Spatio-Temporal Hybrid Attentive Graph Network for Diagnosis of Mental Disorders on fMRI Time-Series Data

Rui Liu, Zhi-An Huang, Yao Hu, *Graduate Student Member, IEEE*, Lei Huang, Ka-Chun Wong, and Kay Chen Tan, *Fellow, IEEE*

***Abstract*—Facing the prevalence of mental disorders around the world, the burden of healthcare services becomes increasingly imminent. To lessen patients' suffering, the timely diagnosis and therapy of mental disorders are particularly essential. Functional magnetic resonance imaging (fMRI), as the de facto non-invasive neuroimaging technique, can effectively examine the spatial and temporal patterns of brain activity. Recently, computer-aided diagnosis (CAD) approaches have emerged to assist doctors in interpreting fMRI images. However, existing CAD methods cannot fully exploit the spatio-temporal dependence in fMRI signals, possibly leading to inaccurate diagnosis. In this study, we propose a spatio-temporal hybrid attentive graph network (ST-HAG) for diagnosing autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) from fMRI data. Specifically, a hybrid graph convolution network is developed to effectively capture complex spatio-temporal dynamics. Meanwhile, a Transformer-based self-attention module helps ST-HAG to extract the full-scale temporal correlation. Finally, we use a gated fusion unit to learn discriminative spatio-temporal graph representations for classification. Cross-validation experiments demonstrate that the proposed ST-HAG achieves state-of-the-art performance with a mean accuracy of 71.9% and 74.8% for ASD and ADHD on ABIDE (1035 subjects) and ADHD-200 (939 subjects) datasets, respectively. Moreover, thanks to the adopted dynamic graph attentive representation, the potent interpretability enables ST-HAG to detect the remarkable temporal association patterns among different brain regions based on dynamic functional connectivity networks.**

***Index Terms*—Attention deficit/hyperactivity disorder, autism spectrum disorder, brain graph construction, computer-aided diagnosis, functional magnetic resonance imaging, graph learning.**

Manuscript received 7 October 2023; revised 4 January 2024; accepted 16 March 2024. Date of publication 17 April 2024; date of current version 23 November 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62202399 and Grant U21A20512, in part by the Research Grants Council of the Hong Kong SAR under Grant PolyU11211521, Grant PolyU15218622, and Grant PolyU15215623, and in part by the Hong Kong Polytechnic University under Grant P0039734 and Grant P0035379. *(Corresponding author: Zhi-An Huang.)*

Rui Liu and Kay Chen Tan are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China (e-mail: ruiliu@polyu.edu.hk; kctan@polyu.edu.hk).

Zhi-An Huang is with the Research Office, City University of Hong Kong (Dongguan), Dongguan 523000, China, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518060, China (e-mail: huang.za@cityu-dg.edu.cn).

Yao Hu, Lei Huang, and Ka-Chun Wong are with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518060, China (e-mail: yaohu4-c@my.cityu.edu.hk; lhuang93-c@my.cityu.edu.hk; kc.w@cityu.edu.hk).

Recommended for acceptance by Y. Wang.

Digital Object Identifier 10.1109/TETCI.2024.3386612

I. Introduction

FUNCTIONAL magnetic resonance imaging (fMRI) has been widely adopted to probe functional brain activities by detecting the variations of blood-oxygen-level-dependent (BOLD) signals in brain blood flow [1]. The interpretation of BOLD signals in the brain provides insights into understanding the etiopathogenesis of complex neurological disorders such as Alzheimer's disease, autism spectrum disorder (ASD), and attention deficit hyperactivity disorder (ADHD) [2]. Specifically, the identification of abnormal functional connectivity (FC) enables the characterization of disease-related neuronal activation patterns among anatomically distinct regions of interest (ROIs) in the brain. Consequently, numerous computer-aided diagnosis (CAD) methods have been developed based on FC analysis, aiming to support medical professionals in the diagnosis of mental disorders [3], [4], [5], [6]. It is worth noting that many of these methods rely on static FC measures computed over the entire scanning duration, assuming that brain activities between different ROIs remain constant over time [7]. However, the recent studies [8] suggest that distinct functional brain activity in complex mental disorders can exhibit dynamic variations within short time intervals, involving different activated ROIs. Therefore, the emergence of dynamic FC analysis offers a promising approach for capturing time-resolved patterns of FC within dynamic spatial brain networks.

In the realm of fMRI-based applications, sequence models such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks [9], and Transformers [10] have demonstrated notable success. These models have been effective in capturing dynamic connectivity features, which can serve as potential dynamic temporal biomarkers [11]. However, despite their ability to model dynamic biomarkers, existing sequence models have failed to fully exploit the advantages of jointly exploring the spatial and temporal dependencies of fMRI signals. This limitation hinders a more precise analysis of brain activity [12], [13]. To overcome this challenge, it is imperative to develop CAD models that effectively investigate the intricate interplay between the spatial and temporal correlations of brain activity.

Fortunately, early attempts have been made to integrate spatial and temporal features for inference in fMRI data analysis. Hartvi et al. [14] proposed two statistical inference models to estimate spatial and temporal activation patterns separately.

Building upon this work, Derado et al. [15] introduced a two-stage autoregressive model that accounted for both spatial dependencies between voxels and temporal dependencies between scanning sessions. Yet, these methods rarely explored the important spatio-temporal interaction, i.e., ignoring how the connectome-scale brain network temporally evolves, possibly due to medical data scarcity in the past. More recently, deep learning models have shown great potential in automatically extracting hierarchical non-linear hidden features with varying levels of complexity. Mao et al. [16] proposed a four-dimensional convolutional neural network (CNN) architecture that simultaneously captured spatial and temporal characteristics from targeted networks in an end-to-end manner. Recognizing the brain network's natural compatibility with graph theory, several spatio-temporal graph neural networks (STGNNs) [17], [18] have been developed to measure intra-subject temporal dynamics and inter-regional spatial associations. Among these STGNNs, the spatio-temporal graph convolutional network (STGCN) [19] stands out as an efficient variant that applies convolutions over time-varying dynamic graphs. By fully exploiting the spatial and temporal characteristics of fMRI data, STGCN models hold great promise for effectively classifying individual patients. Nevertheless, existing STGCN models have struggled to consistently outperform static functional connectivity (FC)-based approaches in terms of classification performance [20]. This issue primarily stems from the construction of current STGCNs, which often model spatial and temporal dependencies alternately, overlooking the interplay between spatial and temporal dynamic variations. Additionally, the limited size of convolution kernels used in STGCNs poses a challenge in capturing comprehensive patterns of temporal dependencies beyond the kernel's receptive field. This limitation can lead to the exclusion of crucial global and subtle properties, which are particularly relevant in the examination of mental disorders [21].

To address the aforementioned limitations, we propose a spatio-temporal hybrid attentive graph network (ST-HAG) for mental disorder diagnosis by capturing the discriminative spatio-temporal interaction dynamics. Specifically, we design a dynamic adaptative-neighbor graph convolutional network (DAN-GCN) that effectively extracts the informative spatial dependencies between brain regions across adjacent timesteps. Based on Transformer, a temporal sliding self-attention module (TSSA) is designed to learn discriminative temporal correlations and provides temporal attention guidance to DAN-GCN during training. Finally, we leverage a spatio-temporal gated fusion block (STGF) to combine learned representations from DAN-GCN and TSSA for further classification. With the DAN-GCN and TSSA modules as a building block, the ST-HAG model can jointly consider the evolving spatial connectivity patterns and the dynamic temporal variations, thereby effectively capture the complex spatio-temporal dynamics of brain activity. The major contributions of this work are summarized as follows:

- We propose an effective network ST-HAG to simultaneously capture discriminative spatio-temporal representation, which is the first attempt to jointly model the spatio-temporal interaction of dynamic FC bran network in this domain.
- Based on the dynamic adaptative-neighbor graph construction and sliding window attention (SWA) technique, the proposed DAN-GCN and TSSA are demonstrated to address the limitations present in standard STGCN, respectively. Moreover, TSSA further strengthens the performance of DAN-GCN by providing the learned temporal attention guidance.
- On two real-world datasets, ST-HAG achieves supreme classification results as compared to the state-of-the-art (SOTA) methods and other STGNN architectures. The findings of remarkable ROIs and reproducible dynamic FC patterns derived by ST-HAG are compatible with the previous literature and statistical sample analysis.

The rest of this paper is organized as follows. Section II reviews the related work. Section III describes the proposed framework including DAN-GCN and TSSA. The experimental results and analysis are discussed in Section IV. Finally, Section V summarizes the key findings and contributions of the study.

II. Related Works

A. Static/Dynamic Graph Representation Learning for FC Analysis

Traditionally, FC analysis research mostly focuses on static brain graphs where the nodes and edges have no change over time. For example, Li et al. [22] designed a dual graph convolutional network (GCN) based on the vanilla GCN to learn discriminative topological features of static brain networks for interpreting and identifying Alzheimer's disease. In [23], hypergraph learning was used to model the high-order relations for static FC detection among multiple ROIs, which was used to calculate a unified hypergraph similarity matrix to estimate the learning ability of individuals. Moreover, for diagnosing ASD from fMRI data, Wang et al. [24] presented a connectivity-based GCN to efficiently extract graph representations of brain activity through FC analysis via k-nearest neighbors. With static graphs, FC analysis can benefit from convenient operations such as storing a fixed graph data structure and performing graph transformations. Since the topological property is consistently revealed, the reliability of graph measures in static brain connectivity can be easily studied.

