Spatio-Temporal Hypergraph Attention Networks for Brain Disease Analysis

Chaojun Li, Peiliang Gong, Shengrong Li, Chunwei Tian, Yinbo Yu, *Member, IEEE*, Ran Wang, *Senior Member, IEEE*, Daoqiang Zhang, *Senior Member, IEEE*, and Qi Zhu, *Member, IEEE*

***Abstract*—Functional brain connectivity networks capture complex relationships and temporal evolution between brain regions, which have become increasingly important for diagnosing neurological disorders. However, existing methods, which are primarily based on vector or graph representations, struggle to adequately characterize the intricate spatio-temporal topological architecture of functional brain networks. Additionally, they predominantly rely on data-driven paradigms and lack priors pertaining to cross-windows network interactions. To address these issues, we propose a spatio-temporal hypergraph attention network framework for brain network analysis. Specifically, we first propose a temporal attention network architecture embedded with temporal similarity-driven prior knowledge, which effectively extracts long-range dependency information from fMRI by combining multi-head self-attention mechanisms and cross-window temporal prior knowledge. Second, we design a hierarchical hypergraph generation module that fuses local and global brain topological information to achieve multi-scale modeling of high-order spatio-temporal structures. Additionally, the spatial attention network, developed based on transformer architecture, employs hypergraph message passing mechanisms to effectively construct multi-level spatial interaction relationships between brain regions. Finally, a multi-layer perceptron (MLP) is adopted for classification. Experiments on the ADNI and PD datasets demonstrate that our method outperforms several state-of-the-art approaches in diagnostic performance and provides discriminative graph features for relevant brain disease diagnosis.**

***Index Terms*—Brain network, brain disease diagnosis, attention mechanism, hypergraph, prior knowledge.**

Received 20 September 2025; revised 17 January 2026; accepted 22 February 2026. Date of publication 12 March 2026; date of current version 17 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62371234, Grant 62076129; in part by the Basic Research Program of the Bureau of Science and Technology under Grant ILF24001; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231438; in part by the National Key Research and Development Program of China under Grant 2023YFF1204803; and in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022842. The associate editor coordinating the review of this article and approving it for publication was Dr. Sicheng Zhao. *(Corresponding author: Qi Zhu.)*

Chaojun Li and Ran Wang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: lichaojun@nuaa.edu.cn; wangran@nuaa.edu.cn).

Peiliang Gong, Shengrong Li, Daoqiang Zhang, and Qi Zhu are with the College of Artificial Intelligence and the Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: plgong@nuaa.edu.cn; lisrong@nuaa.edu.cn; dqzhang@nuaa.edu.cn; zhuqinuaa@163.com).

Chunwei Tian is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: chunweitian@hit.edu.cn).

Yinbo Yu is with the College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with Shenzhen Research Institute, Northwestern Polytechnical University, Shenzhen 518057, China (e-mail: yinboyu@nuaa.edu.cn).

Digital Object Identifier 10.1109/TIP.2026.3671657

I. Introduction

In recent years, interdisciplinary research between brain science and artificial intelligence has opened new pathways for understanding complex brain functions and diagnosing neurological disorders. The brain can be modeled as a functional brain connectivity network, where nodes represent brain regions and edges represent functional interactions [1]. This network reveals brain information processing mechanisms and provides potential biomarkers for early diagnosis [2]. Studies show neurodegenerative diseases are associated with abnormal functional connectivity, as observed in the Alzheimer's Disease Neuroimaging Initiative (ADNI) [3] and Parkinson's disease (PD) [4]. Traditional diagnosis relies on physicians' experience and limited imaging, whereas deep learning automatically extracts important features, improving accuracy. However, given the profound complexity and marked heterogeneity of brain functional network atlases, it is crucial to develop robust functional brain network analysis approaches to facilitate the precise diagnosis of brain diseases.

Functional magnetic resonance imaging (fMRI) examines brain functional connectivity by tracking blood oxygen level-dependent (BOLD) signal variations across brain regions [5], and has become the primary imaging modality for constructing functional brain networks. Research on functional brain networks encompasses two main paradigms: static functional brain networks (SFBN) and dynamic functional brain networks (DFBN) [6]. SFBN analyzes connectivity at discrete time points using methods such as Pearson correlation; however, it assumes stable connectivity throughout the scanning process, thus neglecting temporal variability. In contrast, DFBN captures temporal changes in connectivity patterns, offering a framework to investigate the brain's spatio-temporal dynamics. Compared to SFBN, DFBN better reflects the brain's intrinsic dynamic nature by accounting for time-varying connectivity [7], [8]. Researchers typically construct DFBN using sliding window approaches, which segment fMRI time series into multiple windows and calculate inter-regional connectivity within each window.

Most prior brain network analyses have employed vector-based or graph-based learning for feature extraction. Ming and Kundu [9] integrated Bayesian inference with support vector machines (SVM) [10] for network classification. Hu et al. [11]

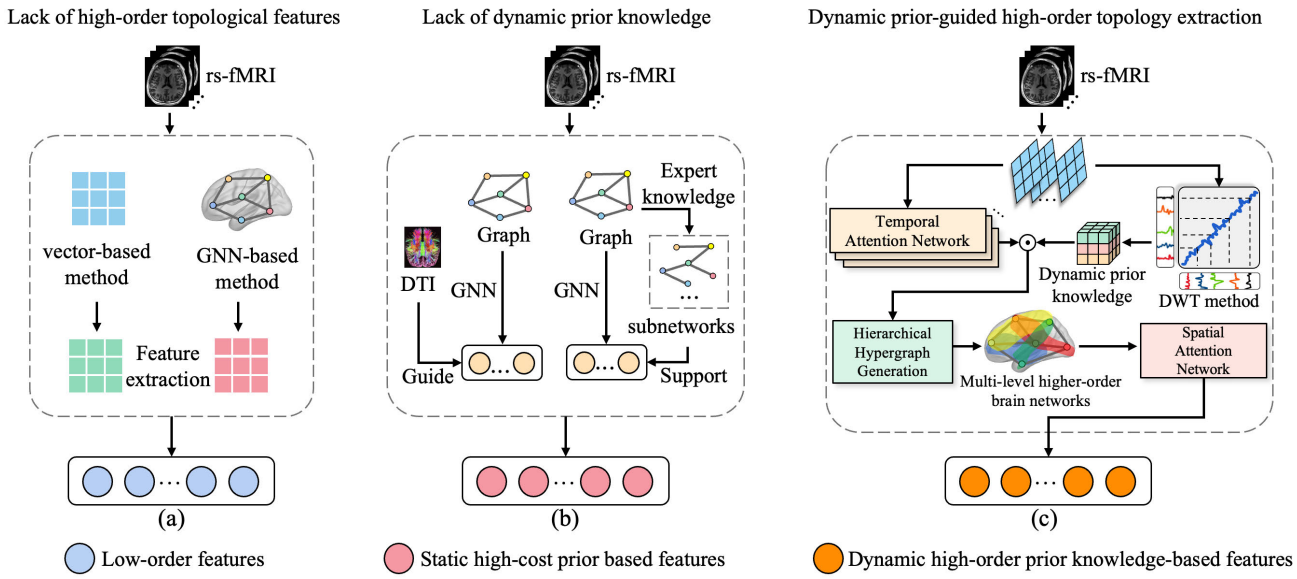


Fig. 1. Comparison between the proposed method and existing brain network feature extraction approaches: (a) vector-based and GNN-based methods, (b) static information and expert knowledge-based prior methods, and (c) our spatio-temporal hypergraph attention networks that construct hypergraph representations of brain networks, embedding dynamic prior knowledge for joint spatio-temporal feature extraction.

employed tensor component analysis to uncover latent network responses. Zhang et al. [12] employed graph convolutional networks (GCN) [13] for functional annotation of human cognitive states using fMRI data. However, vector-based approaches lose topological information, while graph methods only process binary relationships and inadequately capture high-order relationships. Hypergraph-based techniques better extract complex features by representing multi-node relationships. Xiao et al. [14] proposed multi-hypergraph learning with multi-paradigm fMRI data, improving classification. Wang et al. [15] developed a dynamic weighted hypergraph convolutional network enhancing functional connectome representations. However, the spatio-temporal structure of brain functional networks is highly coupled, and individual heterogeneity remains exceptionally high. Therefore, to enable precise diagnosis of brain diseases and the discovery of biomarkers, it is critical to develop more effective feature extraction methods for brain functional networks.

Furthermore, data-driven brain network analysis methods excel in prediction tasks but inadequately incorporate neuroscientific prior knowledge, limiting interpretability [16], [17]. Recent efforts extract prior constraints from fMRI data. Greaves et al. [18] developed structurally informed Bayesian models that use DTI-derived structural priors as regularizing factors for directed connectivity estimation, enhancing both accuracy and interpretability of brain network analysis. Wen et al. [19] developed a multi-view graph convolutional network integrating knowledge-guided structure learning with multi-task embeddings for neurological disorder classification. Li et al. [20] introduced spatio-temporal tensor graph convolutional networks with hierarchical representations and functional subnetwork constraints, yet still rely on predefined static functional parcellations that cannot adapt to individualized dynamic network variation patterns. Static priors fail to capture dynamic brain network changes across tasks or states. Current methods, despite using multi-modal data and expert priors, inadequately address nonlinear temporal interactions and long-range functional coupling phenomena.

To address these limitations, we propose a spatio-temporal hypergraph attention network for fMRI data capturing temporal prior knowledge and high-order brain network features. Specifically, we first embed temporal similarity-driven prior knowledge with a temporal attention network architecture, effectively extracting long-range dependency information from fMRI by fusing multi-head self-attention mechanisms and cross-window temporal prior knowledge. Next, we devise a hierarchical hypergraph generation module that integrates local and global brain topological attributes, thereby enabling multi-scale modeling of high-order spatio-temporal architectures inherent in functional networks. Moreover, the spatial attention network, built on a transformer framework, utilizes hypergraph message-passing paradigms to effectively map multi-level spatial interaction patterns across brain regions. Finally, multi-layer perceptron (MLP) is deployed to generate the diagnosis.

