frontiers in Neuroscience

ORIGINAL RESEARCH
published: 30 June 2020
doi: 10.3389/fnins.2020.00630

Understanding Graph Isomorphism Network for rs-fMRI Functional Connectivity Analysis

*Byung-Hoon Kim and Jong Chul Ye**

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea

Graph neural networks (GNN) rely on graph operations that include neural network training for various graph related tasks. Recently, several attempts have been made to apply the GNNs to functional magnetic resonance image (fMRI) data. Despite recent progresses, a common limitation is its difficulty to explain the classification results in a neuroscientifically explainable way. Here, we develop a framework for analyzing the fMRI data using the Graph Isomorphism Network (GIN), which was recently proposed as a powerful GNN for graph classification. One of the important contributions of this paper is the observation that the GIN is a dual representation of convolutional neural network (CNN) in the graph space where the shift operation is defined using the adjacency matrix. This understanding enables us to exploit CNN-based saliency map techniques for the GNN, which we tailor to the proposed GIN with one-hot encoding, to visualize the important regions of the brain. We validate our proposed framework using large-scale resting-state fMRI (rs-fMRI) data for classifying the sex of the subject based on the graph structure of the brain. The experiment was consistent with our expectation such that the obtained saliency map show high correspondence with previous neuroimaging evidences related to sex differences.

Keywords: graph neural networks, saliency mapping, functional neuroimaging, resting-state, explainable artificial intelligence

OPEN ACCESS

Edited by:
João Manuel R. S. Tavares, University of Porto, Portugal

Reviewed by:
Monica Bianchini, University of Siena, Italy
Regina Júlia Deák-Meszlényi, Hungarian Academy of Sciences (MTA), Hungary

****Correspondence:***
Jong Chul Ye
jong.ye@kaist.ac.kr

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *25 March 2020*
Accepted: *22 May 2020*
Published: *30 June 2020*

Citation:
Kim B-H and Ye JC (2020) Understanding Graph Isomorphism Network for rs-fMRI Functional Connectivity Analysis. Front. Neurosci. 14:630. doi: 10.3389/fnins.2020.00630

1. INTRODUCTION

Graphs provide an efficient way to mathematically model non-regular interactions between data in terms of nodes and edges (Bassett and Bullmore, 2009; He and Evans, 2010; Sporns, 2018). The network of the brain can be modeled as a graph consisting of ROIs as the nodes and their functional connectivity as the edges (Bassett and Sporns, 2017). In classical graph theoretic approaches, various graph metrics including local/global efficiency, average path length, and small-worldedness, are computed to analyze the brain networks (Wang et al., 2010). These metrics could be further used for group comparison to reveal the different network properties, providing insights to the physiological characteristics and the disorders of the brain (Micheloyannis et al., 2006; Tian et al., 2011).

Recently, there have been remarkable progresses and growing interests in Graph Neural Networks (GNNs), which comprise graph operations performed by deep neural networks (see the extensive survey in Wu et al., 2019). The GNNs are suitable for solving tasks such as node classification, edge prediction, graph classification, etc. Usual GNNs typically integrate the features at each layer to embed each node features into a predefined next layer feature vector.

The integration process is implemented by choosing appropriate functions for aggregating features of the neighborhood nodes. As one layer in the GNN aggregates its 1-hop neighbors, each node feature is embedded with features within its k-hop neighbors of the graph after k aggregating layers. The feature of the whole graph is then extracted by applying a readout function to the embedded node features.

Considering the development of GNNs, it is not surprising that there are keen interests in applying GNNs to fMRI data analysis. For example, some works have applied the GNN to classify one's phenotypic status based on the graph structure of the brain functional networks (Ktena et al., 2017, 2018; Ma et al., 2018; Li et al., 2019a,b). Some other works employed the GNN to classify the subjects, not only based on the imaging data, but also including the non-image phenotypic data (Parisot et al., 2017, 2018; He et al., 2019). Despite the early contribution of these works in applying the GNNs for fMRI analysis, there exists a common limitation in that they often fail to provide proper mapping of the ROIs for neuroscientific interpretation. To overcome this limitation, there have been recent attempts to address the issue of neuroscientific interpretability by visualizing the important features of the brain (Arslan et al., 2018; Duffy et al., 2019; Li et al., 2019a). These attempts involved saliency mapping methods of the GNNs, such as class activation mapping (CAM) (Zhou et al., 2016) to delineate the important features, as demonstrated in Arslan et al. (2018).

Here we revisit the Graph Isomorphism Network (GIN) (Xu et al., 2018a), which was recently proposed to implement Weisfeiler-Lehman (WL) graph isomorphism test (Shervashidze et al., 2011) in a neural network. Our classification results on sex classification confirmed that GIN method can provide more powerful classification performance, but the direct calculation of the graph saliency map was not clear.

Therefore, another important contribution of this work is to show that while GIN is similar to spectral-domain approaches such as the graph convolutional network (GCN) in learning the spectral filters from graphs, GIN can be considered as a dual representation of the convolutional neural network (CNN) with two-tab convolution filter in the graph space where the adjacency matrix is defined as a generalized shift operation. With this generalization, we can employ one of the most widely used saliency map visualization technique in CNN, called the gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) that can be applied to any CNN architecture at any layer. We further found that to visualize the important brain regions that are related to a certain phenotypic difference, Grad-CAM should be calculated at the input layer and the one-hot encoding of the graph node is ideally suitable for such saliency map visualization.

Experimental results on sex classification confirm that our method can provide more accurate classification performance and better interpretability of the classification results in terms of saliency maps, which provide some new insights to the topic of sex differences on the resting-state fMRI (rs-fMRI).

1.1. Mathematical Preliminaries

We denote a graph $G = (V, E)$ with a set of vertices $V(G) = \{1, \cdots, N\}$ with $N := |V|$ and edges $E(G) = \{e_{ij}\}$, where an edge e_{ij} connects vertices i and j if they are adjacent or neighbors. The set of neighborhoods of a vertex v is denoted by $\mathcal{N}(v)$. For weighted graphs, the edge e_{ij} has a real value. If G is an unweighted graph, then E is a sparse matrix with elements of either 0 or 1.

When analyzing the fMRI data, the functional connectivity between two regions of the brain is often computed from the Pearson correlation coefficient between the fMRI time series. Specifically, the Pearson correlation coefficient between the fMRI time series $\boldsymbol{y}_i$ at the vertex i and the fMRI time series $\boldsymbol{y}_j$ at the vertex j is given by

$$R_{ij} = \frac{\mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{y}_j)}{\sigma_{\boldsymbol{y}_i}\sigma_{\boldsymbol{y}_j}} \in \mathbb{R}^{N\times N}$$

where $\mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{y}_j)$ is the cross covariance between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, and $\sigma_{\boldsymbol{y}_i}$ denotes the standard deviation of $\boldsymbol{y}_i$. Unweighted graph edge can be derived from the functional connectivity by thresholding the correlation coefficients by a certain threshold.