Recently, dynamic graph representation learning has been demonstrated to effectively characterize the time-evolving brain connectivity at a system level. Gadgil et al. [17] introduced the STGCN in FC analysis for predicting the age and gender of healthy individuals, which is achieved by modeling the non-stationary short sub-sequences segmented from fMRI signals. Kim et al. [25]) proposed two attention-based modules to learn dynamic graph representation of the brain connectome. A novel readout function and Transformer encoder are then employed for temporal attention statistical interpretation. Zhao et al. [26] also designed a dynamic GCN framework to identify ADHD by extracting both static and dynamic properties of subject-level FC patterns. These studies demonstrate that the dynamics of brain connectivity are highly reproducible across repeated scans and

can be regarded as informative biomarkers for developing CAD models [27]. However, current STGCN-based dynamic graph representation learning methods only learn representations of the spatial and temporal levels in an alternating way. The interplay between spatial and temporal dynamics remains largely unexplored, which limits their ability to effectively characterize the temporal dynamics of brain activity and filter out the spurious fluctuations inherent in neuroimaging data [28]. In order to address this issue, we propose an noval graph convolution network DAN-GCN to effectively capture the spatio-temporal interplay by incorporating graph convolutions among brain regions across adjacent timesteps.

B. Attention Mechanism for Neuroimaging Analysis

Attention mechanism in deep learning has achieved impressive results in the analysis and interpretation of neuroimaging data [29]. The intuition behind attention mechanism can be simply regarded as a weighted operation to adaptively highlight important local features in pattern recognition. In this field, attention mechanism is mostly used to emphasize the contribution of various brain regions to a specific downstream task. Therefore, discriminative localization with the attention mechanism is proficient in quantifying the importance of each ROI to the task performance. For instance, Lian et al. [30] applied attention maps for identifying discriminative brain regions from MRI based on an end-to-end CNN model. Huang et al. [31] presented an attention-guided feature learning framework to automatically perform segmentation in MRI. The incorporation of attention mechanism allows their frameworks to generate multi-level attention feature maps to track brain changes during scanning for etiopathogenesis analysis.

Recently, attention models such as Transformers [10] have demonstrated remarkable capability in effectively modeling global temporal dependence. Bedel et al. [32] employed the self-attention mechanism to capture global temporal dependencies in fMRI analysis. Deng et al. [33] designed a spatio-temporal Transformer that models the spatial and temporal dependencies of fMRI data, enabling more accurate diagnoses of mental disorders of ASD and ADHD. Liu et al. [34] developed a co-attention learning framework based on Transformer for mutually modeling the spatial and temporal dependence of fMRI data, with the guidance of spatio-temporal attention matching. Additionally, attention mechanisms have been incorporated into graph neural networks to enhance graph-structured representation learning in neuroimaging data [35], [36]. Chen et al. [37] proposed a novel GNN architecture with attention mechanism to identify critical brain regions and connections contributing to the classification of ASD from structural and functional MRIs. Despite some success, these methods face challenges in striking a balance between complexity and efficiency when capturing both local and global FC patterns of brain activity [21]. In this study, we design a novel approach that combines the sliding window technique with Transformer, enabling effective capture of both local and global patterns throughout the entire fMRI course.

III. Methodology

A. Problem Definition and Preliminaries

To account for the dynamic nature of fMRI data, we formulate the fMRI's ROI-based sequences input (shown in Fig. 1) $X \in \mathbb{R}^{N\times T}$ as a dynamic graph network $G(X) = \mathcal{G}(x_1), \ldots, \mathcal{G}(x_T)$, where T is the entire timestamps and N is the number of ROIs. In this representation, $\mathcal{G}(x_t)$ denotes the graph representation of the input x at the t-th timestamp. The dynamic graph at the t-th timestamp $\mathcal{G}(x_t)$ contains a set of vertices $V(t)$ (N involved ROIs) as well as an adjacency matrix A corresponding to the connections between $V(t)$, i.e., $\mathcal{G}(t) = \{V(t), A(t)\}$. Therefore, our task is to learn a mapping function f that computes the correspondences between the dynamic spatio-temporal graph representation of X and the ground-truth label Y. The optimization problem can be reached as follows:

$$f^* = \arg\min_f \mathbb{E}_{X,Y}[\mathcal{L}(f(G(X)), Y)] \tag{1}$$

where $\mathbb{E}$ is the expectation of the loss function $\mathcal{L}$ over space of (X, Y).

Notably, since the subject samples have different lengths of scanning sessions, we need to fix the same sequence length to construct the dynamic graphs. We propose to apply stepwise sliding window algorithm to sampling T time-series sequences into D scanning sessions with a length of S each ($D = \lfloor T/S \rfloor$), as inputs $V(d) = \{x_d\}_{d=1}^{D} \in \mathbb{R}^{N\times S}$. Based on v_d, the FC measurement between ROIs is calculated by Pearson correlation coefficient as follows:

$$FC_d(i,j) = \frac{\sum_{k=1}^{S}(v_d(i,k) - \overline{v_d(i)})(v_d(j,k) - \overline{v_d(j)})}{\sqrt{\sum_{k=1}^{S}(v_d(i,k) - \overline{v_d(i)})^2 (v_d(j,k) - \overline{v_d(j)})^2}} \tag{2}$$

where i and j represent the ith and jth ROIs, respectively. To detect the remarkable FC patterns and reduce redundancy, the corresponding adjacency matrix $A(d)$ is constructed by including the remarkable FC exhibiting a Pearson correlation coefficient greater than the established threshold $\tau = 0.1$. Therefore, the traditional dynamic graph in this work is defined as $G(X) = \mathcal{G}(x_1), \ldots, \mathcal{G}(x_D)$, $\mathcal{G}(x_d) = \{V(d), A(d)\}$, where $X \in \mathbb{R}^{N\times T}$ and $x_d \in \mathbb{R}^{N\times S}$.

B. Model Design

In this section, we illustrate the architecture of ST-HAG as shown in Fig. 1. First of all, two novel modules DAN-GCN and TSSA are performed for capturing discriminative spatial and temporal representation of fMRI data, respectively. The DAN-GCN module is designed to incorporate the interplay between spatial and temporal dynamics by modeling spatial representations between brain regions across adjacent timesteps. To compensate the absence of dynamic temporal patterns, the TSSA module is proposed by leveraging self-attention mechanisms in conjunction with sliding window algorithm. By employing STGF blocks, DAN-GCN and TSSA are effectively fused together to generate a comprehensive approach that captures spatial and temporal representations, as well as the intricate spatio-temporal dependencies inherent in fMRI data. Finally,

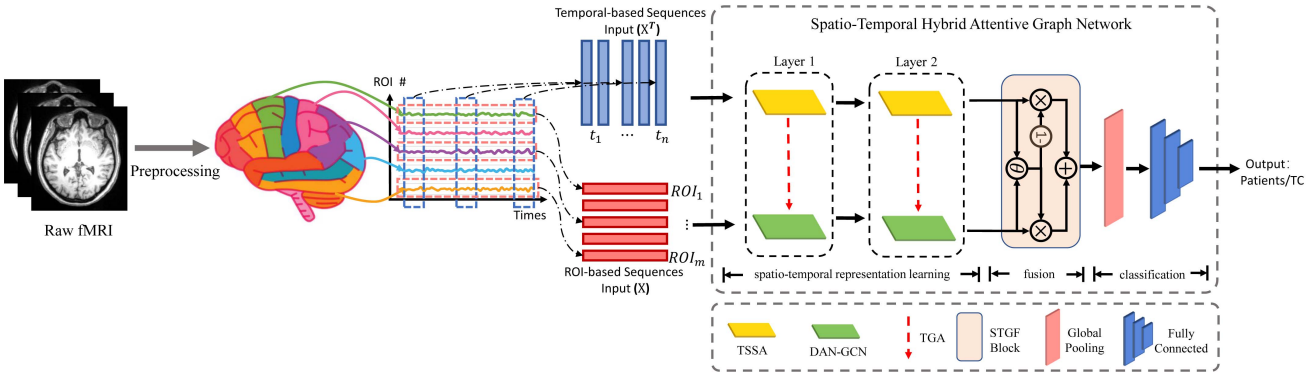


Fig. 1. This figure illustrates the framework of the proposed ST-HAG. Firstly, the raw neuroimaging data is denoised to obtain temporal signal input and dynamic graph input. Next, the preprocessed temporal signals are fed into a stack of TSSA layers to extract full-scale temporal dependency and temporal attention guidance. Subsequently, a stack of DAN-GCNs with temporal attention guidance is utilized to capture local spatio-temporal representations from the dynamic graph x. Finally, the learned graph representations from the two perspectives are fused via a gated fusion module.

global pooling and multilayer perceptron are constructed for further mental disorders diagnosis.

