Multi-Scale Spatial–Temporal Attention Networks for Functional Connectome Classification

Youyong Kong, *Member, IEEE*, Xiaotong Zhang, Wenhan Wang, Yue Zhou, Yueying Li, Yonggui Yuan, and REST-Meta-MDD Consortium

***Abstract*— Many neuropsychiatric disorders are considered to be associated with abnormalities in the functional connectivity networks of the brain. The research on the classification of functional connectivity can therefore provide new perspectives for understanding the pathology of disorders and contribute to early diagnosis and treatment. Functional connectivity exhibits a nature of dynamically changing over time, however, the majority of existing methods are unable to collectively reveal the spatial topology and time-varying characteristics. Furthermore, despite the efforts of limited spatial-temporal studies to capture rich information across different spatial scales, they have not delved into the temporal characteristics among different scales. To address above issues, we propose a novel Multi-Scale Spatial-Temporal Attention Networks (MSSTAN) to exploit the multi-scale spatial-temporal information provided by functional connectome for classification. To fully extract spatial features of brain regions, we propose a Topology Enhanced Graph Transformer module to guide the attention calculations in the learning of spatial features by incorporating topology priors. A Multi-Scale Pooling Strategy is introduced to obtain representations of brain connectome at various scales. Considering the temporal dynamic characteristics between dynamic functional connectome, we employ Locality Sensitive Hashing attention to further capture long-term dependencies in time dynamics across multiple scales and reduce the computational complexity of the original attention mechanism. Experiments on three brain fMRI datasets of MDD and ASD demonstrate the superiority of our proposed approach. In addition, benefiting from the attention mechanism in Transformer, our results are interpretable, which can contribute to the discovery of biomarkers. The code is available at https://github.com/LIST-KONG/MSSTAN.**

***Index Terms*— Graph neural networks, spatial-temporal attention, transformer, brain disorder diagnosis, functional connectivity.**

I. Introduction

NEUROPSYCHIATRIC disorders are characterized by disruptions in cognition, emotion, behavior, or their combination, encompassing a range of conditions that affect the brain and associated psychological functions. They represent a significant global health challenge, accounting for 17% of global deaths and exerting a substantial impact on worldwide disease burden and mortality rates [1]. Major Depressive Disorder (MDD) is a prototypical neuropsychiatric disorder, with a lifetime prevalence of approximately 10%, affecting around 185 million people globally [2]. Autism Spectrum Disorder (ASD), as an early-onset neurodevelopmental disorder, has a global prevalence of around 1% [3]. However, due to incomplete understanding of their underlying mechanisms, current diagnostic methods primarily rely on scale assessments or subjective judgments of clinicians, often leading to misdiagnosis [4], [5]. Hence, there is a pressing need to explore intelligent identification methods to achieve precise clinical diagnosis.

Existing research indicates that the clinical manifestations of patients with neuropsychiatric disorders, including MDD and ASD, are correlated with abnormalities in their brain's functional connectivity network [6]. Functional connectivity can be derived through functional magnetic resonance imaging (fMRI) based on blood oxygen level-dependent (BOLD) signals, which represent the temporal correlation between neural activity patterns in different brain regions. The study of brain functional connectivity is of great importance for the classification of neurological disorders as well as the search for potential biomarkers [7].

Machine learning approaches have been broadly applied in the functional connectivity classification, especially the popular support vector machine (SVM) and random forest

Manuscript received 7 June 2024; revised 15 August 2024; accepted 18 August 2024. Date of publication 22 August 2024; date of current version 2 January 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFE0116700, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant SJCX23_0047, in part by the National Natural Science Foundation of China under Grant 82271570 and Grant 31800825, in part by the Central University Basic Research Fund of China under Grant 2242024K40020, and in part by the Big Data Computing Center of Southeast University. *(Corresponding authors: Youyong Kong; Yonggui Yuan.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Zhongda Hospital, Southeast University, Nanjing, China, under Application No. 2013ZDSYLL021.0.

Youyong Kong, Wenhan Wang, and Yueying Li are with Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, and the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, Southeast University, Nanjing 210096, China (e-mail: kongyouyong@seu.edu.cn; imwhwang@gmail.com; 230228504@seu.edu.cn).

Xiaotong Zhang is with the School of Software Engineering, Southeast University, Nanjing 210096, China (e-mail: xiaotong_zhang@seu.edu.cn).

Yue Zhou and Yonggui Yuan are with the Department of Psychosomatic and Psychiatry, School of Medicine, Zhongda Hospital, Southeast University, Nanjing 210009, China (e-mail: 2725106172@qq.com; yygylh2000@sina.com).

Digital Object Identifier 10.1109/TMI.2024.3448214

methods [8], [9]. Nevertheless, simply adopting traditional machine learning methods is unable to sufficiently explore the complicated connections between various brain regions, which tends to affect the accuracy of classification. Therefore, employing deep learning models to classify functional connectivity is gradually emerging as a research hotspot [10], [11]. Modeling functional connectivity as a graph can provide a better representation of brain networks' properties, where nodes of the graph represent brain regions and edges are functional connections between brain regions. Hence, Graph Neural Networks (GNNs) [12], [13] tailored for graph data have been naturally applied to the study of the classification of functional connectome [14], [15], [16].

However, functional connectivity is not static but fluctuates over time [17]. The studies outlined above using static functional connectome do not adequately reflect the time-varying nature of the brain's intricate neural systems. As a result, dynamic functional connectome-based researches have emerged. Dynamic functional connectome is achieved by partitioning the BOLD time series into different intervals using a sliding window, thereby constructing multiple connectomes. To extract discriminative spatial and temporal features based on dynamic functional connectome, significant efforts have been devoted to studying the patterns embedded in the spatial topology and temporal dynamics. These methods typically use GNNs to extract spatial features and use recurrent neural networks (RNNs) or their variants to obtain temporal features of dynamic functional connectome [18], [19]. With the ascendancy of the Transformer in the field of natural language processing [20], [21], the latest work also extends to using the Transformer instead of RNNs to extract temporal dynamics [22].

Existing spatial-temporal works suffer from two notable limitations in the context of spatial and temporal feature learning. Firstly, the spatial feature learning of brain regions based on the GNNs and variants [23], [24] is constrained by the receptive field, impeding the capture of long-range interactions among brain regions and hindering effective spatial feature acquisition. Secondly, certain spatial-temporal approaches enrich spatial representations by integrating multi-scale information [25], [26]. However, these methods often conclude the learning of temporal characteristics using the original spatial connectome or single fused spatial representation of different scales when performing temporal relationship modeling, neglecting to further explore the temporal features across different scales. Consequently, achieving the learning of rich spatial-temporal features remains a challenging task.

In this work, we propose a novel framework called **M**ulti-**S**cale **S**patial-**T**emporal **A**ttention **N**etworks (MSSTAN) for functional connectome classification, which extracts effective spatial-temporal features. For spatial brain region feature learning, we propose a Topology Enhanced Graph Transformer to guide the attention calculations by incorporating brain network topology priors, thereby enabling a comprehensive extraction of spatial features for brain regions. Furthermore, to explore rich semantic information across diverse scales, we introduce a Multi-Scale Pooling Strategy designed to obtain brain connectome representations at varying scales while retaining significant brain regions. Given the temporal dynamic characteristics in functional connectomes, we employ Locality Sensitive Hashing (LSH) attention to capture long-term dependencies in temporal dynamics across multiple scales and reduce computational complexity, filling the gap in the study of temporal characteristics in multi-scale spaces. Our proposed MSSTAN outperforms on three brain fMRI datasets for diagnosis of two brain disorders, MDD and ASD, and the attention mechanism enhances the interpretability, contributing to biomarker discovery.

Our contributions can be summarized as follows:

- We propose a novel framework MSSTAN to extract multi-scale spatial-temporal features of dynamic functional connectome for functional connectome classification.
- For spatial feature learning, we propose a Topology Enhanced Graph Transformer to guide the brain region feature learning with topological priors. Additionally, we design a Multi-Scale Pooling Strategy to capture spatially rich semantic information of connectome at different scales.
- For temporal feature learning, we introduce the LSH attention mechanism to capture and integrate dynamic temporal characteristics at different spatial scales, ensuring the quality of learned representations while improving computational efficiency.
- Extensive experiments conducted on three real fMRI datasets of MDD and ASD indicate that MSSTAN effectively obtains spatial-temporal representations of functional connectomes, substantially improving diagnostic performance while maintaining excellent interpretability.

II. Related Work

A. Static Functional Connectome Classification

Machine learning and deep learning methods have been widely used to solve functional connectome classification problems. Yang et al. [8] applied SVM to classify autism spectrum disorder patients and typically developing participants and identify the important brain connectivity features. Kamarajan et al. [9] aimed to identify specific features of brain connectivity that can classify adult males with alcohol use disorder from healthy controls using the random forest method. With the help of GNNs, Ktena et al. [27] employed graph convolutional networks to learn the similarity measure between functional connectomes to classify individuals as ASD or healthy control (HC). Li et al. [28] proposed BrainGNN, a novel GNN-based model, featuring innovative strategies for region-of-interests (ROIs) selection and regularization methods. Cui et al. [29] specifically designed an edge-weight-aware message-passing mechanism for brain networks to aggregate brain region features and learned a globally shared edge mask at the group level to capture common patterns associated with specific diseases. Zhang et al. [30] proposed an end-to-end unsupervised graph structure learning method that automatically learns the discriminative structure of functional brain networks through graph generation and topology-aware encoding modules. However, functional connectivity exhibits

