RESEARCH-ARTICLE

Contrasformer: A Brain Network Contrastive Transformer for Neurodegenerative Condition Identification

JIAXING XU, Nanyang Technological University, Singapore City, Singapore

KAI HE, National University of Singapore, Singapore City, Singapore

MENGCHENG LAN, Nanyang Technological University, Singapore City, Singapore

QINGTIAN BIAN, Nanyang Technological University, Singapore City, Singapore

WEI LI, Nanyang Technological University, Singapore City, Singapore

TIEYING LI, Northeastern University, Shenyang, Liaoning, China

View all

Published: 21 October 2024

CIKM '24: The 33rd ACM International Conference on Information and Knowledge Management
October 21 - 25, 2024
ID, Boise, USA

Conference Sponsors:
SIGIR

Contrasformer: A Brain Network Contrastive Transformer for Neurodegenerative Condition Identification

Jiaxing Xu*
Nanyang Technological University
Singapore
jiaxing003@e.ntu.edu.sg

Kai He
National University of Singapore
Singapore
kai_he@nus.edu.sg

Mengcheng Lan
S-Lab
Nanyang Technological University
Singapore
lanm0002@e.ntu.edu.sg

Qingtian Bian
Nanyang Technological University
Singapore
bian0027@ntu.edu.sg

Wei Li
Nanyang Technological University
Singapore
wei014@e.ntu.edu.sg

Tieying Li
Northeastern University
Shenyang, China
tieying@stumail.neu.edu.cn

Yiping Ke
Nanyang Technological University
Singapore
ypke@ntu.edu.sg

Miao Qiao
The University of Auckland
Auckland, New Zealand
miao.qiao@auckland.ac.nz

ABSTRACT

Understanding neurological disorder is a fundamental problem in neuroscience, which often requires the analysis of brain networks derived from functional magnetic resonance imaging (fMRI) data. Despite the prevalence of Graph Neural Networks (GNNs) and Graph Transformers in various domains, applying them to brain networks faces challenges. Specifically, the datasets are severely impacted by the noises caused by distribution shifts across sub-populations and the neglect of node identities, both obstruct the identification of disease-specific patterns. To tackle these challenges, we propose *Contrasformer*, a novel contrastive brain network Transformer. It generates a prior-knowledge-enhanced contrast graph to address the distribution shifts across sub-populations by a two-stream attention mechanism. A cross attention with identity embedding highlights the identity of nodes, and three auxiliary losses ensure group consistency. Evaluated on 4 functional brain network datasets over 4 different diseases, Contrasformer outperforms the state-of-the-art methods for brain networks by achieving up to 10.8% improvement in accuracy, which demonstrates its efficacy in neurological disorder identification. Case studies illustrate its interpretability, especially in the context of neuroscience. This paper provides a solution for analyzing brain networks, offering valuable insights into neurological disorders. Our code is available at https://github.com/AngusMonroe/Contrasformer.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**;
• **Applied computing** → **Health informatics**.

*Corresponding Author

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679560

KEYWORDS

Brain Network, Graph Neural Network, Graph Transformer, Neurological Disorder

ACM Reference Format:
Jiaxing Xu, Kai He, Mengcheng Lan, Qingtian Bian, Wei Li, Tieying Li, Yiping Ke, and Miao Qiao. 2024. Contrasformer: A Brain Network Contrastive Transformer for Neurodegenerative Condition Identification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679560

1 INTRODUCTION

In the field of neuroscience, a central objective is to uncover distinctive patterns associated with neurological disorders (e.g., Alzheimer's, Parkinson's, and Autism) by exploring the brain networks of both healthy individuals and patients with neurological disorders [37]. Among the various techniques for neuroimaging, resting-state functional magnetic resonance imaging (fMRI) is widely used to profile the connectivities among brain regions [46], leading to brain networks where each node is a specific brain region, called a region of interest (ROI), and each edge denotes a pairwise correlation between the blood-oxygen-level-dependent (BOLD) signals of two ROIs [50]. The edge reveals the connectivity between brain regions, showcasing which areas tend to be activated together or exhibit correlated activities. Brain networks model neurological systems as graphs and one could apply graph-based techniques to gain insights into their roles and interactions [23, 29, 45]. Such brain networks could help to decipher distinctive patterns associated with neurological disorders, which can contribute to early diagnosis and effective intervention strategies. It's worth noting that the development of graph analytics for brain networks is still in its early stages.

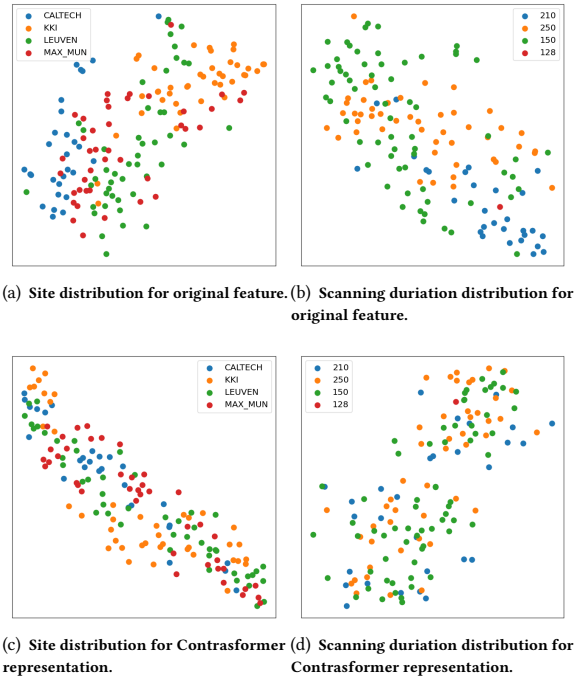
While graph neural networks (GNNs)[18, 26, 43] and graph Transformers [38, 42, 54] have recently been adopted in a wide range of graph-related tasks, applying them to brain networks faces two challenges below in capturing the disease-specific pattern.

Sub-population-specific (SPS) Noise. Analyzing neurological disorders aims to capture disease-specific patterns that are invariant across all populations. However, certain features are common to a sub-population (not generalizable to the entire population) and are unrelated to the disease, constituting sub-population-specific (SPS) noise. For instance, brain network datasets like Autism Brain Imaging Data Exchange (ABIDE) [9] and Alzheimer's Disease Neuroimaging Initiative (ADNI) [10] are collected from multiple sites. Subjects from different sites may exhibit site differences (scanner variability, different inclusion/exclusion criteria)[6]. Such noise could lead the model to focus on the site-specific pattern instead of learning population-invariant information. Besides, the varying scanning duration could result in different periods of region activation recorded for subjects. Furthermore, label inconsistencies may arise due to differences in diagnostic criteria used by doctors labeling these subjects. Fig. 1(a) provides an example of the site distribution of subjects in the ABIDE dataset. Each point in the figure represents a subject, and different colors denote the sites from which these subjects are acquired. From this example, it can be observed that subjects from the same site tend to have similar features, as the site-specific noise dominates the similarity. A similar observation is found in the distribution of sub-populations with different scanning durations. As illustrated in Fig. 1(b), each point represents a subject, and different colors denote the lengths of the BOLD signal for these subjects. Such a conspicuous distribution shift across sub-populations could easily mislead the model to overfit the SPS noise, thereby limiting its performance.

Node-identity awareness. The construction of brain networks requires a specific parcellation method to split the whole brain into several ROIs. The same parcellation method is applied to all subjects, and thus the ROI definition is identical across all brain networks. Such a property does not generally exist in other graph-structured data, necessitating our specialized tailoring for brain networks. Existing general-purposed GNNs are designed to learn the structural pattern of graphs without considering their node identities [43, 51, 52].