1) Dynamic Adaptative-Neighbor GCN: To overcome the limitations of current STGCNs in fMRI analysis, the primary goal of DAN-GCN is to capture discriminative spatial representation, as well as spatio-temporal inter-dependencies across brain regions and scanning sessions. As we can see in Fig. 2, a novel dynamic graph construction method ($\mathcal{G}_{DAN} \in \{V_{DAN}, A_{DAN}\}$) involves establishing connections between nodes within adjacent time sessions is proposed. The scale of $\mathcal{G}_{DAN}$ is determined by the number of adjacent scanning sessions, e.g., we set it to 1 as shown in Fig. 2, resulting in $V_{DAN} = [V(d-1), V(d), V(d+1)] \in \mathbb{R}^{3N \times S}$ (as shown in the middle of Fig. 2 and A_{DAN} with the size of $3\,N \times 3\,N$ (as shown in the bottom of Fig. 2. Inspired by the recent advance in [38], we leverage A_{DAN} to record three kinds of adjacency matrix $\in N \times N$, including the spatial graph connectivity (A_{SG}), temporal graph connectivity (A_{TG}), and padding matrix (A_P). Specifically, A_{SG} is positioned on the diagonal of A_{DAN}, to depict the dynamic graph for each adjacent scanning session. The representation is based on the remarkable FCs with a Pearson correlation coefficient greater than the threshold τ=0.1. The A_{TG} is positioned next to the A_{SG} to capture the changes along the time axis by linking each node to itself. The A_P is used to fill the remaining spaces for padding purposes by setting all its values to 0. In this way, the ROI-based sequences input X can be constructed as $\mathcal{G}_{DAN}$ to extract the local spatio-temporal correlations of dynamic FC matrices by the following dynamic adaptative-neighbor graph convolution module (DAN-GCM).

By expanding such a dynamic FC correlation estimation to the whole time course, we utilize DAN-GCM to capture hybrid spatio-temporal inter-relationships among the whole dynamic brain network as shown in Fig. 3. Based on the constructed $\mathcal{G}_{DAN}$, graph convolution is performed to aggregate the hybrid graph representation of the central node with its neighbors within adjacent scanning sessions. Given the l-th input graph representation as $h_{in}^{(l)} \in \mathbb{R}^{3\,N \times S}$, graph convolution can be formulated as follow:

$$h^{(l)'} = \sigma(A_{DAN}(h_{in}^{(l)} W + b)) \tag{3}$$

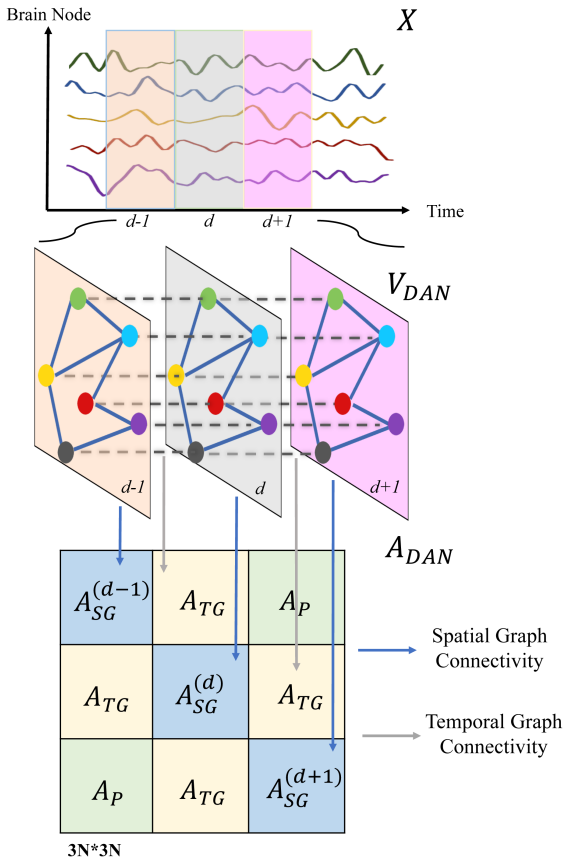


Fig. 2. Figure illustrates the proposed dynamic adaptive-neighbor graph construction. At the top of the figure, the input time-series of fMRI data is shown. In the middle of the figure, an example of the input of a dynamic adaptive-neighbor graph is depicted, which would be generated along the time axis. At the bottom of the figure, the corresponding adjacency matrix of the dynamic adaptive-neighbor graph is shown.

where the variables $W \in \mathbb{R}^{S \times S}$ and $b \in \mathbb{R}^{S}$ are learnable parameters and σ represents the activation function of rectified linear unit ($ReLU$) in this work. As we can see in Fig. 3(a), after graph convolution, we perform the max pooling and then reserve only the graph representation matrix at the current timestamp, resulting in the output graph representation of DAN-GCM $h^{(l)} \in$

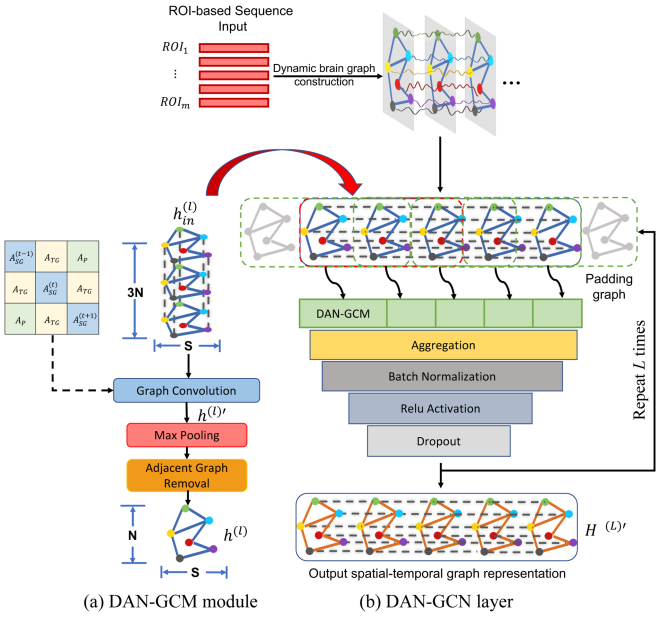


Fig. 3. (a) Workflow process of DAN-GCM at t-th timestamp. (b) The layered architecture of DAN-GCN.

$\mathbb{R}^{N\times S}$. As such, each scanning session of input spatio-temporal series is parallelly processed by DAN-GCM to construct a DAN-GCN layer to generate the local graph representations $\{h_i^{(l)}\}_{i=1}^{T}$ as shown in Fig. 3(b). Note that we need to perform zero-padding to assist the kernel of DAN-GCM in processing the first and last scanning sessions. Then, an aggregation layer is employed to concatenate $h^{(l)}$ as follows,

$$H^{(l)} = \left\langle h_1^{(l)}, \ldots, h_D^{(l)} \right\rangle \in \mathbb{R}^{D\times N\times S} \tag{4}$$

where $\langle . \rangle$ denotes the concatenation operator. In the tail of DAN-GCN, the layers of batch normalization, $ReLU$ activation, and dropout are cascaded to output the learned spatial graph representation $H^{(l)'}$.

2) Temporal Sliding Self-Attention: The previous STGNNs faced a challenge in accurately modeling the time-varying dynamic of fMRI data due to the limited receptive field caused by the kernel size. This limitation hindered the ability to capture complex temporal dependency patterns in fMRI, which is essential for accurate mental disorders diagnosis [39]. To this end, as shown in Fig. 4(a), we develop TSSA layer to extract discriminative temporal representation based on the Transformer encoder [10] as the backbone upon the temporal-based sequences input $X^{\mathsf{T}} \in \mathbb{R}^{S\times N}$ as inputs.

The vanilla Transformer architecture effectively captures the global self-attention by calculating the dependency of local temporal tokens/patterns on the others. However, due to the global attention computation strategy, the standard self-attention mechanism in the Transformer encoder tends to focus on similar temporal patterns with equal peak values while ignoring those similar short to medium-term temporal patterns (closely happen) with different peak values, which is essential for fMRI time-series pattern analysis [40]. Therefore, we design a SWA

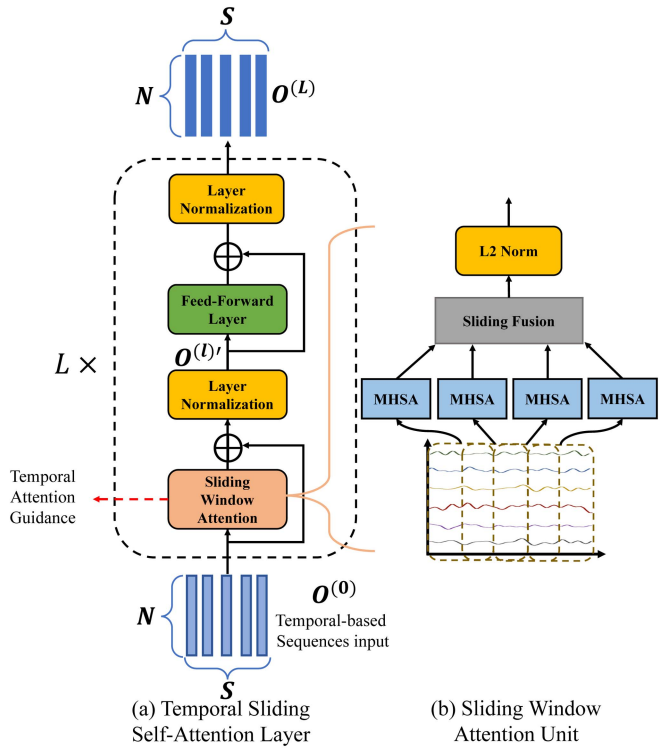


Fig. 4. (a) Architecture of full-scale temporal self-attention module. (b) An example of sliding window attention unit based on multi-head attention mechanism. (MHSA: multi-headed self-attention).

unit to reconstruct the multi-head self-attention module (MHSA) in the Transformer encoder as shown in Fig. 4(b).

