Multi-Hop Spatio-Temporal Graph Attention Networks for Epilepsy Diagnosis and Prognosis

Haoxiang Liu

Department of Integrated Program for Neuroscience

McGill University

April 14, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Neuroscience (Thesis)

Table of Contents

Abstract

Graph Neural Networks (GNNs) represent a formidable breakthrough in feature extraction from non-Euclidean datasets, particularly in medical imaging such as brain networks derived from functional magnetic resonance imaging (fMRI). However, existing GNN frameworks for analyzing spatio-temporal dynamics within brain networks often overlook indirect node connections, leading to suboptimal performance, particularly across heterogeneous datasets from different sites. To address these limitations, we propose a novel approach called Multi-Hop Spatio-Temporal Graph Convolutional Network (MSTGCN) with Reverse Contrastive (RevCon) learning for identifying temporal lobe epilepsy (TLE) and predicting surgical outcomes. Our approach leverages a graph attention mechanism that incorporates both node and edge features to compute edge weights, facilitating information propagation across multiple hops to enhance graph representations. Additionally, we employ the Transformer architecture to effectively handle temporal information. We bolster the model's generalizability across datasets from different sites through RevCon learning. Furthermore, we utilize gradient-based saliency maps to interpret the model's classification and to predict surgical outcomes based on surgery information. Experimental results on two TLE datasets demonstrate the effectiveness of MSTGCN, achieving identification accuracies of up to 85.52% and 78.27% respectively, and 82.07% accuracy on Cross-site datasets, outperforming state-of-the-art methods. Moreover, our model attains an 82% accuracy in predicting surgical outcomes, indicating its potential for future clinical applications.

Abrégé

Les réseaux neuronaux graphiques (GNN) représentent une percée formidable dans l'extraction de caractéristiques à partir de jeux de données non euclidiens, notamment en imagerie médicale comme les réseaux cérébraux dérivés de l'imagerie par résonance magnétique fonctionnelle. Cependant, les cadres GNN existants pour analyser les dynamiques spatio-temporelles au sein des réseaux cérébraux négligent souvent les connexions indirectes des nœuds, entraînant des performances sous-optimales, notamment sur des ensembles de données hétérogènes provenant de différents sites. Pour remédier à ces limitations, nous proposons une approche novatrice appelée Réseau de Convolution Graphique Spatio-Temporel Multi-Sauts (MSTGCN) avec apprentissage contrastif inverse (RevCon) pour identifier l'épilepsie temporale (TLE) et prédire les résultats chirurgicaux. Notre approche exploite un mécanisme d'attention graphique qui intègre à la fois les caractéristiques des nœuds et des arêtes pour calculer les poids des arêtes, facilitant la propagation des informations à travers plusieurs sauts pour améliorer les représentations graphiques. De plus, nous utilisons l'architecture Transformer pour gérer efficacement les informations temporelles. Nous renforçons la généralisabilité du modèle sur des ensembles de données provenant de différents sites grâce à l'apprentissage RevCon. De plus, nous utilisons des cartes de saillance basées sur le gradient pour interpréter la classification du modèle et prédire les résultats chirurgicaux basés sur les informations chirurgicales. Les résultats expérimentaux sur deux ensembles de données TLE démontrent l'efficacité de MSTGCN, atteignant des précisions d'identification allant jusqu'à 85,52% et 78,27% respectivement, et une précision de 82,07% sur des ensembles de données intersites, surpassant les méthodes de pointe. De plus, notre modèle atteint une précision de 82% dans la prédiction des résultats chirurgicaux, indiquant son potentiel pour les futures applications cliniques.

Acknowledgements

I would like to express my sincere gratitude to Prof. Suresh Krishna, Prof. Huafu Chen, and Prof. Rong Li for their invaluable guidance and support throughout my research journey. I am also grateful to the Brain Imaging and Pattern Recognition team and the M2B3 Lab, along with their faculty and fellow students, for their collaboration and contributions.

Special thanks go to my friends, girlfriend, and family for their unwavering encouragement and understanding during this challenging yet rewarding endeavor. Additionally, I extend my appreciation to the two hospitals for providing the necessary data for my study.

Thank you all for being an essential part of my academic and personal growth.

List of Figures

List of Tables

List of Acronyms

TLE	Temporal Lobe Epilepsy
MRI	Magnetic Resonance Imaging
sMRI	Structural Magnetic Resonance Imaging
fMRI	Functional Magnetic Resonance Imaging
DTI	Diffision Tensor Imaging
BOLD	Blood Oxygen Level Dependent
FA	Fractional Anisotropy
mTLE	Mesial Temporal Lobe Epilepsy
nTLE	Neocortex Temporal Lobe Epilepsy
RSNs	Resting State Networks
FC	Functional Connectivity
ROIs	Regions of Interest
DMN	Default Mode Network
SMN	Sensorimotor Network
VN	Visual Netwok
AN	Auditory Network
DAN	Dorsal Attention Network
VAN	Ventral Attention Network
SUB	Subcortical
CNNs	Convolutional Neural Networks
MLP	Multilayer Perceptron
GANs	Generative Adversarial Networks
GNNs	Graph Neural Networks
GCNs	Graph Convolutional Networks
GAT	Graph Attention Networks
ST-GCN	Spatio-temporal Graph Convolutional Network
GARO	Graph-Attention Readout
SERO	Squeeze-Excitation Readout
sFC	Static Functional Connectivity
dFC	Dynamic Functional Connectivity

MSTGCN	Multi-hop Spatio-temporal Graph Convolutional Network
DNNs	Deep Neural Networks
RevCon	Reverse Contrastive
CE	Cross-Entropy
GIN	Graph Isomorphic Networks

1. Background and Related Works

1.1 Statement of Contribution

We are thankful to the Xiangya Hospital of Central South University and the First Affiliated Hospital of Zhengzhou University for providing the non-openly available temporal lobe epilepsy (TLE) datasets. My supervisor, Dr. Suresh Krishna, helped me with every milestone within my academic journey. I received insightful guidance and concrete advice, even if I was working remotely. My supervisor Dr. Huafu Chen and Dr. Rong Li frequently encouraged and motivated me with helpful suggestions. Without all of you, I could not finish my thesis.

The author's contributions are as follows: review of background knowledge in Chapter 1 & 2; building MSTGCN model in Chapter 3; data pre-processing and running all experiments and analyses discussed in the thesis in Chapter 4 & 5.

1.2 Epilepsy and Neural Imaging

Epilepsy is a neurological disorder characterized by recurrent seizures, affecting approximately 70 million people globally [1]. In China alone, there are currently at least ten million epilepsy patients, making it one of the five major neurological and psychiatric disorders targeted for global prevention and treatment by the World Health Organization. Epilepsy not only imposes significant physical, psychological, and economic burdens on individuals, families, and society but also presents a major challenge and public health issue in the current medical landscape. While most epilepsy patients can achieve effective control through certain medical interventions, nearly one-third of patients remain unresponsive to these treatment modalities. Therefore, exploring the pathophysiological mechanisms underlying epilepsy, accurate diagnosis, and effective treatment are of paramount importance.

Firstly, the pathophysiological mechanisms of epilepsy still hold many unresolved mysteries that have yet to be fully understood. There is an urgent need to delve into the localization and network propagation mechanisms of epileptic activity, seeking and developing new neuroimaging methods and technologies. Secondly, epilepsy diagnosis faces certain difficulties due to the lack of effective biomarkers and traditional

electroencephalogram and magnetic resonance imaging (MRI) techniques still encounter issues such as misdiagnosis and subjective interpretation. Additionally, epilepsy symptoms exhibit considerable heterogeneity, with different patients experiencing varying seizure types, frequencies, and severities, along with cognitive and behavioral issues. Therefore, there is a need to find a reliable, non-invasive, and accurate intelligent diagnostic method to provide a more reliable reference for early assessment of epilepsy. Lastly, effective treatments for epilepsy remain elusive. Current epilepsy treatments mainly involve antiepileptic drugs and surgical interventions. However, as mentioned earlier, one-third of patients cannot achieve effective control of seizures through medication and surgery. Therefore, diagnosing epilepsy not only requires detailed medical history inquiries to identify specific clinical manifestations but also necessitates comprehensive neuroimaging examinations to obtain quantitative imaging metrics for more accurate classification and diagnosis of epilepsy patients. Early intervention may help alleviate epilepsy symptoms and frequencies, while precise intelligent treatment methods can increase treatment success rates. In summary, research outcomes addressing the above issues will provide objective neuroimaging evidence for elucidating the pathogenesis of epilepsy, aiding in clinical intelligent diagnosis and personalized precision treatment. Furthermore, these outcomes are expected to facilitate the translation of epilepsy neuroimaging research into clinical applications, ultimately benefiting patients to the greatest extent possible.

In the clinical and neuroscientific fields, researchers have conducted extensive studies on epilepsy. Functional Magnetic Resonance Imaging (fMRI), Structural Magnetic Resonance Imaging (sMRI), and Diffusion Tensor Imaging (DTI) are neuroimaging technologies that have been applied to explore the neural mechanisms of epilepsy. Functional Magnetic Resonance Imaging records changes in the blood oxygen level dependent (BOLD) signal in the brain. BOLD reflects changes in the oxygenated hemoglobin content in surrounding blood when neurons are active, leading to corresponding changes in the fMRI signal recorded. Therefore, the BOLD signal indirectly reflects neural activity and is considered a proxy for brain function depiction. Research based on fMRI has explored changes in brain activity in epilepsy patients, aiding researchers in understanding the mechanisms of the disease. For instance, fMRI-based studies have found differences in spontaneous activity at specific nodes in epilepsy

patients compared to healthy controls. sMRI [2], on the other hand, captures differences in brain anatomy and is commonly used for measuring brain tissue volume or cortical surface morphology. Hippocampal sclerosis is a common histological abnormality in patients with medial temporal lobe epilepsy, observable through sMRI (T2-weighted sequences show high signal, while T1-weighted sequences show low signal). DTI measures water molecule diffusion, enabling observation of brain white matter structure. The basic assumption of DTI imaging is that water molecules near brain white matter are hindered in directions perpendicular to the white matter, while they are relatively unimpeded parallel to the white matter. Based on this characteristic, researchers believe that reduced water diffusion anisotropy Fractional Anisotropy (FA) indicates decreased white matter tract integrity. Through certain algorithms, researchers can also represent white matter tracts in three-dimensional space using DTI. DTI is also an important analysis tool in the field of epilepsy mechanisms, as evidenced by Arfanakis et al.'s 2002 discovery [3] of reduced FA between the external capsule and corpus callosum in patients with temporal lobe epilepsy compared to normal controls.

While the above-mentioned imaging methods can reflect brain structure or functional features from different perspectives, their large and complex data require effective algorithms and advancements in computer technology for relevant discoveries. Machine learning algorithms provide robust support for exploring mechanisms, diagnosis, and prognosis of brain diseases such as Epilepsy. The application of machine learning and deep learning in exploring brain diseases has played a crucial role in tasks such as disease diagnosis in recent years, demonstrating their potential in addressing related issues. The rapid development of artificial intelligence algorithms in recent years is also expected to drive research in brain imaging. Therefore, exploring machine learning and deep learning algorithms for epilepsy brain imaging is an important research problem and direction.

1.3 Temporal Lobe Epilepsy

TLE is the most common form of human focal epilepsy [4]. Temporal lobe epilepsy has two main forms: mesial Temporal Lobe Epilepsy (mTLE), which is believed to originate from the hippocampus or adjacent to the hippocampus, and Neocortex Temporal Lobe Epilepsy (nTLE), where the seizure focus is thought to be in the lateral neocortex of the temporal lobe [5]. In the early stages of TLE seizures, patients often experience

gastrointestinal symptoms such as rising gastric Qi, as well as emotional abnormalities such as fear. During the seizure, a few patients may experience secondary tonic-clonic Convulsions.

Early epilepsy research focused on specific anatomically isolated epileptic foci, where focal seizures were believed to originate from specific, anatomically isolated epileptic foci. However, with the advancement of brain imaging technology, research related to epilepsy has shifted focus from "epileptic foci" to a "network" perspective [6, 7]. This change in perspective is largely attributed to studies on mTLE. In this form of epilepsy, discharges during seizure periods and interictal periods can originate from multiple different lesions, not just limited to the hippocampus itself but also in regions outside the hippocampus [8, 9]. Subsequent studies involving electrocorticography and neuroimaging in humans, as well as experiments conducted on TLE animal models, have led to the concept of a large-scale mTLE network. This network includes not only the hippocampus but also surrounding structures (such as the amygdala), subcortical regions (such as the thalamus), and neocortical areas (such as the frontal lobe, particularly the orbitofrontal region, and the temporal lobe's superior, middle, and inferior gyri). It has been found through research that each brain structure within this network either serves as a source of epileptic discharges or contributes to their propagation [8, 10, 11]. Therefore, unlike the traditional focal perspective, TLE actually involves many parts of the brain. Consequently, the mechanisms of epilepsy have become more complex, and a more comprehensive understanding of network properties is expected to enhance efforts related to the diagnosis, treatment, and prognosis of epilepsy.

Some mTLE patients find it difficult to achieve seizure frequency reduction and a seizure-free state through drug therapy alone, leading them to be classified as drug-resistant epilepsy patients. Surgery is a common and effective treatment option for drug-resistant epilepsy patients. Common surgical approaches include Anterior Temporal Lobotomy and Selective Amygdala Hippocampectomy, with some clinicians and researchers also studying methods involving electrode lesioning, which offer more precise targeting and fewer surgical side effects compared to traditional surgery but are still experimental. However, some post-surgical epilepsy patients still experience relapses. Thus, research suggests that the network causing epilepsy may be the culprit behind poor efficacy and recurrent seizures [12]. Furthermore, functional network analysis indicates

that effective connectivity from the hippocampus to the default mode network serves as a predictive biomarker for drug response in potential TLE patients [13]. Therefore, network analysis provides valuable insights into research related to Temporal Lobe Epilepsy (TLE), which will be elaborated on in the following section.