While some recent GNNs have introduced specialized designs for brain networks and addressed issues related to node identity [34, 35, 56], many of them overlook addressing SPS noise and capturing group-specific patterns. Neglecting these aspects can lead to models that overfit outliers and hinder interpretability [19].

In this paper, we propose *Contrasformer*, a novel contrastive brain network Transformer, that harnesses the distinctive properties of brain networks to fully leverage the capabilities of Transformer-based models for brain network analysis. Specifically, we employ a two-stream attention mechanism to generate a contrast graph that encodes group-specific information. Such two-stream attention mechanism can address the SPS noise by reweighting the original feature across ROIs and subjects. The distribution shift will also be corrected as shown in Fig. 1(c) and (d). Meanwhile, an identity embedding is introduced to incorporate the ROI identities with the input brain network, which highlights the node identity awareness. Moreover, we propose a cross decoder to integrate the contrast graph with the encoded brain network using a cross-attention mechanism. We further take advantage of node identity awareness by introducing a contrastive loss to constrain that identical ROIs across subjects have similar representations. Additionally, a cluster loss is introduced to guarantee group consistency in graph representations and take the graph-level relationship into account. Our contributions can be summarized as follows:



(a) **Site distribution for original feature.** (b) **Scanning duriation distribution for original feature.**

(c) **Site distribution for Contrasformer representation.** (d) **Scanning duriation distribution for Contrasformer representation.**

Figure 1: The distribution of the original feature and Contrasformer representation for subjects from multiple sites and scanning duriation in ABIDE dataset. Each point in the figure represents a subject and different colors denote the sites these subjects are acquired from or their lengths of the BOLD signals. The representation of each subject is obtained by mean pooling and visualized by t-SNE [41]. Compared with (c) and (d), (a) and (b) exhibit obvious distribution shifts.

- We introduce a contrast graph encoder to adaptively assign weights across ROIs and subjects by a two-stream attention mechanism. The generated contrast graph includes the most discriminative ROIs concerning different groups to address sub-population distribution shift and remit overfitting.
- By incorporating the contrast graph with brain network representation learning, we propose a contrastive brain network Transformer named *Contrasformer*. It utilizes an identity encoding and three auxiliary losses to make the model aware of node identity and graph-level relationship.
- We evaluate Contrasformer on 4 fMRI brain network datasets for different neurological disorders. Our results demonstrate the superiority of Contrasformer over 13 state-of-the-art methods on neurological disorder identification.
- We present a case study that underscores the intriguing, straightforward, and highly interpretable patterns extracted by our approach, aligning with domain knowledge found in neuroscience literature.

2 RELATED WORK

2.1 Graph Neural Networks (GNNs)

GNNs have gained prominence as a popular approach in processing and analyzing graph structure data, including social networks [26], molecular data [17], knowledge graphs [33, 47] and user-item graphs [4]. Most GNNs operate on a message-passing scheme [18, 26, 43], wherein they iteratively aggregate information from neighboring nodes and update the representation of each node [52]. In recent years, there has been a surge of interest in employing GNNs for the analysis of brain networks. Ktena et al. [28] utilized graph convolutional networks to learn similarities between pairs of brain networks (subjects). BrainNetCNN [23] introduced edge-to-edge, edge-to-node, and node-to-graph convolutional filters to harness the topological information within brain networks. LiNet [32] presented a two-stage pipeline that uses GNNs to identify brain biomarkers associated with Autism Spectrum Disorder (ASD) from fMRI data. BrainGNN [34] proposed an ROI-selection pooling method to emphasize salient brain regions for each individual. However, these approaches often overlook the unique characteristics of functional brain networks, such as node-identity awareness [29].

To better utilize node-identity awareness of brain networks, PRGNN [35] introduced a graph pooling method with group-level regularization, in the mean time ensuring group-level consistency. GroupINN [53] jointly learnt node grouping and graph feature extraction, though without considering node alignment. Lanciano et al. [29] focused on extracting a dense contrast subgraph to filter relevant information for predictions. Nevertheless, their feature extraction and subject classification treated all brain regions and individuals uniformly, potentially making them susceptible to noisy data. Zhang et al. [56] incorporated both local ROI-GNN and global subject-GNN guided by non-imaging data, but the local ROI-GNN did not account for the node alignment in brain networks. ContrastPool [48], introduced a dual attention mechanism to extract discriminative features across ROIs for subjects within the same group. However, it still neglects the graph-level relationship between different groups, which could suffer from the SPS noise.

In addition, a general issue for GNNs based on the message-passing scheme is called over-smoothing [24], which refers to the loss of discriminative information as the network iteratively aggregates information from a large number of neighbors. In our context of brain networks, correlation-based graphs naturally possess high density that could easily suffer from over-smoothing. Therefore we discard message-passing architecture in our work.

2.2 Graph Transformer

An alternative method for graph representation learning involves Transformer-based models [42], which adapt the attention mechanism to consider global information for each node and incorporate positional encoding to capture graph topological information. Graph Transformers have garnered significant attention due to their impressive performance in graph representation learning. Dwivedi et al. [12] introduced edge information into the attention mechanism and used eigenvectors as positional embeddings. SAN [27] implemented an invariant aggregation of Laplacian's eigenvectors for positional embedding and introduced conditional attention for real and virtual edges within a graph. Graphormer [54] enhanced the attention mechanism with centrality-based positional embedding and introduced pair-wise graph distances to define relative positional encodings. More recently, GPS [38] proposed a hybrid architecture that combines GNN and Transformer components, achieving state-of-the-art results on various datasets by introducing different types of global/local/relative positional/structural embeddings.

Nonetheless, applying these Transformer-based models to brain networks presents challenges [20], primarily due to the correlation-based edges that hinder the use of designs like centrality [14], spatial [54], and edge encoding [38]. Consequently, a series of brain network Transformer methods have emerged. One such method [20] applied Transformers to learn pairwise connection strengths among brain regions across individuals. THC [11] introduced an interpretable Transformer-based model for joint hierarchical cluster identification and brain network classification. DART [21] utilized segmenting BOLD signals to generate dynamic brain networks and then incorporated them with static networks for representation learning. Nevertheless, these approaches often neglect the incorporation of group-level information from subjects into their methodology design. Thus, it would be hard to address the issue of SPS noise. The lack of leveraging node-identity awareness can also limit the power of Transformer.

3 PRELIMINARIES

3.1 Brain Network Construction

In this work, we use the datasets preprocessed and released by Xu et al [50]. We apply Schaefer atlas [40] to parcellate all subjects, which divides the brain into 100 functional ROIs. For each subject, a brain network is represented by a connectivity matrix $\boldsymbol{M} \in \mathbb{R}^{m \times m}$, where m denotes the number of ROIs. The nodes in $\boldsymbol{M}$ are ROIs and the edges are the Pearson's correlation between the region-averaged BOLD signals from pairs of ROIs. Essentially, $\boldsymbol{M}$ captures functional relationships between ROIs. For the detailed preprocessing and brain network construction pipeline, please refer to Xu et al [50].