Figure 1 compares traditional brain network analysis with our approach. Figure 1(a) illustrates vector-based and graph-based methods that neglect topological structures or extract only binary relations, missing high-order information. Figure 1(b) depicts conventional methods incorporating static priors, inadequately representing dynamic interactions. Both approaches overlook dynamic prior knowledge and fail to extract high-order spatio-temporal features. Figure 1(c) presents our spatio-temporal hypergraph attention framework, providing a unified approach for extracting high-order spatio-temporal information while enhancing interpretability. In summary, our proposed method offers the following contributions:

1) We developed a novel spatio-temporal hypergraph attention network framework that effectively extracts high-order spatio-temporal features from fMRI data, significantly

advancing the characterization of functional brain connectivity networks.

2) We integrated cross-window temporal similarity-driven prior knowledge to guide model training, coupled with our temporal attention network architecture that captures complex long-range temporal dependencies and dynamic state transitions in fMRI signals.

3) The proposed hierarchical hypergraph generation module constructs multi-scale hypergraphs from both local and global topological perspectives, while our interpretable spatial attention network efficiently captures high-order connectivity patterns through hypergraph message passing mechanisms.

4) Comprehensive evaluations conducted on the ADNI and PD datasets demonstrate that our proposed approach outperforms existing state-of-the-art brain network analysis methods, providing discriminative graph features and revealing significant potential for identifying neuroimaging biomarkers associated with neurological disorders.

The remainder of the paper is delineated as follows: Section II provides a comprehensive review of the relevant literature. Section III introduces our novel spatio-temporal hypergraph attention networks. In Section IV, we detail the experimental materials and present compared results against existing methodologies for brain disease diagnosis. Section V offers an in-depth analysis and interpretation of the experimental findings. Finally, Section VI concludes with a summary of our work.

II. Related Work

A. Attention-Based Hypergraph Feature Extraction

Traditional brain network topological features are predominantly based on graph convolutional networks and graph attention networks (GAT) [21]. While these methods capture node relationships through neighborhood aggregation operations, they fail to effectively model high-order interaction characteristics inherent in hypergraphs. Consequently, hypergraph-based feature extraction methodologies have progressively been incorporated into brain network analysis. For example, Liu et al. [22] introduced a deep multi-template fusion approach based on spatio-temporal weighted multi-hypergraph convolutional networks, which substantially enhanced brain disease classification and diagnostic performance through the integration of multiple brain templates and implementation of weighting mechanisms. Concurrently, the self-attention mechanism of transformers, capable of capturing global characteristics, has generated significant interest in graph data analysis applications. Recent research attempts to incorporate transformer architectures into graph data analysis, leveraging their global properties to enhance functional connectivity modeling between brain regions [23]. However, existing approaches predominantly focus on traditional graph structures, while the integration of transformers within hypergraph frameworks remains challenging. This study introduces a hierarchical hypergraph generation methodology and an attention-based spatial network that dynamically models multi-level hypergraph interactions through multi-head attention mechanisms.

B. Self-Attention Mechanism

In recent years, transformer architecture has achieved remarkable advancements in deep learning due to its distinctive self-attention mechanism. Vaswani et al. [24] initially introduced this mechanism into neural networks, demonstrating exceptional performance in machine translation tasks. The fundamental advantage of the self-attention mechanism lies in its capacity to establish associations between heterogeneous data nodes through feature space projection, revealing extensive application potential in medical image analysis [25], [26]. When processing time-series signals such as fMRI, this mechanism can effectively model brain regions as data nodes with BOLD signals serving as node features, thereby revealing brain functional connectivity patterns by quantifying correlations between nodes. For topological structure data processing, researchers have proposed improvements to the traditional transformer architecture. Notably, He et al. [27] developed a spatio-temporal graph transformer approach for Alzheimer's disease diagnosis using resting-state fMRI, enhancing the extraction of topological features from brain networks. Since existing methods inadequately leverage the transformer architecture for extracting spatio-temporal features of brain networks, we designed a spatial attention mechanism as illustrated in Figure 3.

III. Proposed Method

In this section, we present a comprehensive overview of our proposed spatio-temporal hypergraph attention network (STHAN) for analyzing brain networks, with the framework architecture illustrated in Figure 2.

A. Temporal Similarity-Driven Prior Knowledge

In this study, we use fMRI signals as input data, recording neural activity across regions of interest (ROIs) during resting state, denoted as $X = (x_1, x_2, \ldots, x_N)^T \in \mathbb{R}^{N\times M}$, where $x_i \in \mathbb{R}^M$ represents the time series signal of the i^{th} brain region, N is the number of brain regions, and M is the total time points. To capture temporal dynamics, we segment the fMRI time series into T subsequences using non-overlapping sliding windows of length L. For the t^{th} window, the time series is $X^{(t)} = \left(x_1^{(t)}, x_2^{(t)}, \ldots, x_N^{(t)}\right)^T \in \mathbb{R}^{N\times L}$, where $x_i^{(t)} \in \mathbb{R}^L (t \in \{1, 2, \ldots, T\})$ is the i^{th} region's time series in the t^{th} window. To characterize dynamic patterns between time windows, we employ dynamic time warping (DTW) to calculate inter-window similarity. For the r^{th} brain region, with series $x_r^{(t_i)}$ and $x_r^{(t_j)}$ from time windows t_i and t_j, the DTW distance is:

$$d_r(t_i, t_j) = \min_{\pi} \sum_{k=1}^{|\pi|} \|x_r^{(t_i)}(p_k) - x_r^{(t_j)}(q_k)\|^2, \tag{1}$$

where $\pi = \{(p_k, q_k)\}_{k=1}^{|\pi|}$ represents the alignment path between time series, p_k, and q_k are time indices in this path, $|\pi|$ denotes the path length, and P represents all possible paths. DTW enables dynamic warping to identify the optimal alignment between time windows, thus accurately characterizing their similarity. To capture inter-window similarity, we compute a

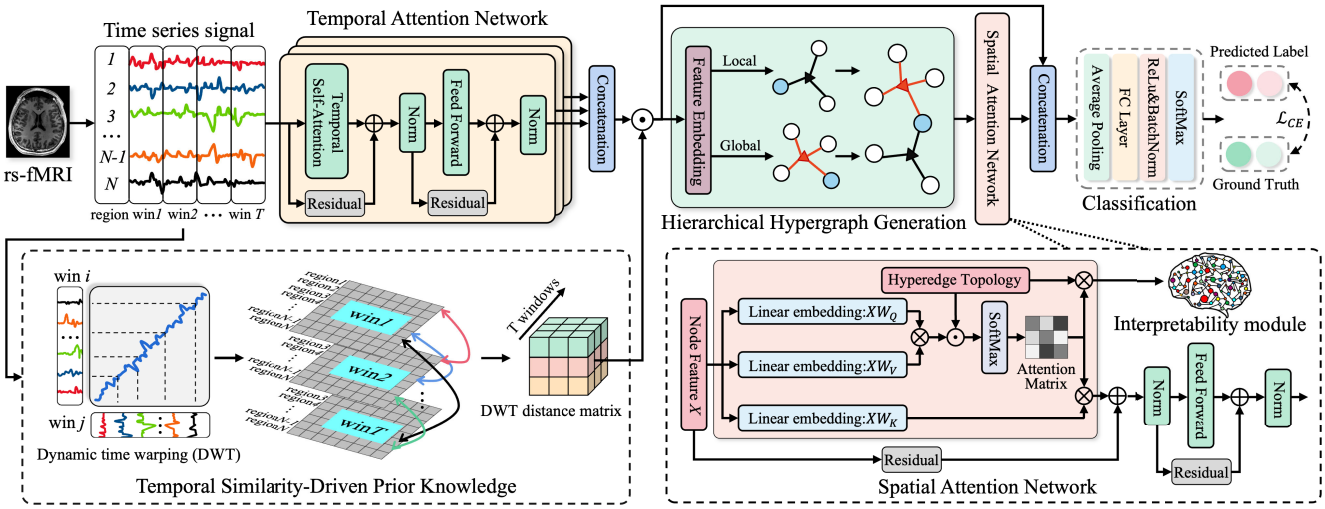


Fig. 2. The architecture of the proposed spatio-temporal hypergraph attention network.

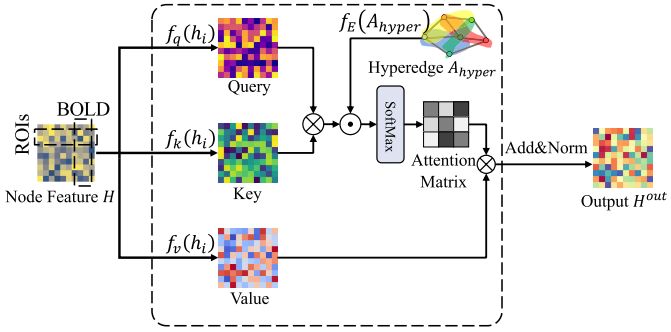


Fig. 3. Schematic diagram of the spatial attention network. Node features H are projected into query, key, and value spaces (f_q, f_k, f_v) and integrated with hyperedge embeddings $f_E(A_{hyper})$. The final output H^{out} is generated via residual connection and layer normalization.

distance matrix for each brain region. For the r^{th} brain region, its distance matrix is represented as:

$$D_r = [d_r(t_i, t_j)]_{1 \le i < j \le T}, \tag{2}$$

where $D_r \in \mathbb{R}^C$, and $C = \frac{T(T-1)}{2}$ represent the number of pairwise combinations between time windows. The temporal window similarity prior knowledge is represented as a distance matrix $D = (D_1, D_2, \ldots, D_N)^T \in \mathbb{R}^{N \times C}$. This similarity-based prior knowledge provides dynamic constraints for brain network modeling, enabling accurate representation of the dynamic functional connectivity between brain regions.

