or not. To achieve this goal, we utilize cross-entropy (CE) loss as the optimization criterion for the prediction outcome. The formulation of the cross-entropy loss can be defined as follows:

$$\mathcal{L}_{cls} = CE\left(softmax(FC(\boldsymbol{G}_r)),\ y\right), \tag{13}$$

where $FC(\cdot)$ is a fully connected layer, which compresses the graph-level representation of the subject to match the dimension of the target classification category. y represents the true label corresponding to the subject.

The overall loss of BSIGNN can be formulated by combining the losses from the two subsections mentioned above, as follows:

$$\mathcal{L}_{total} = \mathcal{L}_G + \mathcal{L}_{sparse} + \mathcal{L}_{cls} + \lambda\mathcal{L}_{rank}, \tag{14}$$

where λ is a trade-off parameter.

IV. Experimental Results

In this section, we first briefly describe the dataset used in this paper in Section IV-A. Then, the details of the experimental settings and the comparison methods are outlined in Sections IV-B and IV-C, respectively. Finally, we present the experimental results, as well as the ablation study, in Sections IV-D and IV-E.

A. Datasets and Preprocessing

In this paper, we evaluated the effectiveness of our proposed method using resting-state functional magnetic resonance imaging (rs-fMRI) data obtained from two publicly available datasets: ABIDE (Autism Brain Imaging Data Exchange)[1] and ADNI (Alzheimer's Disease Neuroimaging Initiative).[2] The objective of the ABIDE dataset is to differentiate between individuals with Autism Spectrum Disorder (ASD) and typically developing (TD) subjects within the population. Similarly, the objective of the ADNI dataset is to differentiate between normal controls (NC) and individuals with mild cognitive impairment (MCI) within the population. The demographics of the recruited subjects are listed in Table I.

TABLE I
Demographic Information of the Datasets Used in This Work

Dataset	Group	Number	Gender (M/F)	Age (mean±std)
ABIDE	TD	571	472/99	17.1±7.8
	ASD	531	467/64	17.1±8.1
ADNI	NC	170	76/94	75.1±6.2
	MCI	365	173/192	71.9±7.2

1) ABIDE Dataset: The ABIDE dataset consists of rs-fMRI data collected from 17 different sites. It includes a total of 531 individuals with ASD and 571 TD subjects. The rs-fMRI data utilized in this study were preprocessed using the Connectomes Analytics Configurable Pipeline (C-PAC) [34]. All data within the ABIDE dataset were processed while ensuring complete anonymity, thus meeting the requirements of the Health Insurance Portability and Accountability Act (HIPAA).

2) ADNI Dataset: The ADNI is a longitudinal cohort dataset designed for the study of Alzheimer's disease (AD). In this study, we utilized a total of 535 subjects' rs-fMRI data from the ADNI dataset to recognize mild cognitive impairment (MCI). The rs-fMRI data of all subjects were preprocessed using the Conn Toolbox.[3] The specific preprocessing protocol involved several steps, including motion correction, registration, normalization to the Montreal Neurological Institute (MNI) space with a resampled voxel size of $3 \times 3 \times 3$ mm^3, outlier detection, and spatial smoothing using a Gaussian kernel with a full width half maximum (FWHM) of 8 mm. The research conducted in this study received approval from the Research Ethics Board of ADNI.[4]

The time-series data for each brain region were extracted from the preprocessed rs-fMRI images using the Automated Anatomical Labeling (AAL) atlas [35] after completing all preprocessing steps for both datasets.