The SWA unit utilizes MHSA to deal with the multiple-scale sliding windows to capture the short-term temporal feature patterns. First, the input sequence is processed by the position embedding layer used in [41] to obtain the hidden feature representation $O \in \mathbb{R}^{T\times N}$ that encodes position information, which can be formulated as follows,

$$O = PositionEmbedding(X^{\mathsf{T}}) = X^{\mathsf{T}} + PE \tag{5}$$

where the position embedding PE is represented as learnable table of vectors, which assigns a unique vector representation to each ROI signals based on its position index of the brain. Then, we split the embeded representation $O^{(l)}$ at the l-th layer into $k = \frac{T}{\lceil w-d \rceil} - 1$ sequence segments $\{o_i^{(l)}\}_{i=1}^{k} \in \mathbb{R}^{T\times w}$. The notations w and d denote the window size and step size, respectively. Then, the MHSA blocks are applied to $\{o_i^{(l)}\}$ in parallel, and their outcomes are aggregated through weighted summation by the sliding fusion block (SF). Finally, we employ the L2-normalization to regulate the contributions of different sequence segments. Mathematically, the learning process of SWA can be formulated as follows,

$$SWA(O^{(l)}) = \frac{SF\left\langle \alpha^1 MHSA(o_1^{(l)}), \ldots, \alpha^k MHSA(o_k^{(l)}) \right\rangle}{||SF\left\langle \alpha^1 MHSA(o_1^{(l)}), \ldots, \alpha^k MHSA(o_k^{(l)}) \right\rangle||_2} \tag{6}$$

where $\alpha^k \in \mathbb{R}^k$ is a parameter vector of SF to be learned.

To finetune the outputs of SWA, the residual connections are used to allow gradients to flow through the whole TSSA layer along with layer normalization (LN). After the first layer normalization, a pointwise feed-forward layer (FFL) is used for further feature extraction. Overall, the TSSA module at the l-th layer can be summarized as follows:

$$O^{(l)'} = LN(SWA(O^{(l)}) + O^{(l)}) \tag{7}$$

$$O^{(l+1)} = LN(FFL(O^{(l)'}) + O^{(l)'}). \tag{8}$$

Besides $O^{(l+1)}$ feeding to the next TSSA layer, the attention scores learned from SWA can also be utilized as temporal attention guidance (TAG) to strengthen the learning capability of DAN-GCN with full-scale temporal dependency. As demonstrated by [34], the application of a sliding window technique based on self-attention mechanism allows the model to focus on more detailed dependencies within local regions of fMRI data. In this context, a smaller window size w enables the model to concentrate on more specific local patterns, while a larger window size allows the model to capture more general global patterns. To preserve the global temporal representation capacity of our method, we adopt a progressive increase in the window size within the SWA module as the number of TSSA layers grows. This incremental adjustment continues until the window size reaches the value of T. By employing this approach, our TSSA module can effectively capture both subtle local temporal patterns and broader global temporal patterns, facilitating the subsequent fusion with the graph representation learned from DAN-GCN.

3) Temporal Attention Guidance: Although DAN-GCN is effective in extracting discriminative spatial representations with spatio-temporal inter-dependency, it may not fully capture specific salient temporal patterns, especially at the global level. To overcome this limitation, we propose TAG module, which leverages the temporal attention scores learned from TSSA to guide the graph representation learning in DAN-GCN. The TAG module benefits from the MHSA mechanism in the Transformer encoder, i.e., queries are compared against key-value pairs to produce attention outputs. The learned attention score at l-th layer can be computed as follows,

$$Q = O^{(l)}W^q; K = O^{(l)}W^k \tag{9}$$

$$Score^{(l)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{10}$$

where $K \in \mathbb{R}^{T\times d}$ and $Q \in \mathbb{R}^{T\times d}$ are the embedding matrices of key and query with dimension d, respectively, and that $\{W^k, W^q\} \in \mathbb{R}^{N\times m}$ are their corresponding learnable weight matrices. Since the proposed SWA processes k sequence segments with k individual MHSA blocks based on the sliding window algorithm, we further compute the mean for the scanning sessions which are overlapped with multiple sliding windows. Finally, the learned temporal attention score can be incorporated into $H^{(l)'}$ for temporal attention guidance as follows,

$$H^{(l+1)'} = H^{(l)'} \odot Score^{(l)} + H^{(l)'} \tag{11}$$

Algorithm 1: Pseudo-Code of the ST-HAG.

Require: Preprocessed data X, and label Y
Ensure: Predicted probabilities of testing set Y^{te}_{pred}

$[X^{tr}, X^{te}, Y^{tr}, Y^{te}] \leftarrow$ Split(X, Y)
Initialize ST-HAG.
for $e = 1, \ldots, epochs$ **do**
  // Full-Scale TSSA
  $O^{(0)} =$ PositionEmbedding $({X^{tr}}^{\mathsf{T}})$
  for $l = 0, \ldots, L-1$ **do** // *l:* # of attention layers
    $(\tilde{O^{(l)}}, Score^{(l)}) =$ SWA$(O^{(l)})$
    ${O^{(l)}}\prime =$ LN$(\tilde{O^{(l)}} + O^{(l)})$
    $O^{(l+1)} =$ LN(FFL$({O^{(l)}}\prime) + {O^{(l)}}\prime$)
  end for
  // Dynamic Adaptative-Neighbor GCN
  $(H^{(0)}, A_{DAN}) = G_{DAN}(X^{tr})$
  for $l = 0, \ldots, L-1$ **do** // *l:* # of DAN-GCN layers
    $H^{(l)\prime} = DAN - GCN(H^{(l)}, A_{DAN})$
    $H^{(l+1)} = H^{(l)\prime} \odot Score^{(l)} + H^{(l)\prime}$
  end for
  $Z = (1-\theta){O^{(L)}}^{\mathsf{T}} \oplus \theta {H^{(L)}}^{\mathsf{T}}$ // Fused representation
  // Tower networks for specific tasks
  $output \leftarrow$ Tower(Z)
  $Loss \leftarrow \mathcal{L}(output, Y^{tr})$
  ST-HAG.update$(Loss)$
end for
$Y^{te}_{pred} \leftarrow ST - HAG.predict(X^{te})$

where $\odot$ is element-wise multiplication. By incorporating the learned temporal attention scores, the TAG module enhances the ability of DAN-GCN to capture and emphasize important temporal patterns.

4) Spatio-Temporal Gated Fusion: To improve the modeling of spatio-temporal dependencies in fMRI data, we fused the final representations of DAN-GCN and TSSA, denoted as $H^{(L)'}$ and $O^{(L)}$ respectively, as shown in Fig. 1. Specifically, a simple yet efficient gated fusion block is leveraged to fuse $H^{(L)'}$ and $O^{(lL}$ by adaptively tuning their weights as follows.

$$Z = (1-\theta){O^{(L)}}^T + \theta H^{(L)'} \tag{12}$$

where the learnable parameter $\theta \in [0, 1]$ controls the fusion weights. At last, a tower network consisting of global pooling and multilayer perceptron is trained to accept Z for classification. To ease the understanding of our whole learning framework, the pseudo-code is shown in Algorithm 1.

IV. Experiments and Results

A. Data Acquisition and Processing

We evaluate the performance of ST-HAG on the diagnosis of two mental disorders: 1) ASD diagnosis on the preprocessed Autism Brain Imaging Data Exchange I (ABIDE I) dataset[1]

[1]http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html

TABLE I
MODEL PARAMETER SETTINGS OF ST-HAG

# of ST-HAG layer	2	# of head in TSSA	10
Window size of SWA	30/60	Max position embeddings	512
Dropout rate	0.1	Learning rate	10^{-5}
Batch size	128	Epochs	30
Configuration of fully connected networks		64-Dropout(0.5)-10-1 (Three layers)	
Activation function to attention layers		Gaussian error linear units (gelu)	
Activation function to output layers		Sigmoid	

and 2) ADHD diagnosis on the preprocessed ADHD-200 Consortium (ADHD-200) dataset.[2] Specifically, ABIDE I contains 1035 valid fMRI samples with 505 ASD subjects and 530 typical controls (TCs) aggregated from 17 different brain imaging sites. ADHD-200 includes 947 valid fMRI samples from 8 international imaging sites involving 362 children and adolescents with ADHD and 585 TCs. For a fair comparison, we downloaded the preprocessed datasets publicly released by ABIDE I and ADHD-200. In addition to the fMRI data, the phenotypic information such as age, gender, handedness, and IQ, is also provided for each subject in these datasets.

B. Experimental Settings

For performance evaluation, 10-fold cross-validation (CV) is conducted on ABIDE I and ADHD-200. Since ABIDE I and ADHD-200 are multi-site datasets with different lengths of time courses, we need to maintain the same sequence length for each subject sample so as to normalize the inputs for model training. Random cropping is commonly used to address this issue, i.e., each sample is randomly cropped into a certain number of sequences with a fixed length. In this work, we crop 10 sequences with a fixed length of 90 for each sample data. For a fair test, the cropped sequences of the same sample cannot be overlapped in both training set and testing set. By adopting the CC200 atlas, the mean time series of each cropped sample can be represented by a matrix of 90×200 where 200 is the total number of ROIs. The model parameter settings of ST-HAG are shown in Table I. All the experiments are conducted under the same runtime environment using one intel core i7-8700K@3.70 GHz, one NVIDIA GeForce RTX 2080 Ti GPU, and 64 GB RAM. The experiment results are compared in terms of accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC).