3.2 Problem Definition

Neurological disorder identification for brain networks aims to predict the distinct class of each subject. Given a dataset of labeled brain networks $\mathcal{D} = (\mathcal{M}, \mathcal{Y}) = \{(\boldsymbol{M}, y_M)\}$, where y_M is the class label of a brain network $\boldsymbol{M} \in \mathcal{M}$, the problem of neurological disorder identification is to learn a predictive function $f\colon \mathcal{M} \to \mathcal{Y}$, which maps input brain networks to the groups they belong to, expecting that f also works well on unseen brain networks. We use ABIDE dataset, which contains groups of typical controls (TC) and individuals diagnosed with ASD, as an example to explain our methodology in the following. Notation-wise, we use calligraphic letters to denote sets (e.g., $\mathcal{M}$), bold capital letters to denote matrices (e.g., $\boldsymbol{M}$), and strings with bold lowercase letters to represent vectors (e.g., $\boldsymbol{hg}$). Subscripts and superscripts are used to distinguish between different variables or parameters, and lowercase letters denote scalars. We use $\boldsymbol{M}(i,:)$ and $\boldsymbol{M}(:,j)$ to denote the i-th row and j-th column of a matrix $\boldsymbol{M}$, respectively. This notation also extends to a 3d matrix.

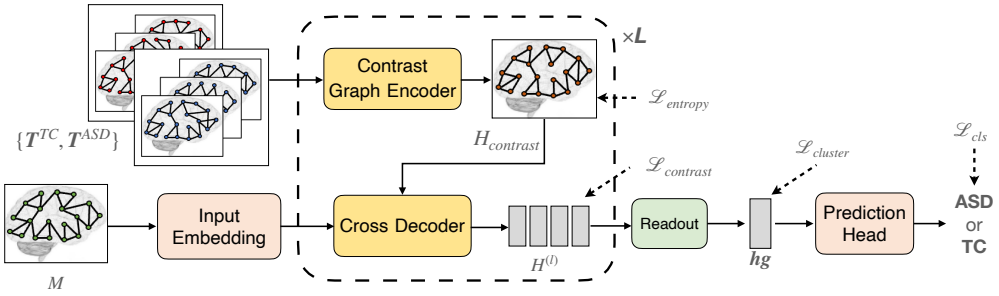


Figure 2: The framework of Contrasformer for neurological disorder identification, using Autism as an example.

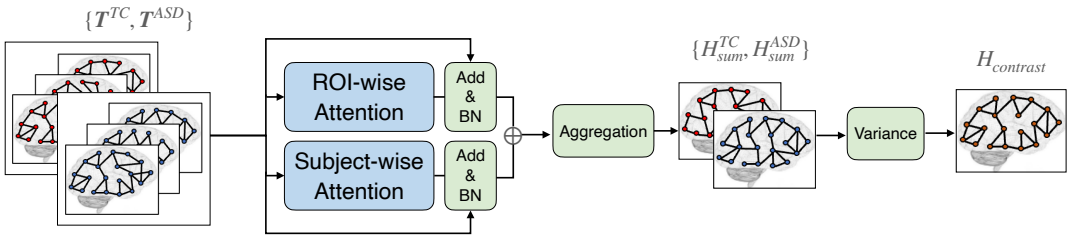


Figure 3: The architecture of contrast graph encoder. Each group of brain networks is fed into the two-stream attention to obtain a summary graph. The contrast graph is generated by contrasting the summary graphs of different groups.

4 METHODOLOGY

In this section, we provide a detailed exposition of the design of our proposed Contrasformer, depicted in Fig. 2. Contrasformer adopts an encoder-decoder architecture, featuring three key components: (1) A contrast graph encoder is introduced (in Section 4.1) to extract the most discriminative task-related features from the training set. To alleviate the SPS noise, a two-stream attention mechanism is employed to generate a contrast graph that captures the invariant information across all sub-population. The learnt contrast graph is then utilized in both training and test stages to be incorporated with the brain network representation learning. (2) A cross decoder is introduced (in Section 4.2) to combine the input brain network with node identity, and subsequently fuse the contrast graph with the identity-embedded brain network by a cross-attention to update node representations. (3) A classification loss $\mathcal{L}_{cls}$ and three auxiliary losses $\mathcal{L}_{entropy}$, $\mathcal{L}_{cluster}$, $\mathcal{L}_{contrast}$ are incorporated (in Section 4.3) to guide the end-to-end training. These losses emphasize the node identity of ROIs and consider group-level relationships.

4.1 Contrast Graph Encoder with Two-stream Attention

To generate a contrast graph with group-specific information, we introduce a two-stream contrast graph encoder. In neuroscience, normally there are only some ROIs in the brain that can reflect the lesion of a neurological disorder. So the aim of this encoder is to extract the most discriminative ROIs for each group while adaptly learning the contribution of each subject. Such a two-stream attention design is able to capture the population-invariant information embedded in subjects and ROIs, which alleviates the impact of SPS noise in the downstream task.

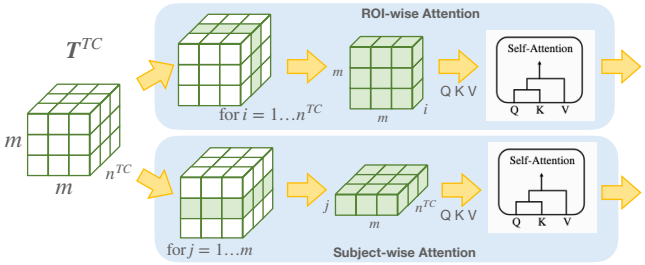


Figure 4: The detail of two-stream attention. The ROI- and subject-wise attention blocks compute self-attention from different views of the input. Parameters of self-attention inside these two branches are independent.

Taking Autism as an example, we use tensors $\boldsymbol{T}^{TC} \in \mathbb{R}^{n^{TC} \times m \times m}$ and $\boldsymbol{T}^{ASD} \in \mathbb{R}^{n^{ASD} \times m \times m}$ to denote all brain networks of TC and ASD groups in the training set, respectively. n^{TC} and n^{ASD} denote the number of subjects in TC/ASD groups and m denotes the number of ROIs. As illustrated in Fig. 3, given the input set of brain networks that belong to different groups, the objective of the two-stream attention is to generate a summary graph for each group. An ROI-wise attention and a subject-wise attention are first computed independently for each group of brain networks. We adopt the self-attention mechanism [42] for the ROI- and subject-wise attention. In general, given a matrix $\boldsymbol{X} \in \mathbb{R}^{k \times d}$, where k and d are arbitrary integers, the self-attention function can be written as:

$$\text{Attn}(\boldsymbol{X}) = \text{norm}\left(\boldsymbol{X} + \phi\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathsf{T}}}{\sqrt{k}}\right)\boldsymbol{V}\right), \tag{1}$$

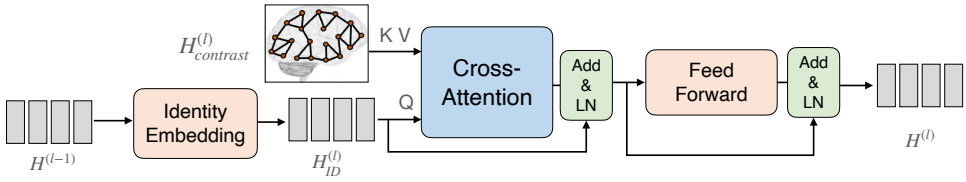


Figure 5: The architecture of the cross decoder. The generated contrast graph is incorporated with the identity-embedded brain network by a cross-attention for the downstream representation learning.

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V, \tag{2}$$

$$\phi(\boldsymbol{z})_i = \text{softmax}(\boldsymbol{z})_i = \frac{e^{\boldsymbol{z}_i}}{\sum_{j=1}^{d} e^{\boldsymbol{z}_j}}, \text{for } i = 1 \dots k, \boldsymbol{z} \in \mathbb{R}^k, \tag{3}$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{d \times d}$ are parameter matrices, e is the Euler's number, and $\text{norm}(\cdot)$ is a normalization function.