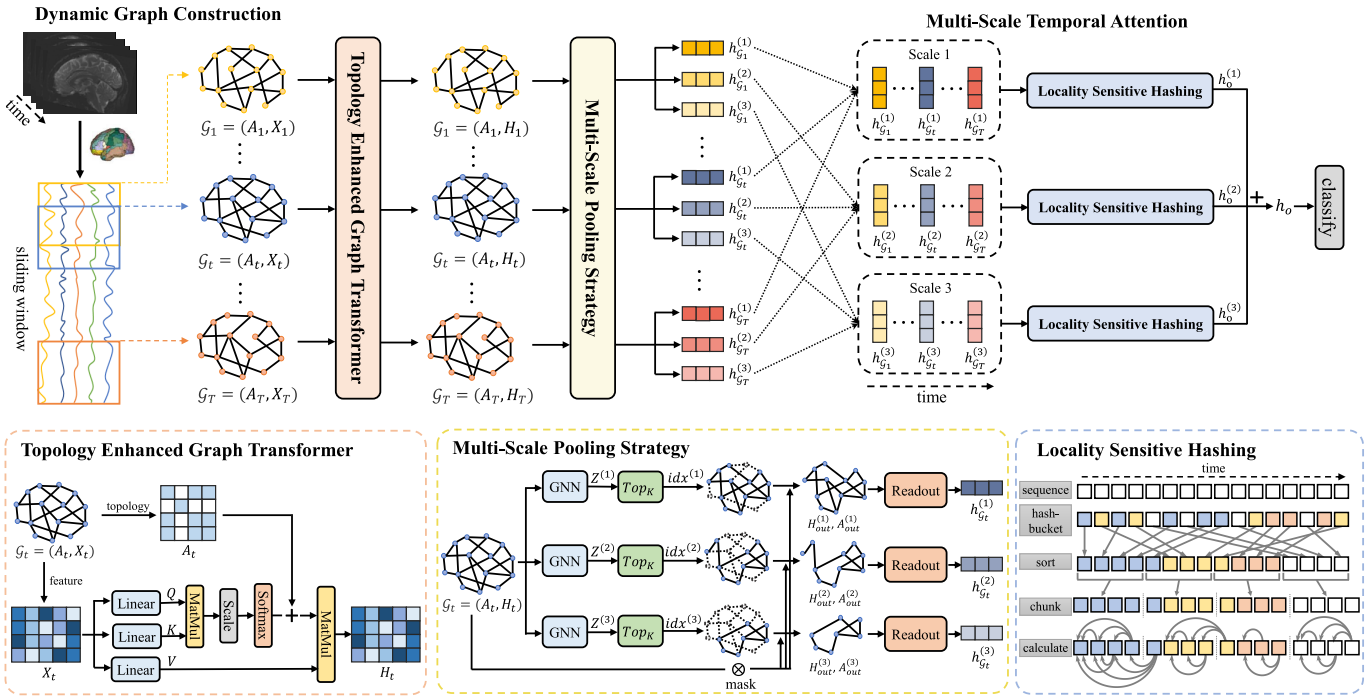


Fig. 1. Overall framework of the proposed Multi-Scale Spatial-Temporal Attention Networks (MSSTAN). The sliding window is used to divide the time series to obtain a dynamic functional connectome set. Each connectome within a window is then input into the Topology Enhanced Graph Transformer for spatial feature learning in brain regions. The Multi-Scale Pooling Strategy captures diverse spatial information at different scales, and the Locality Sensitive Hashing attention mechanism learns temporal characteristics for each scale representation. Finally, the temporal features from multiple scales are fused to obtain the ultimate representation for classification.

temporal dynamics, and the aforementioned static methods fail to unveil the temporal relationships.

B. Dynamic Functional Connectome Classification

There is a growing number of approaches based on dynamic functional connectome. Kong et al. [18] proposed a spatio-temporal graph convolutional network (STGCN) framework for learning dynamic characteristics from functional connectome to achieve MDD diagnosis. Kim et al. [22] and Wang et al. [31] used the combination of GNNs and attention mechanisms to learn dynamic graph representations of functional connectome. Wang et al. [24] and Chen and Zhang [23] employed Transformer to capture temporal patterns of spatial representations learned from GNNs across various time windows. Lee et al. [32] proposed ESTA, which applies attention mechanisms based on feature decomposition separately to temporal and spatial features, extracting fine-grained feature representations. To more effectively model the spatial-temporal features of fMRI, certain studies explored the integration of multi-scale or multi-level information to enhance the performance. Liu et al. [25] integrated multi-scale spatial representations into a unified representation within a time window, employing Long Short-Term Memory (LSTM) to capture temporal relationships between them. Zhang et al. [26] used the node allocation matrix to obtain and fuse multi-level sub-nodes after learning spatial–temporal features through spatial–temporal graph convolution. It is evident that most approaches rely on GNNs for spatial representation learning. However, due to limited receptive fields, they face challenges in facilitating long-distance interaction among brain regions. Few methods consider spatial multi-scale information and eventually transform into a singular representation within a time window. This transformation overlooks the temporal characteristics of functional connectome between adjacent windows at different scales, hindering the modeling of discriminative spatial-temporal patterns to some extent.

III. Proposed Method

This section describes our proposed model MSSTAN in detail, as shown in Fig. 1. Firstly, dynamic functional connectome is constructed utilizing the sliding window technique. Next, Topology Enhanced Graph Transformer module is utilized to learn brain region features with topology prior. After that, a Multi-Scale Pooling Strategy is designed to capture multi-scale spatial information. Then, the Locality Sensitive Hashing attention mechanism is introduced to learn temporal characteristics for representations of various scales. Finally, the learned spatial-temporal features are used to perform classification.

A. Dynamic Functional Connectome Construction

For dynamic functional connectome construction, a common sliding window technique is employed, consisting of the following steps: a) The original fMRI data is first preprocessed according to the Data Processing Assistant for Resting-State Function (DPARSF) [33] preprocessing pipeline. b) The brain is partitioned into $N = 90$ brain regions using the Anatomical

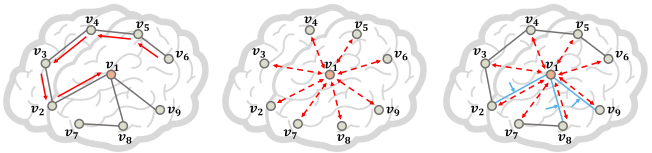


Fig. 2. Comparison of spatial information learning methods: (a) GNN-based Methods, (b) Transformer-based Methods, and (c) Topology Enhanced Graph Transformer (ours).

Automatic Labeling (AAL) [34] atlas which is commonly used in fMRI research. c) The mean time series of BOLD signals are extracted from each brain region. d) The window length w and step size s are then defined, and the sliding window is shifted step by step throughout the time series to capture the signal within each window. e) The strength of the functional connectivity between signals from different brain regions within each window is then calculated using Pearson correlation coefficient, resulting in a symmetrical $N \times N$ matrix. f) The proportional quantization $threshold$ is used to binarize the Pearson correlation matrix and the matrix for delineating the topological connections within the brain is finally derived.

By the aforementioned method, for each subject, we could thus define dynamic brain graphs $\mathcal{G}_t(\mathcal{V}, \mathcal{E})|_{t=1}^{T}$ by node feature matrix $\mathbf{X} = \{X_t | X_t \in \mathbb{R}^{N\times d}, t = 1, 2, \ldots, T\}$ and adjacency matrix $\mathbf{A} = \{A_t | A_t \in \mathbb{R}^{N\times N}, t = 1, 2, \ldots, T\}$. T is the number of sliding windows and d is the dimension of features. To enhance topological information, we utilize original functional connectivity matrix without binarization as node feature.

B. Topology Enhanced Graph Transformer

GNNs play an important role in the application of non-Euclidean data in deep learning and have now become a powerful model for learning graph representations [35]. The study of GNNs primarily delves into information propagation from neighboring nodes. However, the features of each node in a graph are not only related to the features of its neighboring nodes, but also the range of other nodes. In the human brain, functional correlations may also exist between brain regions that are distant from each other. GNNs initially have a limited receptive field, expanding with deeper convolutional layers to connect information across a broader node range. However, nodes become less distinguishable after several layers, leading to the over-smoothing problem [36]. For example, as shown in Fig. 2(a), we construct a brain graph $\mathcal{G}_t$ based on functional connectivity. For a central node (brain region) v_1 to aggregate information from v_6, five convolutional layers are needed, which can easily lead to over-smoothing. In this case, v_1 and v_6 exhibit long-range dependency that is struggle for GNNs to learn. Attention models like Transformer can effectively address this issue by modeling long-term dependencies and taking a global perspective on feature dependency [37]. Self-attention in Transformer can compute the relationship of each node with all others, as shown in Fig. 2(b), but it overlooks the intrinsic connectivity in the graph, the topological property of the brain network in this paper, which is fundamental characteristics of the graph.

To further exploit topological information in spatial feature extraction, we propose Topology Enhanced Graph Transformer, which imports the topological information of the graph by introducing the adjacency matrix in the self-attention calculation, as shown in Fig. 2(c). The merits of this approach are rooted in its augmented ability to exploit the intrinsic structural attributes of the graph, fully considering the functional connectivity characteristics inherent in brain connectome. Consequently, this refinement significantly enhances the overall effectiveness of spatial brain region feature extraction.

Inside the Transformer, a dot product operation is used to calculate the attention of each node. Firstly, the node feature matrix $X_t \in \mathbb{R}^{N\times d}$ is transformed to obtain Q_t, K_t, and V_t through linear layers, which are then fed into the self-attention calculation.

$$\begin{aligned} &Q_t = X_t W_Q, K_t = X_t W_K, V_t = X_t W_V, \\ &S_t = \text{Attention}(Q_t, K_t, V_t) = \text{Softmax}(\frac{Q_t K_t^\top}{\sqrt{d}}) V_t \end{aligned} \tag{1}$$

where W_Q, W_K, $W_V \in \mathbb{R}^{d\times d}$ and $Q_t, K_t, V_t \in \mathbb{R}^{N\times d}$. Self-attention captures the relationships between different nodes by calculating the similarity between any two elements in the sequence, irrespective of the spatial distance separating them. However, directly applying the Transformer to the graph in this way loses the topology information of the graph. Therefore, we incorporate the adjacency matrix, which signifies the topology structure of brain connectome, into the calculation of self-attention. The equation is modified as follows,

$$\begin{aligned} S_t &= \text{Attention}(Q_t, K_t, V_t) \\ &= (\text{Softmax}(\frac{Q_t K_t^\top}{\sqrt{d}}) + A_t) V_t \end{aligned} \tag{2}$$

where A_t represents the adjacency matrix under the $t-th$ window, and Q_t, K_t, V_t are projection matrix in Eq. 1. After calculating the correlations using Q_t and K_t, the scores computed from them are summed with the graph structure A_t. In this approach, the prior knowledge of the topology is involved to enhance connections with strong correlations and facilitate the learning of relationships among all brain regions.

For modeling from multiple dimensions, multi-head self-attention is achieved by performing m projections on the embeddings using m different weight values W_Q, W_K, and W_V to obtain m outputs. These outputs are concatenated together and projected into the d-dimensional space to obtain the final output. The multi-head mechanism can be expressed as follows,

$$\begin{aligned} H_t &= \text{MultiHead}(Q_t, K_t, V_t) \\ &= \text{Concat}(S_{t1}, S_{t2}, \ldots, S_{tm}) \end{aligned} \tag{3}$$

where $H_t \in \mathbb{R}^{N\times d}$ signifies the learned node feature matrix obtained by extracting spatial characteristics from dynamic brain connectomes through the utilization of Transformer architecture and topological constraints.

C. Multi-Scale Pooling Strategy

In order to facilitate temporal feature learning while minimizing the number of parameters, a common practice involves aggregating global information and combining node embeddings into a unified graph representation [24]. However, direct read operations such as MAX or MEAN [13] can discard the graph structure, resulting in the loss of crucial information, especially in real-world networks with intricate topologies like brain networks. Therefore, more ideal alternatives have emerged, restructuring nodes with rich semantic information into super nodes under certain conditions, progressively pooling the graph into a feature vector [28], [38]. Although these pooling methods perform well in capturing hierarchical graph structures compared to global pooling, their typical focus on characteristics within a single-scale space may limit the ability in representing the entirety of brain connectome information.

We propose a Multi-Scale Pooling Strategy capable of obtaining pooled graph representations across various scale spaces. This approach aims to consider different subspace information from different scales while preserving significant brain regions that play an important role in diagnosis, which enhances spatial representation for subsequent multi-scale temporal feature extraction, providing a more comprehensive representation of the dynamic brain connectome. Specifically, the module architecture includes three GNN layers, three pooling layers and three readout layers. In the pooling process, the GNN is first used to compute the self-attention scores, followed by the selection and retention of the Top_K nodes based on these scores to obtain the updated graph.