1.4 Network Theory of Functional Magnetic Resonance Imaging

In Section 1.2, the principles of fMRI data acquisition were discussed. This section will elaborate on the applications of fMRI in exploring, diagnosing, and predicting the prognosis of TLE mechanisms. Due to its non-invasive nature, high spatial resolution, and moderate temporal resolution, fMRI has gradually become an important tool for screening brain diseases.

fMRI scans are typically divided into task-state and rest-state scans. Task-state fMRI refers to fMRI image data obtained from subjects while they perform specific tasks or stimuli during the scan. Task-state fMRI has wide applications in cognitive neuroscience [14], disease research [15], and other fields. On the other hand, rest-state fMRI (rs-fMRI) has attracted increasing attention and research. rs-fMRI focuses on the low-frequency fluctuations in the BOLD signal, and the functional significance of these fluctuations was first proposed by Biswal et al. in 1995 [16]. During rs-fMRI scans, subjects are instructed not to engage in any cognitive, language, or motor tasks and should not enter a sleep state. Therefore, rs-fMRI is considered to record spontaneous brain activity, providing an important observational tool for subsequent brain network studies.

The brain forms a complex and efficient network through information exchange between different regions [17]. To quantify brain networks, functional connectivity (FC) between regions can be calculated based on the similarity of functional fMRI signals, which is an important method in related research. Using rs-fMRI can yield effective and stable FC, which has been widely applied in brain network analysis of neurological and psychiatric disorders [13, 18, 19]. Besides, structural connectivity (SC) is also an effective way to quatify the brain network. DTI uses the map of the white matter fiber bundles to construct SC. SC can reflect the abnormal brain structure of epileptic patients [20].

When calculating FC, the brain can be divided into several Regions of Interest (ROIs) based on specific research questions. Since the minimum unit of fMRI signal acquisition is a voxel, the standard fMRI signal within each ROI is often obtained by averaging the

signals of all voxels in the ROI. Using Pearson's Correlation to calculate the similarity between signals is a common method for computing FC, as defined in Equation (1-1):

$$\rho_{(X,Y)}=\frac{\sum_{t=1}^{T}(X_t-\bar{X})(Y_t-\bar{Y})}{\sqrt{\sum_{t=1}^{T}(X_t-\bar{X})^2\sum_{t=1}^{T}(Y_t-\bar{Y})^2}} \quad (1\text{-}1)$$

The BOLD time series signals of two brain regions, denoted as X and Y, are represented in which T represents the total number of time points in the time series, and $\bar{X}$ and $\bar{Y}$ are the means of the two BOLD signals. The Pearson correlation coefficient has a range of $\rho_{(X,Y)} \in [-1, 1]$. Pearson correlation focuses on the linear correlation between ROIs, and FC calculated using Pearson correlation has been applied in many brain network studies [17]. In addition to Pearson correlation, Spearman's Rank Correlation is also a method for calculating signal correlations. SRC is considered to reflect non-linear information through non-parametric rank correlation. Both methods have their characteristics and reflect the relationships in the network from different perspectives. They have both been applied in brain network analysis research [21-23].

A significant breakthrough in recent TLE research is the demonstration of its impact on a group of brain networks called Resting State Networks (RSNs) [24]. These networks are considered discrete brain network structures in fMRI studies of normal subjects, exhibiting spontaneous synchronized activities that are particularly evident during rest (i.e., resting state) and adjust activity under specific functional tasks. In other words, although RSNs are identified in the resting state, they are associated with many normal brain functions, and their activities are modulated by external or internal stimuli. For example, the activity of the Default Mode Network (DMN) decreases when receiving external stimuli, while the activity of other RSNs increases under similar stimuli [25]. Therefore, based on the response of RSNs during different behaviors or stimuli, RSNs can be broadly divided into two groups. The first group includes networks related to sensory and motor processing. The Sensorimotor Network (SMN), Visual Network (VN), and Auditory Network (AN) are part of this group [17, 26]. The second group mainly comprises networks related to higher-order brain functions. The Default Mode Network (DMN), Dorsal Attention Network (DAN), Ventral Attention Network (VAN), Salience Network (SAN), Executive Control Network (ECN), and Language Network (LN) are part of this group [24, 27, 28]. Networks in subcortical regions (Subcortical, SUB) have also been studied in some articles [29].

In summary, by combining fMRI network theory, quantifying and characterizing brain networks, establishing the association between brain network abnormalities in diseases such as epilepsy and clinical characteristics, researchers can further explore the physiological and pathological mechanisms of brain diseases. This provides important theoretical support for intelligent diagnosis and important analytical tools for more accurate prognosis.

1.5 Review of Current Machine Learning and Deep Learning Approaches for Brain Disease Recognition

The previous sections reviewed the development of brain imaging and brain networks in research related to brain diseases. Although brain imaging and brain network theory provide important research methods for exploring the mechanisms of brain diseases, understanding brain imaging and brain networks remains complex. Using fMRI signals and functional brain networks as examples, their data often have high dimensions, lack intuitiveness, and are prone to noise interference. For tasks such as diagnosis and prognosis, algorithmic analysis is required to effectively utilize brain imaging. The advancement of computer technology has provided powerful tools for understanding brain imaging and brain networks, capable of replacing humans in executing complex algorithms and lengthy computational processes. Therefore, developing effective algorithms and methods can maximize the utilization of brain imaging, helping to improve the accuracy of disease diagnosis and treatment outcomes. In recent years, machine learning and deep learning methods have made significant contributions in this field. For example, Support Vector Machine, Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Autoencoder, and other methods have achieved important results in the diagnosis of brain diseases.

Traditional machine learning methods often rely on feature engineering to achieve effective diagnosis. Feature engineering aims to use domain knowledge to construct sufficiently effective features from complex brain imaging or brain network data. These traditional machine learning methods typically involve two steps. The first step requires feature engineering on complex brain imaging or brain network data. The second step involves using machine learning classification methods such as SVM to obtain the final diagnosis result. Traditional machine learning methods can achieve good results in

scenarios with small datasets. For example, Torlay et al. [30] used changes in MRI signal intensity from specified ROIs to construct features and employed the XGBoost classifier to classify normal controls and TLE patients.

Deep learning, due to its powerful feature extraction capabilities, does not require domain knowledge to construct features. It can automatically extract useful features through iteration, making it suitable for scenarios with large datasets. Models used for disease diagnosis and prognosis mainly include supervised learning models and unsupervised learning models. Supervised learning models require label information, such as healthy and diseased, or good and poor treatment outcomes. Through iterative training of the model, it can implicitly learn the relationship between labels and data, thereby generalizing to judge the corresponding labels of unseen datasets [31]. Unsupervised learning does not require labels and can automatically capture preferences based on the intrinsic features of data. Due to the rapid development of deep learning models and the need to address various types of problems, various architectures of deep learning models with distinct features and functionalities have emerged since the development of deep learning. Work related to brain imaging based on models such as CNNs, recurrent neural networks, generative adversarial networks (GANs), variational autoencoder (VAE), etc., has emerged as a burgeoning and active field. For instance, Zheng et al. [32] utilized VAE to analyze dynamic brain networks and achieved superior results in disease diagnosis across three real brain disease datasets. Dvornek et al. [33] directly applied LSTM to rs-fMRI signals for the diagnosis of autism spectrum disorders. Pan et al. [34] employed GAN to complement missing PET data in their dataset using MRI data and subsequently conducted diagnostic work on patients with mild cognitive impairment using both MRI and PET data.

1.6 Graph Neural Network

In many fields, graphs are the primary form of data derived from nature. This is because most patterns in life can be abstracted into graph structures. Examples include molecular structures, brain networks, social networks, and transportation networks, among others. This potential has been recognized by researchers and practitioners in the scientific and industrial sectors, leading to applications in traffic flow prediction, drug discovery, social network analysis, recommendation systems, and more. While CNNs

have achieved tremendous success in extracting representations from Euclidean data (such as images), the attention has shifted to non-Euclidean data represented by graph structures, thus necessitating effective analysis methods, giving rise to GNNs.

The GNN family includes various variants, among which the Graph Convolutional Neural Network (GCN), inspired by CNNs, has shown outstanding performance on complex graph data. GCN typically takes node features and edge features (as input, often represented as adjacency matrices). Its core idea is to aggregate information from neighboring nodes for each node and update node information through learnable weights. Similar to CNNs, GCN often adopts a stacking approach to enhance its representation power and can be applied to tasks such as node classification, link prediction, and graph classification.

GCNs can be categorized into spectral-based methods and spatial-based methods. Spectral-based methods introduce filters from the perspective of a single operation on the graph to express graph convolution. Spatial-based methods often directly define convolutions on the graph to aggregate neighboring features. The superior performance of attention mechanisms in recent years has attracted attention from researchers, leading to the development of Graph Attention Networks (GAT), which combine spatial graph convolution with attention mechanisms, showcasing robust capabilities in processing graph-structured data. GAT calculates a learnable adjacency matrix based on attention mechanisms using features of neighboring nodes, achieving state-of-the-art results in many GNN-related tasks at that time. Subsequently, many different attention-based GNN works have emerged based on this framework [35-37].

Currently, GNN methods for investigating brain networks can be broadly categorized into two groups: (1) GNN-based methods for static brain networks and (2) GNN-based methods for dynamic brain networks. In GNN-based methods for static brain networks, researchers investigate the topological and spatial properties of static brain networks, such as sFC, which is directly derived from the correlation of the entire time sequence. Ktena et al. [38] used Siamese-GCN to detect similarities between pairs of static FCs and utilized the K-nearest neighbor method to classify healthy controls and individuals with autism spectrum disorder (ASD). Chen et al. [39] developed a node-edge graph attention mechanism to identify ASD patients based on static FC. In GNN-based methods for dynamic brain networks, both temporal and spatial information are typically

considered. Xing et al.[40] proposed a method to extract spatial information from each segment of dFC and then utilized a long short-term memory layer to process temporal information across segments. To simultaneously utilize spatial and temporal information Gadgil et al. [41] proposed a spatio-temporal graph convolutional network (ST-GCN). The work in ST-GCN aggregates information from spatial and temporal neighbors of each ROI. In this case, ST-GCN can aggregate the information from both spatial dimension and temporal dimension. Kim et al. [42] proposed Graph-Attention Readout (GARO) and Squeeze-Excitation Readout (SERO) to learn spatial representation of each fragment of dFC, and utilized Transformer to learn temporal representation over all fragments of dynamic dFC. However, static brain network methods ignore the rich potential dynamic information of the brain network. Dynamic brain network methods require a more complex model to process temporal information, thus limiting the ability to learn spatial representation.

1.7 Multi-hop Graph Neural Network

Hop is defined as the path length from one node to another. If there is only one connection in a path, then the path is one-hop. If there is more than one connection (e.g. there are connections from node A to B, and B to C, then a path from A to C includes 2 connections, which is two hop), the path is multi-hop. Traditional GNNs (including GCNs, GATs etc.) are one-hop and only consider direct neighbors. Deeper stacked application of one-hop graph neural network layers can suffer from Laplacian smoothing (over-smoothing), thereby degrading performance and limiting spatial representation. Recently, some work has explored graph diffusion convolution [37, 43], which aggregates information from multi-hop neighbors rather than just direct one-hop neighbors of nodes, thereby improving performance by broadening the receptive field. However, these approaches have not incorporated attention mechanisms or edge features. Cucurull et al. [44] also explored the extension of attention mechanisms to multi-hop information, their approach requires a higher computational parameter overhead to maintain performance levels.

2. Introduction

2.1 Introduction

Epilepsy is one of the most common serious brain disorders, affecting over 70 million people worldwide [1]. For patients with drug-resistant focal epilepsy, such as TLE, surgical removal of a circumscribed brain area can be beneficial in achieving complete seizure control [45, 46]. In clinical practice, accurate diagnosis of epilepsy and assessment of surgical prognosis remain challenging. Resting-state fMRI is a non-invasive neuroimaging tool that indirectly describes brain activity by measuring blood oxygen level dependence, aiding researchers in identifying brain dysfunctions in diseases. Epilepsy often exhibits altered spatiotemporal patterns of FC derived from fMRI co-activation, suggesting abnormality in brain networks [47, 48]. Therefore, exploring effective brain network representation models is crucial for the precise diagnosis and prognosis of epilepsy.

Several machine learning studies have emerged to analyze brain functional networks in epilepsy [49, 50]. For example, Mazrooyisebdani et al. [49] proposed a support vector machine model to diagnose TLE based on FC features derived from graph theory analysis. These models are typically designed based on static FC (sFC), which is constructed across the entire resting-state fMRI scans to characterize patterns of functional associations between brain regions. Recent physiological evidence suggests that brain functional networks undergo continual reconfiguration and exhibit temporal changes, which cannot not be fully captured by sFC [51]. Dynamic FC (dFC), on the other hand, is constructed from short time segments of resting-state fMRI scans, thereby reflecting changes in brain networks over time. Previous studies have shown that patients with epilepsy exhibit dynamic network reconfiguration , even during interictal periods [4, 7]. Recent advances in the field of deep learning models, such as CNNs [51], which can process complex dynamic spatiotemporal information, have shown great promise in the diagnosis of epilepsy.

In addition, 33-50% of epilepsy patients who undergo epilepsy surgery experience seizure recurrence postoperatively [52]. One important factor contributing to epilepsy surgery failures is inaccurate or incomplete lateralization and localization of seizure onset zones ahead of treatment [7]. Hence, literalization, localization and prognosis are

significant tasks in clinical application. Literalization is helpful for diagnosis and prognosis[53]. Yang et al. analyzed features extracted from resting-state functional magnetic resonance imaging (rs-fMRI) across various scales: local brain regions, interregional connectivity, and the entire brain network. They utilized Random Forest for dimensionality reduction of the data, followed by Support Vector Machine for the classification task, achieving an accuracy rate of 83% in determining the laterality of TLE patients [54]. For localization task, benefitting from development of computer vision, various deep learning methods have been proposed to localize epileptogenic lesions. Nandakumar et al. proposed a GCN based method for automated epileptogenic zone localization from rs-fMRI, allowing clinicians to harness this information from noninvasive imaging that can easily be integrated into the existing clinical workflow [55]. Prognosis prediction usually refers to the classification task of classifying the postoperative state. Gleichgerrcht et al. trained neural network on SC data with corresponding outcome label and achieved promising result.