B. Experimental Settings

1) Implementation Details: The BSIGNN is implemented using the open-source framework PyTorch in Python (Version 3.9.5). The detailed structure of the BSIGNN network is presented below. The BSI module is composed of 90 sub-modules, with each sub-module consisting of an LSTM equipped with a single hidden layer comprising 64 neurons, along with a fully connected layer. The input length for each sub-module is 8, i.e., the time lag length $L = 8$. After the BSI module, we iteratively repeat the combination of a DGC layer and a top-k pooling layer twice within the network. Notably, each of these top-k pooling layers retains 1/3 of the original nodes as important sub-structures. Specifically, the parameter k in Eq. (11) is set as 30 and 10 for the

[1] http://fcon_1000.projects.nitrc.org/indi/abide/
[2] http://adni.loni.usc.edu/
[3] https://web.conn-toolbox.org/
[4] http://adni.loni.usc.edu/study-design/ongoing-investigations/

two instances, respectively. The training of the network is accelerated by a single Nvidia GeForce 1080Ti GPU. The hyperparameters are configured as follows: the number of epochs is set to 100, the learning rate is set to 0.001 for ABIDE and 0.01 for ADNI, the batch size is set to 32, and the balance parameter λ in Eq. (14) is assigned a value of 0.001. The Adam optimizer is utilized for model optimization with a weight decay of 0.01 to avoid overfitting. It is worth noting that the ADNI dataset has an imbalance between the number of positive and negative samples (NC/MCI). To mitigate this issue, we introduced category weights in the cross-entropy loss and set the weight ratio of positive and negative samples to 2:1. By assigning higher weights to the minority class, we aim to alleviate the potential bias caused by imbalanced data distribution and improve the model's performance in recognizing both positive and negative instances.

2) Validation Scheme and Evaluation Metrics: This study evaluates the classification performance of the proposed method using a standard 5-fold cross-validation strategy. However, it is important to note that there are subtle differences in how the two public datasets (ADNI and ABIDE) are partitioned into five subsets. Since the ADNI is a longitudinal dataset, the use of a random partition scheme may result in cases where data from the same subject is distributed across multiple subsets. This situation may affect the fairness of the subsequent assessments. Therefore, for the ADNI dataset, a subject-level split scheme is implemented to ensure that all fMRI data from the same subject are assigned exclusively to one of the five subsets. Conversely, for the ABIDE dataset, a random partition scheme is employed to divide the entire dataset into five equal subsets. The performance is evaluated based on accuracy (ACC), sensitivity (SEN), specificity (SPE), and F1 score. Greater values for these metrics indicate improved performance.

C. Comparison Methods

To validate the effectiveness of the proposed method, we compare it with seven GNN-based methods. These seven methods aim to enhance the performance of GNN models in identifying neurological diseases from different perspectives. Below is a brief introduction to each method:

1) Hi-GCN [10]. The Hi-GCN model utilizes two types of graph convolution layers to automatically extract a diverse range of features from brain networks. These features have been demonstrated to be effective in identifying neurological disorders.
2) BrainGNN [21]. The BrainGNN model employs a novel graph pooling layer to highlight the sub-graph structures that are more crucial in predicting results, thereby enhancing the interpretability of the model.
3) PSCR-GNN [16]. The PSCR-GNN method constructs a novel brain network as input for the graph attention network. The authors conducted extensive experiments on the ABIDE dataset to showcase the superiority of this new approach.
4) MVS-GCN [36]. The MVS-GCN model integrates graph structure learning and graph representation learning within a unified framework. However, the brain network structure is simply defined as the inner product of node representations, this oversimplified definition ignores the complex nonlinear interactions between brain regions.
5) MAHGCN [37]. The MAHGCN model introduces a novel graph convolution layer that captures complementary information from brain networks across various scales, thereby enhancing the model's classification performance. Furthermore, it can explore the hierarchical relationships within the human brain, allowing for a more comprehensive analysis of brain disorders.
6) GRN-GNN [38]. The main contribution of GRN-GNN is the introduction of the graph registration module, which facilitates the learning of shared features among subjects. This module acts as a flexible component that can be incorporated into various graph network models, such as graph classification networks, to improve their performance in downstream tasks.
7) A-GCL [39]. The A-GCL model utilizes the contrastive learning strategy to effectively extract features from fMRI data that are crucial for disease identification. The authors validate the model's performance on three publicly available datasets, all of which demonstrate optimal classification performance.