Self-Attention Graph Pooling (SAGPool) [39] is introduced to conduct the fundamental pooling operation. Attention mechanisms allow to pay more attention to important features and fully learn and absorb this key information, while relegating less significant features to a lower priority. We employs the Top_K strategy [40] to either retain important brain regions and discard relatively insignificant brain regions. Simultaneously, GNN is used to generate self-attention scores, ensuring that the result of pooling operation aligns with the features and topology of the graph. Taking node features H_t and adjacency matrix A_t as input (for simplicity, subscript t is omitted below), the generalized equation for calculating the attention score is as follows,

$$Z = \sigma\left(\mathrm{GNN}\left(H, A\right)\right) \tag{4}$$

where $Z \in \mathbb{R}^{N\times 1}$ is the node attention score matrix. The GNN utilized for calculating the attention score here can be implemented across various frameworks, with the chosen model in this work being the Graph Isomorphism Network (GIN) [13] for its high expressive ability. After obtaining attention scores for all brain regions, SAGPool applies the Top_K mechanism to retain only the $K \times N$ nodes with the highest scores. Nodes are selected based on the score Z, and the corresponding indices are obtained as follows.

$$idx = top\text{-}rank\left(Z, [K \times N]\right) \tag{5}$$

$$Z_{mask} = Z_{idx} \tag{6}$$

where the pooling ratio $K \in (0, 1)$ is a hyperparameter that determines the proportion of nodes to be retained. $top\text{-}rank$ returns the index of the $K \times N$ highest-scoring nodes. $\cdot idx$ denotes the index operation and Z_{mask} is the attention mask. Then we update the pooled feature matrix and adjacency matrix as follows,

$$X_{out} = H_{idx,:} \odot Z_{mask} \tag{7}$$

$$A_{out} = A_{idx,idx} \tag{8}$$

where $H_{idx,:}$ represents the feature matrix indexed by rows (i.e., nodes / brain regions) and $A_{idx,idx}$ is the adjacency matrix indexed by rows and columns. $\odot$ signifies the element-wise product. X_{out} and A_{out} denote the updated feature matrix and its corresponding adjacency matrix, respectively.

Following these feature and adjacency matrix update operations, the number of nodes is effectively reduced. To achieve a fixed representation for each graph, it is necessary to further obtain the global graph representation through readout operation. We employ a combination of MAX and MEAN to aggregate all node features.

$$h_{\mathcal{G}} = mean(X_{out}) \,||\, max(X_{out}) \tag{9}$$

where $||$ is a connection operation and $h_{\mathcal{G}} \in \mathbb{R}^{2d}$ denotes the pooled graph representation. By setting different ratios, the original graph can be transformed to representations with a different number of nodes. The proposed model in this article leverages three scales to derive graph pooling results $h_{\mathcal{G}}^{(1)}, h_{\mathcal{G}}^{(2)}, h_{\mathcal{G}}^{(3)}$ in three distinct scale spaces, which are sent to the subsequent temporal feature learning module to integrate temporal dynamic correlations at different subspace.

D. Multi-Scale Temporal Attention Module

Multi-scale pooling captures the spatial organization characteristics of each graph from different scale spaces [41], and a series of graph representations $\boldsymbol{h_{\mathcal{G}}} = \{h_{\mathcal{G}_t}^{(s)} | h_{\mathcal{G}_t}^{(s)} \in \mathbb{R}^{2d} | s = 1, 2, 3, t = 1, 2, \ldots, T\}$ can be obtained. Due to the dynamic characteristics of functional connectivity, it is also important to learn the relationships between different frames from the sequence and explore the temporal dynamic features at different topological scales, which has been neglected by previous studies. We propose a Multi-Scale Temporal Attention module using Locality Sensitive Hashing (LSH) to learn features in the temporal dimension from different scales and reduce computation complexity.

Traditional Transformer models often excel in achieving state-of-the-art results on sequential tasks, but training these models can incur prohibitive costs [42]. To enhance Transformer efficiency, this work employs the LSH attention method as a substitute for dot product attention, reducing its complexity from $O\left(L^2\right)$ to $O(LlogL)$, where L denotes the length of the sequence. It demonstrates performance on par with the original Transformer while exhibiting enhanced memory efficiency and faster processing on sequence.

In this work, the choice of $Q_{te} = K_{te}$ is employed, where the same linear layer is utilized to project the original features onto both Q_{te} and K_{te}, contributing to a reduction in computational complexity. Moreover, during attention score computation, the Softmax function is applied, directing attention solely to the key in K_{te} that is closest to each query q_i due

to the dominance of the maximum element in Softmax. In contrast to the conventional attention mechanism, where every token computes its association with all others, LSH attention highlights positions in the sequence with a significant impact on all others, thereby enhancing computational efficiency.

In high-dimensional space, efficient identification of neighbors can be facilitated by locality sensitive hashing, where proximate vectors are more likely to share identical hash values, while those farther apart have a diminished probability of sharing the same hash value. Specifically, given a set of graph sequences $\boldsymbol{h}_{\mathcal{G}}^{(s)} = \{h_{\mathcal{G}_t}^{(s)} | h_{\mathcal{G}_t}^{(s)} \in \mathbb{R}^{2d} | t = 1, 2, \ldots, T\}$ of a specific scale s, a hash function $hash(\cdot)$ is employed to allocate each vector to distinct hash buckets. The vectors are then rearranged based on these hash buckets, preserving the original relative positions within the sequence. Subsequently, the sequence is partitioned into distinct chunks, and parallelized attention mechanisms are deployed to train the network on each sequence along with its previous chunk. For a specific query position i, the calculation of LSH attention is expressed as follows:

$$h_{oi} = \sum_{j \in F_i} exp\left(q_i \cdot k_j - f\left(i, F_i\right)\right) v_j,$$
$$F_i = \left\{j : hash\left(q_i\right) = hash\left(k_j\right)\right\} \tag{10}$$

where F_i denotes the set attended to by query q_i at position i, and f represents the normalization factor in the Softmax operation. The use of LSH improves dot product attention in each block, leading to a substantial reduction in memory and computational consumption.

For the graph representation sequences based on different scales output by the multi-scale pooling module, LSH attention for different scales is used to extract temporal features to obtain the spatial-temporal features $h_o^{(1)}, h_o^{(2)}, h_o^{(3)}$. These features model the different spatial-temporal patterns in different scale spaces, and then the fusion module is used to effectively integrate them and summarize the outputs from diverse scales. The fused spatial-temporal representation $h_o \in \mathbb{R}^{2d}$ is finally fed to the fully connected layer for brain disorder diagnosis.

IV. Experiments

A. Experiments Setup

1) Datasets: Two fMRI datasets related to Major Depressive Disorder (MDD) are used: the **zhongdaxinxiang** dataset and the multi-center MDD (**Multi-MDD**) public dataset. Additionally, one fMRI dataset related to Autism Spectrum Disorder (ASD) is used: the Autism Brain Imaging Data Exchange (**ABIDE**) public dataset.

The subject cohort of the **zhongdaxinxiang** dataset is recruited from Zhongda Hospital Affiliated to Southeast University, Henan Provincial Psychiatric Hospital, and Hangzhou Hospital, including 314 MDD patients and 206 Healthy Controls (HCs). The acquired brain MRI included T1 structural images and resting-state functional magnetic resonance imaging (rs-fMRI). Three-dimensional T1-weighted images were acquired using magnetization-prepared rapid gradient echo sequences with the following parameters: TR (Repetition Time) of 1900 ms, echo time (TE) of 2.48 ms, flip angle (FA) of 9°, acquisition matrix of 256 × 256, a field of view of 250 × 250 mm^2, slice thickness of 1.0 mm, no gap between slices, and a total of 176 slices. The parameters for rs-fMRI were set as follows: TR of 2000 ms, TE of 25 ms, a field of view of 240 × 240 mm^2, 36 slices, slice thickness of 3 mm, and 240 time points. The raw fMRI images were preprocessed using the DPARSF [33] pipeline. Primary steps included converting DICOM to NIFTI format, correcting for slice timing, mitigating motion artifacts, detrending, regressing signals from cerebrospinal fluid and white matter, normalizing images to standard space, and applying temporal bandpass filtering. Finally, the average time series of each brain region were extracted. To ensure a stable signal, the first ten time points were discarded during data processing, resulting in a final length of 230 time points.

The **Multi-MDD** dataset is part of the REST-meta-MDD project [43], a collaborative effort involving 17 participating hospitals in China that can be obtained through http://rfmri.org/REST-meta-MDD. The consortium follows standardized procedures for processing MDD data at each site, aiming to minimize heterogeneity in pre-processing methods. Due to the inconsistent length of the time series of data collected from multiple centers, we eliminate data with a time series length less than 230. The final dataset includes 299 HCs and 368 MDD patients.

The **ABIDE** dataset [44] is sourced from 12 different centers and comprises a total of 618 subjects, including 290 individuals diagnosed with ASD and 328 HCs. It is publicly available at https://fcon_1000.projects.nitrc.org/indi/abide. All data used in the experiments follow the official processing protocols. Similarly, we use data with a time series length of more than 230 for validation.

For both datasets mentioned above, the entire brain is parceled using AAL, resulting in a total of 90 brain regions. Subsequently, the average time series for each brain region are obtained as a $N \times T$ matrix. Dynamic functional connectome is then constructed. The number of edges in the brain graph is contingent on the $threshold$ percentage, which determines the retention of only the strongest correlations. Given the lack of consensus in the functional connectivity literature regarding the optimal threshold percentage [45], this study treats the threshold as one of the hyperparameters to be optimized.

2) Implementation Details: In this paper, 10-fold cross-validation and leave-one-site-out validation are used to evaluate the performance of our model. For the leave-one-site-out experiments, we respectively utilize the Henan Provincial Psychiatric Hospital (xinxiang) site in zhongdaxinxiang dataset, the S25 site in Multi-MDD dataset and the UM site in ABIDE dataset as the test set, and train the model using the remaining sites within each dataset to validate the generalization capability. We use five common classification metrics to evaluate the performance: accuracy (ACC), sensitivity (SEN), specificity (SPE), F1-score (F1), and area under the curve (AUC). All experiments are performed ten times, using different random seeds for data partitioning to obtain the final results.