For a simple unweighted graph with vertex set V, the adjacency matrix is a square $|V| \times |V|$ matrix $\boldsymbol{A}$ such that its element A_{uv} is one when there is an edge from vertex u to vertex v, and zero when there is no edge. For the given adjacency matrix $\boldsymbol{A}$, the graph Laplacian $\boldsymbol{L}$ and its normalized version $\boldsymbol{L}_n$ are then defined by

$$\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{A}, \quad \boldsymbol{L}_n = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} \tag{1}$$

where $\boldsymbol{D}$ is the degree matrix with the diagonal element

$$D_{uu} = d(u) = \sum_v A_{uv} \tag{2}$$

and zeros elsewhere.

Graph Laplacian is useful for signal processing on a graph (Shuman et al., 2013; Huang et al., 2018; Ortega et al., 2018). More specifically, the graph convolution for real-valued functions on the set of the graph's vertices, $\boldsymbol{x}, \boldsymbol{y} : V \mapsto \mathbb{R}^{|V|}$ is often defined by

$$\boldsymbol{x} *_G \boldsymbol{y} = \boldsymbol{U}\left(\boldsymbol{U}^\top \boldsymbol{x} \odot \boldsymbol{U}^\top \boldsymbol{y}\right) \tag{3}$$

where the superscript $^\top$ denotes the adjoint operation, $\boldsymbol{U}$ is the matrix composed of singular vectors of the normalized graph Laplacian, i.e.,

$$\boldsymbol{L}_n = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \tag{4}$$

where $\boldsymbol{\Lambda}$ denotes the diagonal matrices with the singular values, which is often referred to as the graph spectrum.

1.2. Graph Neural Networks

The goal of GNNs for the graph classification task is to learn a non-linear mapping g from a graph to a feature vector:

$$g : G \mapsto \boldsymbol{p}_G, \tag{5}$$

where $\boldsymbol{p}_G$ is a feature vector of the whole graph G that helps predicting the labels of the graph. Recent perspective distinguishes the GNNs into two groups based on the neighborhood aggregating schemes (Wu et al., 2019). First group is the spectral-based convolutional GNNs (spectral GNN). This group of GNNs are inspired by the spectral decomposition of the graphs, and aim to approximate the spectral filters in each aggregating layers (Bruna et al., 2013; Kipf and Welling, 2016). The other group of GNNs are the spatial-based convolutional GNNs (spatial GNN). They do not explicitly aim to learn spectral features of graph, but rather implement the neighborhood aggregation based on the nodes' spatial relations. Some well-known examples of the spatial GNNs are the Message Passing Neural Network (MPNN) (Gilmer et al., 2017) and the GIN (Xu et al., 2018a). In this section, we provide a brief review of the these approaches to understand their relationships.

Spectral GNNs are based on the graph convolution relationship (3), in which $\boldsymbol{U}^\top \boldsymbol{y}$ is replaced by the parameterized graph spectrum $\hat{\boldsymbol{y}} := \boldsymbol{U}^\top \boldsymbol{y}$:

$$\boldsymbol{x} *_G \boldsymbol{y} = \boldsymbol{U}\left(\hat{\boldsymbol{y}} \odot \boldsymbol{U}^\top \boldsymbol{x}\right)$$

More specifically, the graph convolutional layer of the spectral GNN is then implemented as follows:

$$\boldsymbol{x}_i^{(k)} = \sigma\left(\sum\nolimits_j \boldsymbol{U}\boldsymbol{Y}_{i,j}^{(k)}\boldsymbol{U}^\top \boldsymbol{x}_j^{(k-1)}\right) \tag{6}$$

where $\sigma(\cdot)$ is an element-by-element non-linearity, $\boldsymbol{x}_i^{(k)}$ is the graph signal at the channel i of k-th layer and $\boldsymbol{Y}_{i,j}^{(k)}$ is a diagonal matrix that parameterized the graph spectrum $\hat{\boldsymbol{y}}$ with learnable parameters.

To realize these ideas, GCN was proposed as the first-order approximation of the spectral GNN (Hammond et al., 2011; Kipf and Welling, 2016). Specifically, the authors of Kipf and Welling (2016) showed that the first order-approximation of the Chebyshev expansion of the spectral convolution operation can be implemented as the spatial domain convolution:

$$\boldsymbol{X}^{(k)} = \sigma\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}^{(k-1)}\boldsymbol{W}^{(k)}\right) \in \mathbb{R}^{N\times C^{(k)}}. \tag{7}$$

where $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$ is the adjacency matrix assuming the recurring loop, $\tilde{\boldsymbol{D}}$ is a degree matrix of $\tilde{\boldsymbol{A}}$, and

$$\boldsymbol{X}^{(k)} = \left[\boldsymbol{x}_1^{(k)} \ \cdots \ \boldsymbol{x}_{C^{(k)}}^{(k)}\right] \in \mathbb{R}^{N\times C^{(k)}} \tag{8}$$

denotes the $C^{(k)}$ channel signals at the k-th layer. This implies that GCN implements the node feature with its neighborhoods by mapping through a layer-specific learnable weight matrix $\boldsymbol{W}^{(k)}$ and non-linearity σ.

Unlike the spectral GNN, spatial-based methods define graph convolutions based on a node's spatial relations. More specifically, this operation is generally composed of the AGGREGATE, and COMBINE functions:

$$\boldsymbol{a}_v^{(k)} = \texttt{AGGREGATE}^{(k)}\left(\left\{\boldsymbol{p}_v^{(k-1)} : u \in \mathcal{N}(v)\right\}\right),$$

$$\boldsymbol{p}_v^{(k)} = \texttt{COMBINE}^{(k)}\left(\boldsymbol{p}_v^{(k-1)}, \boldsymbol{a}_v^{(k)}\right),$$

where $\boldsymbol{p}_v^{(k)} \in \mathbb{R}^{C^{(k)}}$ denotes the k-th layer feature vector at the v-th node. In other words, the AGGREGATE function collects features of the neighborhood nodes to extract aggregated feature vector $\boldsymbol{a}_v^{(k)}$ for the layer k, and COMBINE function then combines the previous node feature $\boldsymbol{p}_v^{(k-1)}$ with aggregated node features $\boldsymbol{a}_v^{(k)}$ to output the node feature of the current k-th layer $\boldsymbol{p}_v^{(k)}$. After this spatial operation, the mapping (5) is defined by

$$\boldsymbol{p}_G = \texttt{READOUT}\left(\left\{\boldsymbol{p}_v^{(k)} | v \in G\right\}\right).$$

Moreover, the AGGREGATE and COMBINE share the similar idea of information propagation/message passing on graphs (Wu et al., 2019).