B. Temporal Attention Network

We developed a temporal attention network for fMRI dynamic feature extraction. The network processes input $X \in \mathbb{R}^{T \times N \times L}$ divided by sliding windows. For the t^{th} window, the series is $X^{(t)} = \left(x_1^{(t)}, x_2^{(t)}, \ldots, x_N^{(t)}\right)^T \in \mathbb{R}^{N \times L}$, with $x_i^{(t)} \in \mathbb{R}^L$ representing the i^{th} region's signal. Positional encoding preserves sequential information, with positions encoded from zero to the subsequence length, serving as input $Z^{(t)} = \left(z_1^{(t)}, z_2^{(t)}, \ldots, z_N^{(t)}\right)^T \in \mathbb{R}^{N \times L}$. A multi-head attention mechanism models global contextual relationships by mapping feature $z_i^{(t)}$ to query $(Q_i^{(t)})$, key $(K_i^{(t)})$, and value $(V_i^{(t)})$ feature spaces according to the following formulas:

$$Q_i^{(t)} = f_Q(z_i^{(t)}), K_i^{(t)} = f_K(z_i^{(t)}), V_i^{(t)} = f_V(z_i^{(t)}), \tag{3}$$

where f_Q, f_K, f_V are linear mapping functions that generate query, key, and value features, respectively, with weights shared across all brain regions. The similarity between queries and keys is calculated through dot products and normalized using the SoftMax function to obtain attention scores:

$$\alpha_i^{(t)} = softmax_j\left(\frac{Q_i^{(t)} \cdot (K_i^{(t)})^T}{\sqrt{d}}\right), \tag{4}$$

where d is a normalization parameter equal to the dimension of $K_i^{(t)}$, i.e., $d = L$. To enhance modeling capability, the temporal attention network employs a multi-head attention mechanism. By computing M attention heads in parallel, each head learns features from different subspaces. Outputs from all heads are concatenated and integrated into a global representation through linear mapping, as formulated below:

$$z_i^{temp} = f_O\left(\|_{m=1}^{M}\left(\alpha_{i,m}^{(t)} V_{i,m}^{(t)}\right)\right), \tag{5}$$

where M is the number of heads of attention, z_i^{temp} represents the attention layer output for the i^{th} region, $\|$ denotes the concatenation operation, and f_O is a linear mapping function. The network applies non-linear transformations through a feedforward neural network with residual connections to enhance training efficiency. This transforms each window's time series into $z_i^{final} \in \mathbb{R}^d$. Features from all windows are concatenated. For implementation with a batch size of B, the resulting feature matrix is denoted as $Z \in \mathbb{R}^{B \times N \times d}$, preserving temporal dynamics while integrating cross-regional patterns. To incorporate prior knowledge of time series similarity, we fuse the feature matrix Z with the prior knowledge matrix D through an element-wise product with broadcasting:

$$H = Z \odot (DW_D + b_D), \tag{6}$$

where $W_D \in \mathbb{R}^{N\times d}$ is the weight matrix, $b_D \in \mathbb{R}^d$ is the bias term, and $\odot$ denotes the element-wise product. Additionally, the feature matrix is $H \in \mathbb{R}^{B\times N\times d}$, where B represents the batch size, N denotes the number of brain regions, and d is the feature dimension. The resulting feature H captures long-range dependencies between time points and generates temporal representations with global dynamic characteristics.

C. Hierarchical Hypergraph Generation

We employ an implicit hypergraph construction strategy leveraging the feature matrix H to robustly model complex brain interactions. Unlike explicit methods reliant on fixed anatomical connections, this data-driven approach dynamically reveals latent functional dependencies, effectively filtering out spurious signals. To reflect the brain's small-world topology, we integrate local hyperedges constructed via the K-nearest neighbors algorithm to capture fine-grained functional specialization with global hyperedges generated via Louvain community detection to extract module-level integration. This dual-perspective design ensures a comprehensive representation of both micro-level details and macro-level network organization. We quantify inter-regional similarity by calculating the correlation matrix $R \in \mathbb{R}^{B\times N\times N}$ from H using normalized covariance, formulated as follows:

$$R_{ij} = \frac{\sum_{k=1}^{d}(H_{ik}-\mu_i)(H_{jk}-\mu_j)}{\sqrt{\sum_{k=1}^{d}(H_{ik}-\mu_i)^2}\cdot\sqrt{\sum_{k=1}^{d}(H_{jk}-\mu_j)^2+\varepsilon}}, \tag{7}$$

where μ_i and μ_j represent the feature means of nodes i and j, and $\boldsymbol{\varepsilon}$ is a regularization term for numerical stability. High values in the correlation matrix R indicate stronger functional similarity between brain regions, providing the foundation for subsequent hyperedge construction.

In the local hyperedge construction phase, we capture neighboring relationships between brain regions using the K-nearest neighbors algorithm based on the correlation matrix R. For each node i, we calculate its similarity with other nodes and select the K most similar neighbors to generate a local incidence matrix $A_{local} \in \mathbb{R}^{B\times N\times N}$, defined as follows:

$$A_{local}(i,j) = \begin{cases} 1, j \in N_k(i) \\ 0, otherwise \end{cases}, \tag{8}$$

where $N_k(i)$ represents the K-nearest neighbors of node i. This local incidence matrix captures neighboring characteristics between brain regions, reflecting local connectivity. In the global hyperedge construction phase, we mine global functional modules by calculating an adjacency matrix based on A_{local} and applying the Louvain community detection algorithm. For each batch, the algorithm divides regions into several functional communities, generating partitioning results $\rho = \{C_1, C_2, \ldots, C_Q\}$, where Q represents the number of communities. Based on these results, we construct a global incidence matrix $A_{global} \in \mathbb{R}^{B\times N\times Q}$, defined as follows:

$$A_{global}(i,c) = \begin{cases} 1, i \in comm(c) \\ 0, otherwise \end{cases}, \tag{9}$$

where $comm(c)$ denotes the community c. This global incidence matrix captures tight connectivity within functional modules. After completing both construction phases, we generate the final hypergraph incidence matrix $A_{hyper} \in \mathbb{R}^{B\times N\times(N+Q)}$ by concatenating the local and global incidence matrices:

$$A_{hyper} = concat(A_{local}, A_{global}), \tag{10}$$

where *concat* represents the concatenation. Matrix A_{hyper} integrates local connectivity and global modular information for comprehensive functional modeling. This hypergraph captures both neighboring relationships between brain regions and structural information of global functional modules. Local hyperedges reflect fine-grained connections, while global hyperedges reveal high-order functional characteristics, providing robust support for classification tasks and brain network analysis.

D. Spatial Attention Network

This study introduces a spatial attention network module integrating the node feature matrix H with the hyperedge feature matrix A_{hyper}, capturing the interactions of the global brain regions through a self-attention mechanism. The architecture of this module is illustrated in Figure 3. As shown in the schematic, the network projects node features into latent spaces while explicitly embedding the hypergraph structure A_{hyper} to guide the attention learning process. By incorporating hypergraph features for high-order structural information, it outperforms traditional graph attention networks. We map A_{hyper} to obtain hyperedge embedding features:

$$E_{hyper} = f_E(A_{hyper}), \tag{11}$$

where f_E is a linear mapping function, $E_{hyper} \in \mathbb{R}^{B\times N\times M'}$ denotes the hyperedge embedding feature, and M' is the dimension after linear transformation. Concurrently, we define the input feature matrix as $H = [h_1, h_2, \ldots, h_N] \in \mathbb{R}^{B\times N\times d}$, where $h_i \in \mathbb{R}^d$ represents the feature vector corresponding to the i^{th} brain region. These node features undergo transformations to generate the query, key, and value feature spaces:

$$q_i = f_q(h_i), k_i = f_k(h_i), v_i = f_v(h_i), \tag{12}$$

where f_q, f_k, f_v are linear mapping functions generating query, key, and value features, respectively. The attention weights between nodes incorporate both node feature relationships and hyperedge-derived structural information to model complex brain region interactions. The attention weight between node i and node j is defined as:

$$\alpha_{ij} = softmax_j\left(\frac{q_i\cdot k_j}{\sqrt{d}}\odot e_{ij}\right), \tag{13}$$

where $\sqrt{d}$ is a scaling factor, $\odot$ denotes element-wise multiplication, and e_{ij} represents the structural connectivity strength derived from the hypergraph incidence matrix E_{hyper}. Specifically, e_{ij} quantifies the topological correlation between brain regions i and j based on their shared hyperedges. This approach incorporates hypergraph structural priors directly into attention weight calculations, enhancing dynamic

modeling of inter-node relationships. The node feature update formula based on attention weight α_{ij} is:

$$h_i^{att} = f_o^h \left(\|_{m=1}^{M} \left(\sum_{j=1}^{N} \alpha_{ij}^m v_j^m \right) \right), \tag{14}$$

where $\|_{m=1}^{M}$ represents the concatenation of M attention heads, and f_o^h is a linear mapping function that fuses features from multiple attention heads, obtaining the updated feature h_i^{att} for node i. To enhance training stability and prevent gradient vanishing, the network incorporates residual connections and layer normalization operations after feature updates. The specific formulas are:

$$H_1 = Norm(H + H^{att}), \tag{15}$$

$$H_2 = W_2 \cdot \sigma(W_1 H_1), \tag{16}$$

$$H^{out} = Norm(H_1 + H_2), \tag{17}$$

where σ represents the activation function, W_1 and W_2 denote feed-forward network weight matrices, and *Norm* refers to layer normalization. These residual connections and normalization enable stable training while enhancing node feature expressiveness. The spatial attention network generates node features H^{out} with spatial interaction characteristics, which are fused with the temporal dynamic features H from the temporal attention network. The integrated features combine spatio-temporal dynamics with hypergraph relationships, effectively representing brain functional characteristics while enhancing classification performance. Algorithm 1 provides a summary of the proposed method in this work.