The experiments are implemented on two NVIDIA RTX 3090 GPUs. We use the Adam optimizer with a batch

TABLE I
COMPARISON BETWEEN OUR PROPOSED MSSTAN MODEL AGAINST BASELINES USING 10-FOLD CROSS-VALIDATION ON ZHONGDAXINXIANG DATASET AND MULTI-MDD DATASET IN MDD DIAGNOSIS. **BOLD** INDICATES THE BEST RESULTS

	zhongdaxinxiang					Multi-MDD				
Methods	ACC (%)	SEN (%)	SPE (%)	F1 (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	F1 (%)	AUC (%)
MSSTAN	**73.19 ± 0.63**	84.69 ± 3.09	**54.01 ± 3.20**	**78.70 ± 1.18**	**69.35 ± 0.80**	**68.65 ± 0.90**	74.67 ± 3.26	**59.45 ± 4.84**	**71.63 ± 1.23**	**67.06 ± 1.39**
STGCN	66.60 ± 0.69	**89.78 ± 2.24**	27.90 ± 2.97	75.74 ± 1.08	48.94 ± 1.89	61.21 ± 1.00	74.65 ± 4.25	40.71 ± 5.03	65.35 ± 2.30	48.81 ± 2.14
STAGCN	69.81 ± 1.05	85.84 ± 3.39	42.90 ± 5.83	76.88 ± 0.89	62.93 ± 0.77	66.04 ± 0.66	74.45 ± 3.92	53.37 ± 5.27	69.55 ± 1.82	58.80 ± 0.61
HFBN	69.23 ± 1.35	87.09 ± 3.14	42.85 ± 6.25	77.14 ± 0.84	64.97 ± 1.82	64.92 ± 1.86	**81.26 ± 4.63**	47.87 ± 7.16	70.44 ± 1.38	61.06 ± 1.67
ESTA	69.65 ± 1.86	86.01 ± 4.13	53.43 ± 3.90	76.89 ± 1.55	66.40 ± 1.96	65.92 ± 1.92	73.83 ± 2.56	57.90 ± 4.69	70.04 ± 1.38	65.22 ± 1.64
SVM	59.35 ± 0.21	66.50 ± 0.30	48.42 ± 0.73	66.09 ± 0.23	57.46 ± 0.46	56.60 ± 0.75	60.87 ± 1.08	51.51 ± 0.98	60.49 ± 0.85	56.19 ± 0.91
RF	58.33 ± 1.86	68.93 ± 1.96	42.05 ± 3.26	66.30 ± 1.52	55.49 ± 2.04	54.67 ± 1.59	62.12 ± 2.13	45.69 ± 2.68	59.97 ± 1.46	53.91 ± 1.58
BrainGNN	70.77 ± 1.17	87.38 ± 1.56	44.17 ± 4.36	78.04 ± 0.70	65.77 ± 1.74	65.18 ± 0.99	73.74 ± 4.69	52.91 ± 5.23	69.11 ± 1.80	63.32 ± 0.90
IBGNN	67.42 ± 1.06	80.85 ± 3.34	46.93 ± 4.73	74.75 ± 1.27	63.89 ± 1.28	63.42 ± 1.24	73.79 ± 3.33	50.66 ± 4.96	68.49 ± 1.22	62.22 ± 1.45
BrainUSL	68.85 ± 2.18	87.96 ± 2.40	38.04 ± 7.50	77.31 ± 0.97	64.00 ± 2.78	65.63 ± 1.73	80.64 ± 3.97	45.34 ± 6.63	71.54 ± 1.40	62.99 ± 1.85

size of 64. We set the window length =100, and window stride =2. The initial learning rate is 0.0003 and the weight_decay is 0.001. Dropout rate 0.1 is applied to the transformer and rate 0 is used in the final fully connected layer for classification. Our proposed model has a strong fitting capability, but continuously reducing the training loss may cause overfitting to the training data features, resulting in poor generalization ability. Therefore, we adopt the proposed flood regularization [46], in which the parameter b is set to 0.5.

3) Competitive Methods: To verify the superiority of the proposed MSSTAN in this paper, 9 methods are selected for comparison. These methods consist of two types of inputs: static functional connectome (FC) and dynamic FC.

Static FC: 1) Traditional machine learning-based methods encompass two approaches: support vector machines (SVM) and random forests (RF). Since the functional connectivity matrix is symmetric, the upper triangle of the matrix is taken to reshape into a vector and fed into the models for classification. 2) Deep learning-based methods include 3 approaches: BrainGNN [28], IBGNN [29] and BrainUSL [30]. These methods use static functional connectome as input.

- SVM employs the Sigmoid function as the kernel function and sets the regularization parameter C to 10.
- RF method configures the number of weak classifiers (n_estimators) to 25 and utilizes Gini Impurity as the criterion for splitting sub-trees.
- BrainGNN [28] adopts two graph convolutional layers and two pooling layers, with a learning rate set to 0.01, weight decay set to 0.005, and pooling rate set to 0.5.
- IBGNN [29] utilizes two layers of graph convolutional layers, two layers of MLP, with a learning rate set to 0.001, and weight decay set to 0.001.
- BrainUSL [30] is configured with a learning rate of 0.001, weight decay of 0.01, pre-training epochs set to 10, remaining training steps epochs set to 60, and a contrastive loss temperature parameter τ set to 0.6.

Dynamic FC: Dynamic FC classification methods include three approaches: STGCN [18], STAGCN [31], HFBN [26] and ESTA [32]. These methods consider the spatial-temporal features of functional connectome.

- STGCN [18] utilizes three layers of STGCN, along with two pooling layers and two fully connected layers. The learning rate is set to 0.01. The window length is set to 100 with a stride of 2, consistent with this study.
- STAGCN [31] employs two layers of STAGCN and two fully connected layers, with a learning rate set to 0.01. Additionally, it also utilizes a window length of 100, and a stride of 2.
- HFBN [26] sets the temporal convolution kernel size to 9, employs six layers of GCN, with eight attention heads, a learning rate of 0.1, and a weight decay of 0.0001.
- ESTA [32] utilizes two layers of temporal convolution and two layers of spatial convolution, and a MLP classifier consisting of two linear layers. The learning rate is initially set to 1e-5 and gradually adjusted, with a weight decay of 5e-6.

The methods described above all utilize the code provided by the authors. Apart from modifying the validation method to ten-fold cross-validation or leave-one-site-out validation to align with the experimental setup of this paper, the rest of the model architectures remain unchanged.

B. Performance Comparison in MDD Diagnosis

For MDD diagnosis, we first perform competitive methods using 10-fold cross-validation on two datasets in MDD diagnosis task, including zhongdaxinxiang and Multi-MDD. The results are shown in Table I. It can be observed that the classification performance of the proposed MSSTAN is better than all baseline methods in the four indicators ACC, SPE, F1 and AUC on two datasets, which indicates that MSSTAN considers from the perspective of multi-scale spatial and temporal fusion Feature extraction in different subspaces can effectively extract comprehensive feature representations of brain connectomes, showing the superiority of the model. Traditional methods like SVM and RF often underperform in classification due to relatively weak representation capabilities. In contrast, deep learning methods exhibit robust nonlinear representation abilities, leading to superior performance. The results highlight the enhanced effectiveness of graph-based approaches, attributed to GNN-based models' superior utilization of brain connectome topological information. Furthermore, the static GNNs solely engage in message passing and aggregation of spatial features within the brain regions. As it neglects the temporal characteristics of dynamic brain connections, which relatively limits its effectiveness. In contrast, MSSTAN

TABLE II

COMPARISON BETWEEN OUR PROPOSED MSSTAN MODEL AGAINST BASELINES USING LEAVE-ONE-SITE-OUT VALIDATION ON ZHONGDAXINXIANG DATASET AND MULTI-MDD DATASET IN MDD DIAGNOSIS. **BOLD** INDICATES THE BEST RESULTS

	zhongdaxinxiang (xinxiang)					Multi-MDD (S25)				
Methods	ACC (%)	SEN (%)	SPE (%)	F1 (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	F1 (%)	AUC (%)
MSSTAN	**70.89 ± 1.89**	**84.62 ± 6.44**	47.61 ± 12.88	**78.47 ± 1.57**	**66.11 ± 3.69**	**68.55 ± 1.97**	**81.75 ± 5.78**	51.50 ± 10.15	**75.83 ± 2.06**	**66.67 ± 2.87**
STGCN	65.46 ± 2.23	77.42 ± 9.33	48.57 ± 13.19	72.16 ± 3.37	62.99 ± 2.61	66.32 ± 1.69	78.65 ± 9.91	48.89 ± 13.01	72.98 ± 3.15	63.77 ± 2.27
STAGCN	65.72 ± 2.75	73.71 ± 8.63	**54.44 ± 13.86**	71.31 ± 2.54	64.08 ± 2.81	67.11 ± 2.05	80.90 ± 7.69	47.62 ± 10.21	74.23 ± 2.03	64.26 ± 2.29
HFBN	65.73 ± 1.09	76.52 ± 3.14	50.48 ± 6.25	72.10 ± 0.84	63.50 ± 1.82	63.27 ± 1.86	77.45 ± 7.28	45.72 ± 8.02	69.09 ± 6.75	61.59 ± 4.63
ESTA	64.18 ±1.98	81.58 ± 7.59	41.38 ± 12.39	72.09 ± 1.99	61.48 ± 3.62	62.82 ± 3.86	72.80 ± 5.43	50.48 ± 11.12	68.35 ± 2.71	61.64 ± 4.38
SVM	59.09 ± 0.00	70.83 ± 0.00	50.00 ± 0.00	60.18 ± 0.00	60.42 ± 0.00	60.81 ± 0.00	62.31 ± 0.00	**58.26 ± 0.00**	67.00 ± 0.00	60.28 ± 0.00
RF	64.92 ± 3.00	77.44 ± 3.04	43.70 ± 5.87	73.52 ± 2.27	60.57 ± 3.40	66.61 ± 2.23	80.90 ± 2.40	42.39 ± 4.48	75.29 ± 1.66	61.64 ± 2.53
BrainGNN	69.76 ± 1.41	82.18 ± 6.19	48.70 ± 9.13	77.28 ± 1.94	65.44 ± 2.00	67.11 ± 2.00	78.65 ± 5.95	50.79 ± 9.40	73.68 ± 1.66	64.72 ± 1.89
IBGNN	65.08 ± 3.31	82.95 ± 7.30	45.65 ± 10.43	75.43 ± 2.24	60.97 ± 2.92	65.13 ± 2.68	73.03 ± 6.15	53.97 ± 9.11	71.69 ± 2.04	63.50 ± 2.84
BrainUSL	67.82 ± 1.27	80.90 ± 7.74	45.65 ± 10.14	75.92 ± 1.66	63.27 ± 2.75	65.20 ± 3.09	80.22 ± 6.34	43.97 ± 11.72	72.74 ± 3.12	62.10 ± 4.00

TABLE III

COMPARISON BETWEEN OUR PROPOSED MSSTAN MODEL AGAINST BASELINES USING 10-FOLD CROSS-VALIDATION AND LEAVE-ONE-SITE-OUT VALIDATION ON ABIDE DATASET IN ASD DIAGNOSIS. **BOLD** INDICATES THE BEST RESULTS