In particular, GIN was proposed by Xu et al. (2018a) as a special case of spatial GNN suitable for graph classification tasks. The network implements the aggregate and combine functions as the sum of the node features:

$$\boldsymbol{p}_v^{(k)} = \texttt{MLP}^{(k)}\left((1+\epsilon^{(k)})\cdot\boldsymbol{p}_v^{(k-1)} + \sum\nolimits_{u\in\mathcal{N}(v)}\boldsymbol{p}_u^{(k-1)}\right) \in \mathbb{R}^{C^{(k)}}, \tag{9}$$

where $\epsilon^{(k)}$ is a learnable parameter, and MLP is a multi-layer perceptron with non-linearity. For graph-level readout, the embedded node features of every layers are summed up and then concatenated to obtain the final graph feature $\boldsymbol{p}_G$ as in (Xu et al., 2018a,b),

$$\boldsymbol{p}_G^{(k)} = \texttt{sum}(\boldsymbol{p}_0^{(k)}, \boldsymbol{p}_1^{(k)}, ..., \boldsymbol{p}_N^{(k)}) \tag{10}$$

$$\boldsymbol{p}_G = \texttt{concatenate}(\{\boldsymbol{p}_G^{(k)}\} | k = 0, 1, ..., K). \tag{11}$$

The authors of Xu et al. (2018a) argue that the proposed network architecture can learn injective mapping of the function g, which makes the model to be possibly as powerful as the WL test for graph classification tasks (Weisfeiler and Lehman, 1968; Shervashidze et al., 2011; Xu et al., 2018a).

2. THEORY

In this section, we mathematically show that the GIN is a dual representation of CNN on the graph space where the adjacency matrix is defined as a generalized shift operation. Along with this finding, we further propose a method for applying the GIN to the rs-fMRI data for graph classification and analysis.

2.1. GIN as a Generalized CNN on the Graph Space

Note that the GIN processing (9) can be decomposed as

$$\boldsymbol{p}_v^{(k)} = \texttt{MLP}^{(k)}(\boldsymbol{r}_v^{(k)}) \ \in \mathbb{R}^{C^{(k)}}, \quad v = 1, \cdots, N, \tag{12}$$

where

$$\boldsymbol{r}_v^{(k)} = c^{(k)}\boldsymbol{p}_v^{(k-1)} + \sum\nolimits_{u\in\mathcal{N}(v)}\boldsymbol{p}_u^{(k-1)} \tag{13}$$

$$= \underbrace{\left[\boldsymbol{p}_1^{(k-1)} \cdots \boldsymbol{p}_N^{(k-1)}\right]}_{\boldsymbol{P}^{(k-1)}} \left(c^{(k)}\boldsymbol{I} + \boldsymbol{A}\right)_{:,v} \in \mathbb{R}^{C^{(k-1)}} \tag{14}$$

where $c^{(k)} := 1 + \epsilon^{(k)}$ and $\boldsymbol{A}$ is the adjacency matrix and $\boldsymbol{M}_{:,v}$ denotes the v-th column of a matrix $\boldsymbol{M}$. This operation is performed for $k = 1, \cdots, K$.

One of the most important observations is that the feature matrix $\boldsymbol{P}^{(k-1)}$ is closely related to the signal matrix $\boldsymbol{X}^{(k-1)}$ in (8). More specifically, we have the following dual relationship:

$$\boldsymbol{X}^{(k-1)} = \boldsymbol{P}^{(k-1)\top} \tag{15}$$

Then, using the observation that $c^{(k)}\boldsymbol{I} + \boldsymbol{A}$ is self-adjoint, the matrix representation of (13) can be converted to a dual representation:

$$\boldsymbol{X}^{(k)} = \sigma\left((c^{(k)}\boldsymbol{I} + \boldsymbol{A})\boldsymbol{X}^{(k-1)}\boldsymbol{W}^{(k)}\right) \in \mathbb{R}^{N\times C^{(k)}} \tag{16}$$

where $\boldsymbol{W}^{(k)} \in \mathbb{R}^{C^{(k-1)}\times C^{(k)}}$ denotes the fully connected network weight from the MLP. Equation (16) shows that aside from the iteration dependent $\epsilon^{(k)}$, the main difference of GIN from GCN is the presence of the $(c^{(k)}\boldsymbol{I} + \boldsymbol{A})$ instead of the normalized adjacency matrix $\tilde{\boldsymbol{A}}$. This implies that GIN can be considered as an extension of the GCN as a first order approximation of the spectral GNN using the unnormalized graph Laplacian.

However, another important contribution of this paper is that the difference is not a minor variation, but that it implies an important difference between the two approaches. More specifically, by exploring the role of $c^{(k)}$ in (16), Theorem 1 shows that (16) is a dual representation of the two tab convolutional neural network without pooling layer on the graph spaces, where the adjacency matrix is defined as a shift operation.

Theorem 1. *The GIN iteration in (13) or (16) is a dual representation of a CNN without pooling layers using two-tab filter on the graph space, where the adjacency matrix* $\boldsymbol{A}$ *is defined as a shift operation.*

Proof: To understand this claim, we first revisit the classical CNN for the 1-D signal. A building block for the CNN is the following multi-channel convolution (Ye and Sung, 2019):

$$\boldsymbol{x}_i^{(k)} = \sigma\left(\boldsymbol{\Phi}^\top \sum_{j=1}^{C^{(k-1)}} \left(\boldsymbol{x}_j^{(k-1)} \circledast \boldsymbol{h}_{i,j}^{(k)}\right)\right) \tag{17}$$

where $C^{(k)}$ is the number of channels at the k-th layer, $\boldsymbol{x}_i^{(k)}$ denotes the i-th channel signal at the k-th layer, and $\boldsymbol{h}_{i,j}^{(k)}$ is the convolution filter that convolves with j-th input channel signal to produce i-th channel output. Finally, $\boldsymbol{\Phi}^\top$ denotes the matrix that represent the pooling operation.

Suppose that the convolution filter $\boldsymbol{h}_{i,j}^{(k)}$ has two tabs. Without loss of generality, the filter can be represented by

$$\boldsymbol{h}_{i,j}^{(k)} = \left[c^{(k)}w_{i,j}^{(k)} \;\; w_{i,j}^{(k)}\right]^\top \in \mathbb{R}^2$$

for some constant $c^{(k)}, w_{i,j}^{(k)}$. Then, the convolution operation can be simplified as

$$\boldsymbol{x}_j^{(k-1)} \circledast \boldsymbol{h}_{i,j}^{(k)} = c^{(k)}w_{i,j}^{(k)}\boldsymbol{x}_j^{(k-1)} + w_{i,j}^{(k)}\boldsymbol{S}\boldsymbol{x}_j^{(k-1)}$$

where $\boldsymbol{S}$ is the shift matrix defined by

$$\boldsymbol{S} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 0 \\ 0 & \cdots & \cdots & 1 & 0 \end{bmatrix} \tag{18}$$

if we assume the periodic boundary condition. Accordingly, for the cases of a CNN with no pooling layers, i.e., $\boldsymbol{\Phi}^\top = \boldsymbol{I}$, (17) with the two-tab filter can be represented in the following matrix form:

$$\boldsymbol{X}^{(k)} = \sigma\left(\left(c^{(k)}\boldsymbol{X}^{(k-1)} + \boldsymbol{S}\boldsymbol{X}^{(k-1)}\right)\boldsymbol{W}^{(k)}\right) \tag{19}$$

where

$$\boldsymbol{X}^{(k)} = \left[\boldsymbol{x}_1^{(k)} \cdots \boldsymbol{x}_{C^{(k)}}^{(k)}\right] \in \mathbb{R}^{N\times C^{(k)}}$$

$$\boldsymbol{W}^{(k)} = \begin{bmatrix} w_{1,1}^{(k)} & \cdots & w_{C^{(k)},1}^{(k)} \\ \vdots & \ddots & \vdots \\ w_{1,C^{(k-1)}}^{(k)} & \cdots & w_{C^{(k)},C^{(k-1)}}^{(k)} \end{bmatrix} \in \mathbb{R}^{C^{(k-1)}\times C^{(k)}}$$

By inspection of the dual representation of GIN in (16) and the CNN operation (19), we can see that the only difference of (16) is the adjacency matrix $\boldsymbol{A}$ instead of the shift matrix $\boldsymbol{S}$ in (19). Therefore, we can conclude that the GIN is a dual representation of CNN with two tab filter in the graph space where adjacency matrix is defined as a shift operation.