D. Classification Results

Table II presents a summary of the average classification results obtained from five repetitions for all methods. The best results are highlighted in bold. As shown in Table II, our method achieves the highest results on both public datasets, demonstrating the effectiveness of our method. The GRN-GNN method exhibits the poorest performance compared to other methods. This is primarily attributed to their training methodology. During the training, subjects from all categories were included to learn shared features, potentially encouraging the model to learn irrelevant features for classification. The absence of discriminative features significantly undermines the performance in subsequent classification tasks. In addition, the performance of the PSCR-GNN method is unsatisfactory. The primary reason for the poor performance of the PSCR-GNN method is that the classification framework is divided into two distinct steps. Initially, the sparse brain network is constructed using a predefined metric. Then, the designed brain network will be inputted into the graph attention network for feature extraction. The quality of the brain network designed in the initial step significantly affects the subsequent classification performance. Consequently, the PSCR-GNN method is likely to yield inferior results compared to other end-to-end methods.

Apart from the two methods mentioned above, the remaining methods primarily focus on exploring how to extract more comprehensive features using graph neural networks to enhance classification performance. For example, Hi-GCN [10] is specifically designed to employ graph neural networks to extract multiple hierarchical representations for the identification of neurological disorders. MAHGCN [37] is specifically designed with multiple branches to extract more comprehensive features from brain networks constructed using multiple brain atlases. However, these methods often

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON TWO PUBLIC DATASETS

Metric	ABIDE (TD vs. ASD)							
	Hi-GCN	BrainGNN	PSCR-GNN	MVS-GCN	MAHGCN	GRN-GNN	A-GCL	BSIGNN
ACC	65.88±0.56	67.57±1.59	64.97±0.71	69.00±1.72	68.82±0.58	57.10±1.52	71.58±0.94	**74.83**±1.07
SEN	67.18±1.78	70.19±2.24	68.44±2.50	73.91±2.58	67.46±2.48	60.04±3.74	75.76±1.56	**77.51**±1.51
SPE	64.48±0.78	64.75±1.55	61.24±1.31	63.73±1.86	70.28±1.53	53.94±1.15	67.08±0.75	**71.94**±0.76
F1	67.10±0.94	69.16±1.69	66.93±1.26	71.18±1.80	69.15±1.17	59.15±2.39	73.41±1.03	**76.14**±1.12
Metric	ADNI (NC vs. MCI)							
	Hi-GCN	BrainGNN	PSCR-GNN	MVS-GCN	MAHGCN	GRN-GNN	A-GCL	BSIGNN
ACC	71.33±0.64	70.21±0.86	68.26±1.13	72.41±2.54	74.99±1.96	59.85±1.81	74.50±1.28	**79.18**±0.90
SEN	61.88±2.48	60.47±3.10	56.35±1.80	66.00±6.15	68.82±4.01	44.59±1.98	71.76±2.52	**75.29**±2.04
SPE	75.73±0.33	74.74±0.40	73.81±1.16	75.40±2.76	77.86±1.59	66.96±1.87	75.78±0.84	**80.99**±0.71
F1	57.84±1.52	56.30±2.58	53.01±1.58	60.28±3.87	63.62±3.07	41.38±2.11	64.12±1.94	**69.70**±1.44

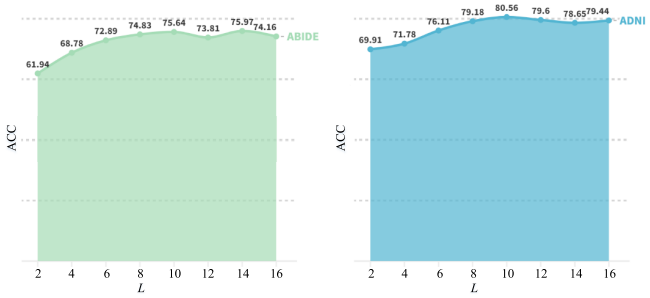


Fig. 3. The classification accuracy of the model on two public datasets when L is set to different values. The accuracy presented in the figure represents the average result obtained from five rounds of measurements.