C. Comparison With State-of-the-art Methods

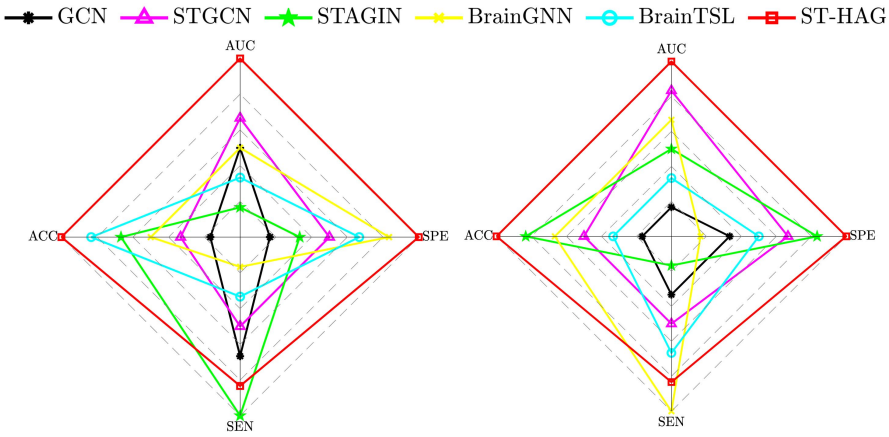
In this section, the proposed ST-HAG is pitted against the SOTA methods on ADHD-200 and ABIDE I, as shown in Table II. It is shown that ST-HAG achieves competitive accuracies of $71.9\%(\pm 3.4\%)$ (SEN: 72.1%, SPE: 68.8%, AUC: 78.2%) and $74.8\%(\pm 3.2\%)$ (SEN: 78.2%, SPE: 68.5%, AUC: 81.2%) on ABIDE I and ADHD-200, respectively. In addition to CV, independent sets of training/testing (IS) were also used for performance validation on ADHD-200 because ADHD-200, as a global competition dataset is officially divided into a training set and a testing set. As we can see, GNN-based methods [37], [42], [45], [48], which only take spatial correlations into consideration, achieved comparable performance to other popular deep learning models like CNN [43] and auto-encoders [44]. The other methods such as STCAL [34], dynamic GCN (dGCN) [26], 4D-CNN [16], and our ST-HAG effectively take advantage of both spatial and temporal information, resulting in superior classification performance compared to those methods that only focus on spatial dependencies. This observation demonstrates the significance of exploring spatio-temporal dependencies in the context of fMRI-based mental disorders diagnosis tasks. Given the cross-site nature of ABIDE I and ADHD-200 datasets, which were collected from different medical institutions using various scanning devices and protocols, training a CAD model using the entire dataset can be more challenging compared to using a subset filtered by specific criteria. Consequently, certain diagnosis models [26], [45], [47], [50] may struggle to improve accuracy when the sample size exceeds a certain threshold (i.e., >800) due to the substantial data heterogeneity. In contrast, our ST-HAG exhibits a strong generalization capability on large cross-site datasets by leveraging the extracted subject-independent extracted hybrid spatio-temporal graph representation.

TABLE II
PERFORMANCE COMPARISON ON ABIDE I AND ADHD-200

Model	Classifier	Validation	Sample #	Accuracy(STD)
For ABIDE I (ASD)				
Ours	ST-HAG	10-fold CV	1035	**71.9(3.4)**
Ktena 2018 [42]	GCN	10-fold CV	871	69.5(N.A.)
Sherkatghanad 2020 [43]	CNN	10-fold CV	871	70.2(4.6)
Wang 2021 [24]	cGCN	10-fold CV	1057	70.7(N.A)
Almuqhim 2021 [44]	SAE	10-fold CV	1035	70.8(N.A.)
Yang 2022 [45]	GIN	10-fold CV	203	70.6(4.9)
Chen 2022 [37]	GAN	10-fold CV	1007	70.8(2.9)
Wen 2023 [46]	BrainGSL	10-fold CV	871	71.3(N.A.)
For ADHD-200 (ADHD)				
Ours	ST-HAG	10-fold CV	939	**74.8(3.2)**
Aradhya 2019 [47]	DTM	10-fold CV	465	70.4(2.0)
Mao 2019 [16]	4D-CNN	IS	788	71.3(N.A)
Yao 2021 [48]	MMTGCN	5-fold CV	627	71.8(1.5)
Ji 2022 [49]	FC-HAT	IS	520	69.2(N.A.)
Zhao 2022 [26]	dGCN	10-fold CV	635	72.0(1.8)
Liu 2023 [34]	STCAL	10-fold CV	939	72.5(4.2)
Li 2023 [50]	BrainSync	10-fold CV	210	73.3(N.A)

The effectiveness of ST-HAG is compared to other STGNN architectures, as no STGNN-based methods have been proposed for ABIDE I or ADHD-200 datasets. We compared several representative STGNN architectures, including STGCN [17], STAGIN [25], BrainGNN [52], and BrainGSL [46], the comparison results are summarized Fig. 5. It is shown that ST-HAG consistently outperforms its competitors by dominating 75% (6/8) of the evaluation metrics, with five of them showing statistical significance ($p \leq 0.05$) according to the Wilcoxon rank-sum test. Compared to the best STGNN framework, STAGIN [25], our proposed ST-HAG achieves a notable improvement in accuracy of 2.5% and 1.6% on the ABIDE I and ADHD-200 datasets, respectively. It is worth noting that the BrianGSL stands out as the only work that utilizes the graph self-supervised learning

[2]http://fcon_1000.projects.nitrc.org/indi/adhd200/



Method	ABIDE I				ADHD-200			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
GCN	67.6(±4.5)	71.9	61.9	74.3	71.0(±4.2)	76.1	63.2	76.7
STGCN	68.2 (±4.7)	71.5	64.2	72.8	72.6(±4.4)	77.5	65.1	78.6
STAGIN	69.5 (±4.1)	**74.6**	64.1	75.6	73.2(±3.2)	75.0	66.0	77.2
BrainGNN	68.7 (±3.9)	69.8	65.8	73.4	72.9(±3.9)	**78.9**	61.3	77.8
BrainGSL	69.7 (±3.7)	70.7	64.8	74.7	72.1(±3.5)	77.6	64.2	76.8
ST-HAG(ours)	**71.9**†(±3.4)	72.1	**68.8**†	**78.2**†	**74.8**(±3.2)	78.2	**68.5**†	**81.2**†

Fig. 5. Performance comparison with the SOTA methods on ABIDE I and ADHD-200. Based on the Wilcoxon rank-sum test with Holm p-value correction ($\alpha = 0.05$), the † indicates the marked method is significantly better than the compared methods.

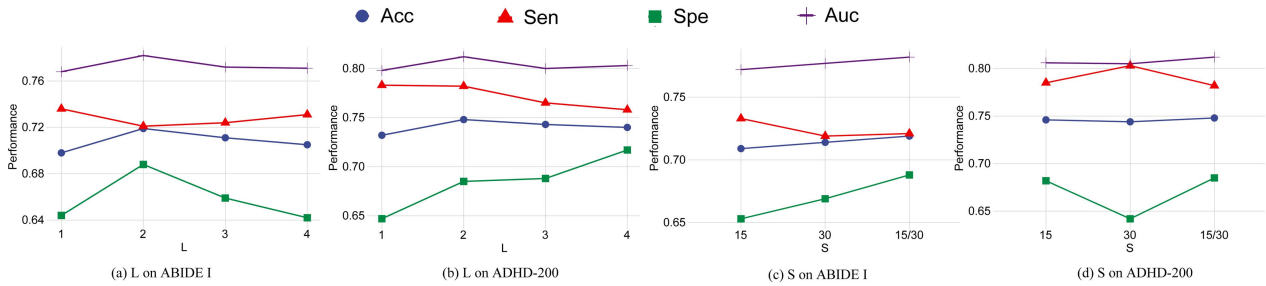


Fig. 6. Performance comparison with the different hyperparameters on ABIDE I and ADHD-200.

paradigm to capture topological patterns in fMRI, achieving consistent and comparable performance on both the ABIDE and ADHD-200 datasets. In addition, there is a significant performance gap between specificity and sensitivity of the compared methods, mostly exceeding 12%. This indicates that these universal STGNN architectures could have a higher risk of misdiagnosis in real-world applications. By comparison, the proposed ST-HAG can manage a decent trade-off between specificity and sensitivity, resulting in the highest AUC values on both datasets. This makes it more suitable for practical assisting diagnosis processes.

D. Parameter Sensitivity Analysis

In this section, we shall show the selection of hyperparameters through comprehensive experiments on ABIDE I and ADHD-200 datasets. The results shown in Table I are evaluated on 10-fold CV and discussed in detail below.

1) Effect of Model Layers L: From the results in Fig. 6(a) and (b), we can see that with increasing L, the performances on both ABIDE I and ADHD-200 datasets first increase, then decreases slightly, and saturates at $L = 2$. Surprisingly, our ST-HAG model with only one layer exhibits comparable performance to the SOTA fMRI-based STGCNs framework on both datasets, demonstrating the excellent spatio-temporal representation ability of our proposed model. The two-layer model has a slightly better performance compared to the three-layer model while also having a smaller model size and faster training speed. The performance saturation at $L = 2$ can be explained by the overfitting problem resulting from a relatively small fMRI training dataset when L increased. This observation is consistent with the study in [17] where the layer of STGCN is set as two.