2.2 Rational, Aims and Hypothesis

The inherent graph structure of brain networks makes them ideal for learning with graph neural networks [56]. GNN-based methods can learn dynamic features of brain networks without disrupting the graph structure information. Gadgil et al. [57] demonstrated that the ST-GCN, which extracts spatial and temporal features simultaneously, achieved promising results in analyzing dynamic brain networks. However, a limitation of the current GNN models is their inability to effectively capture changes in connections with long-distance indirect connections in dynamic brain networks. It is worth noting that the sparsity of a brain network topology determines that the information interaction between two nodes may involve long-distance indirect connections [58]. An increased average path length of brain functional network has been reported in epilepsy patients [59]. Although stacking graph convolutional layers can theoretically extend the receptive field and incorporate information from indirect connections, deeper structures may exacerbate the over-smoothing problem [37, 60]. Li et al. showed that graph convolution operation is a type of Laplacian smoothing, and they proved that after repeatedly applying Laplacian smoothing many times, the features of nodes will converge to similar value [61]. Over-smoothing will do harm to performance

of GNNs [62]. In addition, deeper neural networks may lead to overfitting when applied to limited brain imaging data. Thus, integrating indirect connections into the model without overly complex structures may improve the representational ability of GNNs for epilepsy brain network. It is noted that, SC is also promising for epilepsy diagnosis and prognosis, but we focus on a framework that is able to extract both spatial and temporal information from brain network. Usally there is no temporal information within SC and dynamic of brain network can not be reflected by SC. In this case, this work focus on dynamic FC.

In general, dynamic analysis of brain functional network by GNNs could advance the complicated representation of direct or indirect connections to improve diagnostic performance. However, deep learning-based medical image models remain prior to widespread clinical implementation due to the following challenges. First, data heterogeneity due to differences in patient populations, imaging scanners and acquisition protocols can lead to poor generalizability for representation learning. A few studies have applied the deep learning generalization models to cross-site brain image diagnostic tasks [63, 64]. For example, Chen et al. [65] proposed an adversarial learning-based for autism spectrum disorder identification based on cross-site MRI data [63]. Domain generalization methods, which aim to overcome domain shifts in unseen datasets, also have the potential to improve generalization by overcoming data heterogeneity. Second, deep learning-based diagnosis and individualized treatment prediction are both essential in clinical practice. In the context of epilepsy, 33-50% of patients who undergo epilepsy surgery experience seizure recurrence postoperatively [52]. A unified framework that integrates clinical classifications, such as diagnosis, into the preoperative prediction of epilepsy outcomes would be beneficial for clinical decision making. However, most of the existing artificial intelligence models in epilepsy have been proposed for either diagnostic or prognostic tasks only.

To address the concerns mentioned above, we propose a Multi-hop Spatio-temporal Graph Convolutional Network (MSTGCN) framework, as shown in Fig.1(b). In contrast to the previous GNN models, we not only extract the spatial and temporal features of the dFC, but also utilize graph convolutional operations with a multi-hop spatial attention mechanism to extract features within the indirect connections of the brain network. Specifically, our framework defines the input as a dynamic graph with dFC as the edge

feature, and the feature extracted from the fMRI signal and spatial one-hot encoding as node features. We introduce a multi-hop spatial attention module to calculate graph attention from both node and edge features. Then, we extend the attention matrix to a multi-hop attention matrix and employ a Transformer encoder [66] to extract temporal information over the time-varying dynamic graph feature series. Moreover, we draw inspiration from RevCon loss [67], which originated from the domain generalization research field. RevCon learning aims to improve the generalizability of our model in cross-site heterogeneous datasets by expanding the representation space of the extracted features. The main contributions of this study are summarized as follows:

1) Proposal of a multi-hop spatial attention block and implementation of a multi-hop diffusion process after the attention mechanism to incorporate a wider range of interactions between indirectly connected brain regions.

2) Utilization of Reverse Contrastive Learning to expand the latent space of samples within the same class, thereby improving the model's generalizability in the presence of individual and site heterogeneity.

3) Demonstrating the efficacy of MSTGCN not only in identifying epilepsy but also in predicting surgical outcomes, showcasing its potential clinical applications in both diagnosis and clinical workup and prognosis.

3. Methods

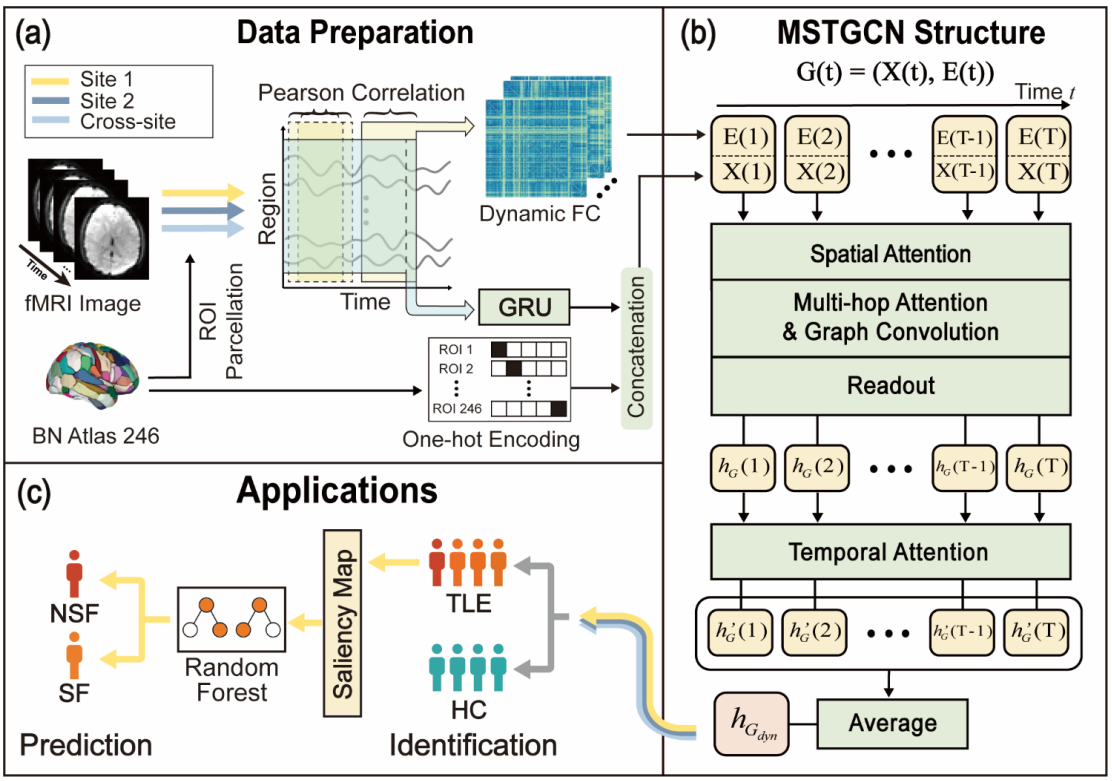

Figure 1: Overview of the framework of our proposed Multi-hop Spatio-temporal Graph Convolutional Network (MSTGCN):

(a) The data preparation process involves several steps. First, ROI-level fMRI signal extraction is conducted, followed by the construction of dynamic functional connectivity (dFC) for the edge feature E(t). Subsequently, the one-hot encoding is concatenated with the output of the Gated Recurrent Unit to generate the node feature X(t). (b) Overall structure of MSTGCN. A sequence of dynamic graphs is first input to the Edge Attention, followed by Attention Diffusion, which produces a sequence of spatially attended graph representation vectors. Temporal attention is computed over and the temporally attended graph representations are averaged to produce the final representation. (c) Application of the extracted representation $h_{G_{dyn}}$ for epilepsy patient identification and surgical outcome prediction. First, we train the model for TLE identification, then we use the saliency maps of patients from the trained model for surgical outcome prediction.

3.1 Datasets

Table I: Demographic and clinical information of the subjects from two sites

Datasets	Subjects	Age (Years)[a]	P-value of age [b]	Gender (M/F)	P-value of gender [c]	Surgical Outcome (SF/NSF)
Site 1	TLE 71	29.99 ±11.34	0.7816	35/36	0.3708	39/24
	HC 74	31.03 ±9.182		31/43		N/A
Site 2	TLE 100	24.30 ±7.671	0.1165	47/53	0.8835	N/A
	HC 79	22.23 ±9.908		38/41		N/A

[a] Data are presented as the mean value ± SD
[b] p value obtained by two-tailed Pearson chi-square test
[c] p value obtained by two-tailed two-sample *t*-test

Xiangya Dataset: The proposed model is validated on two datasets, one of which is the Xiangya Dataset. This dataset includes 145 subjects recruited from the Xiangya Hospital of Central South University (Site 1), comprising 74 patients with TLE and 71 healthy controls. Notably, 63 patients with TLE underwent surgical resection, and their follow-up outcomes were assessed based on the Engel Surgical Outcome Scale as either seizure-free (SF; Engel class IA) or not SF (NSF, Engel class IB to IV). The resting-state fMRI data were acquired using a 3.0 Tesla Siemens Prisma MRI system with a standard 32-channel head coil. Scans were performed using an echo-planar imaging sequence with the following parameters: repetition time (TR) =720 ms, echo time (TE) =37 ms, flip angle=52°, 64 axial slices with 2.5 mm thickness and 2.5 mm spacing, matrix size=90×90, field of view (FOV) =225×255 mm2, and voxel size=2.5×2.5×2.5 mm3. Each resting-state functional sequence lasted 9.456 min, resulting in 788 volumes.

Zhengzhou Dataset: The second dataset contains 179 subjects, including 100 patients with TLE and 79 HC. This dataset was collected from the First Affiliated Hospital of Zhengzhou University (Site 2). Corresponding resting-state fMRI data were acquired on a 3.0 Tesla Siemens Prisma MRI system with a standard 64-channel head coil. Scans were performed using an echo-planar imaging sequence with the following parameters: TR = 1000 ms, TE = 30ms, flip angle = 70°, 52 axial slices with 2.2 mm thickness and 1 mm gap, matrix size = 110 × 110, field of view = 220 × 220 mm2, voxel size = 2.0 × 2.0

× 2.2 mm3. Each resting-state functional sequence lasted 400 seconds, resulting in 400 volumes. The demographic characteristics of these two datasets is shown in Table 1.

3.2 Overview

Fig.1 illustrates the proposed MSTGCN diagnostic and prognostic framework, which consists of Dataset preparation, MSTGCN structure, and Applications. First is Data Preparation, as shown in Fig.1(a), where ROI -level fMRI signals are extracted using the BN atlas template [36], and dynamic graph nodes and edge features are constructed as model inputs. Next is the MSTGCN structure, where the constructed dynamic graph is inputed into the model. As shown in Fig.1(b), our model mainly consists 4 modules, including Spatial Attention module, Multi-hop Attention and Graph Convolution module, Readout module, and Temporal Attention module. The model updates node features based on multi-hop information through the Spatial Attention, Multi-hop Attention and Graph Convolution. After readout for node information as graph-level representations, a temporal attention module extracts dynamic information and aggregates features into to a dynamic graph-level representation by averaging. Finally, for the down-stream applications in Fig.1(c), the dynamic graph-level representation is used for diagnosis by a classification head and a contrastive head, for computing the Cross-entropy loss and RevCon Loss. Additionally, saliency maps of the patient data are used as features and combined with surgical information to further predict the surgical outcome.

3.3 Construction of Dynamic Graph

Fig.1(a) illustrates construction of the input dynamic feature graph. Time-series data for 246 ROIs were extracted from whole brain data using the BN Atlas [68] template. The values of time-series were standardized across time, and a sliding window approach was used to divide them into T windows. For each window, functional connectivity was defined as the Pearson correlation coefficient calculated between the time-series of two ROIs. Each ROI is considered a node in our graph, with functional connectivity serving as the edge between each pair of nodes. The graph for each window creates a dynamic graph series of T graphs for each time step, from step 1 to step T, as is shown in Fig.1(a).

As is shown in Fig.1 (b), $G_{dyn} = (G(1), \ldots, G(T))$ represents a series of dynamic graph with T time steps. At each time t, the graph $G(t) = (X(t), E(t))$ consists of an edge set $E(t) = \{E_1(t), \ldots, E_N(t)\} \in \mathbb{R}^{N*N}$, and vertex set $X(t) = \{X_1(t), \ldots, X_N(t)\} \in \mathbb{R}^{N*D}$. Here, N denotes the number of nodes in each graph, and D is the dimension of node features for each hidden layer (N = 246 according to the BN Atlas template). The temporal variation of the functional connectivity between the nth ROI and all other ROIs, denoted as $E_n(t) = (\mathrm{n} = 1, \ldots, \mathrm{N})$, undergoes changes over time. The feature vector of the nth ROI, denoted as $X_n(t)$(n = 1, …, N). Unlike conventional definitions of node feature vectors at node index n, such as coordinates [69], mean-activation [70], or other handmade features from fMRI [39], the node feature $X_n(t)$ is defined with two concatenated parts:

$$X_n(t) = W[e_n || s_n(t)] \tag{3-1}$$

where e_n represents the spatial one-hot encoding for the nth node, while $s_n(t) \in \mathbb{R}^D$ is a learnable timestamp encoded feature derived from a Gated Recurrent Unit [71], the concatenate operation is denoted by $||$, and $W \in \mathbb{R}^{D*(N+D)}$is learnable parameters for feature linear mapping.