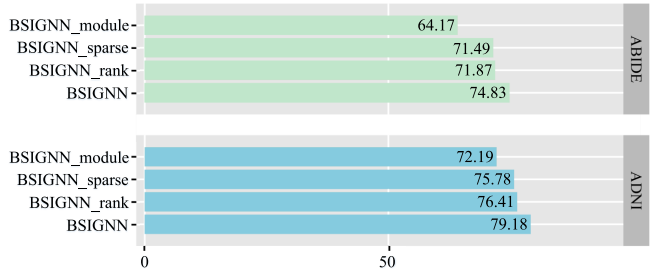


Fig. 4. Recognition performance of different degraded networks. The accuracy presented in the figure represents the average result obtained from five rounds of measurements.

depend on predefined metrics, such as the Pearson correlation coefficient, to construct graphs that serve as inputs to the model. These methods, which rely on predefined graphs as input, exhibit inflexibility. Brain networks constructed using specific metrics may perform effectively for one type of disease but may not yield optimal results for another. Therefore, in this paper, we propose a BSI module to automatically learn the most discriminative graph structures from time series data. This is the reason why other methods do not achieve comparable performance to our method. The graph we employ as input for the model is learnable, enabling co-optimization with downstream tasks.

E. Ablation Study

As illustrated in Fig. 2, our BSI module consists of N sub-modules. Each sub-module takes as input a time series with a length equal to the time lag length L. If the value of L is set to be excessively large, it can result in a model with an overwhelming number of parameters, making it difficult to optimize. Conversely, if the value of L is set too small, it may prove difficult to adequately capture the time dependence between variables. Hence, we conducted an analysis to examine the impact of different L values on model performance. The impact of various L values on the results is presented in Fig. 3.

As shown in Fig. 3, the classification performance of the model on the two public datasets exhibits a gradual improvement as the value of L increases. However, it is worth noting that when the L value exceeds 8, the model's performance improvement on the two public datasets is very limited. In fact, it even oscillates with a slight degradation in performance. Therefore, we believe that an L value of 8 may be the best choice. As the value of L transitions from 8 to 16, the model's performance has stabilized. Using a larger L value requires more computational resources.

The main contribution of this paper is the design of the BSI module, which enables smooth integration with different downstream tasks, such as graph classification tasks. In addition to the BSI module, we also impose two constraints on the potential graph structure: the sparsity constraint $\mathcal{L}_{sparse}$ and the low-rank constraint $\mathcal{L}_{rank}$. To evaluate the contribution of each innovative component to the excellent performance, we conduct an ablation study by designing three degraded networks. These degraded networks include: 1) we remove the BSI module from the network and utilized a brain network constructed from Pearson's correlation coefficient as the input instead, denoted "BSIGNN_module", 2) we remove the sparsity loss $\mathcal{L}_{sparse}$ from the loss function of BSIGNN, denoted "BSIGNN_sparse", and 3) we remove the low-rank loss $\mathcal{L}_{rank}$ from the loss function of BSIGNN, denoted "BSIGNN_rank".

Fig. 4 illustrates the results of the degraded networks on two public datasets. It can be seen that the BSIGNN_module achieves the worst performance among all the degraded networks. This result provides compelling evidence that predefined graphs may not be entirely suitable for downstream tasks. The performance of the other two variants (BSIGNN_sparse