2) Effect of Length of Scanning Sessions S: As the number of DAN-GCM modules used in the DAN-GCN layer is determined by the length of scanning sessions S, we also assessed the effect of S on our model's performance on two-layer setting. In this

TABLE III
EFFECT OF THRESHOLD τ IN ADJACENT MATRIX CONSTRUCTION

Settings	ABIDE I				ADHD-200			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
τ=0	70.4	75.4	64.1	77.4	73.7	78.9	65.7	80.7
τ=0.1	**71.9**	72.1	**68.8**	**78.2**	**74.8**	78.2	**68.5**	**81.2**
τ=0.2	71.0	72.6	65.5	77.8	74.3	78.3	67.8	80.9
τ=0.3	70.1	74.8	63.9	77.2	73.8	78.5	66.4	81.0
τ=0.5	70.3	**76.0**	64.1	76.7	73.8	**79.2**	65.1	81.1

Note: $\tau = 0$ indicates that all FC values were utilized for the construction of the adjacency matrix.

experiment, we evaluate three kinds of length settings including: $S = 15$ time interval (TR), $S = 30$ TR, and $S = 15/30$ TR, where $S = 15/30$ implies we set the value of S differently in the first and second layers as 15 TR and 30 TR, respectively. As depicted in Fig. 6(c) and (d), the performance distinction between these settings is not significant. This confirms that our model is insensitive to the length of the scanning session and can effectively capture both local and global spatio-temporal graph representations across different length settings. Moreover, the performance of $S = 15/30$ slightly outperforms other settings across six evaluation matrics on ABIDE I and ADHD-200 datasets. This can be attributed to the fact that this hierarchical setting (from small to large) is consistent with the way GNNs aggregate information through massage passing algorithm (from near to far neighbors) [53], thus facilitating the extraction of local to global dynamic representations [54].

3) Effect of Threshold τ in Adjacent Matrix Construction: Given the importance of constructing the adjacency matrix in graph modeling, we conducted a series of experiments to assess its impact within our DAN-GCN module. Specifically, we varied the threshold τ used in the adjacency matrix construction and evaluated five different settings: $\tau = 0$, 0.1, 0.2, 0.3, and 0.5. When $\tau = 0$, it indicates that all FC values were utilized for the construction of the adjacency matrix. Analyzing the results shown in Table III, we observed that as the τ increased, the performance of our model initially improved and then plateaued with a slight decrease. Notably, on the ADHD-200 dataset, the performance difference between different τ settings was relatively small. This indicates that our proposed model is robust enough to handle both scenarios of information redundancy ($\tau = 0$) and information scarcity ($\tau = 0.5$). Furthermore, we found that the best performance on both the ABIDE I and ADHD-200 datasets is achieved when $\tau = 0.1$. This suggests that incorporating a moderate threshold for selecting FC values in the adjacency matrix construction can effectively enhance the performance of our model.

E. Ablation Study

In this section, we conducted an ablation study to assess the effectiveness of our proposed major components, including DAN-GCN, TSSA and TAG. To establish a benchmark performance using the standard STGCN, we designed a baseline model that utilized vanilla GCN and a Transformer encoder. Subsequently, we conducted a series of ablation experiments where we individually replaced each component of the baseline model with its corresponding proposed component. For instance, to assess the effectiveness of DAN-GCN, we replaced the vanilla GCN in the baseline model with DAN-GCN to evaluate its performance (denoted as vanilla GCN $\rightarrow$ DAN-GCN).

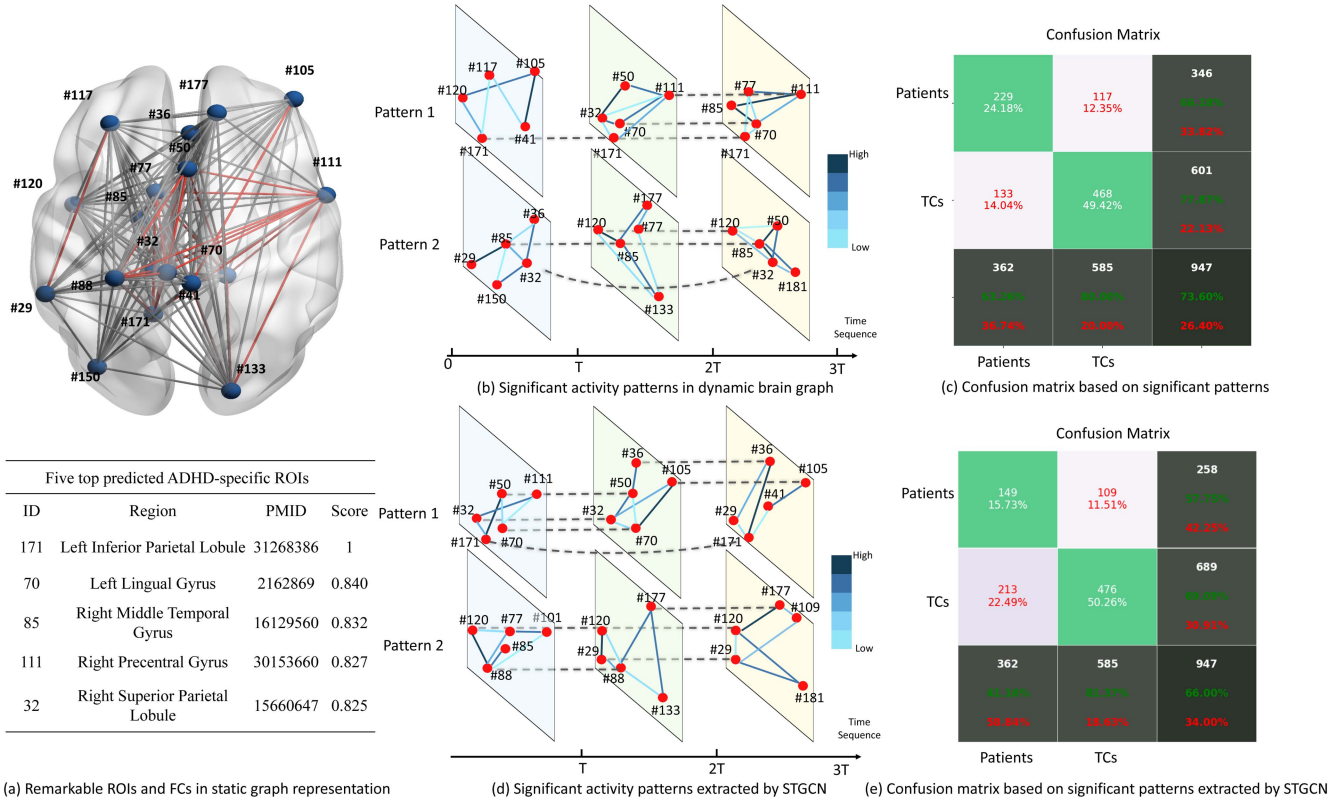
Accordingly, the experiment results are tabulated in Table IV. As observed through the comparison with the baseline model, the proposed ST-HAG outperforms it in 75% (6/8) evaluation matrics with statistical significance ($p \leq 0.05$). To our surprise, the performance of the baseline model is already comparable to, or even superior to, that of some SOTA methods shown in Table II, particularly for the ADHD-200 dataset. This fully demonstrates the effectiveness of STGCN as well as the importance of spatial and temporal feature extraction in fMRI time-series data. By replacing vanilla GCN with DAN-GCN, we can observe a significant accuracy improvement by 2.0% and 1.6% on ABIDE I and ADHD-200, respectively. That is, DAN-GCN can serve as a potent backbone model to extract the informative, dynamic FC correlation within adjacent scanning sessions. Thanks to the extracted full-scale temporal dependency TSSA also makes clear performance contributions to the baseline model. To further enforce the temporal attention in the learning process of hybrid graph representation, we introduce the TAG into the TSSA module, resulting in a more significant performance improvement. To further validate the effectiveness of the proposed TAG module, we conducted additional experiments using GCN and TSSA without the TAG module (denoted as ST-HAG w/o TGA). As we can see, even without the TAG module, our model consistently achieved a performance that was comparable to our ST-HAG model, particularly on the ADHD-200 dataset. This observation further emphasizes the effectiveness of the proposed TAG module in enhancing the overall performance of our model. In the end, ST-HAG obtains the supreme performance as expected by integrating these components into the baseline model.

Furthermore, we conducted additional experiments to evaluate the effectiveness of incorporating phenotypic information and utilizing overlapping sliding windows. The results unequivocally highlight the effectiveness of our ST-HAG approach, even in scenarios where phenotypic information was not utilized (referred to as ST-HAG w/o PI) and overlapping windows were not employed (referred to as ST-HAG w/o overlap). Remarkably, there is only a marginal decrease in diagnostic accuracy for these variations. Specifically, there is a decrease of 0.7% for ASD and 1.0% for ADHD in ST-HAG w/o PI, and a decrease of 0.5% for ASD and 0.6% for ADHD in ST-HAG w/o overlap. Importantly, when compared to other SOTA methods, both ST-HAG w/o PI and ST-HAG w/o overlap still demonstrate competitive classification performance on both the ABIDE and ADHD-200 datasets. These results further highlight the robustness of our proposed ST-HAG model, as it is not overly sensitive to the specific details of the sliding window technique or the inclusion of phenotypic information. This resilience can be attributed to the effectiveness of our dynamic spatio-temporal modeling approach, which enables successful adaptation to different configurations.