Note that the identification of the adjacency matrix as a generalized shift operation is not our own invention, but rather it is a classical observation in graph signal processing literature (Shuman et al., 2013; Huang et al., 2018; Ortega et al., 2018). Accordingly, Theorem 1 confirms that the insight from the classical signal processing plays an important role in understanding the GNN. Based on this understanding, we can now provide a dual space insight of the GIN operations in (10) and (11). More specifically, (10) can be understand as sum-pooling operation, since we have

$$\left(\boldsymbol{p}_G^{(k)}\right)^\top = \boldsymbol{\Phi}_{\text{sum}}^\top \boldsymbol{X}^{(k)}, \tag{20}$$

where the pooling matrix $\boldsymbol{\Phi}_{\text{sum}}^\top$ is given by

$$\boldsymbol{\Phi}_{\text{sum}}^\top = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}. \tag{21}$$

Then, (11) is indeed the multichannel concatenation layer from the pooled feature at each layer as shown in **Figure 1**. Therefore, the GIN operations can be understood as a dual representation of CNN classifier on the graph signal space where the shift operation is defined by the adjacency matrix. In fact, CNN and GIN differs in their definition of the shift operation as shown in **Figures 1**, **2**. We provide an exemplar GIN operation for a more expressive explanation in the **Figure 3** and **Supplementary Material**.

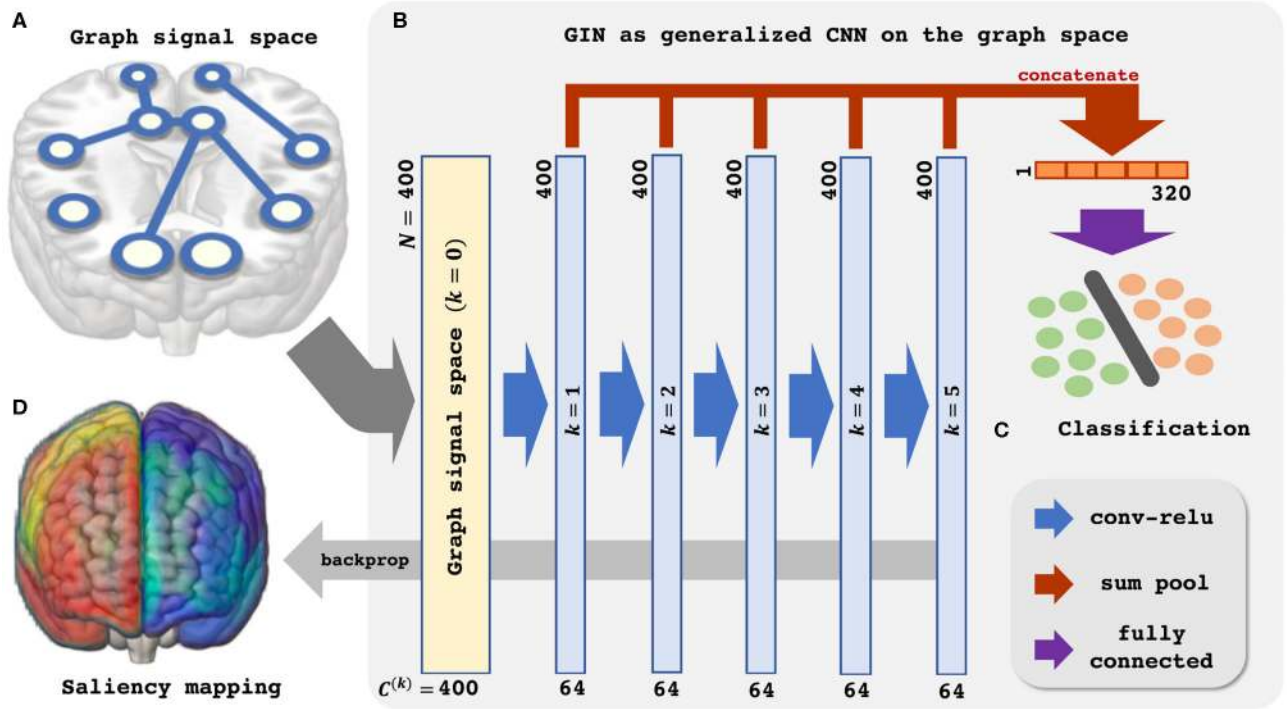

FIGURE 1 | Schematic illustration of the Graph Isomorphism Network based resting-state fMRI analysis. **(A)** Graph signal space. **(B)** GIN as generalized CNN on the graph space. **(C)** Classification. **(D)** Saliency mapping.

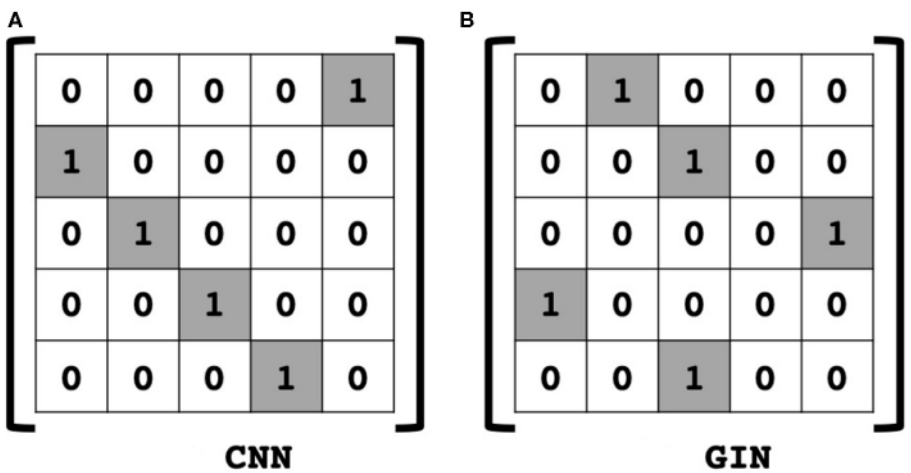

FIGURE 2 | Comparison of shift operation in **(A)** classical CNN, and **(B)** an example of GIN. In the graph space, the adjacency matrix is defined as shift operation.