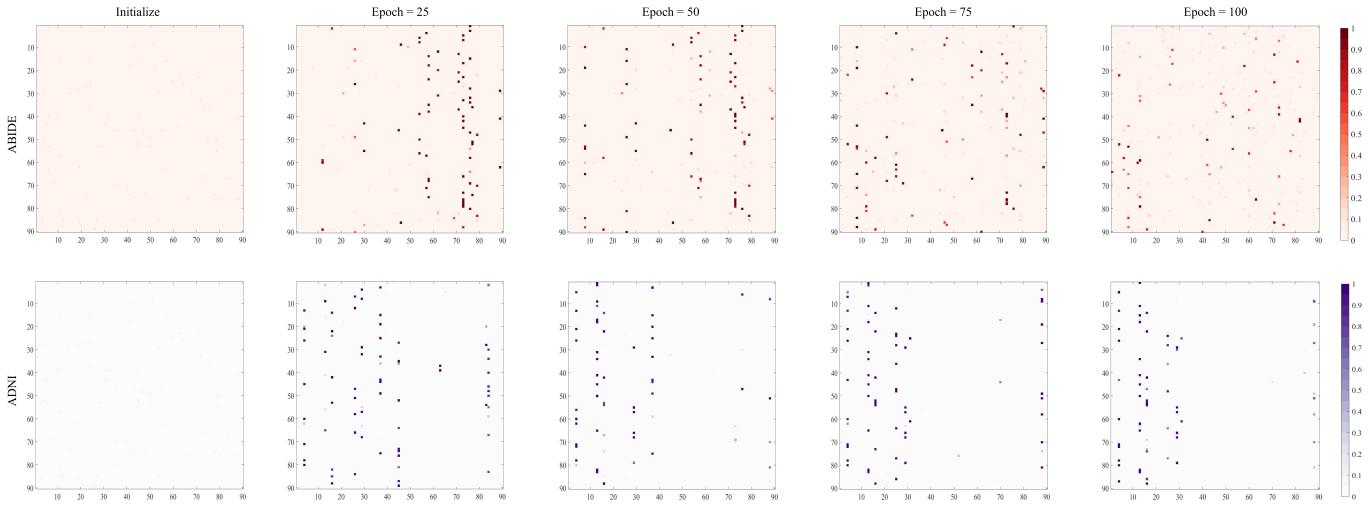
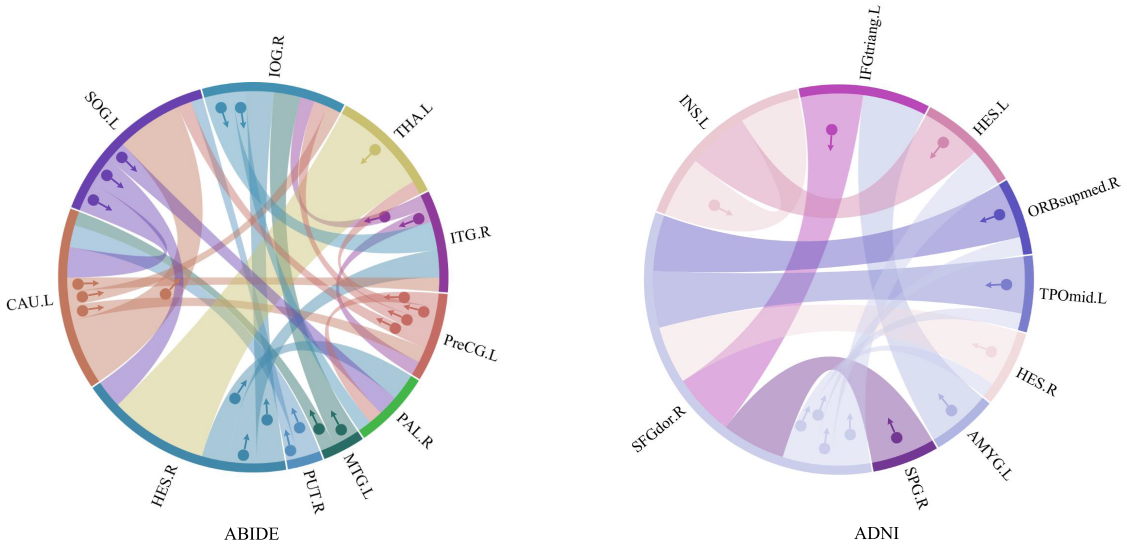


Fig. 5. Visualization results of the average brain network across all subjects as the number of epochs varies. Note that the indices of the x- and y-axes in each map correspond to the brain regions defined by the AAL atlas [35]. The leftmost column illustrates the initial stage of the brain network, starting with a Gaussian distribution (mean = 0, std = 0.01). As the epochs progress sequentially from left to right, the graph's structure transforms into a sparser and lower-rank graph, driven by the training data. By the 100th epoch, the learned brain network has effectively adapted to the assigned task, reinforcing the weights of the irregular connections associated with the particular disease under investigation.