Algorithm 1 The Training Procedure of STHAN

Require: fMRI series $X \in \mathbb{R}^{N \times M}$, window length L, window number T, neighbors K, max epochs E_{max};
Ensure: Trained model parameters;
// **Phase I: Temporal Prior Knowledge Construction**
Calculate DTW distance matrix D_r for each region via Eq. (1);
Construct prior knowledge matrix $D \in \mathbb{R}^{N \times C}$ by Eq. (2)
// **Phase II: Spatio-Temporal Joint Learning**
while $epoch < E_{max}$ **do**
  // **Temporal Attention Network**
  Extract temporal features Z via multi-head attention by Eq. (3)-(5);
  Embed prior knowledge by $H = Z \odot (DW_D + b_D)$;
  // **Hierarchical Hypergraph Generation**
  Calculate correlation matrix R based on H via Eq. (7);
  Construct local incidence A_{local} and global incidence A_{global} by Eq. (8) and Eq. (9);
  Generate unified hypergraph via $A_{hyper} = concat(A_{local}, A_{global})$;
  // **Spatial Attention Network**
  Compute hyperedge embedding by $E_{hyper} = f_E(A_{hyper})$;
  Update node features H^{out} using hypergraph attention α_{ij} via Eq. (12)-(14);
  Apply residuals and norm by Eq. (15)-(17);
  // **Optimization**
  Calculate loss $\mathcal{L}_{CE}$ between predicted label Y_{pred} and Ground Truth Y;
  Back-propagate to update model parameters;
end while
return Model parameter set;

IV. Experiment

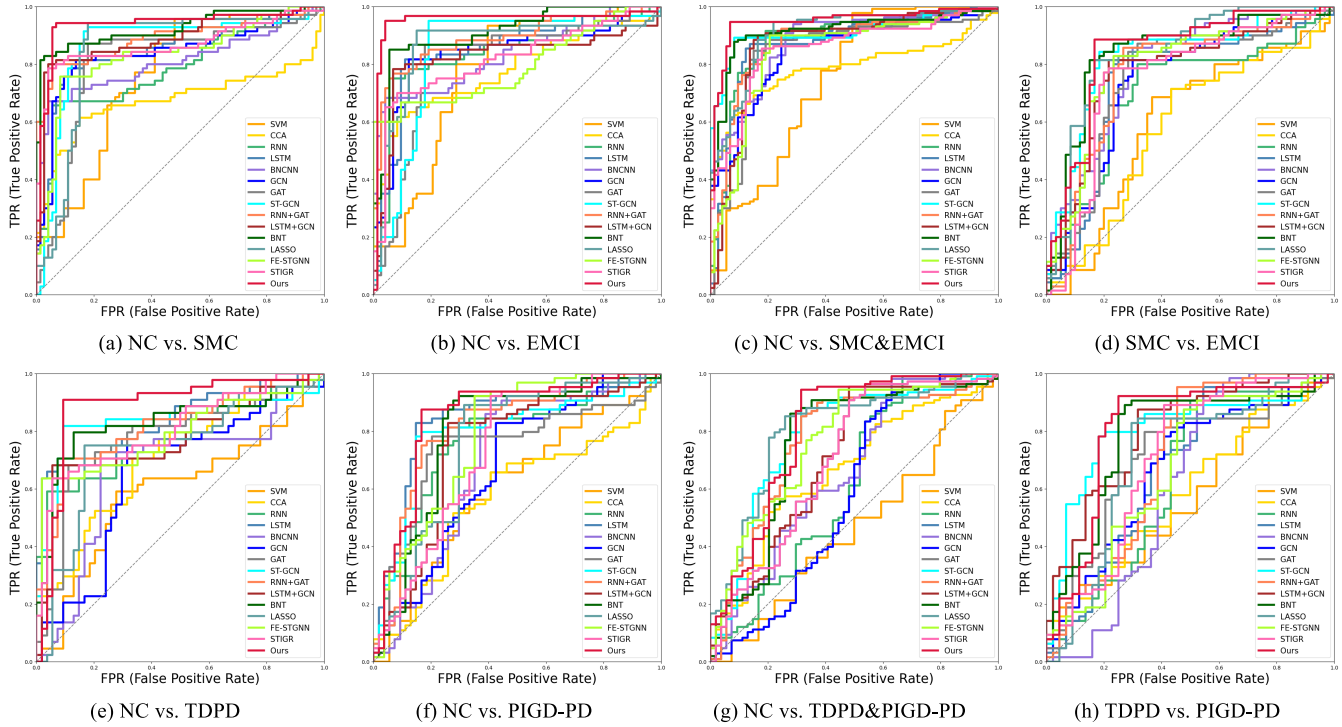
A. Materials

1) Data Acquisition: In this study, we utilize two datasets: the publicly available ADNI dataset (http://adni.loni.usc.edu/) and the PD dataset collected from the Affiliated Brain Hospital of Nanjing Medical University. The ADNI dataset includes fMRI data from 203 subjects, comprising 73 normal control (NC), 70 significant memory concerns (SMC) patients, and 60 early mild cognitive impairment (EMCI) patients. The PD dataset contains 162 samples, including 54 NC, 44 tremor-dominant Parkinson's disease (TDPD) patients, and 64 postural instability and gait disorder-Parkinson's disease (PIGD-PD) patients. The fMRI data from both datasets were collected using a Siemens MAGNETOM Verio 3.0T MRI scanner (repetition time = 2000 ms, echo time = 25 ms, and flip angle = 90°).

2) Data Preprocessing: We applied consistent preprocessing to both datasets using the DPARSF toolbox. The raw fMRI data first underwent slice timing correction to align acquisition times followed by realignment to correct for head motion. The functional images were then spatially normalized to the EPI template and resampled to a uniform voxel size of $3 \times 3 \times 3 \text{mm}^3$. To ensure robust signal quality, we implemented detailed denoising procedures where linear detrending was performed to reduce signal drift and a bandpass filter was applied to eliminate physiological noise. Furthermore, nuisance covariates including head motion parameters along with white matter and cerebrospinal fluid signals were regressed out to minimize non neuronal interference. Following preprocessing, we used the automated anatomical labeling atlas for brain segmentation. For the ADNI dataset, the data were segmented into 90 regions of interest with 197 time points. For the PD dataset, we divided the brain into 90 regions of interest with 220 time points.

B. Comparison Methods

To validate our method's effectiveness, we compared it with fourteen brain network analysis approaches, comprising three traditional machine learning methods and eleven deep learning methods. The traditional methods include support vector machine (SVM) [10], canonical correlation analysis (CCA) [28], and least absolute shrinkage and selection operator (LASSO) [34]. The deep learning methods encompass temporal feature extraction approaches, namely recurrent neural network (RNN) [29] and long short-term memory network (LSTM) [30]. Regarding spatial feature extraction, we selected brain network convolutional neural network (BrainNetCNN) [31], graph convolutional network (GCN) [13], and graph attention network (GAT) [21]. Additionally, we examined spatio-temporal feature extraction techniques, such as spatio-temporal graph convolutional network (ST-GCN) [32], brain



(a) NC vs. SMC (b) NC vs. EMCI (c) NC vs. SMC&EMCI (d) SMC vs. EMCI

(e) NC vs. TDPD (f) NC vs. PIGD-PD (g) NC vs. TDPD&PIGD-PD (h) TDPD vs. PIGD-PD

Fig. 4. ROC curves showing the performance of the proposed method and comparison approaches for brain disease diagnosis on the ADNI and PD datasets.

network transformer (BNT) [23], spatio-temporal graph neural networks with dynamic functional and effective connection (FE-STGNN) [35], and spatio-temporal interactive graph representation framework (STIGR) [36]. Finally, we compared our approach with two hybrid architectures, RNN+GAT [27] and LSTM+GCN [33], which sequentially process temporal and spatial features.

C. Implementation

The proposed method was implemented using Python and the PyTorch framework on an NVIDIA GeForce RTX 3080 GPU. We utilized a single STHAN layer to extract spatio-temporal features, incorporating parallel computing for DTW distance calculations. Specifically, the temporal attention network employed a 5-head self-attention mechanism, whereas the spatial attention network utilized a single-head architecture to capture hypergraph features. The output of STHAN was processed via average pooling and flattening, followed by a multilayer perceptron consisting of layers with 4096, 2048, and 512 neurons, culminating in SoftMax activation. Optimization was performed using the Adam optimizer with a learning rate of 5×10^{-4} and a weight decay of 1×10^{-4}. The model was trained for 300 epochs with a batch size of 20, a dropout rate of 0.5, and a batch normalization parameter of 1×10^{-5}. To facilitate reproducibility and further research, the complete source code is publicly available at https://github.com/xbrainnet/STHAN.git.

We evaluated performance on both datasets using 10-fold cross-validation, with nine subsets for training and one for testing. To strictly prevent data leakage, basic fMRI preprocessing was conducted independently for each subject. Subsequently, the data splitting for cross-validation was performed before any model training or hyperparameter optimization took place. This process was repeated ten times, reporting average metrics with standard deviations while preventing data leakage. Grid search optimized key hyperparameters, including sliding window count {2-10} and local hyperedge degrees {2-10}. We conducted four binary classification tasks for each dataset: NC vs. SMC, NC vs. EMCI, NC vs. SMC&EMCI, and SMC vs. EMCI for the ADNI dataset; and NC vs. TDPD, NC vs. PIGD-PD, NC vs. TDPD&PIGD-PD, and TDPD vs. PIGD-PD for the PD dataset. Performance was assessed using accuracy (ACC), area under the receiver operating characteristic curve (AUC), and F1-score (F1).