2.2. Saliency Map of GIN

Thanks to the mathematical understanding of the similarity between the GIN and the CNN, we can now readily use the saliency map techniques for the CNNs to visualize important brain regions. For example, Arslan et al. (2018) used the CAM to visualize the graph saliency map. Instead, we propose to visualize the salient regions based on the Grad-CAM, which is a generalized version of the CAM without the restriction of the need of the global average pooling layer (Selvaraju et al., 2017). Specifically, the Grad-CAM saliency map at the k-th layer GIN can be calculated by

$$\mathcal{S}(k) = \sum_{j=1}^{N} \alpha_j^{(k)} \boldsymbol{x}_j^{(k)} \tag{22}$$

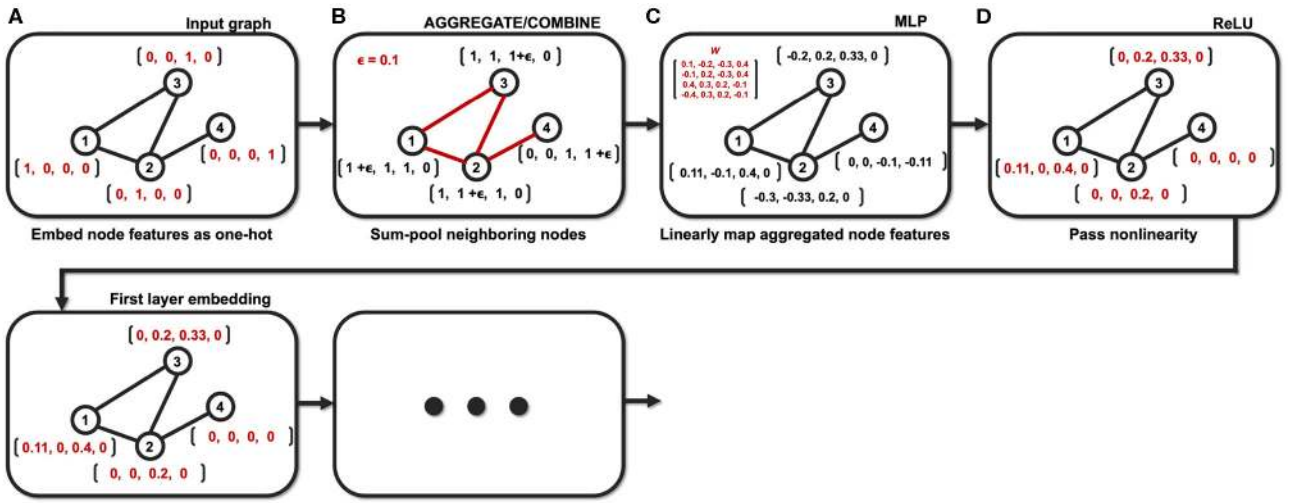

FIGURE 3 | Example of the GIN operation with a small graph ($N = 4$). **(A)** Node features are embedded as one-hot vectors. **(B)** Neighboring nodes are aggregated/combined. **(C)** Aggregated node features are mapped with learnable parameters. **(D)** Mapped node features are passed through nonlinear activation function.

where

$$\alpha_j^{(k)} = \sum_{i=1}^{N} \frac{\partial y}{\partial X_{ij}^{(k)}} \tag{23}$$

where $X_{ij}^{(k)}$ is the (i,j)-th element of $\boldsymbol{X}^{(k)}$ or i-th element of $\boldsymbol{x}_j^{(k)}$. Since we are interested in the input node contribution for the classification, we found that the meaningful Grad-CAM saliency map should be calculated at the input layer, i.e., $k = 0$, in which case the final representation becomes much simpler:

$$\begin{aligned} \mathcal{S}(0) &= \sum_{j=1}^{N} \alpha_j^{(0)} \boldsymbol{x}_j^{(0)} = \sum_{j=1}^{N} \alpha_j^{(0)} \boldsymbol{e}_j \\ &= \left[\sum_{i=1}^{N} \frac{\partial y}{\partial X_{i1}^{(0)}} \cdots \sum_{i=1}^{N} \frac{\partial y}{\partial X_{iN}^{(0)}} \right]^{\top} \in \mathbb{R}^N \end{aligned} \tag{24}$$

where the second equality comes from that $\boldsymbol{x}_j^{(0)}$ is one-hot vector, i.e., $\boldsymbol{x}_j^{(0)} = \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ has one at the j-the elements whereas all the other elements are zero, and the last equality comes from

$$\alpha_j^{(0)} = \sum_{i=1}^{N} \frac{\partial y}{\partial X_{ij}^{(0)}} \tag{25}$$

Note that in contrast to CAM (Zhou et al., 2016) as in Arslan et al. (2018) where sensitivity should be calculated with respect to the last layer, our approach using Grad-CAM provides a direct link from the input nodes to the final classification. Using experimental data, we will show that the resulting saliency map can quantify the sensitivity with respect to the node geometry, which provide a neuroscientific information about the relative importance of the each ROIs related to the class features.

3. MATERIALS AND METHODS

Based on the aforementioned understanding of the GIN, we proceed to apply the GIN to the rs-fMRI data for classification of the subjects' sex and provide neuroscientific interpretation. The **Figure 1** provides schematic illustration of the proposed analysis pipeline.

3.1. Data Description and Preprocessing

The rs-fMRI data was obtained from the Human Connectome Project (HCP) dataset S1200 release (Van Essen et al., 2013). The data was acquired for two runs of two resting-state session each for 15 min, with eyes open fixating on a cross-hair (TR = 720 ms, TE = 33.1 ms, flip angle = 52°, FOV = 208 × 180mm, slice thickness = 2.0mm). Of the total 4 runs, we used the first run of the dataset. Preprocessing of the fMRI volume time-series included gradient distortion correction, motion correction, and field map preprocessing, followed by registration to T1 weighted image. The registered EPI image was then normalized to the standard MNI152 space. Finally, FIX-ICA based denoising was applied to reduce non-neural source of noise in the data (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). Details of the HCP preprocessing pipeline is referred to Glasser et al. (2013).

From the preprocessed HCP dataset, rs-fMRI scans of 1,094 subjects were obtained from the project. To further minimize the unwanted effect of head motion on model training, we discarded the subject scans with framewise displacement (FD) over 0.3mm at any time of the scan. The FD was computed with `fsl_motion_outliers` function of the FSL (Jenkinson et al., 2012). There were 152 discarded scans from filtering out with the FD, and 942 scans were left. The 942 scans consisted of data from 531 female subjects and 411 male subjects. We paired each scan with the sex of the corresponding subject as an input-label for training the neural network.

3.2. Graph Construction From Preprocessed Data

The ROIs are defined from the cortical volume parcellation by Schaefer et al. (2017). We used the 400 parcellations as in Kashyap et al. (2019), Weis et al. (2019). Semantic region labels (e.g., Posterior cingulate cortex) and functional network labels (e.g., Default mode) corresponding to every parcels are provided with the dataset (Schaefer et al., 2017). Vertices are defined as one-hot vectors encoding the semantic region labels of the whole 400 ROIs. It can be said that no actual signal from the fMRI blood oxygen level dependency (BOLD) activity is represented in the vertex of the constructed graph.