TABLE IV
EFFECTIVENESS EVALUATION OF CERTAIN COMPONENTS IN ST-HAG

Method	ABIDE I				ADHD-200			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
Baseline	68.5	72.4	63.4	75.9	72.7	78.7	60.1	78.7
vanilla GCN→DAN-GCN	70.5†	74.4†	64.9	76.5	74.3	**81.2**†	63.5†	80.6†
Transformer-encoder→TSSA	69.8	**76.6**†	62.7	76.4	73.8	77.7	66.0†	80.1
Transformer-encoder→TSSA w/ TAG	70.6†	73.5	66.7†	77.4	74.2	78.1	66.3†	80.3†
ST-HAG w/o TAG	70.9†	74.9†	65.2	77.2	74.0	78.1	65.8†	81.0†
ST-HAG w/o PI	71.2†	71.9	68.0†	77.5†	73.5	79.3	66.6†	80.5†
ST-HAG w/o overlap	71.4†	73.4	66.8†	77.8†	74.2	79.5	65.7†	80.6†
ST-HAG (ours)	**71.9**†	72.1	**68.8**†	**78.2**†	**74.8**†	78.2	**68.5**†	**81.2**†

Based on the Wilcoxon rank-sum test with Holm p-value correction ($\alpha = 0.05$), the † indicates the marked method is significantly better than the baseline. The model layers L is set to 2.



(a) Remarkable ROIs and FCs in static graph representation
(b) Significant activity patterns in dynamic brain graph
(c) Confusion matrix based on significant patterns
(d) Significant activity patterns extracted by STGCN
(e) Confusion matrix based on significant patterns extracted by STGCN

Fig. 7. Identification of remarkable ROIs and reproducible dynamic FC patterns specific to ADHD. The dash lines indicate the involved ROIs are repeatedly highlighted throughout the whole time sequence. The darker color of the link between ROIs we observe, the more highly correlated of the FC it represents. The corresponding confusion matrice is also identified. Here $T = 30$ timestamps.

F. Model Interpretability

Thanks to the foundational concept of message passing [53] in GCNs, the proposed model can facilitate the information exchange among neighboring nodes or edges for graph interpretation analysis, like previous neuroimaging analysis studies based on GNN [55], [56]. As a result, the fused graph representation Z (feature map), obtained after the fusion block for all subjects per class, serves as an importance score matrix to highlight remarkable ROIs and FC patterns. Specifically, based on obtained feature map, the remarkable ROIs are computed by aggregating their corresponding saliency values from the obtained feature maps, while the remarkable FC patterns are calculated using the Pearson correlation coefficient. In this section, we take ADHD as a case study. Fig. 7(a) shows a remarkable static FC graph inferred by ST-HAG, involving a number of ADHD-related ROIs. Amongst these ROIs, we select the five top ADHD-specific ROIs for manual validation, i.e., *Left Inferior Parietal Lobule* (171), *Left Lingual Gyrus* (70), *Right Middle Temporal Gyrus* (85), *Right Precentral Gyrus* (111), and *Right Superior Parietal Lobule* (32). As a result, all of these remarkable

ROIs have been demonstrated to have associations with the neurological manifestations of ADHD by previous literature. For example, [57] demonstrated that in comparison with the typical controls group, ADHD patients showed significantly higher temporal variability in the *Left Inferior Parietal Lobule* (171). This demonstrates that the data-driven outcomes of ST-HAG can be compatible with the previous clinical findings.

Furthermore, in light of these ADHD-related ROIs, we attempt to detect the reproducible dynamic FC patterns specific to ADHD only. That is, they have no significant change in BOLD signals in the TC group. We showcase the two eligible dynamic FC patterns in Fig. 7(b). It is efficient to investigate how these dynamic interaction brain networks evolve along the time axis, thus identifying the reproducible dynamic FC patterns as neuroimaging biomarkers. Through the statistical analysis, 63.3% of ADHD subjects (229/362) on ADHD-200 are found to exhibit similar dynamic FC patterns shown in Fig. 7(b). This result suggests that the dynamic FC patterns identified by ST-HAG could advance the understanding of the disease mechanism of brain dynamics. Our analysis of these patterns (as shown in Fig. 7(c)) reveals that 63.3% of ADHD subjects (229/362) and 20.0% of TC subjects (117/585) on ADHD-200 exhibit these patterns. We further conducted an analysis on the significant dynamic FC patterns extracted by STGCN [17](as shown in 7(d)). As we can see, it is evident that the significant patterns extracted by the STGCN differ from those extracted by our model. However, it is noteworthy that the ROIs deemed significant in our model largely overlap with those identified by STGCN. Besides, our analysis revealed that out of the total ADHD subjects in the ADHD-200 dataset, only 41.2% (149/362) exhibited these identified patterns. Similarly, among the TC subjects, 18.6% (109/585) demonstrated these specific patterns (as shown in 7(e)). These findings highlight the critical role of identifying disease-specific dynamic FC patterns for accurate and robust diagnosis of mental disorders, and emphasiz the sensitivity of these patterns in achieving reliable diagnostic results.

By leveraging the insights provided by these dynamic FC patterns, our framework offers a clearer understanding of the disease mechanisms underlying brain dynamics in ADHD and we achieved an overall classification accuracy of 74.8%. These findings hold great potential for advancing the field of neuroimaging research and improving diagnostic and therapeutic approaches for individuals with ADHD. Future work can further validate and refine these patterns, paving the way for more precise and personalized treatment strategies.

G. Limitation and Future Work

Although our proposed model demonstrates promising performance in diagnosing mental disorders using fMRI data, there are still some limitations that may hinder its overall performance and practical application. As discussed in Section III-B, our current approach uses an identity matrix A_{TG} to capture changes along the time axis, linking each node to itself. However, in reality, brain activity may exhibit spatio-temporal correlations between different regions over time [58]. Each region in the brain graph should be connected to some extent with all other regions in adjacent time sessions. Constructing such an adjacency matrix A_{TG} would be computationally expensive because it requires computing a separate adjacency matrix for each session in each sample. Hence, in this work, we establish connections for each node only between its preceding and subsequent sessions, reducing the computational cost but neglecting some dynamic information between nodes. Consequently, further research attention and investigation are warranted to explore efficient methods for computing the adjacency matrix of a spatio-temporal dynamic graph while fully preserving the dynamic information.

V. Conclusion

In this paper, we proposed a spatio-temporal hybrid attentive graph network ST-HAG to identify mental disorders from fMRI time-series data. As the major components, DAN-GCN extracts the short-term dynamic FC correlations while TSSA characterizes the full-scale temporal dependency. To the best of our knowledge, this is the first attempt to apply spatio-temporal hybrid graph convolution to time-series fMRI. The 10-fold CV demonstrates the effectiveness of ST-HAG and its components on two real-world datasets. The model interpretability of ST-HAG allows for the revelation of reproducible dynamic FC networks along the time-varying patterns. The current work is expected to provide insights into exploiting the dynamic spatio-temporal dependence in fMRI signals for accurate diagnosis of mental disorders.

References

[1] N. K. Logothetis, "The underpinnings of the BOLD functional magnetic resonance imaging signal," *J. Neurosci.*, vol. 23, no. 10, pp. 3963–3971, 2003.

[2] Y. Du, Z. Fu, and V. D. Calhoun, "Classification and prediction of brain disorders using functional connectivity: Promising but challenging," *Front. Neurosci.*, vol. 12, 2018, Art. no. 525.

[3] Z.-A. Huang, R. Liu, Z. Zhu, and K. C. Tan, "Multitask learning for joint diagnosis of multiple mental disorders in resting-state FMRI," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 02, 2022, doi: 10.1109/TNNLS.2022.3225179.

[4] H. Ju, T. Yin, J. Huang, W. Ding, and X. Yang, "Sparse mutual granularity-based feature selection and its application of schizophrenia patients," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 604–614, Feb. 2024, doi: 10.1109/TETCI.2023.3314548.

[5] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, "Identifying autism spectrum disorder from resting-state FMRI using deep belief network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2847–2861, Jul. 2021.

[6] Z.-A. Huang et al., "Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 4, pp. 1137–1149, Apr. 2023.

[7] M. G. Preti, T. A. Bolton, and D. Van De Ville, "The dynamic functional connectome: State-of-the-art and perspectives," *Neuroimage*, vol. 160, pp. 41–54, 2017.

[8] Y. Du et al., "Interaction among subsystems within default mode network diminished in schizophrenia patients: A dynamic connectivity approach," *Schizophrenia Res.*, vol. 170, no. 1, pp. 55–65, 2016.

[9] Y. Hu et al., "Source free semi-supervised transfer learning for diagnosis of mental disorders on FMRI scans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13778–13795, Nov. 2023, doi: 10.1109/TPAMI.2023.3298332.

[10] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[11] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 14, 2022, doi: 10.1109/TNNLS.2022.3219551.

[12] R. A. Poldrack, "Region of interest analysis for FMRI," *Social Cogn. Affect. Neurosci.*, vol. 2, no. 1, pp. 67–70, 2007.

[13] D. S. Bassett and O. Sporns, "Network neuroscience," *Nature Neurosci.*, vol. 20, no. 3, pp. 353–364, 2017.

[14] N. V. Hartvig, "A stochastic geometry model for functional magnetic resonance images," *Scand. J. Statist.*, vol. 29, no. 3, pp. 333–353, 2002.

[15] G. Derado, F. D. Bowman, and C. D. Kilts, "Modeling the spatial and temporal dependence in FMRI data," *Biometrics*, vol. 66, no. 3, pp. 949–957, 2010.

[16] Z. Mao et al., "Spatio-temporal deep learning method for ADHD FMRI classification," *Inf. Sci.*, vol. 499, pp. 1–11, 2019.