D. Classification Performance

Experimental results in Tables I and II show our method's superior performance, achieving accuracies of 93.71%, 95.60%, 94.10%, and 85.38% on the ADNI dataset tasks and 90.78%, 85.68%, 85.88%, and 85.09% on the PD dataset tasks, consistently outperforming all comparative methods. ROC analysis in Figure 4 confirms this superiority, with our method's curves optimally positioned in the upper left corner and showing the highest AUC values. This enhanced performance stems from our framework's ability to capture spatio-temporal dynamics while extracting multi-level high-order features with comprehensive prior information preservation. Unlike traditional approaches (SVM, CCA, LASSO), our method effectively models both temporal and topological characteristics. Compared to neural architectures (RNN, LSTM, BrainNetCNN, GCN, GAT), our temporal similarity-driven prior knowledge module enhances temporal

TABLE I
THE COMPARISON OF CLASSIFICATION PERFORMANCE BETWEEN THE PROPOSED METHOD AND OTHER APPROACHES ON THE ADNI DATASET (%)

Method	NC vs. SMC			NC vs. EMCI		
	ACC	AUC	F1	ACC	AUC	F1
SVM [10]	72.01/09.55	72.50/09.59	77.17/07.28	73.90/13.66	75.37/12.88	76.33/12.01
CCA [28]	70.71/10.58	66.86/13.38	62.96/15.14	75.93/09.55	78.92/12.10	68.98/14.40
RNN [29]	81.33/11.37	83.96/13.99	77.15/14.67	85.10/07.22	85.12/12.71	81.60/10.56
LSTM [30]	83.14/07.31	86.43/09.63	80.96/09.84	84.94/10.58	84.41/13.01	72.52/30.94
BrainNetCNN [31]	81.80/08.38	84.08/08.24	76.78/13.25	83.46/06.51	84.61/10.01	77.43/09.21
GCN [13]	83.14/09.71	79.45/15.03	78.23/17.23	84.94/12.21	81.96/16.23	82.63/12.74
GAT [21]	84.52/07.01	79.65/10.71	82.59/09.97	84.17/05.45	80.68/09.43	81.66/07.56
ST-GCN [32]	87.57/07.13	85.24/09.76	86.53/07.97	87.14/09.08	87.17/09.58	86.65/08.92
RNN+GAT [27]	86.71/08.01	87.34/11.13	82.69/11.39	85.60/13.01	86.83/14.49	81.76/16.57
LSTM+GCN [33]	87.57/07.82	87.97/11.26	83.96/12.01	86.53/07.87	79.75/16.15	75.19/26.65
BNT [23]	90.14/09.67	88.02/12.82	87.44/12.63	90.27/05.84	92.42/04.56	86.70/10.26
LASSO [34]	87.38/05.35	87.64/06.48	86.72/07.07	87.30/08.02	83.94/09.56	85.76/09.51
FE-STGNN [35]	84.05/08.06	86.34/09.71	81.42/10.98	82.09/08.61	79.69/14.43	66.48/26.30
STIGR [36]	87.38/09.48	89.43/10.01	85.52/08.36	82.03/07.47	83.16/13.52	70.52/25.17
STHAN	**93.71/06.62**	**94.82/07.12**	**92.70/06.89**	**95.60/07.08**	**94.92/07.37**	**94.56/07.42**
	NC vs. SMC&EMCI			SMC vs. EMCI		
SVM [10]	78.35/04.86	70.72/05.94	85.24/03.44	66.15/13.84	65.95/14.42	68.62/11.63
CCA [28]	74.42/09.98	79.94/11.63	77.04/11.74	64.61/10.98	64.04/11.62	68.32/09.06
RNN [29]	87.73/06.07	89.86/06.34	90.23/06.06	74.61/09.76	61.53/16.44	76.05/09.80
LSTM [30]	86.71/08.69	85.10/10.21	89.16/07.75	77.69/07.25	69.51/17.59	76.42/10.46
BrainNetCNN [31]	85.73/07.87	86.87/08.74	89.48/06.13	78.46/14.10	73.17/22.42	77.99/17.76
GCN [13]	83.76/07.45	83.22/11.96	87.12/05.88	75.38/08.97	70.01/18.01	67.22/28.17
GAT [21]	85.28/05.99	83.28/07.78	87.81/06.04	78.46/08.28	77.95/08.99	80.86/08.17
ST-GCN [32]	90.09/05.04	91.09/05.37	91.76/04.44	82.30/06.92	82.50/10.06	82.14/10.26
RNN+GAT [27]	86.19/07.69	88.52/07.83	88.65/07.64	79.23/06.01	72.01/10.16	80.07/07.86
LSTM+GCN [33]	87.64/05.53	87.92/06.39	89.84/04.83	80.01/09.85	71.90/18.47	73.58/26.23
BNT [23]	89.14/04.79	90.97/06.57	90.08/07.06	83.07/07.53	**84.70/10.21**	83.11/07.70
LASSO [34]	88.61/06.70	89.46/06.82	91.54/04.95	82.30/08.46	82.97/11.02	82.12/09.50
FE-STGNN [35]	85.67/04.71	85.83/08.52	88.90/03.62	79.23/10.91	75.30/13.80	80.27/12.77
STIGR [36]	84.24/07.66	81.82/12.11	84.87/12.96	78.46/08.97	72.71/19.75	73.12/25.67
STHAN	**94.10/04.32**	**94.07/04.84**	**95.29/03.21**	**85.38/08.03**	82.41/10.93	**85.96/07.26**

feature representation. Beyond existing spatio-temporal frameworks (ST-GCN, RNN+GAT, LSTM+GCN, BNT), our hierarchical hypergraph generation and spatial attention network explore multi-level high-order relationships within brain networks, yielding superior neurological disorder identification. Specifically, on the ADNI dataset, the significant accuracy improvement suggests that our hierarchical hypergraph structure effectively captures the complex, high-order functional deficits associated with Alzheimer's disease, which are often missed by pairwise connectivity models. On the PD dataset, the superior performance is primarily driven by the temporal similarity-driven prior knowledge, which enhances the model's sensitivity to the subtle temporal synchronization disruptions characteristic of Parkinson's disease pathology.

E. Ablation Study

To validate the performance of the three modules in our proposed method, including temporal similarity-driven prior knowledge (TSDPK), hierarchical hypergraph generation (HHG), and spatial attention network (SAN), we conducted ablation studies on both datasets. The results are presented in Table III. The baseline employed only the temporal attention network with MLP for classification. Our findings demonstrate the significant contribution of each module: (1) Removing TSDPK decreased accuracy by 5.43% to 8.38% on the ADNI tasks and 3.79% to 6.11% on the PD tasks, confirming its effectiveness in leveraging fMRI prior knowledge for temporal feature extraction; (2) Eliminating HHG reduced accuracy by 3.93% to 9.13% on the ADNI tasks and 3.64% to 5.66% on the PD tasks, validating its capacity to preserve multi-level high-order brain network features; and (3) Excluding SAN diminished accuracy by 3.85% to 5.52% on the ADNI tasks and 3.49% to 6.29% on the PD tasks, verifying its ability to efficiently extract high-order features from hypergraph brain networks. These results collectively substantiate the essential role of each component in the superior performance of our framework.

V. DISCUSSION

A. Analysis of Parameter Sensitivity

We investigated two critical hyperparameters: local hyperedge degrees (k) and sliding window count (T) using grid search optimization. For k in range {2-10}, classification accuracy exhibited a non-monotonic relationship as shown in Figure 5(a), initially increasing before reaching an optimum and subsequently declining. This pattern reflects a fundamental

TABLE II
THE COMPARISON OF CLASSIFICATION PERFORMANCE BETWEEN THE PROPOSED METHOD AND OTHER APPROACHES ON THE PD DATASET (%)