To define the edges, functional connectivity matrix was constructed as follows. First, mean time-series of cortical parcels were obtained by averaging the preprocessed fMRI data voxels within each ROIs. Functional connectivity is defined as the correlation coefficient of the pearson's correlation between the time-series of the two voxels. Thus, the connectivity matrix is constructed by computing the pearson's correlation coefficient between every other ROIs. Derivation of the mean time-series and the connectivity matrix was performed with the MATLAB toolbox GRETNA (Wang et al., 2015). To derive an undirected, unweighted graph from the connectivity matrix, we threshold the connectivity matrix with sparsity by selecting the top M-percentile elements of the connectivity matrix as connected, and others unconnected.

3.3. Training Details

All following experiments are conducted with PyTorch 1.4.0. We used the GIN (Equation (9)) for our classification experiment. The concatenated graph features from all K layers $\boldsymbol{p}_G$ in (11) is mapped to the classifier output $\boldsymbol{y} = [y[1], \cdots, y[c]]^\top$ for predicting the one-hot vector encoded ground-truth label of the graph $\boldsymbol{y}_{\text{gt}} = [y_{\text{gt}}[1], \cdots, y_{\text{gt}}[c]]^\top$, where $y_{\text{gt}}[i] \in \{0, 1\}$ and c is a set of all possible class labels. Note that we omit the graph feature from the 0-th layer when concatenating since it is the same one-hot embedding of each pre-defined ROIs which have no difference between the subjects. One-dimensional batch normalization was applied after each layers of the network followed by the ReLU activation. The GIN is then trained to minimize the cross-entropy loss $\mathcal{L}_{\texttt{xent}}$:

$$\mathcal{L}_{\texttt{xent}} = -E\left[\sum_{i=1}^{c} y_{\text{gt}}[i] \cdot \log(y[i])\right] \tag{26}$$

where the expectation is taken over the training data. For the sex classification in this paper, the classifier is binary, so we use $c = 2$.

Deep Graph Infomax (DGI) was introduced in Veličković et al. (2018) as an unsupervised method for the representation learning of the graph. The DGI learns the node representation by maximizing the mutual information between the node feature vectors $\boldsymbol{p}_v$ and the corresponding graph feature $\boldsymbol{p}_G$. A discriminator $\mathcal{D}$ that takes a pair of a node feature vector and a graph feature as input is trained to discriminate whether the two embeddings are from the same graph:

$$\mathcal{L}_{\texttt{Infomax}} = \sum \log \mathcal{D}(\boldsymbol{p}_v, \boldsymbol{p}_G) + \sum \log(1 - \mathcal{D}(\tilde{\boldsymbol{p}}_v, \boldsymbol{p}_G)). \tag{27}$$

Here, $\tilde{\boldsymbol{p}}_v$ is a corrupted node feature vector, which is usually obtained by randomly selecting a node feature vector from another sample in the minibatch (Veličković et al., 2018). The DGI was first proposed as an unsupervised representation learning method, but (Li et al., 2019b) has made use of the DGI as a regularizer for the graph classification task.

Following the work by Li et al. (2019b), we added the DGI loss as a regularizer with the expectation that maximizing the mutual information between the node features and the graph features can help extract better representation of the graph. Thus, the final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\texttt{xent}} + \lambda \cdot \mathcal{L}_{\texttt{Infomax}}, \tag{28}$$

where $\mathcal{L}_{\texttt{xent}}$ is the cross entropy loss in (26) and $\mathcal{L}_{\texttt{Infomax}}$ is defined in (27), respectively. In this paper, we coin the term *Infomax regularization* indicating the regularizer $\mathcal{L}_{\texttt{Infomax}}$. To train the network, the Adam optimizer was used for 150 epochs of training with the learning rate of 0.01. Learning rate was decayed by the scale of 0.8 after every 5 epochs of training. We performed 10-fold cross-validation of the 942 graphs following (Varoquaux et al., 2017). The final model hyperparameters are reported in the section 4.1 based on the hyperparameter tuning experiments.

3.4. Comparative Study

To investigate the optimality of the proposed method, we performed comparative study with other methods. The first comparative study was performed to ensure the classification capability of our proposed method over other recent ones. Specifically, we re-implemented and evaluated the performance of the GCN-based method by Arslan et al. (2018) on our HCP dataset to serve as the baseline. Additionally, we compared the results of sex classification accuracy on the same HCP dataset reported by Zhang et al. (2018), Weis et al. (2019). Second comparative study was to find the optimal hyperparameter of our proposed method. We performed several hyperparameter tuning experiments which includes varying the level of sparsity, regularization coefficient λ, number of layers, number of hidden units, learning rate, and the dropout rate with the same dataset and the same GIN model. Lastly, we compared the classification performance when the input features were not encoded in one-hot vectors. Instead of embedding the input feature as a one-hot vector of each parcellation ROIs, we embedded the input features as mean BOLD activation of the ROI or its centroid coordinates (Ktena et al., 2017, 2018; Li et al., 2019a,b), and trained the proposed model with same model hyperparameters. The centroid coordinates are defined as a three-dimensional vector with each vector element representing the location of the axis R, A, and S. To exclude the possibility that the difference in classification performance comes from the first layer width of the model, we performed an additional experiment that the embedded centroid coordinate node features are first linearly mapped into the same dimension as in the one-hot encoded case, which is 400.

3.5. Saliency Mapping

The proposed saliency mapping was applied for visualizing the brain regions that are related to each class of sexes. We

computed the saliency map using (24) for each test subject. To obtain the group-level map, each subject-level saliency map was averaged across all subjects, and then was normalized to the range [0.0, 1.0]. Here, we specifically focus on the regions within the top 5-percentile values, which correspond to top 20 regions of the 400. To clarify the validity and advantages of our method, we compare the robustness and mapping results with the CAM-based saliency mapping method by Arslan et al. (2018). We evaluate how many top 5-percentile salient regions from only a subset of the subject groups match those from the whole group to demonstrate the robustness of the methods. Specifically, we compute the ratio of matching top salient regions between the maps derived by aggregating the full fold results and the maps derived from each fold of the cross-validation tests. Each cross validation fold consisted of around one tenth ($n = 95$ or $n = 94$) of the whole subjects ($n = 942$). The final robustness is calculated as the average of the matching ratios of the each 10-fold maps. Comparison of the full fold aggregated result and the five-fold aggregated result ($n = 470$ or $n = 472$) was additionally done.

4. RESULTS

4.1. Classification Results

The classification accuracy, precision, and recall are reported in **Table 1** along with other methods on the same first run of the HCP dataset. Highest accuracy of 84.61% was achieved by the proposed method, whereas the baseline GCN-based method achieved 83.98% accuracy. Other recent approaches with non GNN-based methods reported the classification performance lower than the baseline.

Results of the experiments to find the optimal hyperparameters of our method are as follows. We first compared the classification performance given the sparsity 5, 10, 15, 20, 30, 40% to find the optimal level of sparsity of the graph edges. The level of sparsity vs. classification accuracy was also tested with the GCN-based baseline method, which showed similar trend to the proposed method with slightly lower accuracy (**Figure 4**). The best performance was achieved with the sparsity 30%, so we report the results with sparsity 30% from here. Results of the other hyperaparameter tuning experiments, including the regularization coefficient λ, dropout rate, learning rate, number of layers, and number of hidden units in each layers are summarized in the **Table 2**. Based on theses hyperparameter experiments, the final GIN model was implemented 5 layers deep with 64 hidden units in each layers. Dropout was applied at the final linear layer with dropout rate of 0.5 during the training phase, and the regularization cofficient λ of (28) was set to 0.05.