For the brain networks of TC group $\boldsymbol{T}^{TC}$, the detail of two-stream attention is shown in Fig. 4. In ROI-wise attention $\text{Attn}_{ROI}(\cdot)$, for each subject $i \in \{1 \dots n^{TC}\}$, we compute the self-attention across all ROIs in this subject. The most informative ROIs inside each subject will be extracted in this way. Similarly, in subject-wise attention $\text{Attn}_{subject}(\cdot)$, for each ROI $j = 1 \dots m$, we compute the self-attention across all subjects in this group. The input of subject-wise attention is implemented by simply transposing the subject and ROI dimensions. This operation helps our model highlight the most discriminative subjects of a certain ROI. The rationale of such attention mechanism is that we intend to extract the group-invariant feature and filter out the SPS noise.

Note that the normalization function we are using here (in Fig. 3) is batch normalization (BN) instead of the commonly used layer normalization (LN) in Transformer [42]. It is because (1) the input of the two-stream attention is constant for each training step; (2) the lengths of input sequences are consistent (m for ROI-wise attention, n^{TC} for subject-wise attention); (3) we aim to highlight the consistency within each group. We conduct an ablation study in Section 5.6 to verify the effectiveness of our model architecture.

Afterward, the summary graph of the TC group $\boldsymbol{H}^{TC}_{sum} \in \mathbb{R}^{m \times m}$ is generated as:

$$\boldsymbol{H}^{TC}_{sum} = \frac{1}{n^{TC}} \sum_{i=1}^{n^{TC}} (\boldsymbol{H}^{TC}_{ROI} + \boldsymbol{H}^{TC}_{subject})(i, :, :), \tag{4}$$

$$\boldsymbol{H}^{TC}_{ROI} = \left[\text{Attn}_{ROI} \left(\boldsymbol{T}^{TC}(i, :, :) \right) : i = 1, \dots, n^{TC} \right], \tag{5}$$

$$\boldsymbol{H}^{TC}_{subject} = \left[\text{Attn}_{subject} \left(\boldsymbol{T}^{TC}(:, j, :) \right) : j = 1, \dots, m \right], \tag{6}$$

where $\boldsymbol{H}^{TC}_{ROI}, \boldsymbol{H}^{TC}_{subject} \in \mathbb{R}^{n^{TC} \times m \times m}$ denote two 3-dimensional matrices to store the outputs of the ROI-wise and subject-wise attention functions, and $[\cdot]$ denotes the stack function to combine a set of matrices to a single higher-dimensional matrix. The summary graphs of other groups can also be obtained in a similar way. Once we obtain the summary graphs $\mathcal{H}_{sum} = \{\boldsymbol{H}^{TC}_{sum}, \boldsymbol{H}^{ASD}_{sum}\}$ for TC and ASD groups, the contrast graph is generated by computing the variance of all summary graphs:

$$\boldsymbol{H}_{contrast} = \frac{1}{|\mathcal{H}_{sum}|} \sum_{\boldsymbol{H}_{sum} \in \mathcal{H}_{sum}} (\boldsymbol{H}_{sum} - \bar{\boldsymbol{H}}_{sum})^2, \tag{7}$$

$$\bar{\boldsymbol{H}}_{sum} = \frac{1}{|\mathcal{H}_{sum}|} \sum_{\boldsymbol{H}_{sum} \in \mathcal{H}_{sum}} \boldsymbol{H}_{sum}. \tag{8}$$

The generated contrast graph $\boldsymbol{H}_{contrast}$ contains the discriminative information about the specific disease, which can be incorporated with the cross decoder to boost the downstream brain network representation learning. The contrast graph generation also works for datasets with multiple groups by making contrast among all groups. Note that we only use subjects in the training set to train the contrast graph encoder. The generated contrast graph is used in the test stage as prior knowledge to avoid data leakage.

4.2 Cross Decoder with Identity Embedding

By leveraging the group-discriminative information captured in the contrast graph, the cross decoder of Contrasformer aims to produce a high-quality representation for each input brain network (Fig. 5).

In graph transformer models, positional embedding is commonly used to encode the topological information of the graph. However, existing designs for general graph representation learning, such as distance-based, centrality-based and eigenvector-based positional embedding [30, 44, 54], can hardly migrate to brain network due to its high density (always fully connected). The correlation-based brain networks naturally contain sufficient positional information for ROIs. Therefore the general-purposed positional embedding is not only expensive but also redundant in our case.

Instead of positional embedding that captures the topological information of the graph structure, we propose a learnable identity embedding to adaptively learn the unique identity for each ROI. Such embedding attaches the same identity for nodes that belong to the same ROI. As shown in Eq. (9), we introduce a parameter matrix $\boldsymbol{W}^{(l)}_{ID} \in \mathbb{R}^{m \times m}$ to encode the identity of nodes. $\delta(\cdot)$ denotes a multilayer perceptron (MLP).

$$\boldsymbol{H}^{(l)}_{ID} = \boldsymbol{H}^{(l-1)} + \delta(\boldsymbol{H}^{(l-1)} + \boldsymbol{W}^{(l-1)}_{ID}). \tag{9}$$

After identity embedding, we combine the contrast graph with the encoded brain network by a cross-attention function followed by a layer normalization. The encoded brain network $\boldsymbol{H}^{(l)}_{ID}$ serves as $\boldsymbol{Q}$ while the contrast graph $\boldsymbol{H}^{(l)}_{contrast}$ is treated as $\boldsymbol{K}$ and $\boldsymbol{V}$ in Eq. (1). Each input brain network is fed into the cross-attention module to query the task-specific information in the contrast graph. The intuition here is to use the contrast graph as prior knowledge to guide the brain network representation learning. By hiding the non-indicative ROIs/connections and emphasizing the indicative ones, task-specific domain knowledge is introduced to the embedded brain networks.

In addition to the cross-attention sub-layer, a position-wise feed-forward network (FFN) with a layer normalization is applied to each position to get the output node representations $\boldsymbol{H}^{(l)}$ of the l-th Contrasformer layer. The FFN is applied to each position separately and identically [42]. After L layers of Contrasformer, a readout function Readout$(\cdot)$ is applied to the node representations to generate a graph representation: $\boldsymbol{hg} = \text{Readout}(\boldsymbol{H}^{(l)})$. The graph representation is then passed to the prediction head for classification.

4.3 Loss Functions

In order to introduce domain knowledge and make model optimization easier to converge, we utilize 4 loss functions to guide the end-to-end training. (1) A commonly-used cross-entropy loss $\mathcal{L}_{cls}$ [8] for graph classification; (2) an entropy loss $\mathcal{L}_{entropy}$ for contrast graph sparsification; (3) a cluster loss $\mathcal{L}_{cluster}$ to take the group relationship (i.e., similarity and discrepancy) into account; (4) a contrastive loss $\mathcal{L}_{contrast}$ to constrain node-identity awareness. The total loss is computed by:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 * \mathcal{L}_{entropy} + \lambda_2 * \mathcal{L}_{cluster} + \lambda_3 * \mathcal{L}_{contrast}, \quad (10)$$

where λ_1, λ_2 and λ_3 are trainable trade-off hyperparameters.