HES.R: right heschl gyrus
SOG.L: left superior occipital gyrus
THA.L: left thalamus
PreCG.L: left precental gyrus
MTG.L: left middle temporal gyrus
CAU.L: left caudate nucle
IOG.R: right inferior occipital gyrus
ITG.R: right inferior temporal gyrus
PAL.R: right lenticular nucleus
PUT.R: right Lenticular nucleus
SFGdor.R: right superior frontal gyrus
IFGtriang.L: left inferior frontal gyrus
ORBsupmed.R: right superior frontal gyrus
HES.R: right heschl gyrus
SPG.R: right superior parietal gyrus
INS.L: left insula
HES.L: left heschl gyrus
TPOmid.L: left middle temporal gyrus
AMYG.L: left amygdala

Fig. 6. The most discriminative connections identified by our method. The motion direction of the ball in each arc represents the causal relationship between two brain regions (from cause to effect).

and BSIGNN_rank) is degraded compared to BSIGNN, indicating that the two losses we introduced are contributing to the model's performance.

V. Discussion

A. Visualization of Learned Brain Networks

Fig. 5 presents the changes in the average brain network across all subjects as the number of epochs varies. The first column shows the initialization of the brain network, where a Gaussian initialization with a mean of 0 and a standard deviation of 0.01 was used. The last column, i.e., epoch=100, represents the final brain network obtained after 100 epochs. As can be seen from Fig. 5, the graph structures obtained at different epochs exhibit sparsity and low rank. This phenomenon is attributed to the two constraints we introduced: the sparsity constraint $\mathcal{L}_{sparse}$ and the low-rank constraint $\mathcal{L}_{rank}$.

Comparing the graph structures learned from the two datasets, it is apparent that our method exhibits different convergence rates for each dataset. Specifically, our method achieves convergence after approximately 50 epochs on the ADNI dataset, whereas it takes 75 epochs to achieve convergence on the ABIDE dataset. The main reason for this phenomenon is the difference in demographic characteristics

between the two datasets. The ABIDE dataset primarily consists of adolescents, while the majority of individuals in the ADNI dataset are older adults. Since the adolescent population is in the growth and development stage, they are more susceptible to age-related factors. As a result, individual differences are more pronounced in this group compared to the older population.

B. Most Discriminative Connections

In Fig. 6, we present the most discriminative connections across brain regions identified by our method. There are three points that require clarification regarding the Circos graph. First, the significance of a brain region in recognizing neurological disorders is indicated by the size of the corresponding arc, with larger arcs representing greater importance. Then, the color of each arc is randomly assigned and does not convey any specific meaning. Finally, the motion direction of the ball in each arc represents the causal relationship between two brain regions (from cause to effect).

In the left figure, it can be seen that the brain region left caudate nucleus (CAU.L) has the highest number of abnormal connections among all brain regions, with a total of seven. Among these seven connections, the connection from CAU.L to the left superior occipital gyrus (SOG.L) exhibits the highest weight, suggesting its potential as a reliable biomarker for ASD diagnosis. Similarly, in the right figure, the brain region right superior frontal gyrus (SFGdor.R) has the highest number of abnormal connections among all brain regions, with a total of five. Among these five connections, the connection from the right superior frontal gyrus (ORBsupmed.R) to SFGdor.R possesses the highest weight, indicating its potential as an effective biomarker for the diagnosis of MCI.

In fact, the abnormal brain regions identified in Fig. 6, such as CAU.L [2], [9], [16] and SFGdor.R [1], [14], [37], have been extensively documented in numerous previous studies. This finding not only highlights the effectiveness of our approach but also demonstrates the high interpretability of our approach.

VI. Conclusion

In this paper, we propose a BSI module to estimate potential graph structures that are highly beneficial for downstream tasks. Along with the development of this new module, we have introduced two additional constraints on graph structures to improve the generalization performance of our model. The designed module can be seamlessly integrated with downstream tasks within a unified framework for joint optimization. Remarkably, our method achieved excellent performance on both public datasets, which fully demonstrates the effectiveness of our approach.