The last comparative study was on classification performance of different node embeddings. It was found from the experiments that embedding the node feature as the centroid coordinate or the mean BOLD activity resulted in a significantly lower classification accuracy (**Table 3**). To evaluate the latent space of the model trained with differently embedded node features, we visualized the latent space of the model with the t-SNE (Maaten and Hinton, 2008), and computed the silhouette score between the two classes (Rousseeuw, 1987). The silhouette score represents how each subjects are well-clustered to its class in the latent space. The t-SNE visualization of the latent space in **Figure 5** was found to be more linearly separable when trained with one-hot embedded node features, while other embedding methods showed highly entangled latent space. The mean silhouette score of the test data across the 10-folds was 0.123 with the one-hot node features, while the BOLD mean, centroid coordinate, and the dimension matched centroid coordinate node features resulted in lower scores with 0.007, 0.014, 0.017, respectively.

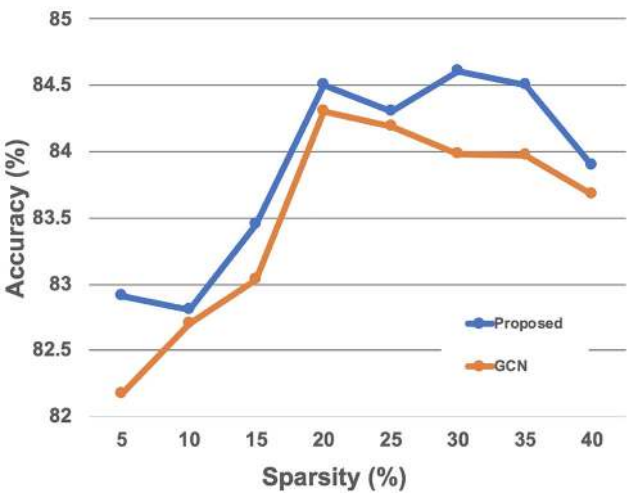

FIGURE 4 | Classification accuracy with respect to the edge sparsity.

TABLE 1 | Comparison of various methods for sex classification with the HCP dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	Subjects	Parcellation	Validation	Author	Year
GIN + Infomax	**84.61 ± 2.9**	86.19 ± 3.3	86.81 ± 4.9	942	Schaefer 400	10-fold	Ours	2020
GIN	84.41 ± 2.8	85.39 ± 2.6	87.60 ± 7.5	942	Schaefer 400	10-fold	Ours	2020
SVM-RBF	68.7 ± 2.6	-	-	434	Schaefer 400 + Fan 39	10-fold	Weis et al.	2019
SVM-RBF	64.3 ± 2.6	-	-	310	Schaefer 400 + Fan 39	Separate	Weis et al.	2019
GCN* (baseline)	83.98 ± 3.2	84.59 ± 3.1	87.78 ± 6.4	942	Schaefer 400	10-fold	Arslan et al.	2018
PLS	79.9 ± 0.9	-	-	820	Dosenbach 160	10-fold	Zhang et al.	2018

**Re-implemented to test for the HCP dataset. Bold value indicates the saliency with respect to the input (24).*

TABLE 2 | Hyperparameter tuning experiments.

Model	λ	Dropout	Learning rate	Layers	Hidden units	Accuracy (%)	Precision (%)	Recall (%)
GCN (Baseline)	None	0.5 (2,4,5 layer)	0.005	5	32/32/64/64/128	83.98 ± 3.2	84.59 ± 3.1	87.78 ± 6.4
GIN+Infomax	0.05	0.5	0.005	5	64	**84.61 ± 2.9**	86.19 ± 3.3	86.81 ± 4.9
GIN	0.0	-	-	-	-	84.41 ± 2.8	85.39 ± 2.6	87.60 ± 7.5
-	0.01	-	-	-	-	84.08 ± 2.2	86.72 ± 4.4	85.31 ± 5.5
-	0.1	-	-	-	-	84.51 ± 2.1	86.85 ± 4.5	86.06 ± 5.5
-	-	0.0	-	-	-	83.99 ± 3.4	85.78 ± 4.4	86.26 ± 6.1
-	-	-	0.01	-	-	83.13 ± 3.4	85.89 ± 3.4	84.01 ± 5.2
-	-	-	0.001	-	-	81.54 ± 3.3	85.45 ± 3.4	81.37 ± 7.3
-	-	-	-	4	-	83.11 ± 3.2	84.62 ± 2.8	85.70 ± 4.2
-	-	-	-	-	32	83.13 ± 3.4	85.20 ± 4.3	85.14 ± 5.5

Bold value indicates the saliency with respect to the input (24).

TABLE 3 | Comparison of different node feature embeddings.

Node feature	Accuracy (%)	Precision (%)	Recall (%)
One-hot	84.61 ± 2.9	86.19 ± 3.3	86.81 ± 4.9
BOLD mean	67.73 ± 2.9	69.90 ± 4.1	76.46 ± 8.3
Coordinate	72.19 ± 4.4	76.06 ± 6.8	75.88 ± 7.2
Coordinate*	70.90 ± 4.1	72.94 ± 4.9	78.33 ± 8.6

**Dimension matched to one-hot.*

4.2. Saliency Mapping

First, we demonstrate the robustness of the proposed saliency mapping method. Experiment on the robustness of the proposed method showed average of 63.5 and 65.5% top region match on one-fold aggregated saliency maps for female and male classes, respectively (**Table 4**). The robustness was higher for five-fold aggregated result as expected, showing 92.5 and 87.5% top region match. Significantly lower top region match with high standard deviation was found with the CAM-based saliency mapping method under same conditions. This was especially notable for the saliency maps of the female class, which showed 46.5% top region match on the one-fold aggregated maps and 70.0% top region match on the five-fold aggregated maps.

Plotted image and the full list of ROIs of the top 5-percentile salient regions from the proposed method are reported in the **Figure 6**, and the **Table 5**. The brain regions shown to be salient to the female class were the left prefrontal cortex (PFC), the right medial PFC, the right orbitofrontal cortex, the right cingulate cortex, the left frontal operculum, the left frontal eye field, the left temporal pole, the left temporal and parietal lobe regions, the bilateral visual cortex, and the bilateral somatomotor area. The functional networks that these brain regions comprise include all seven networks from the Yeo 7 networks (Thomas Yeo et al., 2011), which are the default mode network (DMN), the saliency/ventral attention network (SVN), the cognitive control network (CCN), the dorsal attention network (DAN), the limbic network (LN), the somatomotor network (SMN), and the visual network (VN). Among the seven networks, regions within the DMN was the most prominent taking up 30% of the 20 regions, followed by the SMN (25%), and the SVN (20%). Between the two hemispheres, salient regions were dominant in the left hemisphere (65%) when compared to the right hemisphere (35%).