Method	NC vs. TDPD			NC vs. PIGD-PD		
	ACC	AUC	F1	ACC	AUC	F1
SVM [10]	66.33/10.89	66.41/13.01	60.52/13.00	62.65/07.30	62.52/06.70	64.18/11.23
CCA [28]	65.56/13.12	71.48/13.01	63.95/14.67	61.21/15.98	61.59/15.59	63.42/13.67
RNN [29]	79.33/12.64	78.15/15.01	66.03/27.82	75.37/07.22	72.84/09.65	77.19/07.98
LSTM [30]	81.67/14.31	81.64/14.05	71.92/22.07	82.95/10.08	81.84/14.18	81.53/13.41
BrainNetCNN [31]	75.66/06.29	64.67/13.03	69.70/11.70	73.78/09.61	66.67/20.75	77.63/07.73
GCN [13]	71.33/07.77	55.38/15.51	53.75/36.05	70.45/12.29	59.52/16.85	64.46/27.05
GAT [21]	79.44/09.37	78.28/09.51	70.40/16.92	78.93/08.36	73.54/11.54	78.35/10.82
ST-GCN [32]	85.88/07.95	85.13/12.23	83.41/08.72	81.43/07.14	78.42/07.01	81.73/07.51
RNN+GAT [27]	81.78/13.31	75.15/24.53	73.69/28.27	79.01/11.77	71.34/18.61	69.64/30.67
LSTM+GCN [33]	82.55/09.44	74.92/15.31	74.99/13.74	78.86/08.41	74.70/13.96	79.07/10.25
BNT [23]	83.99/09.16	82.69/11.35	80.72/12.01	82.12/07.18	**82.05/08.82**	83.09/08.64
LASSO [34]	79.55/12.06	73.85/19.73	76.34/14.17	78.03/07.46	74.66/12.67	75.82/09.76
FE-STGNN [35]	81.67/10.25	81.58/18.55	68.52/27.98	79.70/15.03	73.17/24.02	83.19/11.21
STIGR [36]	79.56/06.39	79.29/11.25	66.61/18.95	76.21/05.23	70.82/21.66	78.47/08.74
STHAN	**90.78/05.44**	**86.57/08.46**	**88.86/08.84**	**85.68/09.21**	81.04/18.54	**85.04/11.94**
	NC vs. TDPD&PIGD-PD			TDPD vs. PIGD-PD		
SVM [10]	57.35/10.76	48.37/10.15	69.21/10.09	60.18/9.22	53.07/09.04	73.22/06.44
CCA [28]	65.29/13.97	64.31/19.61	73.77/12.23	59.27/11.65	59.96/18.11	66.80/09.89
RNN [29]	77.13/09.64	63.78/07.65	84.67/7.67	74.18/09.72	60.15/25.86	76.17/17.17
LSTM [30]	79.59/07.99	64.55/15.94	85.92/05.92	75.01/11.12	59.37/21.78	78.85/12.39
BrainNetCNN [31]	77.75/05.03	68.76/09.95	85.31/03.67	72.99/13.14	60.78/17.06	79.64/11.61
GCN [13]	74.74/12.46	68.63/23.25	76.11/26.59	70.45/08.67	60.01/17.89	65.11/28.70
GAT [21]	79.63/04.82	74.84/09.21	84.80/03.89	74.18/06.44	72.36/10.83	78.60/04.18
ST-GCN [32]	80.29/04.39	77.30/09.78	84.59/05.81	79.72/10.81	78.03/14.39	80.77/10.11
RNN+GAT [27]	79.63/04.81	74.84/09.21	84.80/03.89	77.72/07.44	73.41/15.06	75.14/25.51
LSTM+GCN [33]	80.88/09.79	63.61/17.15	85.49/09.29	78.72/10.88	63.39/28.46	72.56/27.58
BNT [23]	81.39/06.40	78.10/11.06	85.65/05.90	81.45/10.15	77.56/12.02	83.22/11.77
LASSO [34]	81.54/08.18	77.87/09.23	76.12/08.72	77.63/09.82	61.19/20.03	66.67/14.12
FE-STGNN [35]	81.54/05.20	77.75/21.89	86.56/05.46	73.01/08.15	64.36/19.84	79.23/07.53
STIGR [36]	79.60/07.99	64.55/15.94	85.92/05.92	76.91/08.21	63.30/17.61	81.38/07.24
STHAN	**85.88/08.18**	**81.14/15.34**	**88.68/07.93**	**85.09/06.29**	**81.52/15.26**	**84.74/12.08**

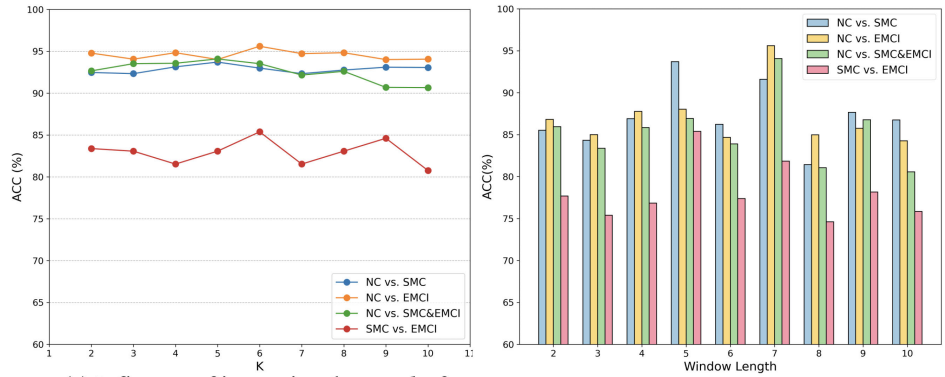
TABLE III
THE ABLATION STUDY RESULT OF GLOBAL AND LOCAL ATTENTION NETWORK(%)

Dataset	Method	NC vs. SMC			NC vs. EMCI			NC vs. SMC&EMCI			SMC vs. EMCI		
		ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
ADNI	baseline	81.85	83.17	82.02	83.45	90.27	81.03	84.21	86.87	86.92	74.62	72.48	73.41
	w/o TSDPK	86.80	89.10	82.28	87.22	93.28	85.95	88.67	90.11	90.77	78.46	75.49	77.51
	w/o HHG	87.42	90.92	87.56	86.47	92.53	84.48	90.17	90.65	91.43	80.76	80.34	82.47
	w/o SAN	88.19	87.70	86.58	90.22	93.83	88.89	89.71	91.21	90.96	81.53	76.79	81.59
	STHAN	**93.71**	**94.82**	**92.70**	**95.60**	**94.92**	**94.56**	**94.10**	**94.07**	**95.29**	**85.38**	**82.41**	**85.96**
		NC vs. TDPD			NC vs. PIGD-PD			NC vs. TDPD&PIGD-PD			TDPD vs. PIGD-PD		
PD	baseline	78.67	75.79	74.49	76.96	69.13	67.55	77.75	68.96	81.02	75.72	65.21	80.25
	w/o TSDPK	84.67	84.05	79.71	81.13	79.45	82.73	82.09	75.10	86.74	79.54	68.64	82.10
	w/o HHG	85.55	79.30	82.86	80.53	72.33	79.50	80.22	73.09	84.93	81.45	73.91	83.69
	w/o SAN	86.77	85.01	79.53	82.19	73.29	83.10	79.59	78.51	83.50	80.54	68.36	81.36
	STHAN	**90.78**	**86.57**	**88.86**	**85.68**	**81.04**	**85.04**	**85.88**	**81.14**	**88.68**	**85.09**	**81.52**	**84.74**

trade-off: insufficient degrees constrain hyperedge interactions and limit node relationship modeling, while excessive degrees introduce redundancy through overly dense connections. For T in range {2-10}, we identified task-dependent optimal values as illustrated in Figure 5(b): $T = 5$ for NC vs. SMC and SMC vs. EMCI tasks, and $T = 7$ for NC vs. EMCI and NC vs. SMC&EMCI tasks. Performance diminished at both extremes of T, as too few windows limited temporal ROI interactions, while excessive partitioning generated redundant information, both adversely affecting classification accuracy.

B. Visualization of Embedding Features

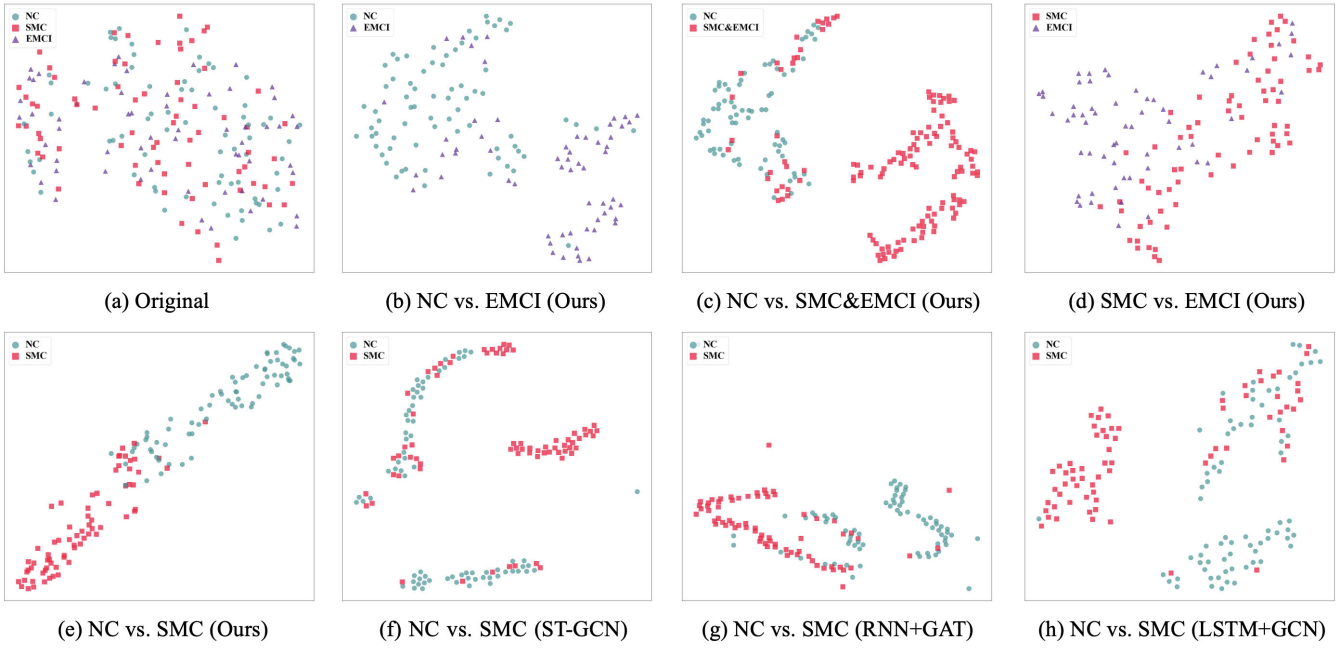
To validate our model's discriminative capability, we performed t-SNE visualization of features extracted from



(a) Influence of increasing degrees k of local hyperedge

(b) Influence of sliding window length T

Fig. 5. The influence of hyperparameters on different classification tasks on the ADNI dataset: (a) is the effect analysis of degrees of local hyperedge (from 2 to 10). (b) is the effect analysis of sliding window length (from 2 to 10).