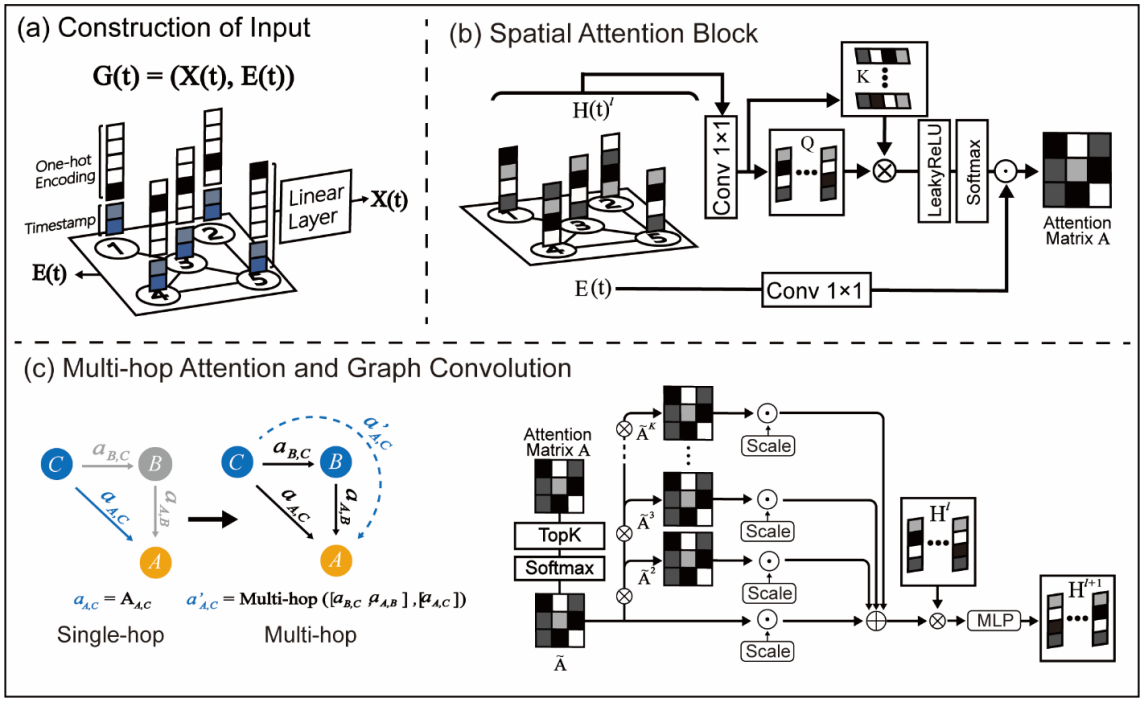

Figure 2: Details of the input features and the Spatial Attention and Multi-hop Attention modules

(a) Input includes edge feature E(t) and node feature X(t). The node feature X(t) is the concatenation of the one-hot encoding and the output of the Gated Recurrent Unit at time-point t (input with ROI level fMRI signal time-point by time-point). (b) Computation of the weighted adjacent matrix with attention. (c) Illustration of attention diffusion and computation. The Left part of (c) is a comparison between one-hop attention and multi-hop attention. According to the information propagation method, one-hop only aggregates information from the one-hop direct edge of the adjacent matrix, while multi-hop aggregates information from both the direct and indirect edges. The right part of (c) is the multi-hop attention diffusion process and graph convolution operation.

3.4 Spatial Attention Mechanism

Fig.2(b) illustrates our spatial attention module, which calculates an attention-weighted adjacent matrix. The conventional methods define edges as remaining static during the graph convolution operation. To determine which node to aggregate in the graph convolution operations, functional connectivity is simply processed by thresholding [70], top-k [42] or similarity [72] methods for binarization. However, these approaches ignored significant precise information of the connections. Velickovic et al. [73] proposed GAT to aggregate neighboring nodes with different weights, enabling more effective information propagation. However, original GAT only considered connectivity relationship and ignored edge features. Our Spatial Attention mechanism can aggregate node information with emphasis by computing an attention-weighted adjacent matrix $A(t)$ that incorporates both node similarity and functional connectivity information of edges. Following the inspiration of node-edge attention [39], we define the feature representation of nodes, $H^l(t)$ (where $H^0(t) = X(t)$) as Q, K $(Q = K)$, and $E(t)$ as V, as set in the self-attention mechanism. The superscript $l \in [1, L]$ denotes the l layer of our model, and L is the total number of layers of MSTGCN. We compute the attention-weighted adjacency matrix as follows:

$$Q = K = \mathrm{Conv}\Big(H^l(t)\Big), V = \mathrm{Conv}(E(t)) \tag{3-2}$$

$$A^l(t) = \mathrm{Softmax}\big(\mathrm{LeakyReLU}(QK^T)\big) \odot V \tag{3-3}$$

The feature of $X^l(t)$ and $E(t)$ in Eq. 3-2 is augmented using 1×1 convolution operation. Next, the similarity matrix of nodes is computed using QK^T. The Softmax

function is then used to project each line of the similarity matrix into an attention matrix, where the summation of each line is 1. The FC matrix undergoes a 1×1 convolution operation and is defined as V. Finally, the projected FC matrix is then used to obtain the attention-weighted adjacent matrix $A^l(t)$ by product. Therefore, $A^l(t)$ considers both node and edge features and is effective in learning graph representation.

3.5 Multi-hop Attention and Graph Convolution

The Multi-hop Attention method uses attention diffusion to expand the attention-weighted adjacent matrix $A^l(t)$ into a multi-hop attention form P. Firstly, a sparsification operation is performed on the adjacent matrix $A^l(t)$. Performing TopK and Softmax operations on each row of the attention matrix $A^l(t)$ pushes the values of edges with smaller attention weights to nearly 0, while ensuring that the sum of each row equals 1, as shown in Eq. 3-4:

$$\tilde{A} = \mathrm{Softmax}\Big(\mathrm{TopK}(A^l)\Big) \tag{3-4}$$

P considers attention between nodes that are not directly connected. The graph diffusion operation is utilized on the attention matrix to help nodes aggregate neighboring nodes beyond one hop. The diffusion process is formulated as follows:

$$P = \sum_{i=0}^{K} \theta_i \big(\tilde{A}\big)^i \ , \sum_{i=0}^{\infty} \theta_i = 1 \ \ and \ \ \theta_i > 0 \tag{3-5}$$

$$\theta_i = \alpha(1-\alpha)^i, \alpha \in (0,1] \tag{3-6}$$

The K-th power of $\tilde{A}$ represents an adjacency matrix that includes the number of K-hop paths, extending the receptive field. The scale factor θ_i is used to weight the K-hop adjacent matrix, with $\theta_i > \theta_{i+1}$. As described in [74], we utilize the geometric distribution θ_i to scale the weight of the adjacent matrix with more hops, as shown in Eq. 3-6. This selection is based on the inductive bias that nodes that are farther away should carry less weight in the message aggregation process. Scale factors are assigned independently in a sequential manner to nodes with varying path lengths to the target node. The whole updating process of node features uses the graph convolution operation as follows:

$$H^{l+1} = \mathrm{MLP}(PH^l;\ \Theta) \tag{3-7}$$

where $H^l \in \mathbb{R}^{N\times D}$, the PH^l is information propagation operation. And updating process of node features is different from traditional method, which uses one fully-connected layer, as is used in [75]. Instead, we use a Multi-layer Perceptron (MLP) with batch normalization for aggregation, as in previous work [43, 74]. Θ represents learnable parameters of the MLP. However, to compute the P in Eq.3-5 can be costly, as it involves computing the power of matrix [76], which can have a complexity of up to $O(N^3)$, thereby slowing down the training and inference process. Building on the methodology of Approximate Personalized Propagation of Neural Prediction [76], we utilize an iteration manner to approximate the information propagation process PH^l as shown in Eq. 3-8. Corresponding updating process is formulized in Eq. 3-9:

$$H^l_{(k+1)} = (1-\alpha)\tilde{A}H^l_{(k)} + \ \alpha H^l_{(0)}, k \in [1, K] \tag{3-8}$$

$$H^l_{(0)} = \ H^l, H^{l+1} = \mathrm{MLP}\left(H^l_{(K)}\right) \tag{3-9}$$

In Eq. 3-8 , the $H^l_{(k)}$ is an intermediate variable of H^l with k hop attention information and α is the teleport probability, set as shown in Eq. 3-6. As K approach to infinity, the $H^l_{(K)}$ converges to $\tilde{A}H^l$ [74]. The aggregated node features will be updated with MLP in Eq. 3-9. The proof process is shown in Eq. 3-10:

$$H^l_{(K)} = ((1-\alpha)^K\tilde{A}^K + \alpha\sum_{i=0}^{K-1}(1-\alpha)^i\tilde{A}^i)\ H^l_{(0)} \tag{3-10}$$

The term $(1-\alpha)^K\tilde{A}^K$ converges to 0 when $K \to \infty$, because $\alpha \in (0,1)$ and elements of $\tilde{A}$ is also in $(0,1)$. Thus, $\lim_{K\to\infty} H^l_{(K)} = (\sum_{i=0}^{\infty}\alpha(1-\alpha)^i\tilde{A}^i)H^l$, which is consist with Eq. 3-5 and Eq. 3-6 By this way of approximation, the complexity of multi-hop attention computation is $O(N^2)$.

3.6 Readout Block

MSTGCN utilizes a readout block for the representation of all nodes into a graph representation feature vector. To derive the representation of the whole graph, we employ an average pooling operation combined with a max pooling operations to capture the multi-view features of the graph representation. Average pooling averages over all node representations to obtain a fixed size graph level representation, and max pooling takes the maximum feature over all nodes. Finally we define our readout function as $h^l_G =$

$\text{Avg}(H^l) + \gamma\ \text{Max}(H^l)$, $h_G^l \in \mathbb{R}^D$, where γ is the weight of max pooling. For each time point we have a corresponding graph representation $h_G^l(t)$.

3.7 Temporal Attention Block

To extract temporal information, we utilize a widely used Transformer encoder [66] to create our graph representation sequence $\boldsymbol{S}^l = (h_G^l(1), h_G^l(2), \dots, h_G^l(T))$ with one head attention, where $\boldsymbol{S} \in \mathbb{R}^{D\times T}$. For each layer of our model, we compute a dynamic graph representation by averaging the output feature of the Transformer encoder across time, which we define as $h_{G_{dyn}}^l$ for layer l. The Transformer encoder in our work is formulated as follow for each layer l:

$$(\boldsymbol{S}^l)' = \text{SelfAttention}(\boldsymbol{S}^l) + \boldsymbol{S}^l \tag{3-11}$$

Where SelfAttention is identical to the Transformer encoder, and $(\boldsymbol{S}^l)' \in \mathbb{R}^{D\times T}$. The feed-forward network calculation process is as follows:

$$(\boldsymbol{S}^l)'' = \text{LN}(\text{MLP}(\text{LN}((\boldsymbol{S}^l)');\ \Theta) + (\boldsymbol{S}^l)') \tag{3-12}$$

Where LN represents layer-normalization operation, MLP stands for multi-layer perceptron, Θ is its trainable parameter, and $(\boldsymbol{S}^l)'' \in \mathbb{R}^{D\times T}$. For the final step with the classification and RevCon loss, we define the input as follow:

$$h_{G_{dyn}}^l = \text{Avg}(\boldsymbol{S}^l)'' \tag{3-13}$$

$$h_{G_{dyn}} = concat\left(h_{G_{dyn}}^1, h_{G_{dyn}}^2 \dots, h_{G_{dyn}}^L\right) \tag{3-14}$$

where Avg stands for average operation, $concat$ represents concatenation operation, L is the total number of layers in our model. $h_{G_{dyn}}^l \in \mathbb{R}^D$ is the output of layer l. $h_{G_{dyn}}$ represents the concatenation of the output of each layer. The output will be processed with a classification head and a contrastive head, respectively. The two heads consist of one layer of fully-connected layer with different output dimensions. In our work, the output dimension of classification head corresponds to the number of categories, which is 2. On the other hand, the output dimension of the contrastive head is D_c.

3.8 Reverse Contrastive Learning

Deep neural networks (DNNs) are effective in capturing correlations between patterns and labels. However, when it comes to DNN classification, there is a tendency to heavily rely on the simplest and most predictive patterns [77, 78]. It is important to note that patterns learned during the training process may turn out to be misleading, leading to potential oversight of authentic patterns within the test dataset and compromising the network's generalizability. Furthermore, data collection can introduce misleading patterns [79, 80], including subtle and specific cues that are unique to certain hospitals [81]. When using these patterns for classification tasks, the focus is on maximizing inter-class specificity while neglecting intra-class variability [81]. While these patterns may be highly effective for classification, they also run the risk of being misleading and inauthentic. Therefore, one promising approach is to encourage neural networks to identify patterns with higher intra-class variability using RevCon learning [67]. The primary objective of RevCon learning is to expand the intra-class representation of neural networks. In this work, we posit that the significant inter-individual and collection site differences within fMRI data may contribute to the learning of misleading patterns. To address this issue, we introduce the concept of RevCon learning to our approach. The output feature vector of the contrastive head is defined as $f \in \mathbb{R}^{D_c}$. RevCon loss function is outlined as follows:

$$\mathcal{L}_{RC} = -\frac{1}{N_p}\sum_{n=1}^{N_p} d(f_a, f_p), where\ \ p\ \in Pos(a) \tag{3-15}$$

The function involves the use of feature vectors f_a and f_p, where $f_a \in \mathbb{R}^{D_c}$ and $f_p \in \mathbb{R}^{D_c}$ are feature vectors of sample a and its positive sample p, respectively. Positive samples are defined as samples within the same class, and a positive sample pair is a pair of samples within the same class. $Pos(a)$ represents the set of indices of all positive samples of sample a in the batch. N_p represents the number of positive sample pairs in a mini-batch, and $d(f_a, f_p)$ denotes the cosine similarity between f_a and f_p. The loss function is calculated by multiplying with -1 to maximize the intra-class representation. Combined with classification loss, we have our final loss function as Eq. 3-16:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{RC} \tag{3-16}$$

Where $\mathcal{L}_{CE}$ represents the Cross-entropy loss for classification and $\mathcal{L}_{RC}$ is RevCon loss defined in Eq. 3-10. λ is the weight of $\mathcal{L}_{RC}$ and can adjust our loss function.

3.9 Saliency Map

To enhance the interpretability of our novel model's performance, we use a gradient-based saliency map method [82]. This method reveals significant brain regions and connections that play a pivotal role in influencing the classification outcome, potentially valuable biomarkers. The approach based on gradients enables the creation of saliency maps for individuals in each category by calculating the gradients of inference scores that correspond to input features. This encapsulates the importance associated with each feature [13]. The derivation of node and edge saliency maps is detailed in the following formulation.

$$M_X = \frac{\partial p_d}{\partial X}, M_E = \frac{\partial p_d}{\partial E} \tag{3-17}$$

where M_X and M_E are the saliency map of node and edge features, respectively. p_d is the predicted score of category c corresponding to input. The magnitude of elements of M_X and M_E reflects the importance of the corresponding input feature of X and E for classification.

4. Results

4.1 Preprocessing

All the resting-state fMRI data preprocessing was conducted using a combination of DPARSF software (http://www.rfmri.org/DPARSF) [83] and customized MATLAB scripts. To begin, the initial volumes (18 and 10 volumes for two datasets respectively) were discarded to ensure signal stability. Volume is defined as the fMRI image of each time point. Spatial realignment was performed to correct for motion artifacts. Subsequently, the functional images were normalized to the Montreal Neurological Institute (MNI) space and smoothed using a Gaussian kernel with a full width at half maximum (FWHM) of 6 mm. To address confounding factors, such as linear trends, head motion parameters [84], white matter signal, and cerebrospinal fluid signal, multiple linear regression was utilized. Temporal filtering was then applied using a bandpass filter with a range of 0.01-0.1 Hz. The purpose of filtering is to improve the signal-to-noise ratio of the data by filtering out the data in irrelevant frequency bands. Participants exhibiting excessive head motion (translation > 3mm or rotation > 3 degrees, or micro movement quantified by mean frame-wise displacement exceeding 0.3mm) were excluded from subsequent analyses.