Entropy Loss. To prevent a smooth contrast graph that treats all ROIs equally, risking the loss of discriminative ability, we introduce a sparsity constraint. To achieve this, we employ an entropy loss, compelling the model to prioritize the most task-specific ROI connections. The entropy loss is formulated as follows:

$$\mathcal{L}_{entropy} = \frac{1}{m}\sum_{i=1}^{m} \text{entropy}\left(\boldsymbol{H}_{contrast}(i,:)\right), \quad (11)$$

$$\text{entropy}(\boldsymbol{p}) = -\sum_{j=1}^{m} \boldsymbol{p}_j \log(\boldsymbol{p}_j). \quad (12)$$

Cluster Loss. Most existing GNN/Transformer architectures treat individual input graphs independently during training. Neglecting the relationships between classes could lead to a significant compromise in model effectiveness for downstream classification tasks [49]. For our application, we want to find the common patterns/biomarkers for a certain neurological disorder identification task. Thus we propose a cluster loss to leverage graph-level similarity and make the graph representations more sepratable:

$$\mathcal{L}_{cluster} = \log \frac{\exp(\sum_{c\in C}\sigma_c^2)}{\exp(\sum_{c\in C}\sum_{i\in C}||\mu_c - \mu_i||_2)}, \quad (13)$$

$$\mu_c = \sum_{k\in\mathcal{S}^c}\frac{\boldsymbol{hg}^k}{|\mathcal{S}^c|}, \quad \sigma_c^2 = \sum_{k\in\mathcal{S}^c}\frac{(\boldsymbol{hg}^k - \mu_c)^2}{|\mathcal{S}^c|}, \quad (14)$$

where μ_c and σ_c denote the mean and standard deviation of graph representations belonging to group c, $\mathcal{S}^c$ denotes the subject indices that belong to group c, C denotes the set of classes, and $\boldsymbol{hg}^k$ denotes the graph representation with index k in $\mathcal{S}^c$. As in Fig. 6(a), the cluster loss aims to pull the graph representations within a group close to each other and push the centers of groups as far as possible. By using such cluster loss, the group-level relationship of all classes is considered equally no matter how many subjects it contains.

Contrastive Loss. Existing graph contrastive learning technologies [55, 57] require data argumentation by modifying graph structure or dropping node/edge features. Such contrast is still limited to each individual graph. It also cannot be migrated to our brain networks because the connectivity matrix naturally contains its structural and positional information.

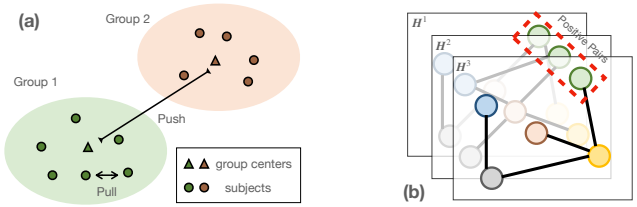


Figure 6: (a) The cluster loss enforces subjects that belong to the same group to get similar representations, the subjects from different groups less similar. (b) The contrastive loss treats the nodes belonging to the same ROI as positive pairs, and all the other node pairs are considered negative pairs.

Herein, we design an ROI-level contrastive loss to further leverage the node identity of ROIs. Thanks to the unique property of brain networks' node-identity awareness, we are able to conduct contrastive learning by aligning ROIs across subjects. To the best of our knowledge, this is the first attempt to bring in contrastive constraints at the ROI level for brain network analysis. The contrastive loss $\mathcal{L}_{contrast}$ shown in Eq. (15) is defined to enforce maximizing the consistency between positive pairs compared with negative pairs [7]. We use $\boldsymbol{H}_j^i$ to denote the node representation for the j-th node in the i-th subject, where $j = 1,\ldots,m$, $i = 1,\ldots,n$, and n is the total number of subjects in the training set. We denote the set of positive pairs as $\mathcal{P}^{pos} = \{(\boldsymbol{H}_j^i, \boldsymbol{H}_j^p)\}$ and the set of negative pairs as $\mathcal{P}^{neg} = \{(\boldsymbol{H}_j^i, \boldsymbol{H}_r^q)\}$, where $p = 1,\ldots,n, p \neq i$, $q = 1,\ldots,n$, $r = 1,\ldots,m, r \neq j$. A temperature hyper-parameter τ [55] is introduced to control the smoothness of the probability distribution, and sim$(\cdot)$ denotes the cosine similarity function. As elaborated in Fig. 6(b), we treat the same ROI of all subjects as positive pairs, and different ROIs from the same/different subjects as negative pairs to emphasize the node-identity awareness of brain networks.

$$\mathcal{L}_{contrast} = -\log \frac{\sum_{(\boldsymbol{H}_j^i, \boldsymbol{H}_j^p)\in\mathcal{P}^{pos}} \exp(\text{sim}(\boldsymbol{H}_j^i, \boldsymbol{H}_j^p)/\tau)}{\sum_{(\boldsymbol{H}_j^i, \boldsymbol{H}_r^q)\in\mathcal{P}^{neg}} \exp(\text{sim}(\boldsymbol{H}_j^i, \boldsymbol{H}_r^q)/\tau)}. \quad (15)$$

5 EXPERIMENTAL STUDY

Table 1: Statistics of Brain Network Datasets.

Dataset	Condition	Site#	Graph#	Avg BOLD length	Class#
Mātai	mTBI	1	60	198	2
PPMI	PD	21	209	198	4
ADNI	AD	89	1326	344	6
ABIDE	ASD	17	1025	201	2

5.1 Brain Network Datasets

We use four brain network datasets from different data sources for various disorders, which are Mātai for mild traumatic brain injury (mTBI) [50], PPMI [3] for Parkinson's disease (PD), ADNI

Table 2: Graph Classification Results (Average Accuracy ± Standard Deviation) over 10-fold-CV. The best result is highlighted in bold. The second-best result is underlined.

	Model	Mātai	PPMI	ADNI	ABIDE
Conventional ML methods	LR	60.00 ± 20.00	56.48 ± 6.76	61.97 ± 4.24	64.81 ± 3.70
	SVM	56.67 ± 17.00	63.21 ± 8.62	61.52 ± 4.95	64.41 ± 5.09
General-purposed GNNs	GCN	56.67 ± 18.56	54.02 ± 9.06	60.40 ± 4.89	60.19 ± 2.96
	GraphSAGE	61.67 ± 10.67	55.00 ± 12.89	59.35 ± 3.39	61.75 ± 4.35
	GAT	66.67 ± 18.33	54.98 ± 8.03	59.73 ± 2.85	60.10 ± 4.13
	GatedGCN	58.33 ± 8.33	52.60 ± 11.51	64.55 ± 1.87	61.66 ± 3.36
	GPS	63.33 ± 14.53	57.50 ± 7.83	65.17 ± 3.50	63.04 ± 3.36
Neural networks tailored for brain networks	BrainNetCNN	61.67 ± 13.33	57.33 ± 10.32	60.48 ± 3.29	65.75 ± 3.24
	LiNet	51.67 ± 13.84	60.71 ± 10.61	61.91 ± 1.02	54.05 ± 4.50
	PRGNN	55.00 ± 16.75	58.83 ± 6.89	62.51 ± 3.36	59.71 ± 4.54
	BrainGNN	53.33 ± 24.49	61.71 ± 6.05	61.05 ± 1.23	62.88 ± 2.46
	BNTF	61.67 ± 11.17	51.60 ± 6.15	65.49 ± 3.25	63.70 ± 4.84
	ContrastPool	61.67 ± 13.02	64.00 ± 6.63	65.67 ± 6.64	65.01 ± 3.84
Ours	Contrasformer	**68.33** ± 18.93	**67.00** ± 4.58	**69.33** ± 3.63	**66.27** ± 3.67

Table 3: Results of more evaluation metrics on ABIDE. The best result is highlighted in bold. The second-best result is underlined.