[17] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, "Spatio-temporal graph convolution for resting-state FMRI analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2020, pp. 528–538.

[18] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[19] H. Wang, Y. Cheng, C. P. Chen, and X. Wang, "Broad graph convolutional neural network and its application in hyperspectral image classification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 2, pp. 610–616, Apr. 2023.

[20] B.-H. Kim and J. C. Ye, "Understanding graph isomorphism network for RS-FMRI functional connectivity analysis," *Front. Neurosci.*, vol. 14, 2020, Art. no. 630.

[21] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 813–824.

[22] H. Li, X. Shi, X. Zhu, S. Wang, and Z. Zhang, "FSNet: Dual interpretable graph convolutional network for alzheimer's disease analysis," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 15–25, Feb. 2023.

[23] L. Xiao et al., "Multi-hypergraph learning-based brain functional connectivity analysis in FMRI data," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1746–1758, May 2020.

[24] L. Wang, K. Li, and X. P. Hu, "Graph convolutional network for FMRI analysis based on connectivity neighborhood," *Netw. Neurosci.*, vol. 5, no. 1, pp. 83–95, 2021.

[25] B.-H. Kim, J. C. Ye, and J.-J. Kim, "Learning dynamic graph representation of brain connectome with spatio-temporal attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 4314–4327.

[26] K. Zhao, B. Duka, H. Xie, D. J. Oathes, V. Calhoun, and Y. Zhang, "A dynamic graph convolutional neural network framework reveals new insights into connectome dysfunctions in ADHD," *Neuroimage*, vol. 246, 2022, Art. no. 118774.

[27] R. F. Betzel et al., "The modular organization of human anatomical brain networks: Accounting for the cost of wiring," *Netw. Neurosci.*, vol. 1, no. 1, pp. 42–68, 2017.

[28] S. L. Warren and A. A. Moustafa, "Functional magnetic resonance imaging, deep learning, and Alzheimer's disease: A systematic review," *J. Neuroimaging*, vol. 33, no. 1, pp. 5–18, 2023.

[29] E. Eldele et al., "ADAST: Attentive cross-domain eeg-based sleep staging framework with iterative self-training," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 210–221, Feb. 2023.

[30] C. Lian, M. Liu, L. Wang, and D. Shen, "Multi-task weakly-supervised attention network for dementia status estimation with structural MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 4056–4068, Aug. 2022.

[31] W. Huang et al., "Feature pyramid network with level-aware attention for meningioma segmentation," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1201–1210, Oct. 2022.

[32] H. A. Bedel, I. Sivgin, O. Dalmaz, S. U. Dar, and T. Çukur, "BOLT: Fused window transformers for FMRI time series analysis," *Med. Image Anal.*, vol. 88, 2023, Art. no. 102841.

[33] X. Deng, J. Zhang, R. Liu, and K. Liu, "Classifying ASD based on time-series FMRI using spatial–temporal transformer," *Comput. Biol. Med.*, vol. 151, 2022, Art. no. 106320.

[34] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state FMRI data," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 17, 2023, doi: 10.1109/TNNLS.2022.3154755.

[35] Y. Ma, D. Yan, C. Long, D. Rangaprakash, and G. Deshpande, "Predicting autism spectrum disorder from brain imaging data by graph convolutional network," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[36] W. Yin, L. Li, and F.-X. Wu, "A graph attention neural network for diagnosing ASD with FMRI data," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2021, pp. 1131–1136.

[37] Y. Chen et al., "Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 14, 2022, doi: 10.1109/TNNLS.2023.3243000.

[38] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 914–921.

[39] M. Gadgil, E. Peterson, J. Tregellas, S. Hepburn, and D. C. Rojas, "Differences in global and local level information processing in autism: An FMRI investigation," *Psychiatry Res.: Neuroimaging*, vol. 213, no. 2, pp. 115–121, 2013.

[40] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the FMRI signal," *Nature*, vol. 412, no. 6843, pp. 150–157, 2001.

[41] J.D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics-HLT*, 2019, pp. 4171–4186.

[42] S. I. Ktena et al., "Metric learning with spectral graph convolutions on brain connectivity networks," *NeuroImage*, vol. 169, pp. 431–442, 2018.

[43] Z. Sherkatghanad et al., "Automated detection of autism spectrum disorder using a convolutional neural network," *Front. Neurosci.*, vol. 13, 2020, Art. no. 1325.

[44] F. Almuqhim and F. Saeed, "ASD-SAENet: A sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using FMRI data," *Front. Comput. Neurosci.*, vol. 15, 2021, Art. no. 654315.

[45] S. Yang, D. Jin, J. Liu, and Y. He, "Identification of young high-functioning autism individuals based on functional connectome using graph isomorphism network: A pilot study," *Brain Sci.*, vol. 12, no. 7, 2022, Art. no. 883.

[46] G. Wen et al., "Graph self-supervised learning with application to brain networks analysis," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 8, pp. 4154–4165, Aug. 2023, doi: 10.1109/JBHI.2023.3274531.

[47] A. M. Aradhya, A. Joglekar, S. Suresh, and M. Pratama, "Deep transformation method for discriminant analysis of multi-channel resting state FMRI," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 2556–2563.

[48] D. Yao et al., "A mutual multi-scale triplet graph convolutional network for classification of brain disorders using functional or structural connectivity," *IEEE Trans. Med. Imag.*, vol. 40, no. 4, pp. 1279–1289, Apr. 2021.

[49] J. Ji, Y. Ren, and M. Lei, "FC–Hat: Hypergraph attention network for functional brain network classification," *Inf. Sci.*, vol. 608, pp. 1301–1316, 2022.

[50] J. Li, Y. Liu, J. L. Wisnowski, and R. M. Leahy, "Identification of overlapping and interacting networks reveals intrinsic spatiotemporal organization of the human brain," *NeuroImage*, vol. 270, 2023, Art. no. 119944.

[51] P. Khan, P. Ranjan, and S. Kumar, "Data heterogeneity mitigation in healthcare robotic systems leveraging the Nelder–Mead method," in *Artificial Intelligence for Future Generation Robotics*. New York, NY, USA: Elsevier, 2021, pp. 71–82.

[52] U. Mahmood, Z. Fu, V. Calhoun, and S. Plis, "Attend to connect: End-to-end brain functional connectivity estimation," in *Proc. Workshop Geometrical Topological Representation Learn.*, 2021, pp. 1–8.

[53] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1263–1272.

[54] K. Bose and S. Das, "Can graph neural networks go deeper without over-smoothing? Yes, with a randomized path exploration!," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 5, pp. 1595–1604, Oct. 2023.

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

[56] X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, "Graph neural network for interpreting task-fmri biomarkers," in *Proc. Med. Image Comput. Comput. Assist. Interv.: 22nd Int. Conf.*, 2019, pp. 485–493.

[57] H. Zou and J. Yang, "Temporal variability–based functional brain lateralization study in ADHD," *J. Attention Disord.*, vol. 25, no. 6, pp. 839–847, 2021.

[58] S. L. Bressler and J. S. Kelso, "Cortical coordination dynamics and cognition," *Trends Cogn. Sci.*, vol. 5, no. 1, pp. 26–36, 2001.

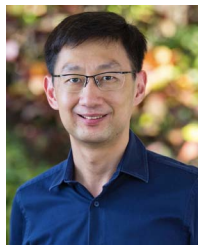
Rui Liu received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2023. He is currently a Postdoc Fellow with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. His current research interests include machine learning, multi-task/modality learning, and medical images diagnosis and applied deep learning.

Zhi-An Huang received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2021. He is currently a Research Fellow with the City University of Hong Kong (Dongguan). He has authored or coauthored more than 30 papers in esteemed journals and conference proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, PLoS CB, Brief Bioinform, Bioinformatics, BMC Bioinformatics, BIBM, and IJCNN. His research interests include AI in healthcare, machine learning, bioinformatics, and medical imaging analysis. He is currenty an Associate Editor for IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTA SYSTEMS.

Yao Hu (Graduate Student Member, IEEE) received the B.S. degree in mining engineering and the M.Sc. degree in control science and engineering from the China University of Mining and Technology, Xuzhou, China, in 2017 and 2020, respectively. He is currently working toward the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong. His current research interests include machine learning, transfer learning, and federated learning.

Lei Huang received the bachelor's degree in information management and information system from Wuhan University, Wuhan, China. He is currently working toward the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong. His current research interests include machine learning for drug discovery, bioinformatics, BioNLP, and applied machine learning.

Ka-Chun Wong received the B.Eng. degree in computer engineering and the M.Phil. degree from the Chinese University of Hong Kong, Hong Kong, in 2008 and 2010, respectively, and the Ph.D. degree from the Department of Computer Science, University of Toronto, Toronto, ON, Canada, in 2015. He was an Associate Professor with the City University of Hong Kong, Hong Kong. His current research interests include bioinformatics, computational biology, evolutionary computation, data mining, machine learning, and interdisciplinary research.



Kay Chen Tan (Fellow, IEEE) received the B.Eng. (first class Hons.) degree and the Ph.D. degree from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively. He is currently a Chair Professor with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. He was the Editor-in-Chief of IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, and currently serves on the Editorial Board member of more than ten journals. He is currently the Vice-President (Publications) of IEEE Computational Intelligence Society, an Honorary Professor with the University of Nottingham, Nottingham, U.K., and the Chief Co-Editor of Springer Book Series on *Machine Learning: Foundations, Methodologies, and Applications*.