However, it is worth noting that the algorithm employed in this study has not fully take into account the hierarchical organization of the brain. This limitation leads to the inadequacy of the algorithm in capturing cross-scale features, potentially overlooking some crucial discriminative factors. Therefore, we intend to further improve the algorithm in future studies to thoroughly analyze brain interactions across multiple scales. If this issue is solved, it will not only improve the accuracy of neurological disease recognition but also contribute to a better understanding of the intricate dynamic interaction mechanisms within the brain.

References

[1] X. Jiang, L. Qiao, R. De Leone, and D. Shen, "Joint selection of brain network nodes and edges for MCI identification," *Comput. Methods Programs Biomed.*, vol. 225, Oct. 2022, Art. no. 107082.

[2] M. Wang, J. Huang, M. Liu, and D. Zhang, "Modeling dynamic characteristics of brain functional connectivity networks using resting-state functional MRI," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102063.

[3] X. Sheng, J. Chen, Y. Liu, B. Hu, and H. Cai, "Deep manifold harmonic network with dual attention for brain disorder classification," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 1, pp. 131–142, Jan. 2023.

[4] C. Ding et al., "Global, regional, and national burden and attributable risk factors of neurological disorders: The global burden of disease study 1990–2019," *Frontiers Public Health*, vol. 10, Nov. 2022, Art. no. 952161.

[5] T. Wadhera and M. Mahmud, "Brain functional network topology in autism spectrum disorder: A novel weighted hierarchical complexity metric for electroencephalogram," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 4, pp. 1718–1725, Apr. 2023.

[6] J. Cai et al., "Dynamic graph theoretical analysis of functional connectivity in Parkinson's disease: The importance of Fiedler value," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1720–1729, Jul. 2019.

[7] S. Avvaru and K. K. Parhi, "Effective brain connectivity extraction by frequency-domain convergent cross-mapping (FDCCM) and its application in Parkinson's disease classification," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 8, pp. 2475–2485, Aug. 2023.

[8] W. Wang, L. Xiao, G. Qu, V. D. Calhoun, Y.-P. Wang, and X. Sun, "Multiview hyperedge-aware hypergraph embedding learning for multisite, multiatlas fMRI based functional connectivity network analysis," *Med. Image Anal.*, vol. 94, May 2024, Art. no. 103144.

[9] W. Cui et al., "Dynamic multi-site graph convolutional network for autism spectrum disorder identification," *Comput. Biol. Med.*, vol. 157, May 2023, Art. no. 106749.

[10] H. Jiang, P. Cao, M. Xu, J. Yang, and O. Zaiane, "Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction," *Comput. Biol. Med.*, vol. 127, Dec. 2020, Art. no. 104096.

[11] H. Zhang et al., "Classification of brain disorders in rs-fMRI via local-to-global graph neural networks," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 444–455, Feb. 2023.

[12] J. Runge et al., "Inferring causation from time series in earth system sciences," *Nature Commun.*, vol. 10, no. 1, p. 2553, Jun. 2019.

[13] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.

[14] Z. Xia, T. Zhou, S. Mamoon, A. Alfakih, and J. Lu, "A structure-guided effective and temporal-lag connectivity network for revealing brain disorder mechanisms," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 6, pp. 2990–3001, Jun. 2023.

[15] J. Ji and Y. Yao, "Convolutional neural network with graphical lasso to extract sparse topological features for brain disease classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2327–2338, Nov. 2021.

[16] C. Yang, P. Wang, J. Tan, Q. Liu, and X. Li, "Autism spectrum disorder diagnosis using graph attention network based on spatial-constrained sparse functional brain networks," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104963.

[17] C.-Y. Wee, S. Yang, P.-T. Yap, and D. Shen, "Sparse temporally dynamic resting-state functional connectivity networks for early MCI identification," *Brain Imag. Behav.*, vol. 10, no. 2, pp. 342–356, Jun. 2016.

[18] G. Zhang et al., "Detecting abnormal connectivity in schizophrenia via a joint directed acyclic graph estimation model," *NeuroImage*, vol. 260, Oct. 2022, Art. no. 119451.