(a) Original (b) NC vs. EMCI (Ours) (c) NC vs. SMC&EMCI (Ours) (d) SMC vs. EMCI (Ours)

(e) NC vs. SMC (Ours) (f) NC vs. SMC (ST-GCN) (g) NC vs. SMC (RNN+GAT) (h) NC vs. SMC (LSTM+GCN)

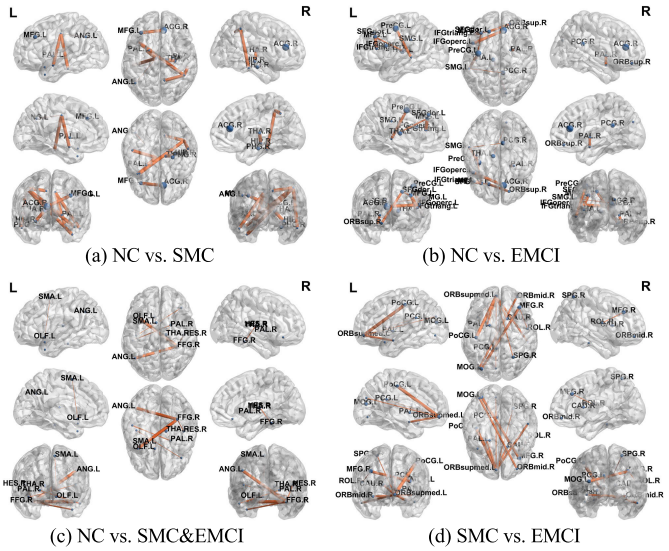
Fig. 6. T-SNE visualization of learned feature representations across different methods. (a) distribution of original samples in the representation space. (b)-(e) the representation spaces generated by our proposed method on multiple diagnostic tasks. (f)-(h) the representation spaces produced by compared methods on the NC vs. SMC classification task.

STHAN's final pre-softmax layer for the NC vs. SMC task on the ADNI dataset, as shown in figures 6(b)-(e). We also provide visualizations of original data and comparative methods in Figure 6(a) and 6(f)-(h), respectively. Analysis of Figure 6 revealed the following findings: (1) STHAN effectively clusters same-class samples while establishing clearer inter-class boundaries compared to raw data, (2) while healthy and patient groups demonstrated effective separation, SMC and EMCI classifications exhibited less distinct boundaries, potentially due to shared pathological brain regions across cognitive impairment levels, and (3) STHAN outperformed alternative spatio-temporal feature extraction methods in both intra-class cohesion and inter-class separation, corroborating our superior ACC and AUC metrics from experimental evaluations.

C. Discriminative and Interpretable Brain Connections

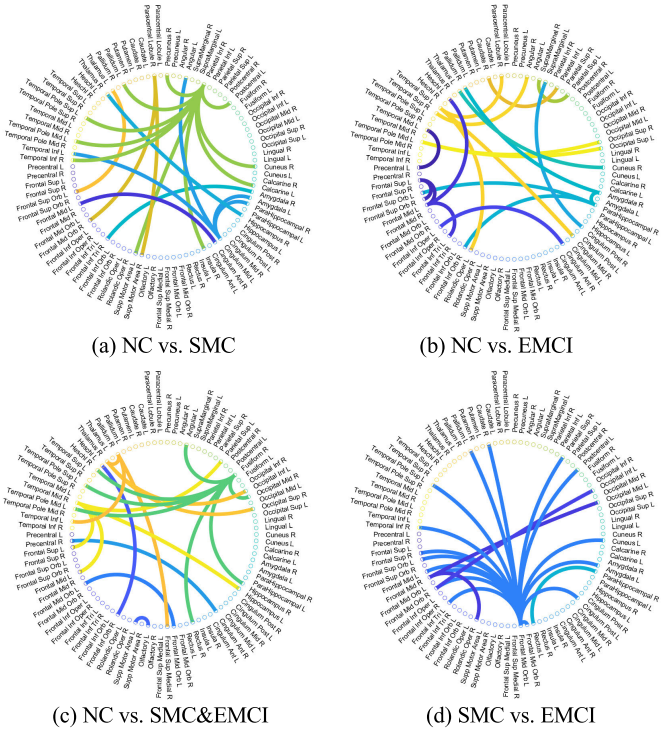
We evaluated our method's effectiveness in identifying interpretable and discriminative high-order brain connections for the ADNI. To enhance interpretability, we transformed hyperedge features through residual connections, layer normalization, and feed-forward neural networks after multiplying hyperedge topology and attention matrices in SAN. To assess discriminative capability, we employed non-negative elastic net regression to link transformed features with diagnostic labels [37]. Through 10-fold cross-validation, we identified the top 20 most discriminative connections, visualized in Figure 7 and 8.

In NC vs. SMC, significant connections concentrated in the Right Inferior Parietal Lobule, Left Mid Cingulate, and Right Paracentral Lobule. These neural substrates have been



(a) NC vs. SMC (b) NC vs. EMCI (c) NC vs. SMC&EMCI (d) SMC vs. EMCI

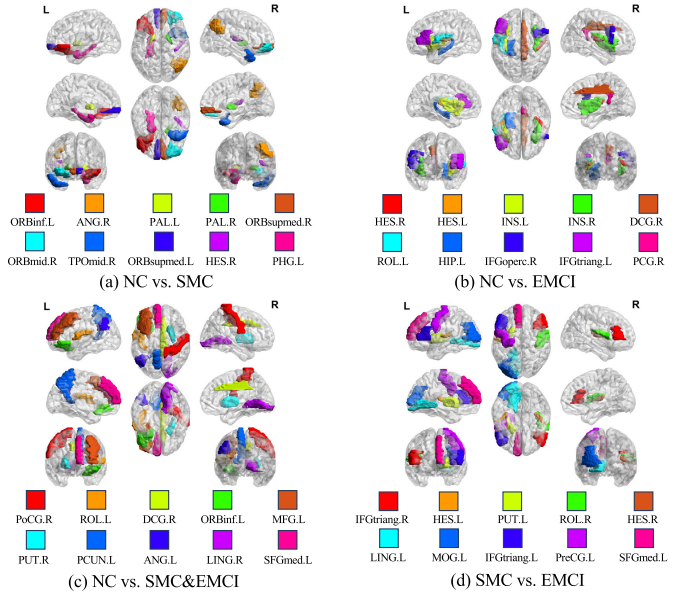
Fig. 7. Visualization of brain views highlighting important brain connections under different tasks: (a) NC vs. SMC, (b) NC vs. EMCI, (c) NC vs. SMC&EMCI, and (d) SMC vs. EMCI.



(a) NC vs. SMC (b) NC vs. EMCI (c) NC vs. SMC&EMCI (d) SMC vs. EMCI

Fig. 8. The top 20 most crucial brain connections identified across multiple classification tasks.

extensively documented to subserve spatial cognition, memory integration, and emotional regulation processes [38], [39], [40]. These are aligned with reported SMC symptoms including spatial orientation deficits and memory impairments. For NC vs. EMCI, key connections appeared in the Right Parahippocampal Gyrus, Left Superior Temporal Gyrus, and Left Middle Frontal Gyrus. These regions are critically implicated in memory processing and emotional regulation mechanisms [41], [42], corresponding to EMCI symptoms of memory



(a) NC vs. SMC (b) NC vs. EMCI (c) NC vs. SMC&EMCI (d) SMC vs. EMCI

Fig. 9. The spatial distribution of the top 10 most discriminative brain regions across different classification tasks.

loss and executive dysfunction. In NC vs. SMC&EMCI, important connections are concentrated on the Right Fusiform Gyrus, Right Middle Temporal Gyrus, and Right Thalamus. These neural structures are fundamentally involved in visual recognition, memory processing, and perceptual integration [43], [44], [45], representing common functional processing areas in Alzheimer's disease. For SMC vs. EMCI, significant connections appeared in the Left and Right Middle Frontal Orbital Gyri and Right Middle Frontal Gyrus. These regions are integral to emotional regulation and executive function [40], [44], highlighting differential impacts on these functions between SMC and EMCI patients.

D. Discriminative Brain Regions

Additionally, we evaluated our method's effectiveness in identifying discriminative brain regions. Using non-negative elastic net regression, we linked SAN-transformed hypergraph features with diagnostic labels. The regression coefficients represent brain region importance. To ensure robust identification despite variations in cross-validation, we aggregated the regression coefficients for each brain region by summing them across all ten validation folds. Based on these accumulated coefficient values, we ranked and determined the top 10 most discriminative regions for each classification task. Figure 9 highlights distinct regional patterns across different stages of cognitive impairment.

For the NC vs. SMC task, discriminative power is concentrated in the Left Inferior Frontal Gyrus and the Right Angular Gyrus. The former is associated with cognitive functions and emotional regulation [46], while the latter contributes to language comprehension and information integration [47]. These findings align with the subtle cognitive complaints and emotional dysregulation often reported in SMC patients. In contrast, for the NC vs. EMCI task, the identification of the Left Hippocampus as a critical region corroborates its central role in memory encoding [48], reflecting the overt memory

loss characteristic of EMCI. Concurrently, the Right Median Cingulate Gyrus emerges as significant, potentially reflecting altered attention allocation demands during cognitive decline. Regarding commonalities across tasks, we observe overlapping patterns in sensory and perceptual regions, validating the robustness of our biomarkers. Specifically, the Right Heschl's Gyrus and Right Postcentral Gyrus appear as critical nodes for auditory processing and sensory integration [49], suggesting widespread sensory network disruption serves as a shared feature. Furthermore, regions like the Right Lingual Gyrus, which is implicated in visual memory [50], and the Left Superior Frontal Gyrus, which relates to emotional regulation [51], demonstrate consistent discriminative roles. These specific regional alterations, driven by long-term cognitive demands [52], confirm that our method can isolate reliable and interpretable biomarkers for ADNI diagnosis. Notably, in the NC vs. SMC&EMCI task, we observe overlapping patterns that further validate these findings.