	Model	Precision	Recall	micro-F1	ROC-AUC
General-purposed GNNs	GCN	59.69 ± 5.50	57.56 ± 7.01	58.49 ± 5.78	61.08 ± 4.92
	GraphSAGE	60.33 ± 4.91	58.20 ± 7.85	58.99 ± 5.45	61.59 ± 4.36
	GAT	59.57 ± 4.63	55.11 ± 7.89	56.96 ± 5.16	60.43 ± 3.88
	GatedGCN	61.65 ± 4.12	55.74 ± 11.58	58.05 ± 8.20	62.31 ± 4.32
	GPS	59.97 ± 5.36	68.75 ± 11.22	63.48 ± 5.98	63.34 ± 5.15
Neural networks tailored for brain networks	BrainNetCNN	**63.98** ± 3.47	61.25 ± 5.66	62.39 ± 3.13	64.78 ± 2.52
	LiNet	56.34 ± 6.94	30.37 ± 9.23	38.66 ± 8.32	54.39 ± 3.29
	PRGNN	60.83 ± 7.44	53.68 ± 8.36	56.77 ± 7.15	61.01 ± 5.74
	BrainGNN	62.48 ± 5.92	57.15 ± 4.66	59.42 ± 3.23	62.74 ± 3.58
	BNTF	60.34 ± 5.40	70.19 ± 8.66	64.64 ± 5.65	64.15 ± 5.42
	ContrastPool	63.56 ± 3.62	61.45 ± 5.43	62.28 ± 2.81	64.52 ± 2.30
Ours	Contrasformer	62.59 ± 4.40	**73.07** ± 4.69	**67.25** ± 3.01	**66.62** ± 3.47

[10] for Alzheimer's disease (AD), and ABIDE [9] for Autism (ASD). Statistics of the brain network datasets are summarized in Table 1.

Mātai Mātai is a longitudinal single site, single scanner study designed for detecting subtle changes in the brain due to a season of playing contact sports. This dataset consists of the brain networks preprocessed from the data collected from Gisborne-Tairāwhiti area, New Zealand, with 35 contact sport players imaged at pre-season (N=35) and post-season (N=25) with subtle brain changes confirmed using diffusion imaging study due to playing contact sports.

PPMI The PPMI[1] is a comprehensive study aiming to identify biological markers associated with Parkinson's risk, onset, and progression. PPMI comprises multimodal and multi-site MRI images. The dataset consists of subjects from 4 distinct classes: normal control (NC), scans without evidence of dopaminergic deficit (SWEDD), prodromal, and PD.

ADNI The ADNI raw images used in this paper were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. We include subjects from 6 different stages of AD, from cognitive normal (CN), significant memory concern (SMC), mild cognitive impairment (MCI), early MCI (EMCI), late MCI (LMCI) to AD.

ABIDE The ABIDE initiative supports the research on ASD by aggregating functional brain imaging data from laboratories worldwide. ASD is characterized by stereotyped behaviors, including irritability, hyperactivity, and anxiety. Subjects in the dataset are classified into two groups: TC and individuals diagnosed with ASD.

5.2 Baseline Models

We employ diverse baseline models to benchmark our approach, including (1) Conventional machine learning models: Logistic Regression (**LR**) and Support Vector Machine Classifier (**SVM**) from scikit-learn [36]. These models take the flattened connectivity matrix as vector input, instead of using the brain network. (2)General-purposed GNNs: **GCN** [26], a mean pooling baseline with a graph convolution network as a message-passing layer; **GraphSAGE** [18] is a mean pooling baseline, which adopts sampling to obtain a fixed number of neighbors for each node; **GAT** [43], a mean pooling

[1] https://www.ppmi-info.org/accessdata-specimens/download-data

baseline utilizing an attention mechanism to learn relative node-neighbor weights; **GatedGCN** [5] incorporates gate mechanisms to selectively control information flow during message passing, enabling improved modeling of long-range dependencies in graph; and **GPS** [38], a state-of-the-art Transformer-based model incorporating various positional/structural embeddings. (3) Neural networks tailored for brain networks: **BrainNetCNN** [23], the pioneering CNN regressor for connectome data; **LiNet** [32], an inductive GNN for task-fMRI properties identification; **PRGNN** [35], a graph pooling method ensuring ROI-selection coherence; **BrainGNN** [34], a GNN method with ROI-aware convolutional layers for fMRI information integration; Brain Network Transformer (**BNTF**) [20], a specialized graph Transformer with orthonormal clustering for brain network analysis; and **ContrastPool** [48], a node clustering pooling using a dual-attention block for domain-specific information capturing.

Note that, to prevent over-smoothing in GNNs, we sparsify the adjacency matrices of all input graphs by retaining the top 20% edges with the highest correlations in $\boldsymbol{M}$.

5.3 Implementation Details

In Contrasformer, we adopt a Linear layer for input embedding, a mean pooling layer as the readout function Readout$(\cdot)$, and a two-layer MLP with ReLU as the prediction head. The temperature hyper-parameter τ in Eq. (15) is set to 0.02. The settings of our experiments mainly follow those in [13]. We split each dataset into 8:1:1 for training, validation and test, respectively. We evaluate each model with the same random seed under 10-fold cross-validation and report the average accuracy. The whole network is trained in an end-to-end manner using the Adam optimizer [25]. We use the early stopping criterion, i.e., we stop the training once there is no further improvement on the validation loss during 25 epochs. All experiments were conducted on a Linux server with an Intel(R) Core(TM) i9-10940X CPU (3.30GHz), a GeForce GTX 3090 GPU, and a 125GB RAM.

5.4 Main Results

We report the accuracy on 4 brain network datasets in Table 2. Our proposed Contrasformer consistently outperforms all 13 baselines on all datasets. In particular, Contrasformer improves over all networks specifically designed for brain networks on these four datasets by up to 10.8% ((68.33%-61.67%)/61.67% = 10.8% on Mātai). These experimental results demonstrate the effectiveness of our brain network oriented model design. The improvement may result from two reasons. First, the participation of the contrast graph in brain network representation learning provides reasonable and discriminative information about certain conditions. Second, the properties of fMRI and group constraints are introduced to the model training by dedicated loss functions.

Apart from accuracy, we also report other evaluation metrics, including precision, recall, micro-F1, and ROC-AUC, of all the models on the ABIDE dataset. As shown in Table 3, Contrasformer performs the best over all these metrics except for precision. We can also discover that compared with other baselines, our Contrasformer can dramatically improve recall without sacrificing precision. Besides, in medical diagnostics, it's crucial to ensure that all individuals with a certain condition are correctly identified, even if it means some false positives. Missing a true positive (failing to diagnose a disease) can have severe consequences, while false positives can be further examined or retested. Therefore, models with higher recall rates, like our Contrasformer, are more suitable for application in real-life medical auxiliary diagnosis.

5.5 Model Interpretation

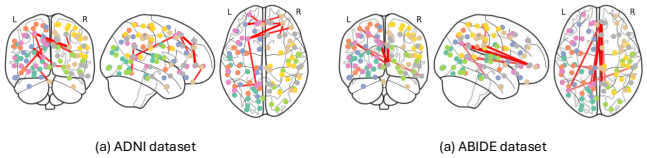


Figure 7: Contrast graph visualization by highlighting the top 10 edges with the highest strength.