	ABIDE (10-fold)					ABIDE (UM)				
Methods	ACC (%)	SEN (%)	SPE (%)	F1 (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	F1 (%)	AUC (%)
MSSTAN	**72.28 ± 0.91**	**65.33 ± 3.07**	77.02 ± 2.30	**68.06 ± 1.77**	**71.17 ± 1.17**	**67.64 ± 3.43**	56.88 ± 12.43	75.97 ± 6.22	**59.84 ± 7.48**	**66.42 ± 4.23**
STGCN	67.80 ± 1.17	59.66 ± 5.70	73.01 ± 4.71	61.59 ± 3.44	66.33 ± 1.25	65.18 ± 2.60	54.58 ± 9.94	73.39 ± 7.26	57.32 ± 5.70	63.99 ± 2.94
STAGCN	69.05 ± 1.13	56.20 ± 4.14	75.73 ± 2.89	61.53 ± 1.15	65.97 ± 1.12	65.91 ± 1.96	58.54 ± 13.90	71.61 ± 9.90	59.12 ± 6.77	65.08 ± 2.77
HFBN	66.67 ± 1.61	58.76 ± 4.93	69.88 ± 4.00	59.50 ± 3.13	64.32 ± 1.68	64.57 ± 3.50	51.03 ± 16.38	76.55 ± 12.45	55.99 ± 10.60	63.79 ± 3.19
ESTA	66.36 ± 0.82	60.72 ± 3.00	70.33 ± 2.16	62.33 ± 1.66	66.03 ± 0.88	63.09 ± 2.64	49.38 ± 15.98	73.71 ± 12.28	51.98 ± 12.20	61.54 ± 3.24
SVM	63.30 ± 1.39	58.01 ± 1.49	68.06 ± 1.96	59.25 ± 1.38	63.03 ± 1.48	62.13 ± 0.00	55.21 ± 0.00	68.19 ± 0.00	57.41 ± 0.00	61.70 ± 0.00
RF	61.63 ± 1.50	52.80 ± 2.63	69.79 ± 2.64	55.83 ± 1.90	61.30 ± 1.59	61.64 ± 3.70	61.04 ± 7.04	62.10 ± 3.83	58.01 ± 4.96	61.57 ± 3.95
BrainGNN	68.83 ± 0.83	57.98 ± 2.42	77.88 ± 2.32	62.82 ± 1.37	67.93 ± 1.00	64.45 ± 1.11	**61.88 ± 13.47**	66.45 ± 10.67	59.63 ± 5.40	64.16 ± 1.79
IBGNN	65.59 ± 0.97	62.10 ± 4.30	68.66 ± 4.20	62.26 ± 1.98	65.38 ± 0.95	60.64 ± 3.52	51.25 ± 14.83	67.90 ± 11.75	52.17 ± 8.34	59.58 ± 3.87
BrainUSL	70.10 ± 1.20	56.05 ± 3.84	**81.34 ± 1.88**	60.93 ± 2.97	68.70 ± 1.29	66.36 ± 5.69	47.71 ± 24.36	**80.81 ± 10.03**	50.55 ± 22.91	64.26 ± 7.69

employs topology enhanced graph transformer for spatial feature extraction and incorporates temporal attention to capture temporal dependencies, endowing it with the ability to depict spatial-temporal features from dynamic brain connectomes.

Moreover, comparisons among dynamic FC classification models are conducted. Compared with STGCN, STAGCN, HFBN, and ESTA, MSSTAN demonstrates an increase in classification performance on the two datasets. This improvement can be attributed to the fact that MSSTAN considers multi-scale spatial-temporal information, thereby affirming the importance of understanding the spatial-temporal characteristics of different subspaces for MDD diagnosis. Furthermore, despite the greater heterogeneity among data sources in the Multi-MDD dataset due to the larger number of sites, which resulted in lower classification performance compared to zhongdaxinxiang dataset, MSSTAN still achieves the highest classification performance among all methods. The fact that our model outperforms all other models is a good indication that the proposed MSSTAN can be applied as an effective method for classifying functional connectome.

To further evaluate the generalization performance and robustness of our proposed MSSTAN, we conduct leave-one-site-out validation experiments on two datasets to compare the performance of different methods. The final results are illustrated in Table II. It can be observed that MSSTAN achieves the best classification performance on both datasets, owing to the proposed multi-scale spatial-temporal fusion strategy, which enables the model to focus on spatial-temporal patterns in different subspaces, effectively learning discriminative spatial-temporal representations that are not specific to the dataset. Additionally, we find that satisfactory results are obtained even when training on the Multi-MDD dataset, which includes more sites, indicating the relative robustness of MSSTAN across different sites. Overall, MSSTAN not only exhibits good classification ability but also demonstrates sound generalization performance.

C. Performance Comparison in ASD Diagnosis

We conduct comparative experiments on the ABIDE dataset for ASD diagnosis to validate the model's performance on different functional connectome classification tasks. The results are shown in Table III. Similar results can be observed: MSSTAN achieves the best diagnostic performance across both validation methods and all comparison methods, fully demonstrating its effectiveness in various functional connectome classification tasks. Additionally, it is reasonable to speculate that BrainUSL, being an unsupervised graph structure learning method, inherently has higher generalization performance, hence showing competitive results here. It is noteworthy that HFBN, which performs spatial hierarchical feature fusion after spatial-temporal feature learning, shows poorer performance compared to our multi-scale spatial-temporal fusion. This further demonstrates the advantage of our approach in effectively extracting comprehensive representations of brain connectomes by learning dynamic characteristics in different scale subspaces.

D. Ablation Study

We conduct a series of ablation studies on three modules incorporated in MSSTAN across two datasets: zhongdaxinxiang and Multi-MDD, including Topology Enhanced Graph

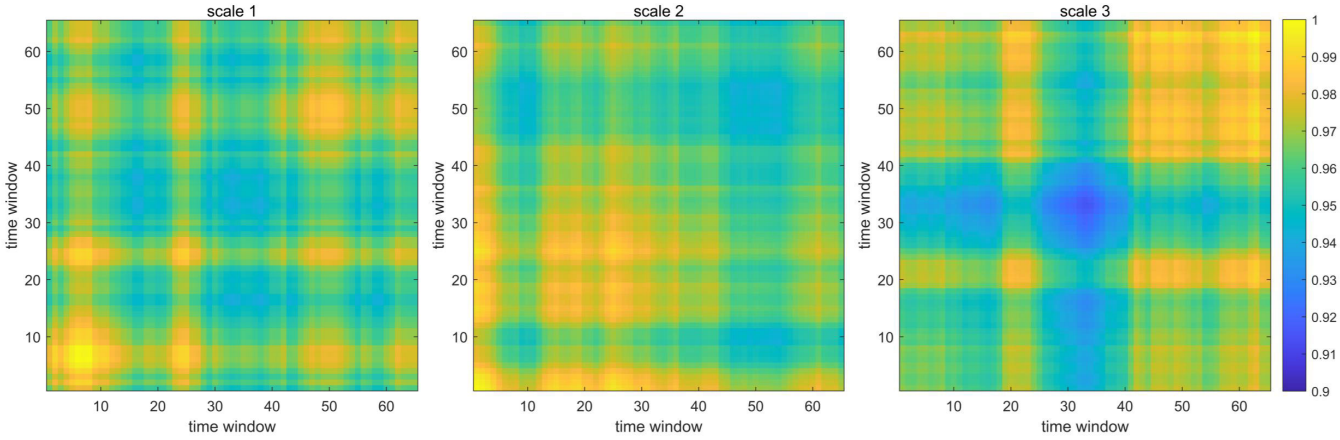


Fig. 3. Temporal information at different scales on zhongdaxinxiang.

TABLE IV
ABLATION STUDY OF DIFFERENT COMPONENTS ON ZHONGDAXINXIANG AND MULTI-MDD. **BOLD** INDICATES THE BEST RESULTS AND UNDERLINING SIGNIFIES SECONDARY OUTCOMES

Dataset	Component	ACC (%)	F1 (%)	AUC (%)
zhongdaxinxiang	MSSTAN	**73.19 ± 0.63**	**78.70 ± 1.18**	**69.35 ± 0.80**
	w/o S-Trans	71.60 ± 1.07	77.79 ± 1.07	67.35 ± 1.36
	w/o S-Topo	72.52 ± 0.73	78.18 ± 0.56	68.48 ± 0.89
	w/o MS-Pool	72.40 ± 0.96	78.48 ± 0.76	67.86 ± 1.49
	w/o MS-Temp	72.65 ± 0.93	78.23 ± 1.29	68.65 ± 1.30
	w/o LSH-Att	72.85 ± 0.97	78.50 ± 1.19	68.58 ± 1.26
Multi-MDD	MSSTAN	**68.65 ± 0.90**	71.63 ± 1.23	**67.06 ± 1.39**
	w/o S-Trans	67.03 ± 0.87	70.00 ± 1.08	66.07 ± 0.89
	w/o S-Topo	67.44 ± 1.25	71.59 ± 0.93	65.39 ± 1.43
	w/o MS-Pool	67.53 ± 1.14	71.06 ± 1.84	65.74 ± 1.25
	w/o MS-Temp	68.08 ± 0.76	72.05 ± 1.52	66.12 ± 0.65
	w/o LSH-Att	68.52 ± 0.83	**72.29 ± 0.94**	66.70 ± 1.04

Transformer, Multi-Scale Pooling Strategy and Multi-Scale Temporal Attention. The results are shown in Table IV.

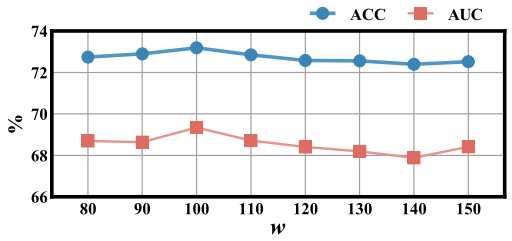
Topology Enhanced Graph Transformer utilizes the adjacency matrix to guide self-attention calculations in Transformer for topological considerations. We replace the transformer with a GCN, defined as "w/o S-Trans", which only uses local message aggregation. We also removes the spatial topology guidance defined as "w/o S-Topo", exclusively utilizing the original transformer and employing full connection to calculate attention correlation across all nodes, which will not strengthen the attention scores of existing connections. The results show that, compared to using only GCN, the transformer better captures the information interaction between brain regions with long-distance dependencies, which is beneficial for global spatial representation. Compared to the version without topology enhancement, adding topology information to the transformer improves performance. MSSTAN introduces the adjacency matrix into the transformer calculation, addressing the issue of the original transformer ignoring the structural knowledge within graph, effectively advancing spatial modeling of brain connectomes.

Multi-Scale Pooling Strategy aims at acquiring brain connectome representations across various scale spaces. To validate the effectiveness of it, we only preserve the spatial pooling results from a single scale, employing an optimal multi-scale parameter with a node retention ratio of 0.4, denoted as "w/o MS-Pool". It can be seen from the results that compared with single-scale pooling, the design of multi-scale pooling strategy has a positive effect on the model. This is attributed to the fact that configuring distinct node retention ratios enhances the richness of feature representation extracted from the spatial space. The features of different scales are then sent to multi-scale temporal attention, ensuring a more comprehensive exploration of learned spatial-temporal features.