VI. Conclusion

In this paper, we introduce a novel spatio-temporal hypergraph attention network framework that addresses key challenges in functional brain connectivity networks analysis. Our approach captures high-order spatio-temporal topological patterns and incorporates dynamic prior knowledge effectively. The framework extracts long-range dependencies from fMRI time series by introducing temporal similarity-driven prior knowledge with attention mechanisms, enabling adaptive representation of brain state dynamics. We implement a hierarchical hypergraph generation module that integrates both local and global topological information, facilitating multi-scale modeling of complex spatio-temporal structures. Additionally, our transformer-based spatial attention network utilizes hypergraph message passing to establish multi-level spatial relationships among brain regions, enhancing the representation of intricate functional network structures. Evaluations on real-world neurological datasets demonstrate that our method substantially enhances diagnostic accuracy over existing approaches and successfully isolates salient biomarkers. Importantly, by unveiling the underlying mechanisms linking high-order spatio-temporal patterns to cognitive dysfunction, our framework delivers high clinical value, ultimately supporting more precise diagnosis and personalized treatment strategies.

References

[1] J. Ji and Y. Zhang, "Functional brain network classification based on deep graph hashing learning," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2891–2902, Oct. 2022.

[2] X. Song, K. Wu, and L. Chai, "Brain network analysis of schizophrenia patients based on hypergraph signal processing," *IEEE Trans. Image Process.*, vol. 32, pp. 4964–4976, 2023.

[3] J. Zhou et al., "DCLNet: Double collaborative learning network on stationary-dynamic functional brain network for brain disease classification," *IEEE Trans. Image Process.*, vol. 34, pp. 4026–4039, 2025.

[4] K. Szewczyk-Krolikowski et al., "Functional connectivity in the basal ganglia network differentiates pd patients from controls," *Neurology*, vol. 83, no. 3, pp. 208–214, 2014.

[5] M. H. Lee, C. D. Smyser, and J. S. Shimony, "Resting-state fMRI: A review of methods and clinical applications," *Amer. J. Neuroradiol.*, vol. 34, no. 10, pp. 1866–1872, Oct. 2013.

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

[7] R. M. Hutchison et al., "Dynamic functional connectivity: Promise, issues, and interpretations," *NeuroImage*, vol. 80, pp. 360–378, Oct. 2013.

[8] C. Li et al., "Multi-channel spatio-temporal graph attention contrastive network for brain disease diagnosis," *NeuroImage*, vol. 307, Feb. 2025, Art. no. 121013.

[9] J. Ming and S. Kundu, "Flexible Bayesian support vector machines for brain network-based classification," 2022, *arXiv:2205.12143*.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.

[11] G. Hu et al., "Discovering hidden brain network responses to naturalistic stimuli via tensor component analysis of multi-subject fMRI data," *NeuroImage*, vol. 255, Jul. 2022, Art. no. 119193.

[12] Y. Zhang, L. Tetrel, B. Thirion, and P. Bellec, "Functional annotation of human cognitive states using deep graph convolution," *NeuroImage*, vol. 231, May 2021, Art. no. 117847.

[13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[14] 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.

[15] J. Wang et al., "Dynamic weighted hypergraph convolutional network for brain functional connectome analysis," *Med. Image Anal.*, vol. 87, Jul. 2023, Art. no. 102828.

[16] X. Zhang, U. Braun, H. Tost, and D. S. Bassett, "Data-driven approaches to neuroimaging analysis to enhance psychiatric diagnosis and therapy," *Biol. Psychiatry, Cognit. Neurosci. Neuroimag.*, vol. 5, no. 8, pp. 780–790, Aug. 2020.

[17] E. A. Maguire, C. D. Frith, and R. G. Morris, "The functional neuroanatomy of comprehension and memory: The importance of prior knowledge," *Brain*, vol. 122, no. 10, pp. 1839–1850, 1999.

[18] M. D. Greaves, L. Novelli, S. Mansour L.,A. Zalesky, and A. Razi, "Structurally informed models of directed brain connectivity," *Nature Rev. Neurosci.*, vol. 26, no. 1, pp. 23–41, Jan. 2025.

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

[20] S. Li, Q. Zhu, C. Tian, W. Shao, and D. Zhang, "Interpretable dynamic brain network analysis with functional and structural priors," *IEEE Trans. Med. Imag.*, vol. 44, no. 12, pp. 4878–4889, Dec. 2025.

[21] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *Stat*, vol. 1050, no. 20, p. 10, 2017.

[22] J. Liu et al., "Deep fusion of multi-template using spatio-temporal weighted multi-hypergraph convolutional networks for brain disease analysis," *IEEE Trans. Med. Imag.*, vol. 43, no. 2, pp. 860–873, Feb. 2024.

[23] H. Cui, W. Dai, Y. Guo, X. Kan, C. Yang, and Z. Zhang, "Brain network transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 25586–25599.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[25] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, "Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4023–4033, Dec. 2020.

[26] Z. Ji, S. Ye, and X. Ma, "Sparse coding inspired LSTM and self-attention integration for medical image segmentation," *IEEE Trans. Image Process.*, vol. 33, pp. 6098–6113, 2024.

[27] P. He, Z. Shi, Y. Cui, R. Wang, and D. Wu, "A spatiotemporal graph transformer approach for Alzheimer's disease diagnosis with rs-fMRI," *Comput. Biol. Med.*, vol. 178, Aug. 2024, Art. no. 108762.

[28] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. Cham, Switzerland: Springer, 1992, pp. 162–190.

[29] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[31] J. Kawahara et al., "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, Feb. 2017.

[32] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.

[33] C. Tang, L. Yang, G. Cao, J. Du, and Q. Zhang, "ST-GCN: EEG emotion recognition via spectral graph and temporal analysis with graph convolutional networks," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2024, pp. 3748–3751.

[34] D. Zhu, X. Li, X. Jiang, H. Chen, D. Shen, and T. Liu, "Exploring high-order functional interactions via structurally-weighted lasso models," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2013, pp. 13–24.

[35] 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. Interv.* Cham, Switzerland: Springer, 2023, pp. 67–76.

[36] R. Liu et al., "Dynamic graph representation learning for spatio-temporal neuroimaging analysis," *IEEE Trans. Cybern.*, vol. 55, no. 3, pp. 1121–1134, Mar. 2025.

[37] Q. Zhu, J. Huang, and X. Xu, "Non-negative discriminative brain functional connectivity for identifying schizophrenia on resting-state fMRI," *Biomed. Eng. OnLine*, vol. 17, no. 1, p. 32, Dec. 2018.

[38] S. J. Greene and R. J. Killiany, "Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's disease," *Neurobiol. Aging*, vol. 31, no. 8, pp. 1304–1311, Aug. 2010.

[39] S. Tolomeo et al., "A causal role for the anterior mid-cingulate cortex in negative affect and cognitive control," *Brain*, vol. 139, no. 6, pp. 1844–1854, 2016.

[40] S. Brunton, C. Gunasinghe, N. Jones, M. J. Kempton, E. Westman, and A. Simmons, "A voxel-based morphometry comparison of the 3.0T ADNI-1 and ADNI-2 volumetric MRI protocols," *Int. J. Geriatric Psychiatry*, vol. 30, no. 5, pp. 531–538, May 2015.

[41] W.-Y. Wang et al., "Impacts of CD33 genetic variations on the atrophy rates of hippocampus and parahippocampal gyrus in normal aging and mild cognitive impairment," *Mol. Neurobiol.*, vol. 54, no. 2, pp. 1111–1118, Mar. 2017.

[42] N. Franzmeier, M. Duering, M. Weiner, M. Dichgans, and M. Ewers, "Left frontal cortex connectivity underlies cognitive reserve in prodromal Alzheimer disease," *Neurology*, vol. 88, no. 11, pp. 1054–1061, Mar. 2017.

[43] S. Risacher, A. Saykin, J. Wes, L. Shen, H. Firpi, and B. McDonald, "Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort," *Current Alzheimer Res.*, vol. 6, no. 4, pp. 347–361, Aug. 2009.

[44] J. P. Aggleton, A. Pralus, A. J. Nelson, and M. Hornberger, "Thalamic pathology and memory loss in early Alzheimer's disease: Moving the focus from the medial temporal lobe to papez circuit," *Brain*, vol. 139, no. 7, pp. 1877–1890, 2016.

[45] J. Zhang et al., "Asynchronous functional brain network construction with spatiotemporal transformer for MCI classification," *IEEE Trans. Med. Imag.*, vol. 44, no. 3, pp. 1168–1180, Mar. 2025.

[46] A. Hampshire, S. R. Chamberlain, M. M. Monti, J. Duncan, and A. M. Owen, "The role of the right inferior frontal gyrus: Inhibition and attentional control," *NeuroImage*, vol. 50, no. 3, pp. 1313–1319, Apr. 2010.

[47] Y. Zhang, H. Zhang, E. Adeli, X. Chen, M. Liu, and D. Shen, "Multiview feature learning with multiatlas-based functional connectivity networks for MCI diagnosis," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6822–6833, Jul. 2022.

[48] T. Paus et al., "Human cingulate and paracingulate sulci: Pattern, variability, asymmetry, and probabilistic map," *Cerebral Cortex*, vol. 6, no. 2, pp. 207–214, 1996.

[49] M. Roll, "Heschl's gyrus and the temporal pole: The cortical lateralization of language," *NeuroImage*, vol. 303, Dec. 2024, Art. no. 120930.

[50] J. Bogousslavsky, J. Miklossy, J. P. Deruaz, G. Assal, and F. Regli, "Lingual and fusiform gyri in visual processing: A clinico-pathologic study of superior altitudinal hemianopia," *J. Neurol. Neurosurg. Psychiatry*, vol. 50, no. 5, pp. 607–614, 1987.

[51] G. N. Bratman, J. P. Hamilton, K. S. Hahn, G. C. Daily, and J. J. Gross, "Nature experience reduces rumination and subgenual prefrontal cortex activation," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 28, pp. 8567–8572, 2015.

[52] B. Voytek, M. Davis, E. Yago, F. Barcelo, E. K. Vogel, and R. T. Knight, "Dynamic neuroplasticity after human prefrontal cortex damage," *Neuron*, vol. 68, no. 3, pp. 401–408, 2010.