4.2 Experimental Settings

We set three datasets, including Site 1, Site 2 and the combination of Site 1 and Site 2, which is named as Cross-site. All three datasets underwent 5-fold cross-validation, with data divided into 5 subsets of approximately equal size, maintaining a consistent ratio of HC to TLE. Additionally, a stricter Leave-one-site-out validation strategy was employed to test the generalizability of the proposed method. In Leave-one-site-out validation strategy, the test set uses data from one site while the training set uses data from other sites. The classification performance is evaluated using two metrics: accuracy (ACC) and the area under the receiver operating characteristic (AURCO). ACC represents the ratio of correctly classified samples to the total number of samples, indicating the overall classification performance. The receiver operating characteristic (ROC) curve displays the True Positive Rate versus the False Positive Rate at various classification thresholds.

The area under the ROC curve provides a more reliable evaluation performance under category distribution.

The model was implemented using PyTorch. Experiments were accelerated using an NVIDIA GeForce RTX 3090 GPU. To extract fMRI time sequence, we utilized BN Atlas [68] with 246 ROIs. To compute dFC, we employed windows of 50 TRs, 36 TRs, and 50 TRs, with corresponding strides set to 3 TRs, 2 TRs, and 3 TRs for Site 1 (TR = 0.72 seconds), Site 2 (TR = 1 second), and Cross-site, respectively. The number of layers was set to $L = 2$, with an embedding dimension of $D = 128$,and $D_c = 68$. The value of γ in the Readout module as 0.0001 . One cycle learning rate method is employed, gradually increasing the learning rate from 0.0005 to 0.001 during the first 20% of training, and then decreasing it to 5.0×10-7. Each training on all datasets consisted of 30 epoch with a mini-batch size of 16. During the training stage, the time dimension of fMRI sequence was randomly clipped to a fixed length of 300, 200, 200 for Site 1, Site 2 and Cross-site, respectively. This stochastic augmentation of training data relieves computational overload and aligns datasets with different collection lengths, as inspired by previous work [41, 42]. Additionaly, ROIs are randomly flipped from left to right (or vice versa) to augment the training data, as all TLE patients in our study are unilateral. During the testing stage, full-length fMRI sequences are performed, and no flip operation is performed.

4.3 Comptitive Methods

1) **GCNN** (Static) [75]. The Graph Convolutional Neural Networks (GCNN) method was designed for semi-supervised learning and classification of graph-structured data in static graphs. GCNN is motivated by local first-order approximation of spectral graph convolutions, which justifies the choice of convolutional architecture. Batch normalization was added to the original structure for better training performance. In this study, four GNN model for static brain network are included as comparative to demonstrate benefits of utilizing the dynamic characteristic.

2) **GIN** (Static) [70] Graph Isomorphic Networks (GIN) is powerful GNN for graph classification, and Kim et al. developed a framework for analyzing static brain network fMRI data using GIN. They use the saliency map technique for GIN with one-hot encoding to visualize important brain regions, which inspired us.

3) **ChebNet** (Static) [85]. Arslan et al. proposed approach that explores the role of GCNs in ROI recognition. They utilize an activation-based approach using ChebNet, a high-order GCNN, to identify salient graph nodes after a gender classification task with static brain network. We utilize their model for our task of diagnosing TLE.

4) **GraphSAGE** (Static) [86]. Li et al. designed a regularized pool layers that highlight ROIs, allowing the model to infer which ROIs are important for identifying specific diseases based on node pooling scores. In this work, we focus on the classification performance of GraphSAGE with our datasets.

5) **ST-GCN** [57]. The core idea of the ST-GCN method is to develop a framework for analyzing rs-fMRI data based on spatio-temporal graph convolution. The model is constructed based on dFC, defining neighboring nodes not only in spatial, but also in temporal aspects. This allows the model to learn the importance of both spatial and temporal features through spatial and temporal convolution operations.

6) **ST-fMRI** [87]. ST-fMRI models the long-range spatio-temporal dynamics of dFC by introducing a bone-based motion recognition method named MS-G3D. To account for inter-subject cortical heterogeneity, they use double regression ICA maps to explain the intersubjective variability in functional organization. For a fair comparison, we utilize the same ROI-level parcelation with BN Atlas as our proposed framework.

7) **STAGIN-SERO** and **STAGIN-GARO** [28]. Spatio-temporal attention graph network isomorphism (STAGIN) is proposed based on attention mechanism for dynamic brain network representation. STAGIN includes two attention-based readout methods for spatial information, namely SERO and GARO, and uses Transformer encoders for temporal information of dFC. STAGIN is evaluated on a task of TLE diagnosis, and both of the readout methods are tested.

4.4 Performance on Epilepsy Classification

The performance of our model was evaluated on three datasets, Site 1, Site 2 and Cross-site. We compared it with several brain network analysis methods, including sFC-based methods such as GCNN [75], GIN [70], ChebNet [85] and GraphSAGE [86], and dFC-base methods such as ST-GCN [57], ST-fMRI[87], STAGIN with GARO (STAGIN-GARO) [42] and STAGIN with SERO (STAGIN-SERO) [42]. Table II summarizes the results, indicating that our model achieved the highest classification accuracy of 85.52%

and the highest AUROC of 0.913 on Site 1. Similarly, our model outperformed alternatives on Site 2, achieving the highest classification accuracy of 78.27% and an AUROC of 0.857. The merged Cross-site dataset was also tested, and MSTGCN outperformed other compared method with an ACC of 82.09% and an AUROC of 0.869.

Table II: The classification results of different methods on HC vs. TLE

Method	Site 1 (n = 145)		Site 2 (n = 179)		Cross-site (n = 324)	
	ACC (%)	AUROC	ACC (%)	AUROC	ACC (%)	AUROC
GCNN [75] (Static)	66.90 ± 5.770	0.728 ± 0.071	62.56 ± 11.27	0.673 ± 0.093	61.12 ± 6.707	0.683 ± 0.063
GIN [70] (Static)	66.90 ± 14.343	0.730 ± 0.085	61.43 ± 13.88	0.635 ± 0.162	65.43 ± 3.998	0.712 ± 0.042
ChebNet [85] (Static)	70.34 ± 9.317	0.788 ± 0.073	64.79 ± 8.959	0.645 ± 0.151	64.20 ± 2.277	0.702 ± 0.035
GraphSAGE [86] (Static)	71.03 ± 9.317	0.781± 0.114	58.68 ± 9.974	0.606 ± 0.136	60.47 ± 7.561	0.655 ± 0.069
ST-GCN [57]	71.18 ±6.783	0.805 ± 0.069	69.24 ± 9.176	0.712 ± 0.126	70.83 ± 6.783	0.787± 0.045
ST-fMRI [87]	77.93 ±16.98	0895 ± 0.051	66.43 ± 6.322	0.775 ± 0.078	75.90 ± 4.915	0.856± 0.044
STAGIN-SERO [42]	78.62 ±5.115	0.884 ± 0.053	76.00 ± 6.602	0.809 ± 0.089	77.15 ± 4.580	0.860 ± 0.024
STAGIN-GARO [42]	82.76 ±5.452	0.852 ± 0.039	68.14 ± 6.471	0.777 ± 0.074	75.93 ± 5.134	0.850 ± 0.057
Proposed	85.52 ± 5.115	0.913 ± 0.048	78.27 ± 8.092	0.857 ± 0.077	82.09 ± 3.907	0.869 ± 0.036

4.5 Generalization of Model

We compared the generalizability performance using Cross-entropy (CE) loss, Adversarial (Adv) loss and RevCon loss on the Cross-site dataset, a merged dataset. For Adv loss, we used FGSM [88] method, which has been proven effective in enhancing generalizability with a novel independent dataset [39] with static FC. The adversarial sample and the final adversarial loss function $\mathcal{L}_{Adv}$ follow previous work and are described in the following:

$$\mathcal{L}_{ADV} = (1-\lambda)\mathcal{L}_{CE}(x,y,\theta) + \lambda\mathcal{L}_{CE}(x_{ADV},y,\theta) \quad (4\text{-}1)$$

$$x_{ADV} = x + \epsilon Sign\big(\nabla_x L_{CE}(x,y,\theta)\big) \quad (4\text{-}2)$$

where x, y and θ represent the input normal sample, label and model parameters, respectively. The λ is set to 0.5, following Chen et al. [39]. The magnitude of the deviation of the adversarial sample from the normal sample is represented by ϵ. The comparison results are shown in Fig.3.

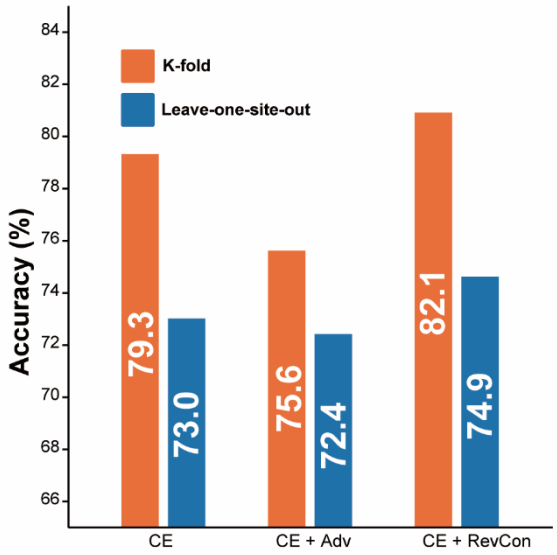

Figure 3: Model generalizability of TLE classification performance on Cross-site dataset validation:

Results are reported for three training methods, using 5-fold and Leave-one-site-out cross-validation strategies.

Fig.3 illustrates the comparison of classification performance CE loss, Adv loss and RevCon loss. When using only CE loss, the performance of ACC is 79.3% and 73.0% for K-fold and Leave-one-site-out validations, respectively. When adding adversarial loss, there is a decrease in accuracy for both validation methods (75.6% and 72.4%, respectively). The results differ from those reported in [39]. The difference may be caused by the discrepancy between inputted dynamic and static FC and model structure. Our RevCon loss improved the classification accuracy on both validation strategies (82.1% and 74.9%, respectively), indicating its effectiveness in improving generalizability. It should be noted that our result of adversarial training resulted in the best performance with $\epsilon \in [0.001,0.005,0.01,0.02,0.05,0.1,0.2]$.

4.6 Ablation Study

Ablation studies were conducted to validate the effectiveness of the proposed model structure on classification performance. It should be noted that the ablation study was performed based on three datasets to increase its dependability.

Multi-hop Mechanism: The effectiveness of the multi-hop mechanism in our model was evaluated by removing the attention diffusion block and directly performing the

graph convolution operation on the calculated spatial attention matrix. The attention matrix is calculated by Eq.3-2 and Eq.3-3, which is as one-hop attention. The accuracy in Table III decreases when using traditional one-hop attention, which proves that our multi-hop attention diffusion operation enhances the capacity of spatial representation.

Spatial Attention: In our proposed model, we utilized spatial attention and expanded it to multi-hop attention. To validate the effectiveness of the spatial attention module, we replaced the spatial attention adjacency matrix calculated by Eq. 3-2 and Eq. 3-3with the traditional threshold method. Therefore the adjacent matrix $A^l = D^{-(1/2)}\hat{A}^l_{\text{thrd}}D^{-(1/2)}$, and $\hat{A}^l_{\text{thrd}} = A^l_{\text{thrd}} + I$, where A^l_{thrd} is a binarized adjacency matrix with threshold, I is identity matrix and D is the degree matrix of $\hat{A}^l_{\text{thrd}}$. The result reported in Table III demonstrates that the attention mechanism effectively enhances the classification performance.

Temporal Attention: Temporal processing module is adopted to handle dynamic characteristics within the dynamic brain network. In this part we removed Transformer block and replaced it with an average operation on the graph representation sequence $(h_G(1), h_G(2) \dots, h_G(T))$, which is calculated by Eq.3-9. Table III indicates a decrease in accuracy, indicating the significance of the dynamic feature of the brain network and the effectiveness of the temporal attention block for dynamic feature representation.

Reverse Contrastive Loss: To validate the effectiveness, we removed RevCon loss $\mathcal{L}_{RC}$ and only used traditional Cross-entropy loss. The results in Table III depict that the $\mathcal{L}_{RC}$ enhances the classification performance. The effectiveness of $\mathcal{L}_{RC}$ was tested on Site 1 and Site 2, resulting in an increase in classification performance.

Readout module: As is stated in section 3.5, the readout module in our proposed model includes both max pooling and average pooling. In the ablation study we removed the max pooling to evaluate the effect of Avg+Max pooling on our results. Table III shows that our Avg+Max pooling improves classification performance.

Overall, the modules utilized in our proposed framework are effective in extracting discriminative patterns from dFC, thereby improving the performance.

Table III: Ablation study

Model Structures	Ablation Settings	Site 1 (n = 145)		Site 2 (n = 179)		Cross-site (n = 324)	
		ACC (%)	AUROC	ACC (%)	AUROC	ACC (%)	AUROC
Multi-hop Spatial Attention	No Multi-hop	80.00±6.633	0.869±0.055	74.90±6.660	0.843±0.090	79.01±5.068	0.846±0.043
	No Attention	83.45±10.74	0.912±0.070	74.33±6.797	0.826±0.094	79.31±5.626	0.875±0.0035
	No Multi-hop & Attention	82.76±9.753	0.908±0.058	75.44±7.097	0.807±0.091	76.85±3.926	0.839±0.042
Temporal Attention	Mean	83.45±8.233	0.884±0.085	76.54±3.546	0.840±0.034	76.54±3.546	0.845±0.051
Loss Function	Only Cross-entropy	80.69±6.264	0.899±0.075	75.98±3.667	0.829±0.072	80.55±5.621	0.865±0.039
Readout	Only Average Pooling	84.14±5.230	0.856±0.046	78.83±8.378	0.861±0.079	79.93±6.345	0.856±0.046
Proposed	-	85.52 ± 5.115	0.913 ± 0.048	78.27 ± 8.092	0.857 ± 0.077	82.09 ± 3.907	0.869 ± 0.036

4.7 Impact of Model Hyperparameters

This study examines impact of model hyperparameters on classification performance. We focus on three main hyperparameters. The first hyperparameter, denoted as K in Eq.3-5, represents the number of hops. Our results indicate that the best performance for Site 1 and Site 2 is achieve with 4 hops, while Cross-site achieves the best classification performance with 5 hops. We found that the optimal multi-hop number K depend on different datasets, which is consistent with prior research [43, 62]. I note that the performance of single hop is in ablation study. The teleport probability α is the second parameter in Eq. 3-8, The optimal performance is achieved when α is set to 0.15. Beyond this value, there is a slight decrease in performance for Site 1 and Cross-site. Prior research has shown that a small value of α increases the low-pass effect, directing the model's focus towards large-scale graphs and eliminating noisy high-frequency information [74]. Conversely, an excessively small value of α can cause the model to overly focus on large-scale graphs, resulting in performance degradation. The third hyperparameter is the weight of the RevCon loss $\mathcal{L}_{RC}$. Performance shows an upward

trend as λ increases, up until a certain point. An excessively large value of λ will lead to a collapse in validation accuracy. The results are depicted in Fig.4.