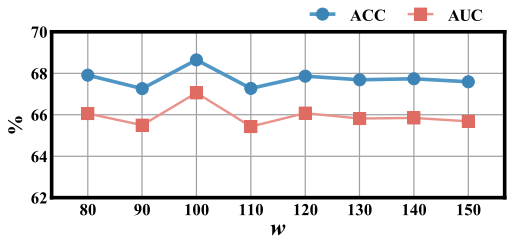
Multi-Scale Temporal Attention employs multi-scale LSH attention to compute temporal dynamic characteristics. To validate the effectiveness of multi-scale temporal modeling, we directly integrate the multi-scale spatial representations into a unified feature representation for temporal attention computation, denoted as "w/o MS-Temp". As observed, there is a decrease in performance. Furthermore, we visualize the temporal relationships information at different scales, as shown in Fig.3, revealing disparities in temporal relationships across scales. Therefore, our MSSTAN can capture the diversity of these spatial-temporal patterns for enhanced representation. Additionally, we replace LSH attention with the original transformer for feature extraction in the temporal dimension, denoted as "w/o LSH-Att". LSH attention employs locality sensitive hashing to prioritize features that are closely situated in the feature space, reducing the computational complexity from $O\left(L^2\right)$ to $O(LlogL)$. This approach not only effectively decreases computational complexity but also contributes to the improvement in MDD classification. Furthermore, the utilization of LSH attention leads to a notable reduction in the number of parameters. The original Transformer model comprises 868,386 parameters, whereas the use of LSH attention reduces the parameter count to 376,986. This represents a reduction of nearly two-thirds, consequently enhancing computational efficiency.

E. Parameter Experiments

1) Dynamic Functional Connectome Construction Parameters: In the process of constructing dynamic functional connectome from input time series, the sliding window

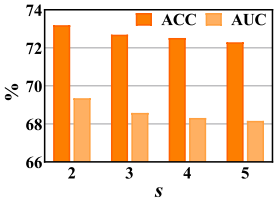


(a) zhongdaxinxiang

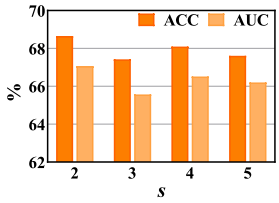


(b) Multi-MDD

Fig. 4. Results of different window lengths w on (a) zhongdaxinxiang and (b) Multi-MDD.



(a) zhongdaxinxiang



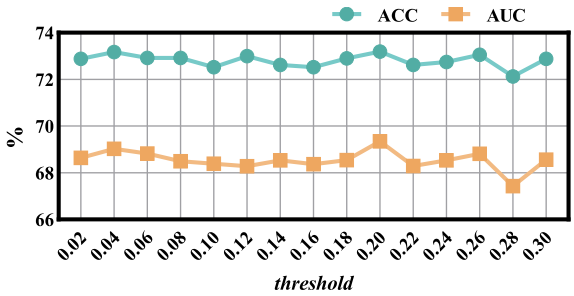
(b) Multi-MDD

Fig. 5. Results of different step sizes s on (a) zhongdaxinxiang and (b) Multi-MDD.

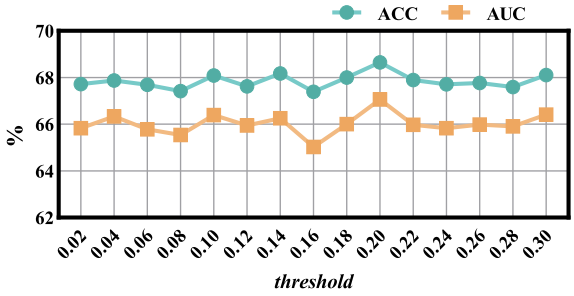
technique is employed. Two parameters, specifically the sliding window length w and the step size s, significantly impact the construction of dynamic graphs in this process. Moreover, proportional quantization $threshold$ is employed for generating the adjacency matrix in the construction of graph. Applying different $threshold$ produces diverse levels of sparsity in the adjacency matrix, indicating varying degrees of richness in the underlying brain connectivity information. Subsequent experiments are conducted on these three parameters.

a) The effect of sliding window length $\boldsymbol{w}$*:* The sliding window primarily defines the length of the average time series of BOLD signals for each graph. A larger w allows for a longer time series within each window. Correspondingly, the length of the graph sequences involved in dynamic graph construction becomes shorter. To achieve a better balance between window size and sequence length for better performance, we experimented with window sizes ranging from 80 to 150 with the interval as 10. The results are illustrated in Fig. 4, indicating a trend of initially increasing followed by decreasing. Overall, the relatively optimal window size is found to be 100. Utilizing a sliding window of this size to slide on the time series ensures that each window captures appropriate spatial representations while maintaining an optimal length for the graph sequence. It indicates that the balance between spatial representations and temporal features is crucial for preserving the integrity of the data and enhancing performance.

b) The effect of step size $\boldsymbol{s}$*:* The step size s also significantly impacts the performance of the MSSTAN. Different step



(a) zhongdaxinxiang



(b) Multi-MDD

Fig. 6. Results of different proportional quantization *threshold* on (a) zhongdaxinxiang and (b) Multi-MDD.

sizes s are evaluated in our experiments, and the results are presented in Fig. 5. It can be observed that as the step size increases, the model's performance decreases. Notably, the results indicate that a step size of 2 yields the most effective outcome. A shorter step size will make the length of the graph sequence longer, which can facilitate the extraction of multi-scale temporal features. By capturing more temporal dynamics across different scales, the model gains a richer understanding of the underlying dynamics of the connectome, ultimately contributing to enhanced performance and accuracy.

c) The effect of proportional quantization ***threshold****:* When constructing dynamic graphs, a proportional quantization $threshold$ is used to generate the adjacency matrix. The choice of different thresholds leads to different levels of sparsity in the adjacency matrix, reflecting variations in the richness of brain connectivity information. Given the inherently sparse nature of the brain connectome, the $threshold$ in this experiment ranged from 0.02 to 0.30, with an interval of 0.02, to select the optimal $threshold$. As illustrated in Fig. 6, it is determined that the overall performance remains relatively stable, with the most optimal results achieved when the threshold value is 0.2. It can be attributed to the balance it strikes between capturing sufficient connectivity information and maintaining an appropriate level of sparsity in the adjacency matrix.

2) Model Parameters: For model parameters, this section primarily investigates the impact of the number of pooling layers and pooling ratios of the multi-scale pooling module, as well as the influence of the number of attention heads of the topology enhanced graph transformer module.

a) The impact of the multi-scale pooling parameters: We conducted experiments with pooling layers set at 2 and 3, using pooling ratios with intervals of 0.2. The experimental results are shown in Fig. 7. It can be seen that the optimal performance is achieved when employing three layers of

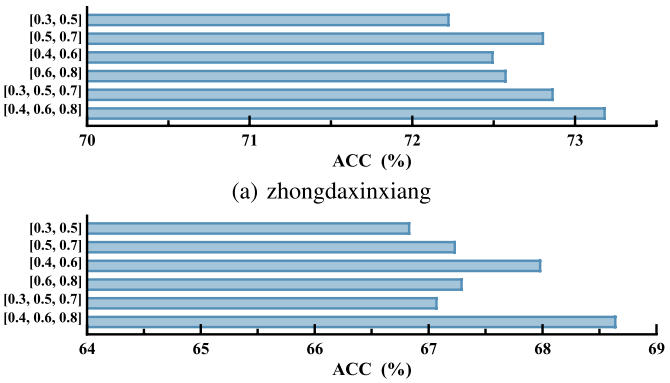


Fig. 7. Performance of different pooling parameters on (a) zhongdaxinxiang and (b) Multi-MDD.

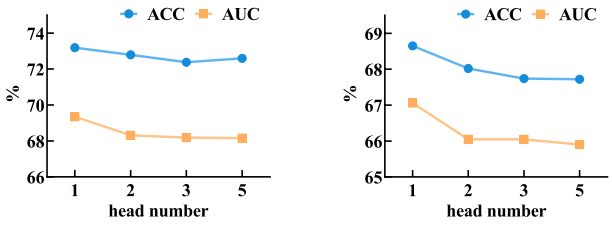


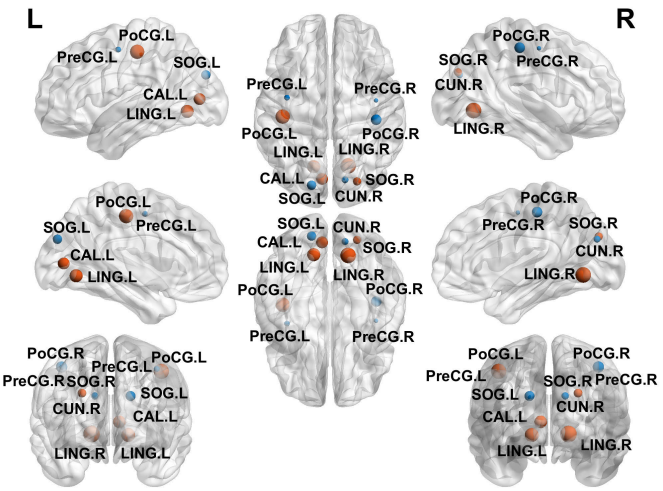
Fig. 8. Accuracy across various attention heads on (a) zhongdaxinxiang and (b) Multi-MDD.

multi-scale pooling with ratios [0.4, 0.6, 0.8]. One plausible explanation lies in the better ability to capture complementary information at different scales with the chosen pooling ratios. By incorporating pooling ratios at 0.4, 0.6, and 0.8, MSSTAN effectively captures and fuses richer spatiotemporal information at different levels of scales within the brain connectome, thereby enriching the model's representation capacity and enhancing its overall performance.

b) The impact of attention heads: To further investigate how varying attention head numbers influence model performance, we compared performance across different attention head numbers, including 1, 2, 3, and 5. As illustrated in Fig. 8, the model performs best when the number of attention heads is set to 1. This can be attributed to the trade-off between model complexity and performance. A single-head attention model inherently possesses fewer parameters compared to models with multiple attention heads. Consequently, it offers a simpler architecture, which may help prevent overfitting and lead to a more generalized model fit. In contrast, models with multiple attention heads may introduce redundancy or noise due to their increased complexity, potentially reducing the effectiveness of the attention mechanism.

F. Attention Analysis for MDD Diagnosis

For the functional connectome classification task intending to assist diagnosis, high accuracy is not the only goal, but also requires the model to be interpretable. With the help of the self-attention mechanism in the topology enhanced graph Transformer, we obtain the weights accounted for brain regions and functional connectivities (FCs). Fig. 9 (a)(b) and Fig. 10 (a)(b) present the top ten rated brain regions and



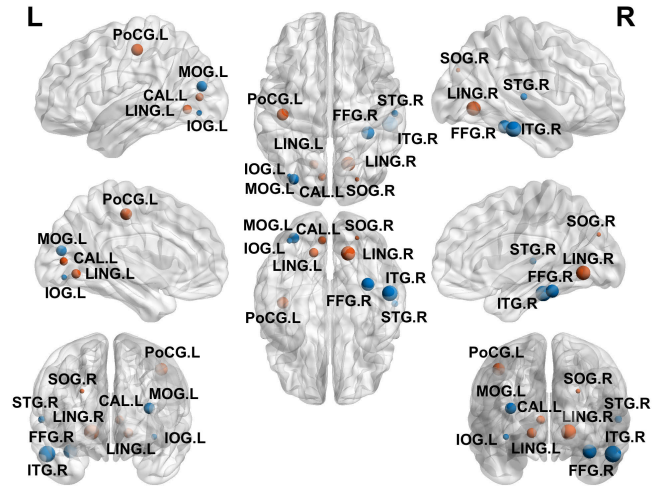
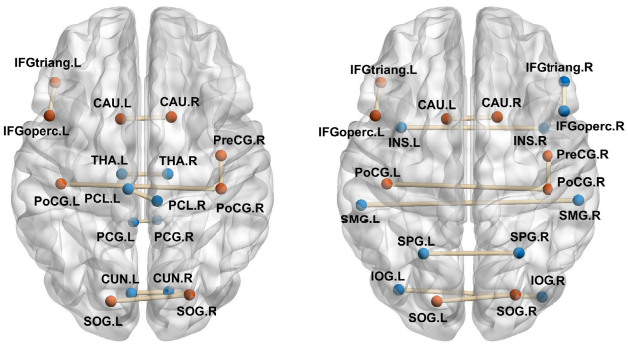


Fig. 9. Top ten discriminative brain regions on (a) zhongdaxinxiang and (b) Multi-MDD in MDD diagnosis, with red indicating consistent brain regions and blue indicating specific brain regions.