Contrast Graph Visualization. Despite the high accuracy achieved by our model, a critical concern is the interpretability of their decision-making process. In the context of brain biomarker detection, identifying salient ROIs associated with predictions as candidate/potential biomarkers is crucial. In this study, we leverage built-in model interpretability to explore disease-specific biomarker analysis. To interpret Contrasformer's reasoning, we visualize the learnt contrast graphs for Alzheimer's Disease and Autism on the ADNI and ABIDE datasets, by using the Nilearn toolbox [1]. We select edges with the top 10 attention weights. As depicted in Fig. 7(a), highlighted connections between the lateral prefrontal cortex, prefrontal cortex, and dorsal prefrontal cortex medial prefrontal cortex in the ADNI dataset suggest potential AD-specific neural mechanisms [31]. These regions have been recognized as key areas in previous AD studies [16, 39]. Similar ROI-wise interpretations are found in Autism. In Fig. 7(b), highlighted ROIs related to precuneus posterior cingulate cortex, cingulate, and dorsal prefrontal cortex medial prefrontal cortex on ABIDE align with domain knowledge from prior Autism research [2, 15, 22].

Generalization Ability. While task-specific biomarkers are valuable for identifying disease-relevant features, it is crucial to determine whether these biomarkers are invariant over the entire population, i.e., whether they generalize well across sub-populations and other diverse populations. To assess the generalization ability of Contrasformer, we conduct evaluations on subjects from previously unseen sites. Specifically, we designate two sites from the ABIDE dataset as the test set, while the remaining subjects are split into training and validation sets, maintaining an 8:1:1 ratio. The test set remains constant across 10 experiments with different train-validation splits, and the average results are reported in Table 4. Notably, Contrasformer consistently outperforms all baseline models, demonstrating its robustness against SPS noise. The baseline models exhibit a significant performance reduction compared to Table 2, indicating the detrimental impact of SPS noise on their generalization ability. In contrast, Contrasformer, with its two-stream attention mechanism, effectively extracts and emphasizes task-related features, mitigating the adverse effects of SPS noise.

Table 4: Results when generalizing to unseen sites. The best result is highlighted in bold. The second-best result is underlined.

	Accuracy	Precision	Recall	micro-F1	ROC-AUC
GCN	55.15 ± 3.63	48.82 ± 3.54	56.67 ± 7.04	52.33 ± 4.69	55.32 ± 3.76
GatedGCN	59.03 ± 1.55	52.87 ± 1.67	55.33 ± 8.52	53.79 ± 4.96	58.61 ± 2.17
BrainNetCNN	61.55 ± 2.31	54.44 ± 2.10	74.44 ± 5.19	62.81 ± 2.45	63.00 ± 2.28
LiNet	53.50 ± 2.89	45.99 ± 4.07	35.11 ± 3.69	39.71 ± 3.26	51.43 ± 2.70
PRGNN	55.63 ± 2.34	49.21 ± 2.77	49.56 ± 3.86	49.35 ± 3.08	54.95 ± 2.43
ContrastPool	56.18 ± 2.01	48.60 ± 3.91	42.27 ± 10.77	44.61 ± 8.18	54.50 ± 2.61
Contrasformer (ours)	**64.61** ± 1.83	**56.52** ± 1.56	**77.95** ± 3.53	**65.50** ± 1.93	**66.22** ± 1.88

Table 5: Ablation study on important modules/loss functions in Contrasformer on ABIDE dataset.

Attn_{ROI}	$\text{Attn}_{subject}$	batch norm	ID enc	Accuracy	Precision	Recall	micro-F1	ROC-AUC
	✓	✓	✓	63.14 ± 4.71	59.18 ± 5.36	75.76 ± 7.83	66.04 ± 4.11	63.77 ± 4.78
✓		✓	✓	63.63 ± 4.28	62.42 ± 6.02	60.00 ± 6.37	60.91 ± 4.79	63.44 ± 4.03
✓	✓		✓	65.69 ± 3.90	**62.83** ± 4.36	68.90 ± 10.24	65.22 ± 5.47	65.83 ± 3.91
✓	✓	✓		64.22 ± 3.87	59.97 ± 5.41	**76.80** ± 4.51	67.12 ± 3.31	64.84 ± 4.60
✓	✓	✓	✓	**66.27** ± 3.67	62.59 ± 4.40	73.07 ± 4.69	**67.25** ± 3.01	**66.62** ± 3.47
$\mathcal{L}_{cls}$	$\mathcal{L}_{entropy}$	$\mathcal{L}_{cluster}$	$\mathcal{L}_{contrast}$	Accuracy	Precision	Recall	micro-F1	ROC-AUC
✓		✓	✓	64.61 ± 3.93	61.79 ± 4.63	68.51 ± 9.36	64.45 ± 4.80	64.80 ± 3.69
✓	✓		✓	64.41 ± 4.34	60.36 ± 4.90	**75.57** ± 5.44	66.84 ± 2.86	64.99 ± 4.00
✓	✓	✓		59.51 ± 4.09	57.66 ± 6.84	58.25 ± 13.77	56.98 ± 8.30	59.50 ± 5.72
✓	✓	✓	✓	**66.27** ± 3.67	**62.59** ± 4.40	73.07 ± 4.69	**67.25** ± 3.01	**66.62** ± 3.47

5.6 Ablation Study

In this subsection, we empirically validate the design of (1) the important modules; and (2) the loss functions. All experiments in this subsection are conducted on ABIDE dataset.

Important Modules. To inspect the effect of the important modules, we conduct experiments by disabling each of them without modifying other settings. The results are reported in Table 5. For ROI-wise attention Attn_{ROI}, subject-wise attention $\text{Attn}_{subject}$, and identity encoding (denoted as "ID enc" in the table), we disable them by simply removing these modules. When disabling "batch norm", we replace the batch normalization functions in the two-stream attention by layer normalizations. The results demonstrate that Contrasformer with all important modules enabled achieves the best performance. Besides, the experiment of disabling $\text{Attn}_{subject}$ indicates that the outstanding recall of Contrasformer is mainly contributed by the subject-wise attention, demonstrating the effectiveness of extracting discriminative ROIs across subjects.

Loss Functions. To verify the effectiveness of our proposed losses, we test our design of the loss functions by disabling them one by one. As shown in Table 5, the results demonstrate that all of those three auxiliary losses are effective in boosting the model performance. Besides, we find that the most important one is the contrastive loss. This observation indicates the necessity of introducing the constraint of node awareness.

6 CONCLUSION AND FUTURE WORKS

To overcome the hurdles of SPS noise and node-identity awareness, we introduce Contrasformer, a contrastive brain network Transformer. Through a contrast graph encoder with two-stream attention and a cross decoder with identity embedding, Contrasformer adaptly handles SPS noise, enhances node identity awareness, and captures group-specific patterns. Our model outperforms state-of-the-art methods in identifying neurological disorders across diverse datasets. The improvement over all the best models specifically designed for brain networks is up to 10.8%. Beyond superior performance, Contrasformer provides interpretable insights aligned with neuroscience literature. This work marks a significant advancement in harnessing Transformer models for fMRI-based brain network analysis, opening avenues for deeper understanding and diagnosis of neurological conditions. In the future, we plan to extend our model to various modalities of medical imaging, such as Diffusion Tensor Imaging (DTI), and explore the multi-modal solution for neurological disorder identification.

7 ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative, and the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20220-0006), and MBIE Catalyst: Strategic Fund NZ-Singapore Data Science Research Programme UOAX2001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, and the Ministry of Education, Singapore.

Part of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

REFERENCES

[1] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. 2014. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics* 8 (2014), 14.

[2] Michal Assaf, Kanchana Jagannathan, Vince D Calhoun, Laura Miller, Michael C Stevens, Robert Sahl, Jacqueline G O'Boyle, Robert T Schultz, and Godfrey D Pearlson. 2010. Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. *Neuroimage* 53, 1 (2010), 247–256.

[3] Liviu Badea, Mihaela Onu, Tao Wu, Adina Roceanu, and Ovidiu Bajenaru. 2017. Exploring the reproducibility of functional connectivity alterations in Parkinson's disease. *PLoS One* 12, 11 (2017), e0188196.