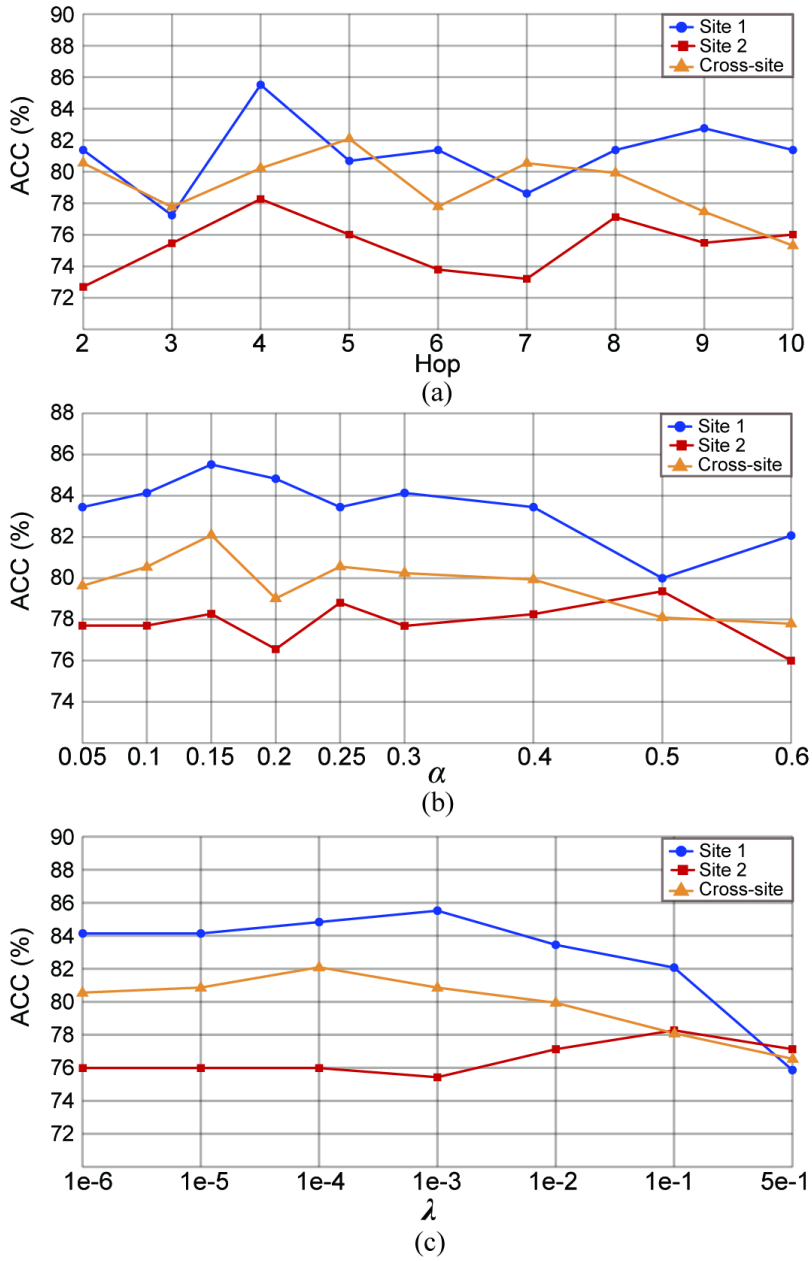

Figure 4: Impact of model hyper parameters:

(a) Classification accuracy with different values of number of hop. (b) Classification accuracy of different teleport probability α. (c) Classification accuracy of different weight of Reverse Contrastive Loss, which is represented by λ.

4.8 Discriminative Brain Networks to Epilepsy

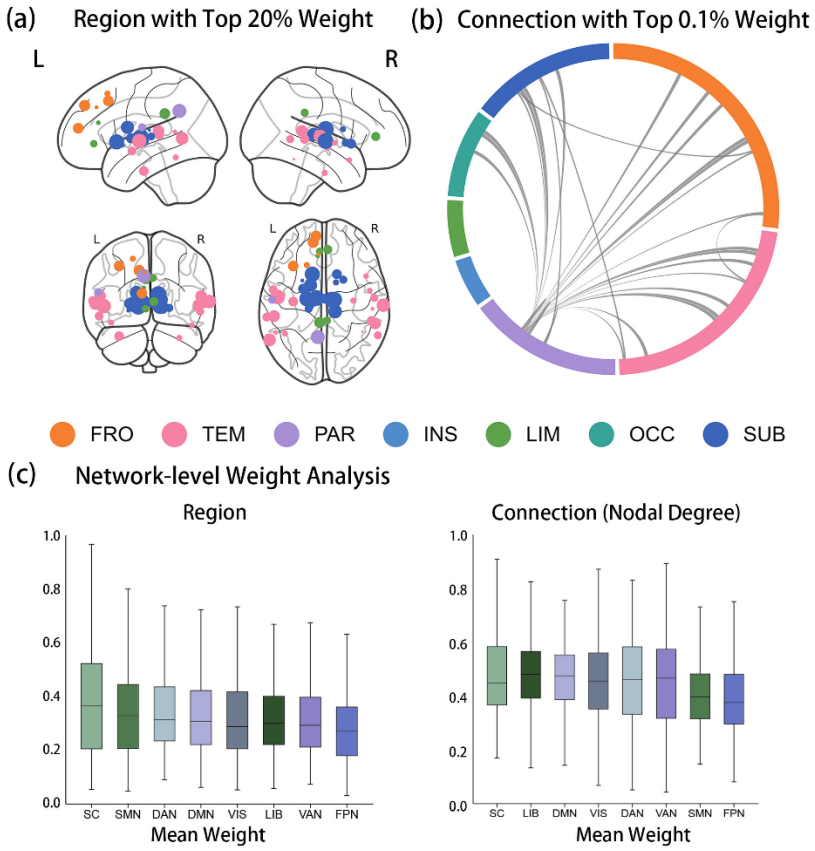

Figure 5: Visualization of the group-level features learned by the proposed model for TLE diagnosis:

(a) displays the top 20% influential ROIs (49 nodes). (b) shows the top 0.1% functional connectivity (30 edges) in a chord map. The size of the nodes and edges denotes the weight of classification. (c) represents the distribution of classification weights in the functional network derived from the BN Atlas (246 ROIs) parcellation. Noted that for the weight of edge feature, i.e. weight of functional connectivity, is included in the calculation of degree to facilitate division into subnetworks.

We implemented the aforementioned saliency map method for model interpretation by selecting the brain regions and connections with significant contribution to the classification. Note that the best performance model trained on cross-site is adopted, as data from two sites indicate better generalizability. We first calculated the node and edge saliency map of the patient group and then selected significant features. Specifically, we

calculated the saliency map of TLE patients on test samples of all folds, and averaged saliency map of M_X and M_E across all time points and normalized the gradients across features. A summation operation was then performed for a saliency map at the TLE group level. We selected the top 20% (49) nodes with the largest absolute values in the node saliency map and visualized these discriminative brain nodes as shown in Fig. 5 (a). The most discriminative brain nodes are concentrated in the subcortical structures and temporal lobe, including the thalamus, basal ganglia, temporal sulcus, superior and middle temporal gyrus, and fusiform gyrus. These identified subcortical and cortical regions play a central role in the initiation and propagation of temporal lobe seizures [89]. In addition, the cingulate gyrus and frontal lobe (medial superior frontal areas) also contribute significantly to the classification.

Similarly, we selected the top 0.1% (30) edges with the largest absolute values in the edge saliency map at the TLE group level. The results are visualized in Fig.5(b). Edge features showed that the connections related to the precuneus, which is a ROI in the parietal lobe, contributed most to the identification of TLE patients. The precuneus, as a core hub of the posterior default mode network (DMN), has been typically been implicated in epileptogenic networks [90, 91]. In addition, connections related to the temporal lobe, subcortical nuclei, and frontal lobe have also made important contributions to TLE identification.

It has been shown that TLE is related to a group of resting-state networks. In this case, we further performed a statistical result of regions and connections, which are divided into 7 classical subnetworks, according to the work of Yeo et al. [92]: DMN, limbic network (LIB), ventral attention network (VAN), dorsal attention network (DAN), frontoparietal network(FPN), visual network (VIS), somatomotor network (SMN). We also included the subcortical (SC) network in this statistical analysis. Firstly, to analyze the region-level saliency map in subnetworks, we computed the mean value of the region-level saliency map within each subnetwork for each subject. As depicted in the left panel of Fig. 5(c), the distributions of region-level feature importance are mostly concentrated in the SC network, followed by the SMN, DAN, and DMN. Second, to assess the connection weight in subnetworks, we used the degree of nodes as a representation of connection weight. We then calculated the averaged nodal degree within each subnetwork for each subject. The result is shown in the right panel of Fig. 5(c). SC, LIM and DMN

are the top 3 subnetworks that made significant contributions to the identification of epilepsy from the HC in our proposed model.

4.9 Epilepsy Surgical Prognosis

The saliency map of the models can be seen as a reflection of the potential of biomarker mining. We implemented the aforementioned saliency map result for model interpretation to perform the surgical outcome prediction. First, we obtained the surgical resection template of each patient by inspecting the difference between the preoperative T1 images and the postoperative T1 images. Then, we defined standard resection mask at the group level by merging each patient's individual resection template (T1 images of right-sided TLE patient were flipped), thus creating 21 brain ROIs. Consequently, the features were divided into surgically resected features and spared features, as shown in Fig. 6(a). Resected features are within the resected ROIs while spared features are features that has no connections with the resected ROIs. We use feature selection (F-score) and random forest to classify patients into SF and NSF. As can be seen from Fig. 6(b), the models based on spared features achieve better surgical outcome prediction performance than the models based on resected features. In particular, the AUROC and ACC are best with the spared node saliency map, which are 0.829 and 82.0%, respectively. We also present importance of top 10 node features in Fig.5(c). Surgically spared nodes including the middle frontal gyrus, parahippocampus and hippocampus are important for the surgical outcome prediction.

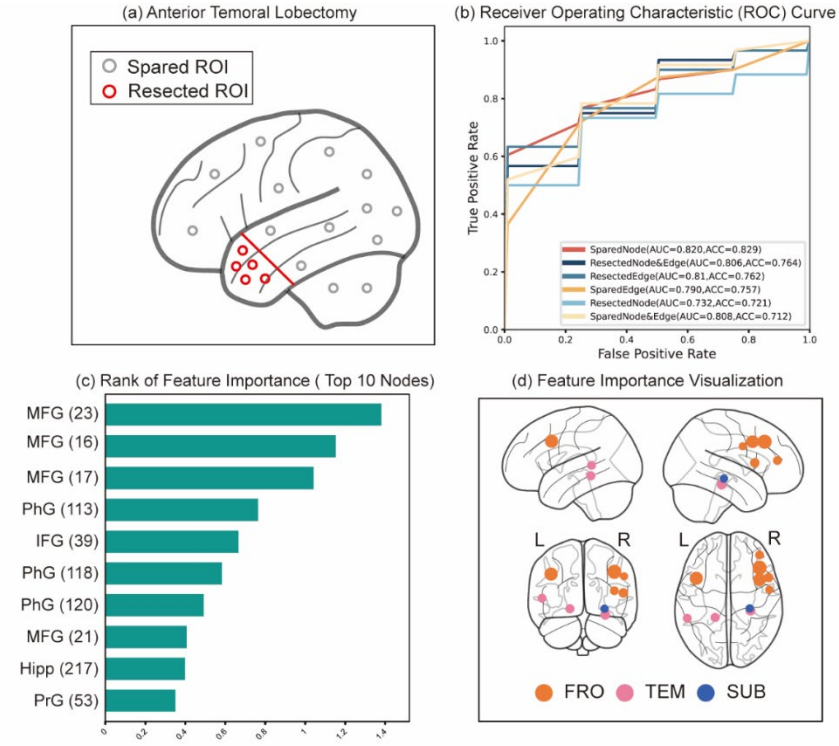

Figure 6: Results of surgical outcome prediction based on the saliency map:

(a) illustrates the anterior temporal lobectomy. With surgery information, ROIs are divided into resected ROIs and spared ROIs. (b) shows the ROC of surgical outcome prediction using six different way of input feature construction. (c) displays the top 10 feature importance of the spared node with the best prediction performance. (d) Visualization of the importance of node feature.

5. Discussion

5.1 Effect of Multi-hop Attention

In this study, we proposed the MSTGCN framework to provide complex multi-hop spatio-temporal representations of dynamic brain functional networks. The multi-hop mechanism has the objective alleviating the over-smoothing problem that arises when performing traditional one-hop graph attention. Furthermore, the multi-hop mechanism does not introduce additional trainable parameters while expanding the receptive field, thus enhancing the capacity of graph representation without increasing the model complexity. The experimental results of our method and comparative methods on epilepsy diagnosis are presented in Table II. Our method achieved the highest ACC values of 85.52% (Site 1), 78.27% (Site 2), and 82.09% (Cross-site) in the classification of TLE patients vs. HC, which are higher than the other methods. Furthermore, our method also demonstrated the highest area under the receiver operating characteristic curve (AUROC) values, indicating its superior performance.