[19] S. Parisot et al., "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and Alzheimer's disease," *Med. Image Anal.*, vol. 48, pp. 117–130, Aug. 2018.

[20] X. Song et al., "Graph convolution network with similarity awareness and adaptive calibration for disease-induced deterioration prediction," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101947.

[21] X. Li et al., "BrainGNN: Interpretable brain graph neural network for fMRI analysis," *Med. Image Anal.*, vol. 74, Dec. 2021, Art. no. 102233.

[22] P. Li et al., "Robust brain causality network construction based on Bayesian multivariate autoregression," *Biomed. Signal Process. Control*, vol. 58, Apr. 2020, Art. no. 101864.

[23] N. Chen, M. Guo, Y. Li, X. Hu, Z. Yao, and B. Hu, "Estimation of discriminative multimodal brain network connectivity using message-passing-based nonlinear network fusion," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 4, pp. 2398–2406, Jul./Aug. 2023.

[24] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica, J. Econ. Soc.*, vol. 37, no. 3, pp. 424–438, Aug. 1969.

[25] B. Liu et al., "PRTA: Joint extraction of medical nested entities and overlapping relation via parameter sharing progressive recognition and targeted assignment decoding scheme," *Comput. Biol. Med.*, vol. 176, Jun. 2024, Art. no. 108539.

[26] 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 Recognit.*, vol. 90, pp. 220–231, Jun. 2019.

[27] M. J. Rosa et al., "Sparse network-based models for patient classification using fMRI," *NeuroImage*, vol. 105, pp. 493–506, Jan. 2015.

[28] J. Xu et al., "Large-scale functional network overlap is a general property of brain functional organization: Reconciling inconsistent fMRI findings from general-linear-model-based analyses," *Neurosci. Biobehavioral Rev.*, vol. 71, pp. 83–100, Dec. 2016.

[29] P. Lennie, "The cost of cortical computation," *Current Biol.*, vol. 13, no. 6, pp. 493–497, Mar. 2003.

[30] X. Liao, A. V. Vasilakos, and Y. He, "Small-world human brain networks: Perspectives and challenges," *Neurosci. Biobehavioral Rev.*, vol. 77, pp. 286–300, Jun. 2017.

[31] Z. Li et al., "Disrupted brain network topology in chronic insomnia disorder: A resting-state fMRI study," *NeuroImage, Clin.*, vol. 18, pp. 178–185, Jan. 2018.

[32] Z. Tong, Y. Liang, C. Sun, D. S. Rosenblum, and A. Lim, "Directed graph convolutional network," 2020, *arXiv:2004.13970*.

[33] D. Bacciu and L. Di Sotto, "A non-negative factorization approach to node pooling in graph convolutional neural networks," in *Proc. 18th Int. Conf. Italian Assoc. Artif. Intell.*, Rende, Italy. Springer, Nov. 2019, pp. 294–306. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-35166-3_21

[34] S. Sharad et al., "Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC)," *Frontiers Neuroinform.*, vol. 8, pp. 45–46, Jul. 2013.

[35] N. Tzourio-Mazoyer et al., "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, Jan. 2002.

[36] G. Wen, P. Cao, H. Bao, W. Yang, T. Zheng, and O. Zaiane, "MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis," *Comput. Biol. Med.*, vol. 142, Mar. 2022, Art. no. 105239.

[37] M. Liu, H. Zhang, F. Shi, and D. Shen, "Hierarchical graph convolutional network built by multiscale atlases for brain disorder diagnosis using functional connectivity," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 20, 2023, doi: 10.1109/TNNLS.2023.3282961.

[38] Z. Gürler, M. A. Gharsallaoui, and I. Rekik, "Template-based graph registration network for boosting the diagnosis of brain connectivity disorders," *Computerized Med. Imag. Graph.*, vol. 103, Jan. 2023, Art. no. 102140.

[39] S. Zhang et al., "A-GCL: Adversarial graph contrastive learning for fMRI analysis to diagnose neurodevelopmental disorders," *Med. Image Anal.*, vol. 90, Dec. 2023, Art. no. 102932.