[4] Qingtian Bian, Jiaxing Xu, Hui Fang, and Yiping Ke. 2023. CPMR: Context-Aware Incremental Sequential Recommendation with Pseudo-Multi-Task Learning. *arXiv preprint arXiv:2309.04802* (2023).

[5] Xavier Bresson and Thomas Laurent. 2017. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553* (2017).

[6] Yi Hao Chan, Wei Chee Yew, and Jagath C Rajapakse. 2022. Semi-supervised Learning with Data Harmonisation for Biomarker Discovery from Resting State fMRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 441–451.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[8] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232.

[9] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics* 7 (2013), 27.

[10] Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, Gaël Varoquaux, Alzheimer's Disease Neuroimaging Initiative, et al. 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage* 192 (2019), 115–134.

[11] Wei Dai, Hejie Cui, Xuan Kan, Ying Guo, Sanne van Rooij, and Carl Yang. 2023. Transformer-based hierarchical clustering for brain network analysis. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.

[12] Vijay Prakash Dwivedi and Xavier Bresson. 2020. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699* (2020).

[13] Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2020. Benchmarking Graph Neural Networks. *arXiv preprint arXiv:2003.00982* (2020).

[14] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2021. Graph neural networks with learnable structural and positional representations. *arXiv preprint arXiv:2110.07875* (2021).

[15] Christine Ecker, Andre Marquand, Janaina Mourão-Miranda, Patrick Johnston, Eileen M Daly, Michael J Brammer, Stefanos Maltezos, Clodagh M Murphy, Dene Robertson, Steven C Williams, et al. 2010. Describing the brain in autism in five dimensions—magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *Journal of Neuroscience* 30, 32 (2010), 10612–10623.

[16] Stefan Frisch, Juergen Dukart, Barbara Vogt, Annette Horstmann, Georg Becker, Arno Villringer, Henryk Barthel, Osama Sabri, Karsten Müller, and Matthias L Schroeter. 2013. Dissociating memory networks in early Alzheimer's disease and frontotemporal lobar degeneration-a combined study of hypometabolism and atrophy. *PloS one* 8, 2 (2013), e55251.

[17] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*. PMLR, 1263–1272.

[18] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[19] Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. 2022. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International Conference on Medical Imaging with Deep Learning*. PMLR, 618–637.

[20] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. 2022. Brain network transformer. *arXiv preprint arXiv:2210.06681* (2022).

[21] Xuan Kan, Antonio Aodong Chen Gu, Hejie Cui, Ying Guo, and Carl Yang. 2023. Dynamic Brain Transformer with Multi-level Attention for Functional Brain Network Analysis. *arXiv preprint arXiv:2309.01941* (2023).

[22] Rajesh K Kana, Emma B Sartin, Carl Stevens Jr, Hrishikesh D Deshpande, Christopher Klein, Mark R Klinger, and Laura Grofer Klinger. 2017. Neural networks underlying language and social cognition during self-other processing in Autism spectrum disorders. *Neuropsychologia* 102 (2017), 116–123.

[23] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage* 146 (2017), 1038–1049.

[24] Nicolas Keriven. 2022. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *Advances in Neural Information Processing Systems* 35 (2022), 2268–2281.

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[27] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. 2021. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems* 34 (2021), 21618–21629.

[28] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. 2017. Distance metric learning using graph convolutional networks: Application to functional brain networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 469–477.

[29] Tommaso Lanciano, Francesco Bonchi, and Aristides Gionis. 2020. Explainable classification of brain networks via contrast subgraphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3308–3318.

[30] Pan Li, Yanbang Wang, Hongwei Wang, and Jure Leskovec. 2020. Distance encoding: Design provably more powerful neural networks for graph representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 4465–4478.

[31] Rui Li, Xia Wu, Adam S Fleisher, Eric M Reiman, Kewei Chen, and Li Yao. 2012. Attention-related networks in Alzheimer's disease: A resting functional MRI study. *Human brain mapping* 33, 5 (2012), 1076–1088.

[32] Xiaoxiao Li, Nicha C Dvornek, Yuan Zhou, Juntang Zhuang, Pamela Ventola, and James S Duncan. 2019. 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*. Springer, 485–493.

[33] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*. 2572–2580.

[34] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. 2021. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis* 74 (2021), 102233.

[35] Xiaoxiao Li, Yuan Zhou, Nicha C Dvornek, Muhan Zhang, Juntang Zhuang, Pamela Ventola, and James S Duncan. 2020. 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*. Springer, 625–635.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[37] Russell A Poldrack, Yaroslav O Halchenko, and Stephen José Hanson. 2009. Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological science* 20, 11 (2009), 1364–1372.

[38] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. 2022. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems* 35 (2022), 14501–14515.

[39] SARB Rombouts, F Barkhof, CS Van Meel, and P Scheltens. 2002. Alterations in brain activation during cholinergic enhancement with rivastigmine in Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry* 73, 6 (2002), 665–671.

[40] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral cortex* 28, 9 (2018), 3095–3114.

[41] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[43] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[44] Haorui Wang, Haoteng Yin, Muhan Zhang, and Pan Li. 2022. Equivariant and stable positional encoding for more powerful graph neural networks. *arXiv preprint arXiv:2203.00199* (2022).

[45] Xinlei Wang, Jinyi Chen, Bing Tian Dai, Junchang Xin, Yu Gu, and Ge Yu. 2023. Effective Graph Kernels for Evolving Functional Brain Networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*.

150–158.

[46] Keith J Worsley, Chien Heng Liao, John Aston, V Petre, GH Duncan, F Morales, and Alan C Evans. 2002. A general statistical analysis for fMRI data. *Neuroimage* 15, 1 (2002), 1–15.

[47] Jialun Wu, Kai He, Rui Mao, Chen Li, and Erik Cambria. 2023. MEGACare: Knowledge-guided multi-view hypergraph predictive framework for healthcare. *Information Fusion* 100 (2023), 101939.

[48] Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. 2024. Contrastive Graph Pooling for Explainable Classification of Brain Networks. *IEEE Transactions on Medical Imaging* (2024), 1–1. https://doi.org/10.1109/TMI.2024.3392988

[49] Jiaxing Xu, Jinjie Ni, and Yiping Ke. 2024. A class-aware representation refinement framework for graph classification. *Information Sciences* 679 (2024), 121061. https://doi.org/10.1016/j.ins.2024.121061

[50] Jiaxing Xu, Yunhan Yang, David Tse Jung Huang, Sophi Shilpa Gururajapathy, Yiping Ke, Miao Qiao, Alan Wang, Haribalan Kumar, Josh McGeown, and Eryn Kwon. 2023. Data-Driven Network Neuroscience: On Data Collection and Benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[51] Jiaxing Xu, Aihu Zhang, Qingtian Bian, Vijay Prakash Dwivedi, and Yiping Ke. 2024. Union subgraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 16173–16183.

[52] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

[53] Yujun Yan, Jiong Zhu, Marlena Duda, Eric Solarz, Chandra Sripada, and Danai Koutra. 2019. Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 772–782.

[54] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems* 34 (2021), 28877–28888.

[55] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems* 33 (2020), 5812–5823.

[56] Hao Zhang, Ran Song, Liping Wang, Lin Zhang, Dawei Wang, Cong Wang, and Wei Zhang. 2022. Classification of Brain Disorders in rs-fMRI via Local-to-Global Graph Neural Networks. *IEEE Transactions on Medical Imaging* (2022).

[57] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*. 2069–2080.