The ablation experiments indicate that the multi-hop diffusion operation improves the cross-site classification ACC by more than 2% compared to a model without a multi-hop attention mechanism. Furthermore, the multi-hop strategy and attention mechanism are validated in an ablation study. The result of a model without a multi-hop strategy shows that the multi-hop strategy is effective in TLE diagnosis and better than a one-hop situation in our framework. The results also indicated the effectiveness of the attention mechanism, which is consistent with prior work [62]. We also conducted an ablation study on Sites 1 and 2, and obtained an ACC improvement of 5% and 3.37%, respectively. These results suggest that the multi-hop spatial attention mechanism is a superior strategy to enhance the graph spatial representation ability without the need to stack additional layers or parameters. Since the average operation can ignore the temporal information within dynamic brain networks, the ablation result of the temporal attention module demonstrates the contribution of temporal information for TLE diagnosis. Prior works have shown that the dynamics of brain network of epilepsy patients can be abnormal [4, 7], which supports the performance increase of MSTGCN. The ablation study on the proposed readout function is inspired by works in GNN-related research [93]. The results demonstrate that the readout function in MSTGCN is capable of enhancing performance.

Furthermore, the efficacy of RevCon loss is evident in the improved accuracy of the Site 1 and Cross-site datasets. Notably, the addition of RevCon loss to the Cross-site dataset resulted in an increase of over 2% in accuracy.

Additionally, an experiment was conducted to assess the impact of hyperparameters associated with multi-hop attention in our work. The results demonstrated that employing the multi-hop mechanism contributes to enhancing classification performance. Moreover, it was found that the optimal multi-hop number K depended on different datasets. This finding aligns with previous research [62]. But if the number of hop is too large, the performance declined. This phenomennon may be attributed to the presence of spurious connections within dFC, as previously investigated by Huang et al. [51]. The accumulation of spurios connections during multi-hop diffusion process will is a key factor in this regard. The outcome of the impact analysis with teleport probability aligns with the findings of previous research [62]. Prior studies have demonstrated that a low value of α increases the low-pass effect, directing the model's focus towards large-scale graphs and eliminating noisy high-frequency information [74]. Conversely, an excessively small value of α can cause the model to overly focus on large-scale graphs, resulting in a decline in performance. Therefore, a reasonable increase of teleport probability will enhance the impact of indirect connections in attention diffusion, thereby improving performance. With regard to the impact of the weight of RevCon loss, it functions as a regularization item in loss function. Furthermore, the impact of the weight of RevCon loss is consistent with that of a regularization item in a loss function. That is to say, a reasonable value will be beneficial to the performance, but an overly large weight will degenerate the performance.

5.2 Effect of Reverse Contrastive Learning

One of the objectives of computer-aided diagnosis is to develop models that can generalize to data from different sites or other sources. In this study, we utilized the RevCon loss method to generalize the model to cross-site fMRI data. To assess the efficacy of the RevCon loss method, we compared our results with those obtained using adversarial learning and solely CE loss. We employed the K-fold and Leave-one-site-on validation strategies in section 4.5 to analyze the performance of these methods. The results of the introduction of RevCon loss revealed an improvement in ACC of 2.8% and

1.9% on two validation strategies, respectively, indicating better generalizability with RevCon loss. Additionally, the effectiveness of RevCon loss is supported by the classification performance on Site 1 and Site 2. The promising adversarial learning may not be suitable for our framework, according to the results. One potential explanation for this discrepancy is that the multi-hop mechanism may result in the accumulate of perturbation introduced by adversarial samples. Another possible reason is that multi-modal features have been considered in previous work. Multi-modal features, including features derived from fMRI and MRI, may assist the model in more effectively capturing the complexity of the information and providing the model with a multi-perspective understanding. In contrast, the RevCon loss can enhance the representation of intra-class data and achieve superior classification performance in our MSTGCN framework. The results of the ablation study also provide evidence of the effectiveness of the RevCon loss. The performance of the model is enhanced by the inclusion of the RevCon loss. In addition, we also tested the impact of the weight of RevCon loss. The result have been discussed in section 5.1.

It is noted that the margin and easy-positive and hard-negative sampling strategy have been removed. According to the original work of RevCon loss, they tested that removal of the margin (setting it to infinity) will lead to better performance. They inferred that it is CE loss that keeps the distance between anchor samples and negative samples. Without the margin, hard-negative selection does not work. In our own work, we tested the performance of setting the margin. Our results indicated that performance was not as good as without margin. This result is consistent with the inference made by the original authors. We also compared the performance of easy-positive sample selection. This strategy resulted in a further deterioration in performance. Therefore, we calculated the mean value of the distance between all pairs of positive samples and removed the sample selection. In light of the CLIP work [94], we employed the cosine similarity to quantify the distance between samples, diverging from the original RevCon approach, which utilized the L1-distance. To ascertain the efficacy of the L1-distance in comparison to the cosine similarity, we conducted a test. Our findings indicated that the L1-distance did not demonstrate superior performance to the cosine similarity in our study.

5.3 Model Interpretability and Surgical Outcome Prediction

We investigated the interpretability of the MSTGCN framework to identify the potential biomarkers for epilepsy diagnosis and surgical prognosis. The gradient-based saliency map strategy was utilized at the node, and edge, and network levels, respectively, to recognize important brain areas and connections contributing to the TLE vs. HC classification. The regions and networks with maximal importance for classification included the subcortico-cortical structures such as thalamus and basal ganglia, temporal lobe structures, cingulate gyrus, frontal and parietal association cortices. Among the resting-state network markers, the DMN is of particular importance, as it is intricately connected to the mesial temporal regions and is often involved in the propagation and clinical expression of seizures. These experimental results are in line with previous epilepsy fMRI studies [95, 96] and confirm that the proposed method successfully identifies a spatially distributed network associated with TLE.

Exploring the optimal spatio-temporal representations of brain graphs via MSTGCN is not only beneficial for the diagnosis of epilepsy, but also has a potential application in predicting the surgical outcome of drug-resistant epilepsy. To further validate the interpretability and enhance the potential clinical value of the proposed model, we performed a surgical outcome prediction task. By combining the result of the saliency map and the clinical prior surgery knowledge, we showed that SF patients and NSF patients can be classified effectively using the random forest, with an AUROC of 0.829 and an ACC of 82.0%. This result further supports that the potential of biomarker mining of the trained model, especially in surgical planning and prediction of surgical outcomes.

5.4 Limitations and Future Research

Although the proposed MSTGCN framework has achieved outstanding performance in epilepsy diagnosis and surgical prognosis, there are still several limitations that need to be addressed in future work. First, we use dFC derived from fMRI as input for our model. However, short-time segments may introduce spurious fluctuations in the observed scans, which increases the spurious connections in dFCs [62]. Innovative developments with multi-modal deep learning methods have emerged. Expanding the input into multi-modal data may improve the performance of computer-aided diagnosis model by alleviating their shortcomings.

Second, our approach only learns FC features based on a single brain Atlas, i.e. BN Atlas. As for the number of voxels in fMRI signal is large, most of diagnosis works with FC will perform the template the on fMRI signal to integrate information with ROIs, which is parcellation. The parcellation can be treated as down-sampling method. Different template will have different way of designing the ROIs, some are based on the function and some are based on the structure. Some of them even use independent component analysis to parcelate the ROIs. Hence multi-scale brain ROI segmentation provides more information on functional brain activity. The classification performance is expected to be further improved by integrating FC features learned from multi-scale brain maps [63].

Finally, except for the diagnosis and prognosis, there are vital downstream tasks like literalization, localization and so on. For most of the artificital intelligence methods, they only consider one of the downstream tasks. Also, each task needs a specialized training process, which constrains the scalability of model application. Recently, large language models emerge and attracts attention of researchers. The prompt engineering, pre-training and fine-tuning are significant components and technologies in large language models. Such technologies are promising to design a method with different different downsteam tasks.

6. Conclusion

In this study, we introduce a novel Multi-hop Spatio-temporal Graph Convolutional Network model designed for epilepsy diagnosis and surgical outcome prognosis. Our approach involves several key innovations. First, we incorporate indirect connections between ROIs within the brain network using multi-hop spatial attention, thereby enhancing the model's spatial representation capability. Second, we introduce the RevCon loss, computed from extracted features, to regularize the CE loss, thus improving the model's generalizability across different fMRI datasets. Experimental evaluations on two real epilepsy datasets demonstrate that our method surpasses other GNN-based diagnosis methods in terms of accuracy and reliability. Furthermore, our proposed model can effectively identify discriminative brain regions and connections in Temporal Lobe Epilepsy patients. By leveraging saliency maps, our model showcases promising potential in discovering biomarkers and facilitating clinical applications in epilepsy prognosis.

References

[1] R. D. Thijs, R. Surges, T. J. O'Brien, and J. W. Sander, "Epilepsy in adults," *The Lancet,* vol. 393, no. 10172, pp. 689-701, 2019.

[2] Z. Zhang, G. Lu, Y. Zhong, Q. Tan, H. Chen, W. Liao, L. Tian, Z. Li, J. Shi, and Y. Liu, "fMRI study of mesial temporal lobe epilepsy using amplitude of low‐frequency fluctuation analysis," *Human brain mapping,* vol. 31, no. 12, pp. 1851-1861, 2010.

[3] K. Arfanakis, B. P. Hermann, B. P. Rogers, J. D. Carew, M. Seidenberg, and M. E. Meyerand, "Diffusion tensor MRI in temporal lobe epilepsy," *Magnetic resonance imaging,* vol. 20, no. 7, pp. 511-519, 2002.

[4] R. Li, C. Deng, X. Wang, T. Zou, B. Biswal, D. Guo, B. Xiao, X. Zhang, J. L. Cheng, and D. Liu, "Interictal dynamic network transitions in mesial temporal lobe epilepsy," *Epilepsia,* vol. 63, no. 9, pp. 2242-2255, 2022.

[5] J. D. Kennedy and S. U. Schuele, "Neocortical temporal lobe epilepsy," *Journal of Clinical Neurophysiology,* vol. 29, no. 5, pp. 366-370, 2012.

[6] M. P. Richardson, "Large scale brain models of epilepsy: dynamics meets connectomics," *Journal of Neurology, Neurosurgery & Psychiatry,* 2012.

[7] J. Courtiol, M. Guye, F. Bartolomei, S. Petkoski, and V. K. Jirsa, "Dynamical mechanisms of interictal resting-state functional connectivity in epilepsy," *Journal of Neuroscience,* vol. 40, no. 29, pp. 5572-5588, 2020.

[8] S. S. Spencer and D. D. Spencer, "Entorhinal‐hippocampal interactions in medial temporal lobe epilepsy," *Epilepsia,* vol. 35, no. 4, pp. 721-727, 1994.

[9] F. Bartolomei, P. Chauvel, and F. Wendling, "Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG," *Brain,* vol. 131, no. 7, pp. 1818-1830, 2008.

[10] S. S. Spencer, "Neural networks in human epilepsy: evidence of and implications for treatment," *Epilepsia,* vol. 43, no. 3, pp. 219-227, 2002.

[11] J. P. Lieb, R. M. Dasheiff, J. Engel, Genton, and Genton, "Role of the frontal lobes in the propagation of mesial temporal lobe seizures," *Epilepsia,* vol. 32, no. 6, pp. 822-837, 1991.

[12] E. Van Dellen, L. Douw, A. Hillebrand, P. C. de Witt Hamer, J. C. Baayen, J. J. Heimans, J. C. Reijneveld, and C. J. Stam, "Epilepsy surgery outcome and functional network alterations in longitudinal MEG: a minimum spanning tree analysis," *Neuroimage,* vol. 86, pp. 354-363, 2014.

[13] K. Wang, F. Xie, C. Liu, L. Tan, J. He, P. Hu, M. Zhang, G. Wang, F. Chen, and B. Xiao, "Abnormal functional connectivity profiles predict drug responsiveness in patients with temporal lobe epilepsy," *Epilepsia,* vol. 63, no. 2, pp. 463-473, 2022.

[14] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant, "A continuous semantic space describes the representation of thousands of object and action categories across the human brain," *Neuron,* vol. 76, no. 6, pp. 1210-1224, 2012.

[15] K. S. Lee, C. N. Hagan, M. Hughes, G. Cotter, E. M. Freud, K. Kircanski, E. Leibenluft, M. A. Brotman, and W.-L. Tseng, "Systematic review and meta-analysis: Task-based fMRI studies in youths with irritability," *Journal of the American Academy of Child & Adolescent Psychiatry,* vol. 62, no. 2, pp. 208-229, 2023.

[16] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo‐planar MRI," *Magnetic resonance in medicine,* vol. 34, no. 4, pp. 537-541, 1995.

[17] M. P. Van Den Heuvel and H. E. H. Pol, "Exploring the brain network: a review on resting-state fMRI functional connectivity," *European neuropsychopharmacology,* vol. 20, no. 8, pp. 519-534, 2010.

[18] T. Xu, K. R. Cullen, B. Mueller, M. W. Schreiner, K. O. Lim, S. C. Schulz, and K. K. Parhi, "Network analysis of functional brain connectivity in borderline personality disorder using resting-state fMRI," *NeuroImage: Clinical,* vol. 11, pp. 302-315, 2016.

[19] K. Wang, M. Liang, L. Wang, L. Tian, X. Zhang, K. Li, and T. Jiang, "Altered functional connectivity in early Alzheimer's disease: A resting‐state fMRI study," *Human brain mapping,* vol. 28, no. 10, pp. 967-978, 2007.