FCs respectively on two datasets, zhongdaxinxiang and Multi-MDD.

From Fig. 9 (a)(b), it is evident that among the ten highly discriminative brain regions obtained from two datasets of MDD, five of them are consistent. These regions include the right lingual gyrus, left postcentral gyrus, left lingual gyrus, left calcarine fissure and surrounding cortex, and right superior occipital gyrus, highlighting the consistency of the output results. Among them, the lingual gyrus is a part of the visual recognition network and plays a role in word processing and facial perception [47]. Studies have indicated a reduction in activity in the visual areas of MDD patients, such as occipital lobe, lingual gyrus, and fusiform gyrus, providing objective confirmation of the visual system dysfunction as a potential primary feature of MDD [48]. The postcentral gyrus is a major sensorimotor cortex, serving as a crucial hub in the auditory and sensorimotor networks [49]. This aberrant functionality in the postcentral gyrus affects the auditory and sensorimotor networks, potentially leading to perceptual impairments in MDD patients [50]. The occipital cortex is involved in integrating information into visual working memory. Yan et al. [51]



(a) zhongdaxinxiang (b) Multi-MDD

Fig. 10. Top ten discriminative FCs on (a) zhongdaxinxiang and (b) Multi-MDD in MDD diagnosis, with red indicating consistent brain regions and blue indicating specific brain regions.

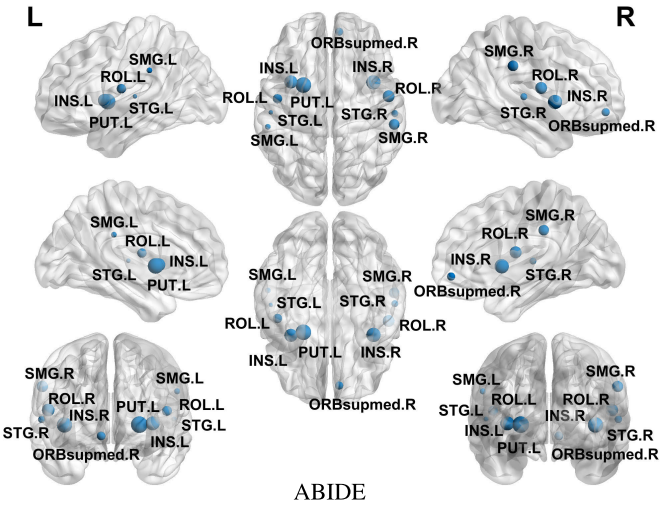
discovered a decrease in Regional Homogeneity (ReHo) in the right superior occipital gyrus and middle occipital gyrus of MDD patients, which also corroborates that the abnormal activity in the right superior occipital gyrus might be a stable and distinctive biological marker of MDD.

We further visualize the top 10 discriminative FCs identified by MSSTAN in Fig. 10 (a)(b). Among them, changes in symmetric inter-hemispheric functional connections are observed in brain regions such as the caudate nucleus (CAU.L—-CAU.R), postcentral gyrus (PoCG.L—-PoCG.R), and occipital gyri (SOG.L—-SOG.R), indicating disruptions or abnormal exchanges in information pathways related to emotion processing, motor control, and self-awareness processing. This inter-hemispheric dysregulation may represent a potential biological feature of patients with MDD [52]. Functional connectivity changes observed in the left inferior frontal gyrus (IFGtriang.L—-IFGoperc.L) of the Default mode network (DMN) network can still serve as a primary target for understanding the pathophysiology of MDD, closely related to the etiology of MDD [43]. Additionally, the consistent finding of altered functional connectivity between the precentral gyrus (PreCG.R) and postcentral gyrus (PoCG.R) reflects abnormal cooperation in the brain's sensorimotor functions, which may reveal related pathological changes in MDD patients.

Additionally, due to the heterogeneity of MDD itself [53] and the demographic differences between the two datasets, we also observed some changes in potential pathological mechanisms that might be related to symptoms. In the Multi-MDD dataset, we identified functional abnormalities in the right inferior temporal gyrus (ITG.R) and abnormal connectivity between the left and right insula (INS.L-INS.R), which are associated with suicidal tendencies [54], [55]. However, this may not be prominent in the zhongdaxinxiang dataset, as suicidal tendencies are not evident. This also suggests that the discriminative brain regions identified by MSSTAN in specific populations can provide new biomarker insights into the symptom heterogeneity of MDD.

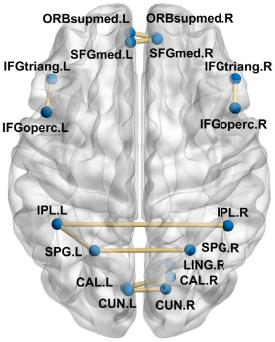
G. Attention Analysis for ASD Diagnosis

For the ASD diagnosis task, we also visualize significant discriminative brain regions and FCs, as shown in Fig. 11 and Fig. 12.



ABIDE

Fig. 11. Top ten discriminative brain regions on ABIDE in ASD diagnosis.



ABIDE

Fig. 12. Top ten discriminative FCs on ABIDE in ASD diagnosis.

Among these, as shown in Fig. 11, the putamen (PUT.L), which exhibits notable differences, is responsible for cognitive behaviors such as motor control and reward learning. Existing research has identified morphological changes in the putamen of ASD [56], and our study reveals its functional neural alterations, which further elucidate the pathogenesis of ASD. The insula (INS.R, INS.L) is involved in subjective body representation, while the frontal lobe (ORBsupmed.R), temporal lobe (STG.R, STG.L), parietal lobe (SMG.R, SMG.L) and Rolandic operculum (ROL.R, ROL.L) are related to sensory, motor, and language networks. Abnormal functional activation in these regions leads to deficits in corresponding functions in ASD patients and is closely related to the severity of ASD [57].

From Fig. 12, it can be observed that abnormal functional connectivities are found between the superior frontal gyrus (SFGmed.L-SFGmed.R, ORBsupmed.L-ORBsupmed.R) and the inferior frontal gyrus (IFGoperc.L-IFGtriang.L, IFGoperc.R-IFGtriang.R) located in the prefrontal cortex, which reflects irregularities in the regulation of self-related and external information processing in ASD patients, leading to certain social impairment symptoms [58]. The pericalcarine cortex and surrounding areas, such as the cuneus and lingual gyrus, are responsible for visual signal processing. Abnormal activity in this local network (CAL.L-CAL.R, CUN.L-CUN.R, CAL.R-LING.R) may be the potential

reason for the poor performance of ASD patients in tasks related to visual perception. The parietal lobe plays a crucial role in processing various sensory and perceptual information [59]. The abnormal functional connectivities in SPG.L-SPG.R, IPL.L-IPL.R, and SPG.L-IPL.L might explain the neural mechanism of the decreased eye contact response ability accompanied by the decreased social ability of ASD patients [60].

V. Discussion and Future Work

In this study, we evaluate the performance of the proposed MSSTAN using ten-fold cross-validation and leave-one-site-out validation, and further enhance the interpretability of the model by leveraging the attention mechanism of the transformer to analyze discriminative brain regions and FCs. However, there are still some limitations that need further exploration. Firstly, due to restrictions such as data collection and availability, the samples used in the experiments for MDD diagnosis are concentrated in China, which may limit the evaluation of different racial groups [61]. Secondly, this study explores the abnormal brain function based on fMRI, whereas DTI or T1 modalities could also provide useful structural information. Future research will explore the fusion representation of multimodal data to enrich the comprehensive understanding of the brain. Lastly, considering practical application scenarios, incorporating additional validation methods such as hierarchical validation in the future could enhance the robustness of the model. Additionally, strategies like federated learning could be considered to protect data privacy [62].

VI. Conclusion

In this work, we propose a novel framework called Multi-Scale Spatial-Temporal Attention Networks (MSSTAN) for functional connectome classification. In order to comprehensively model the spatial characteristics of brain functional connectomes and make full use of the connection relationships between brain regions, we propose a Topology Enhanced Graph Transformer module to facilitate the learning of the spatial features of brain regions by incorporating topological guidance in attention calculations. Furthermore, a Multi-Scale Pooling Strategy is proposed to fully capture brain connectome representations in different subspaces while retaining significant brain regions. To address temporal dependencies, we present a Multi-Scale Temporal Attention module to enhance the performance and computational efficiency through the use of LSH, which efficiently captures the dynamic characteristics of functional connectome in multi-scale spaces. Extensive experiments demonstrate the superiority of our proposed MSSTAN. At the same time, MSSTAN is interpretable, providing valuable insights for diagnostic assistance.

Acknowledgment

Members of the REST-Meta-MDD Consortium: Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, Chang Cheng, Yu-Qi Cheng, Xi-Long Cui, Jia Duan, Yi-Ru Fang, Qi-Yong Gong, Wen-Bin Guo, Zheng-Hua Hou, Lan Hu, Li Kuang, Feng Li, Kai-Ming Li, Tao Li, Yan-Song Liu, Zhe-Ning Liu, Yi-Cheng Long, Qing-Hua Luo, Hua-Qing Meng, Dai-Hui Peng, Hai-Tang Qiu, Jiang Qiu, Yue-Di Shen, Yu-Shu Shi, Chuan-Yue Wang, Fei Wang, Kai Wang, Li Wang, Xiang Wang, Ying Wang, Xiao-Ping Wu, Xin-Ran Wu, Chun-Ming Xie, Guang-Rong Xie, Hai-Yan Xie, Peng Xie, Xiu-Feng Xu, Hong Yang, Jian Yang, Jia-Shu Yao, Shu-Qiao Yao, Ying-Ying Yin, Yong-Gui Yuan, Ai-Xia Zhang, Hong Zhang, Ke-Rang Zhang, Lei Zhang, Zhi-Jun Zhang, Ru-Bai Zhou, Yi-Ting Zhou, Jun-Juan Zhu, Chao-Jie Zou, Tian-Mei Si, Xi-Nian Zu, Jing-Ping Zhao, and Yu-Feng Zang.