[20] M. G. Preti and D. Van De Ville, "Decoupling of brain function from structure reveals regional behavioral specialization in humans," *Nature communications,* vol. 10, no. 1, p. 4747, 2019.
[21] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for FMRI," *Neuroimage,* vol. 54, no. 2, pp. 875-891, 2011.
[22] P. Garces, E. Pereda, J. A. Hernández‐Tamames, F. Del‐Pozo, F. Maestu, and J. Ángel Pineda‐Pardo, "Multimodal description of whole brain connectivity: A comparison of resting state MEG, fMRI, and DWI," *Human brain mapping,* vol. 37, no. 1, pp. 20-34, 2016.
[23] J. Richiardi, H. Eryilmaz, S. Schwartz, P. Vuilleumier, and D. Van De Ville, "Decoding brain states from fMRI connectivity graphs," *Neuroimage,* vol. 56, no. 2, pp. 616-626, 2011.
[24] D. Tomasi and N. D. Volkow, "Association between functional connectivity hubs and brain networks," *Cerebral cortex,* vol. 21, no. 9, pp. 2003-2013, 2011.
[25] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle, "The human brain is intrinsically organized into dynamic, anticorrelated functional networks," *Proceedings of the National Academy of Sciences,* vol. 102, no. 27, pp. 9673-9678, 2005.
[26] J. Xiong, L. M. Parsons, J. H. Gao, and P. T. Fox, "Interregional connectivity to primary motor cortex revealed using MRI resting state images," *Human brain mapping,* vol. 8, no. 2, pp. 151-156, 1999.
[27] A. B. Protzner and M. P. McAndrews, "Network alterations supporting word retrieval in patients with medial temporal lobe epilepsy," *Journal of Cognitive Neuroscience,* vol. 23, no. 9, pp. 2605-2619, 2011.
[28] R. Li, X. Wu, K. Chen, A. Fleisher, E. Reiman, and L. Yao, "Alterations of directional connectivity among resting-state networks in Alzheimer disease," *American Journal of Neuroradiology,* vol. 34, no. 2, pp. 340-345, 2013.
[29] R. Li, H. Wang, L. Wang, L. Zhang, T. Zou, X. Wang, W. Liao, Z. Zhang, G. Lu, and H. Chen, "Shared and distinct global signal topography disturbances in subcortical and cortical networks in human epilepsy," *Human Brain Mapping,* vol. 42, no. 2, pp. 412-426, 2021.
[30] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain informatics,* vol. 4, no. 3, pp. 159-169, 2017.
[31] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.
[32] K. Zheng, B. Ma, and B. Chen, "DynBrainGNN: Towards Spatio-Temporal Interpretable Graph Neural Network Based on Dynamic Brain Connectome for Psychiatric Diagnosis," in *International Workshop on Machine Learning in Medical Imaging*, 2023: Springer, pp. 164-173.
[33] N. C. Dvornek, P. Ventola, K. A. Pelphrey, and J. S. Duncan, "Identifying autism from resting-state fMRI using long short-term memory networks," in *Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8*, 2017: Springer, pp. 362-370.
[34] Y. Pan, M. Liu, C. Lian, T. Zhou, Y. Xia, and D. Shen, "Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, 2018: Springer, pp. 455-463.
[35] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D. Y. Yeung, "GaAN: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs," in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018.
[36] K. K. Thekumparampil, C. Wang, S. Oh, and L.-J. Li, "Attention-based graph neural network for semi-supervised learning," *arXiv preprint arXiv:1803.03735,* 2018.
[37] Z. Liu, C. Chen, L. Li, J. Zhou, X. Li, L. Song, and Y. Qi, "Geniepath: Graph neural networks with adaptive receptive paths," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 4424-4431.
[38] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert, "Metric learning with spectral graph convolutions on brain connectivity networks," *NeuroImage,* vol. 169, pp. 431-442, 2018.

[39] Y. Chen, J. Yan, M. Jiang, T. Zhang, Z. Zhao, W. Zhao, J. Zheng, D. Yao, R. Zhang, and K. M. Kendrick, "Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification," *IEEE Transactions on Neural Networks and Learning Systems,* 2022.

[40] X. Xing, Q. Li, M. Yuan, H. Wei, Z. Xue, T. Wang, F. Shi, and D. Shen, "DS-GCNs: Connectome Classification using Dynamic Spectral Graph Convolution Networks with Assistant Task Training," *Cerebral Cortex,* vol. 31, no. 2, pp. 1259-1269, 2020.

[41] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, "Spatio-temporal graph convolution for resting-state fmri analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, 2020: Springer, pp. 528-538.

[42] B.-H. Kim, J. C. Ye, and J.-J. Kim, "Learning dynamic graph representation of brain connectome with spatio-temporal attention," *Advances in Neural Information Processing Systems,* vol. 34, pp. 4314-4327, 2021.

[43] X. Fan, M. Gong, and Y. Wu, "Markov clustering regularized multi-hop graph neural network," *Pattern Recognition,* vol. 139, p. 109518, 2023.

[44] G. Cucurull, K. Wagstyl, A. Casanova, P. Veličković, E. Jakobsen, M. Drozdzal, A. Romero, A. Evans, and Y. Bengio, "Convolutional neural networks for mesh-based parcellation of the cerebral cortex," in *Medical imaging with deep learning*, 2022.

[45] P. Ryvlin, J. H. Cross, and S. Rheims, "Epilepsy surgery in children and adults," *The Lancet Neurology,* vol. 13, no. 11, pp. 1114-1126, 2014.

[46] C. Barba, S. Giometto, E. Lucenteforte, S. Pellacani, G. Matta, A. Bettiol, S. Minghetti, L. Falorni, F. Melani, and G. Di Giacomo, "Seizure outcome of temporal lobe epilepsy surgery in adults and children: a systematic review and meta-analysis," *Neurosurgery,* vol. 91, no. 5, pp. 676-683, 2022.

[47] D. J. Englot, L. B. Hinkley, N. S. Kort, B. S. Imber, D. Mizuiri, S. M. Honma, A. M. Findlay, C. Garrett, P. L. Cheung, and M. Mantle, "Global and regional functional connectivity maps of neural oscillations in focal epilepsy," *Brain,* vol. 138, no. 8, pp. 2249-2262, 2015.

[48] H. Burianová, N. L. Faizo, M. Gray, J. Hocking, G. Galloway, and D. Reutens, "Altered functional connectivity in mesial temporal lobe epilepsy," *Epilepsy Research,* vol. 137, pp. 45-52, 2017.

[49] R. D. Bharath, R. Panda, J. Raj, S. Bhardwaj, S. Sinha, G. Chaitanya, K. Raghavendra, R. C. Mundlamuri, A. Arimappamagan, and M. B. Rao, "Machine learning identifies “rsfMRI epilepsy networks” in temporal lobe epilepsy," *European radiology,* vol. 29, pp. 3496-3505, 2019.

[50] M. Mazrooyisebdani, V. A. Nair, C. Garcia-Ramos, R. Mohanty, E. Meyerand, B. Hermann, V. Prabhakaran, and R. Ahmed, "Graph theory analysis of functional connectivity combined with machine learning approaches demonstrates widespread network differences and predicts clinical variables in temporal lobe epilepsy," *Brain connectivity,* vol. 10, no. 1, pp. 39-50, 2020.

[51] J. Huang, M. Wang, H. Ju, Z. Shi, W. Ding, and D. Zhang, "SD-CNN: A static-dynamic convolutional neural network for functional brain networks," *Medical Image Analysis,* vol. 83, p. 102679, 2023.

[52] S. Spencer and L. Huh, "Outcomes of epilepsy surgery in adults and children," *The Lancet Neurology,* vol. 7, no. 6, pp. 525-537, 2008.

[53] J. Yuan, X. Ran, K. Liu, C. Yao, Y. Yao, H. Wu, and Q. Liu, "Machine learning applications on neuroimaging for diagnosis and prognosis of epilepsy: A review," *Journal of neuroscience methods,* vol. 368, p. 109441, 2022.

[54] Z. Yang, J. Choupan, D. Reutens, and J. Hocking, "Lateralization of temporal lobe epilepsy based on resting-state functional magnetic resonance imaging and machine learning," *Frontiers in neurology,* vol. 6, p. 110661, 2015.

[55] N. Nandakumar, D. Hsu, R. Ahmed, and A. Venkataraman, "DeepEZ: a graph convolutional network for automated epileptogenic zone localization from resting-state fMRI connectivity," *IEEE Transactions on Biomedical Engineering,* vol. 70, no. 1, pp. 216-227, 2022.

[56] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open,* vol. 1, pp. 57-81, 2020.

[57] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, "Spatio-Temporal Graph Convolution for Resting-State fMRI Analysis," in *Medical image computing and computer-assisted intervention: MICCAI... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, vol. 12267, pp. 528-538.

[58] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, "Mapping the structural core of human cerebral cortex," *PLoS biology,* vol. 6, no. 7, p. e159, 2008.
[59] S. Ponten, F. Bartolomei, and C. Stam, "Small-world networks and epilepsy: graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures," *Clinical neurophysiology,* vol. 118, no. 4, pp. 918-927, 2007.
[60] K. Oono and T. Suzuki, "Graph Neural Networks Exponentially Lose Expressive Power for Node Classification," in *International Conference on Learning Representations*, 2019.
[61] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32, no. 1.
[62] G. Wang, R. Ying, J. Huang, and J. Leskovec, "Multi-hop Attention Graph Neural Networks," International Joint Conferences on Artificial Intelligence Organizatio n, 2021/8//, pp. 3089-3096.
[63] Y. Chen, J. Yan, M. Jiang, T. Zhang, Z. Zhao, W. Zhao, J. Zheng, D. Yao, R. Zhang, K. M. Kendrick, and X. Jiang, "Adversarial Learning Based Node-Edge Graph Attention Networks for Autism Spectrum Disorder Identification," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 1, no. 1, pp. 1-12, 2022 2022.
[64] W. Cui, J. Du, M. Sun, S. Zhu, S. Zhao, Z. Peng, L. Tan, and Y. Li, "Dynamic multi-site graph convolutional network for autism spectrum disorder identification," *Computers in Biology and Medicine,* vol. 157, p. 106749, 2023.
[65] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 4, pp. 4396-4415, 2022.
[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.
[67] T. Duboudin, E. Dellandréa, C. Abgrall, G. Hénaff, and L. Chen, "Encouraging intra-class diversity through a reverse contrastive loss for single-source domain generalization," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 51-60.
[68] L. Fan, H. Li, J. Zhuo, Y. Zhang, J. Wang, L. Chen, Z. Yang, C. Chu, S. Xie, A. R. Laird, P. T. Fox, S. B. Eickhoff, C. Yu, and T. Jiang, "The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture," *Cerebral Cortex,* vol. 26, no. 8, pp. 3508-3526, 2016.
[69] X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, "Graph neural network for interpreting task-fmri biomarkers," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, 2019: Springer, pp. 485-493.
[70] B.-H. Kim and J. C. Ye, "Understanding graph isomorphism network for rs-fMRI functional connectivity analysis," *Frontiers in neuroscience,* vol. 14, p. 630, 2020.
[71] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555,* 2014.
[72] A. Kazi, S. Shekarforoush, S. Arvind Krishna, H. Burwinkel, G. Vivar, K. Kortüm, S.-A. Ahmadi, S. Albarqouni, and N. Navab, "InceptionGCN: receptive field aware graph convolutional network for disease prediction," in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, 2019: Springer, pp. 73-85.
[73] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903,* 2017.
[74] G. Wang, R. Ying, J. Huang, and J. Leskovec, "Multi-hop attention graph neural network," *arXiv preprint arXiv:2009.14332,* 2020.
[75] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907,* 2016.
[76] J. Gasteiger, A. Bojchevski, and S. Günnemann, "Predict then propagate: Graph neural networks meet personalized pagerank," *arXiv preprint arXiv:1810.05997,* 2018.
[77] S. Singla, B. Nushi, S. Shah, E. Kamar, and E. Horvitz, "Understanding failures of deep networks via robust feature extraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12853-12862.
[78] K. Hermann and A. Lampinen, "What shapes feature representations? exploring datasets, architectures, and training," *Advances in Neural Information Processing Systems,* vol. 33, pp. 9995-10006, 2020.

[79] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," *Domain adaptation in computer vision applications,* vol. 1, no. 1, pp. 37-55, 2017.
[80] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, and I. Kompatsiaris, "A survey on bias in visual datasets," *Computer Vision and Image Understanding,* vol. 223, p. 103552, 2022.
[81] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence,* vol. 3, no. 7, pp. 610-619, 2021.
[82] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014: ICLR.
[83] C. Yan and Y. Zang, "DPARSF: a MATLAB toolbox for" pipeline" data analysis of resting-state fMRI," *Frontiers in systems neuroscience,* vol. 4, p. 1377, 2010.
[84] K. J. Friston, S. Williams, R. Howard, R. S. Frackowiak, and R. Turner, "Movement ‐ related effects in fMRI time ‐ series," *Magnetic resonance in medicine,* vol. 35, no. 3, pp. 346-355, 1996.
[85] S. Arslan, S. I. Ktena, B. Glocker, and D. Rueckert, "Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity," in *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities: Second International Workshop, GRAIL 2018 and First International Workshop, Beyond MIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 2*, 2018: Springer, pp. 3-13.
[86] X. Li, Y. Zhou, N. C. Dvornek, M. Zhang, J. Zhuang, P. Ventola, and J. S. Duncan, "Pooling regularized graph neural network for fmri biomarker analysis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, 2020: Springer, pp. 625-635.
[87] S. Dahan, L. Z. Williams, D. Rueckert, and E. C. Robinson, "Improving phenotype prediction using long-range spatio-temporal dynamics of functional connectivity," in *Machine Learning in Clinical Neuroimaging: 4th International Workshop, MLCN 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4*, 2021: Springer, pp. 145-154.
[88] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICML*, 2015.
[89] G. Assenza, J. Lanzone, A. Insola, G. Amatori, L. Ricci, M. Tombini, and V. Di Lazzaro, "Thalamo-cortical network dysfunction in temporal lobe epilepsy," *Clinical Neurophysiology,* vol. 131, no. 2, pp. 548-554, 2020.
[90] J. Royer, B. C. Bernhardt, S. Larivière, E. Gleichgerrcht, B. J. Vorderwülbecke, S. Vulliémoz, and L. Bonilha, "Epilepsy and brain network hubs," *Epilepsia,* vol. 63, no. 3, pp. 537-550, 2022.
[91] W. Liao, Z. Zhang, Z. Pan, D. Mantini, J. Ding, X. Duan, C. Luo, Z. Wang, Q. Tan, and G. Lu, "Default mode network abnormalities in mesial temporal lobe epilepsy: a study combining fMRI and DTI," *Human brain mapping,* vol. 32, no. 6, pp. 883-895, 2011.
[92] B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, and J. R. Polimeni, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of neurophysiology,* vol. 106, pp. 1125-1165, 2011.
[93] J. Du, S. Wang, H. Miao, and J. Zhang, "Multi-Channel Pooling Graph Neural Networks."
[94] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Clip: Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020,* 2021.
[95] Z. Haneef, A. Lenartowicz, H. J. Yeh, J. Engel Jr, and J. M. Stern, "Network analysis of the default mode network using functional connectivity MRI in temporal lobe epilepsy," *JoVE (Journal of Visualized Experiments),* no. 90, p. e51442, 2014.
[96] K. Lee, H. M. Khoo, J.-M. Lina, F. Dubeau, J. Gotman, and C. Grova, "Disruption, emergence and lateralization of brain network hubs in mesial temporal lobe epilepsy," *NeuroImage: Clinical,* vol. 20, pp. 71-84, 2018.