References

[1] R. M. Gibbs et al., "Toward precision medicine for neurological and neuropsychiatric disorders," *Cell Stem Cell*, vol. 23, no. 1, pp. 21–24, Jul. 2018.
[2] W. Marx et al., "Major depressive disorder," *Nature Rev. Disease Primers*, vol. 9, no. 1, p. 44, 2023.
[3] C. Lord et al., "Autism spectrum disorder," *Nature Rev. Disease Primers*, vol. 6, no. 1, pp. 1–23, 2020.
[4] J. Yang et al., "Landscapes of bacterial and metabolic signatures and their interaction in major depressive disorders," *Sci. Adv.*, vol. 6, no. 49, 2020, Art. no. eaba8555.
[5] L. Fusar-Poli, N. Brondino, P. Politi, and E. Aguglia, "Missed diagnoses and misdiagnoses of adults with autism spectrum disorder," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 272, no. 2, pp. 187–198, 2022.
[6] Y. Zhang et al., "Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography," *Nature Biomed. Eng.*, vol. 5, no. 4, pp. 309–323, Oct. 2020.
[7] C. Hohenfeld, C. J. Werner, and K. Reetz, "Resting-state connectivity in neurodegenerative disorders: Is there potential for an imaging biomarker?" *NeuroImage: Clin.*, vol. 18, pp. 849–870, Jan. 2018.
[8] X. Yang, M. S. Islam, and A. A. Khaled, "Functional connectivity magnetic resonance imaging classification of autism spectrum disorder using the multisite ABIDE dataset," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, May 2019, pp. 1–4.
[9] C. Kamarajan et al., "Random forest classification of alcohol use disorder using fMRI functional connectivity, neuropsychological functioning, and impulsivity measures," *Brain Sci.*, vol. 10, no. 2, p. 115, Feb. 2020.
[10] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *Frontiers Neurosci.*, vol. 14, p. 779, Oct. 2020.
[11] Y. Li, J. Liu, Z. Tang, and B. Lei, "Deep spatial–temporal feature fusion from adaptive dynamic functional connectivity for MCI identification," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2818–2830, Sep. 2020.
[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[13] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=ryGs6iA5Km
[14] B.-H. Kim and J. C. Ye, "Understanding graph isomorphism network for rs-fMRI functional connectivity analysis," *Frontiers Neurosci.*, vol. 14, p. 630, Jun. 2020.
[15] 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.
[16] B. Lei et al., "Multi-scale enhanced graph convolutional network for mild cognitive impairment detection," *Pattern Recognit.*, vol. 134, Feb. 2023, Art. no. 109106.
[17] R. Hindriks et al., "Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI?" *NeuroImage*, vol. 127, pp. 242–256, Feb. 2016.
[18] Y. Kong et al., "Spatio-temporal graph convolutional network for diagnosis and treatment response prediction of major depressive disorder from functional connectivity," *Human Brain Mapping*, vol. 42, no. 12, pp. 3922–3933, Aug. 2021.
[19] X. Xing et al., "DS-GCNs: Connectome classification using dynamic spectral graph convolution networks with assistant task training," *Cerebral Cortex*, vol. 31, no. 2, pp. 1259–1269, Jan. 2021.
[20] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[21] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for NLP tasks," in *Proc. 15th Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2020, pp. 179–183.

[22] 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.*, vol. 34, 2021, pp. 4314–4327.

[23] D. Chen and L. Zhang, "FE-STGNN: Spatio-temporal graph neural network with functional and effective connectivity fusion for mci diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Vancouver, BC, Canada: Springer, 2023, pp. 67–76.

[24] Q. Wang et al., "Leveraging brain modularity prior for interpretable representation learning of fMRI," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 8, pp. 2391–2401, Aug. 2024.

[25] M. Liu, H. Zhang, F. Shi, and D. Shen, "Building dynamic hierarchical brain networks and capturing transient meta-states for early mild cognitive impairment diagnosis," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Strasbourg, France: Springer, 2021, pp. 574–583.

[26] J. Zhang, Y. Guo, L. Zhou, L. Wang, W. Wu, and D. Shen, "Constructing hierarchical attentive functional brain networks for early AD diagnosis," *Med. Image Anal.*, vol. 94, May 2024, Art. no. 103137.

[27] S. I. Ktena et al., "Distance metric learning using graph convolutional networks: Application to functional brain networks," in *Proc. 20th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Quebec City, QC, Canada: Springer, 2017, pp. 469–477.

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

[29] H. Cui, W. Dai, Y. Zhu, X. Li, L. He, and C. Yang, "Interpretable graph neural networks for connectome-based brain disorder analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Singapore: Springer, 2022, pp. 375–385.

[30] P. Zhang et al., "BrainUSL: Unsupervised graph structure learning for functional brain network analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Vancouver, BC, Canada: Springer, 2023, pp. 205–214.

[31] W. Wang, Y. Kong, Z. Hou, C. Yang, and Y. Yuan, "Spatio-temporal attention graph convolution network for functional connectome classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1486–1490.

[32] J. Lee, E. Kang, J. Maeng, and H.-I. Suk, "Eigendecomposition-based spatial–temporal attention for brain cognitive states identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 1921–1925.

[33] C.-G. Yan, X.-D. Wang, X.-N. Zuo, and Y.-F. Zang, "DPABI: Data processing & analysis for (resting-state) brain imaging," *Neuroinformatics*, vol. 14, no. 3, pp. 339–351, Jul. 2016.

[34] 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.

[35] 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.

[36] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, "Measuring and relieving the over-smoothing problem for graph neural networks from the topological view," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 3438–3445.

[37] Z. Wu, P. Jain, M. Wright, A. Mirhoseini, J. E. Gonzalez, and I. Stoica, "Representing long-range context for graph neural networks with global attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13266–13279.

[38] H. Cui et al., "BrainGB: A benchmark for brain network analysis with graph neural networks," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 493–506, Feb. 2023.

[39] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3734–3743.

[40] C. Liu, Y. Zhan, X. Ma, D. Tao, B. Du, and W. Hu, "Masked graph auto-encoder constrained graph pooling," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Grenoble, France: Springer, 2022, pp. 377–393.

[41] R. F. Betzel and D. S. Bassett, "Multi-scale brain networks," *NeuroImage*, vol. 160, pp. 73–83, Oct. 2017.

[42] N. Kitaev, L. Kaiser, and A. Levskaya, "ReFormer: The efficient transformer," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–12.

[43] C.-G. Yan et al., "Reduced default mode network functional connectivity in patients with recurrent major depressive disorder," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 18, pp. 9078–9083, 2019.

[44] A. Di Martino et al., "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[45] K. A. Garrison, D. Scheinost, E. S. Finn, X. Shen, and R. T. Constable, "The (in)stability of functional brain network measures across thresholds," *NeuroImage*, vol. 118, pp. 651–661, Sep. 2015.

[46] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, "Do we need zero training loss after achieving zero training error?" in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 4604–4614.

[47] D. Simeonova, R. Paunova, K. Stoyanova, A. Todeva-Radneva, S. Kandilarova, and D. Stoyanov, "Functional MRI correlates of stroop N-back test underpin the diagnosis of major depression," *J. Integrative Neurosci.*, vol. 21, no. 4, p. 113, 2022.

[48] X. Zhong, W. Pu, and S. Yao, "Functional alterations of fronto-limbic circuit and default mode network systems in first-episode, drug-Naïve patients with major depressive disorder: A meta-analysis of resting-state fMRI data," *J. Affect. Disorders*, vol. 206, pp. 280–286, Dec. 2016.

[49] Z. Yao, R. Yan, M. Wei, H. Tang, J. Qin, and Q. Lu, "Gender differences in brain activity and the relationship between brain activity and differences in prevalence rates between male and female major depressive disorder patients: A resting-state fMRI study," *Clin. Neurophysiol.*, vol. 125, no. 11, pp. 2232–2239, Nov. 2014.

[50] Y. Yuan et al., "Abnormal neural activity in the patients with remitted geriatric depression: A resting-state functional magnetic resonance imaging study," *J. Affect. Disorders*, vol. 111, nos. 2–3, pp. 145–152, 2008.

[51] M. Yan et al., "Disrupted regional homogeneity in melancholic and non-melancholic major depressive disorder at rest," *Frontiers Psychiatry*, vol. 12, Feb. 2021, Art. no. 618805.

[52] Z. Hou et al., "Prognostic value of imbalanced interhemispheric functional coordination in early therapeutic efficacy in major depressive disorder," *Psychiatry Res., Neuroimag.*, vol. 255, pp. 1–8, Sep. 2016.

[53] B. S. Jermy, K. P. Glanville, J. R. I. Coleman, C. M. Lewis, and E. Vassos, "Exploring the genetic heterogeneity in major depression across diagnostic criteria," *Mol. Psychiatry*, vol. 26, no. 12, pp. 7337–7345, Dec. 2021.

[54] C. Hu et al., "The amplitude of low-frequency fluctuation characteristics in depressed adolescents with suicide attempts: A resting-state fMRI study," *Frontiers Psychiatry*, vol. 14, Jul. 2023, Art. no. 1228260.

[55] L. Hu, M. Xiao, J. Cao, Z. Tan, M. Wang, and L. Kuang, "The association between insular subdivisions functional connectivity and suicide attempt in adolescents and young adults with major depressive disorder," *Brain Topography*, vol. 34, no. 3, pp. 297–305, May 2021.

[56] M. Schuetze, M. T. M. Park, I. Y. Cho, F. P. Macmaster, M. M. Chakravarty, and S. L. Bray, "Morphological alterations in the thalamus, striatum, and pallidum in autism spectrum disorder," *Neuropsychopharmacology*, vol. 41, no. 11, pp. 2627–2637, Oct. 2016.

[57] Y. Zhou, L. Shi, X. Cui, S. Wang, and X. Luo, "Functional connectivity of the caudal anterior cingulate cortex is decreased in autism," *PLoS ONE*, vol. 11, no. 3, Mar. 2016, Art. no. e0151879.

[58] B. Qin, L. Wang, J. Cai, T. Li, and Y. Zhang, "Functional brain networks in preschool children with autism spectrum disorders," *Frontiers Psychiatry*, vol. 13, Jul. 2022, Art. no. 896388.

[59] E. T. Rolls, Y. Zhou, W. Cheng, M. Gilson, G. Deco, and J. Feng, "Effective connectivity in autism," *Autism Res.*, vol. 13, no. 1, pp. 32–44, Jan. 2020.

[60] J. Hirsch et al., "'Neural correlates of eye contact and social function in autism spectrum disorder," *PLOS ONE*, vol. 17, no. 11, pp. 1–32, 2022, doi: 10.1371/journal.pone.0265798.

[61] M. M. Hasan, J. Phu, A. Sowmya, E. Meijering, and M. Kalloniatis, "Artificial intelligence in the diagnosis of glaucoma and neurodegenerative diseases," *Clin. Exp. Optometry*, vol. 107, no. 2, pp. 130–146, Feb. 2024.

[62] M. M. Hasan, C. N. Watling, and G. S. Larue, "Validation and interpretation of a multimodal drowsiness detection system using explainable machine learning," *Comput. Methods Programs Biomed.*, vol. 243, Jan. 2024, Art. no. 107925.