For the male class, salient regions were the left PFC, the right medial and lateral PFC, the left orbitofrontal cortex, the bilateral posterior cingulate cortex (PCC), the right precuneus, the bilateral cingulate cortex, the left temporal pole, the right temporal lobe region, the right intraparietal sulcus, the right visual cortex, and the bilateral somatomotor area. The DMN was also predominant of all the functional networks as in the female class. While ratio of the dominant networks in the male class showed a similar trend to the female class, the left hemisphere dominance was not present as in the female class (See pie charts of the **Figure 6**).

Next, we explore the saliency mapping result from the CAM-based method (Arslan et al., 2018) and compare it with our method (**Figure 7**, **Table 6**). From the CAM-based methods, salient regions from both the female and the male class overlapped with our proposed method, including areas such as the PFC, the orbitofrontal cortex, the cingulate cortex, the PCC, the precuneus, and the temporal/parietal lobe regions. The most notable difference was the absence of the regions from the SMN and the VN in both classes. There were five functional networks that included the salient regions, the DMN, the SVN, the CCN, the DAN, and the LN. The dominance ratio of these five functional networks were similar to that found in our proposed saliency mapping results. In the male class, not only the regions from the SMN and the VN were missing, but also from the SVN, the DAN, and the LN. The only salient regions in the male class were the left PFC, the right medial/lateral PFC, the left PCC, the left precuneus, the temporal lobe and the parietal lobe regions from the DMN and the CCN. Hemisphere dominance showed a similar trend to the proposed method in that the female class clearly showed left hemisphere dominance (75%), while the male class did not show any hemisphere dominance (50%).

5. DISCUSSION

In this study, we proposed a framework for analyzing the fMRI data with the GIN. The framework suggests on first constructing the graph from the semantic region labels and

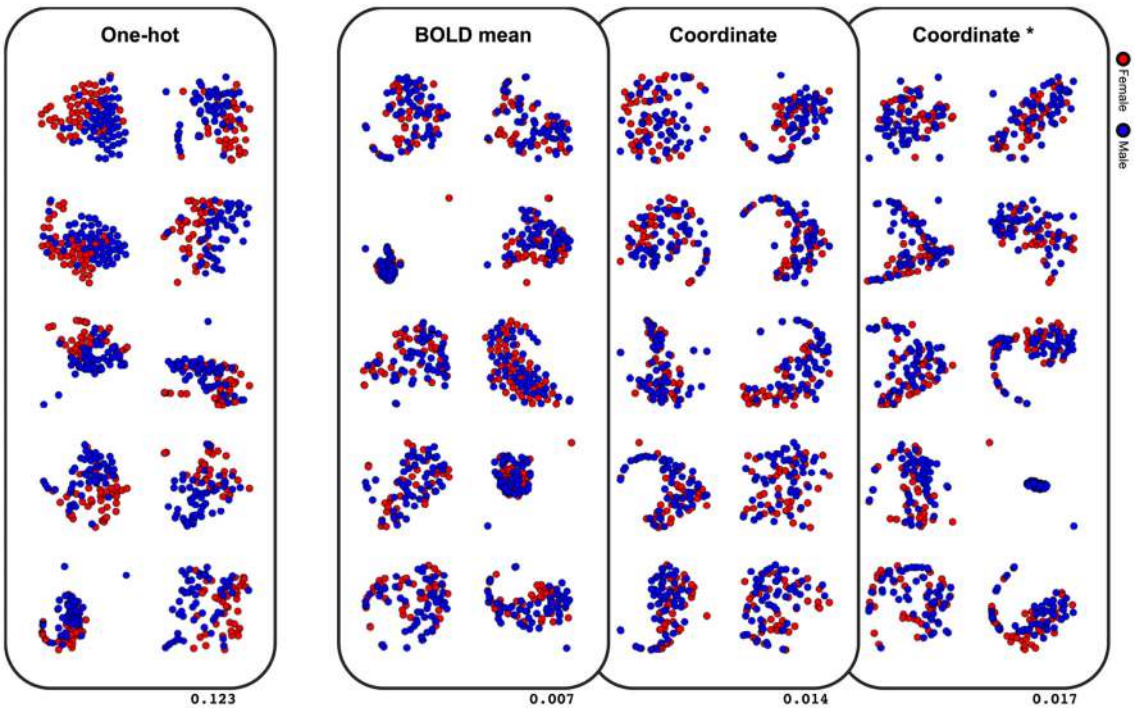

FIGURE 5 | Visualization of the latent space with t-SNE. Values on the lower right indicate the mean silhouette score of each embedding methods. Results of the 10-fold cross-validation are plotted in separate spaces with perplexity 50. Asterisk indicate that the dimension is matched to the one-hot embedding.

TABLE 4 | Robustness of the saliency mapping methods.

Method	Proposed		CAM	
	One-fold (%)	Five-folds (%)	One-fold (%)	Five-folds (%)
Female	63.5±6.7	92.5±2.5	46.5±16.9	70.0±5.0
Male	65.5±5.2	87.5±7.5	62.0±25.4	92.5±2.5

the functional connectivity between them. We train a GIN for classifying the subject phenotype based on the whole graph properties. After training, we can classify the subject with the trained GIN, or visualize the regions related to the classification by backpropagating through the trained GIN. An important theoretical basis that we found which underlie in this proposed method is that the GIN is not just a black-box operation that aggregates the graph structure with the MLP, but is actually a dual representation of a CNN on the graph space where the adjacency matrix is used as a generalized shift operator.

Classification of sex based on the rs-fMRI data resulted in the accuracy, precision, and recall of 84.61, 86.19, and 86.81%, respectively. The performance of the classifier is at least comparable, if not outperforming, to other recent methods for classifying sex based on the rs-fMRI data of the HCP dataset (Arslan et al., 2018; Zhang et al., 2018; Weis et al., 2019) (**Table 1**). Through the comparative studies, we have shown the validity of our proposed method that it can accurately classify the sex of the subjects with the rs-fMRI data. When training the GNN, adding the Infomax regularization had improved the classification performance of the GIN (**Table 2**). We have not gone through extensive experiment regarding the role of the Infomax regularization, but suggest to add it when training the neural network based on the results of our experiment. One interesting finding in our comparative experiments was that embedding the node feature as vectors of centroid coordinate or mean BOLD activity results in a significantly lower classification performance (**Table 3**). We expect that this comes from the linear dependence of the node features when embedded with centroid coordinate or mean BOLD activity. Further discussion regarding this topic is covered in the **Supplementary Material**.

After fully training the GIN for the sex classification task, we could map the salient regions related to the classification by the saliency mapping method. From the saliency mapping result, we could find that the regions within the DMN takes the most prominent role in classifying both the female and the male subjects. Importance of the DMN in the sex classification based on rs-fMRI data has been consistently reported (Zhang et al., 2018; Weis et al., 2019). In the study by Zhang et al. (2018), there were seven features involving the DMN of the top twenty important regions (35%) for sex classification, which is similar to our result (30% for the female class and 35% for the male class). This importance of the DMN for the sex classification task is known to be related to the difference of the DMN functional connectivity between the two sexes during the resting-state (Mak et al., 2017). Considering the difference of the DMN between the two sexes, it has been found consistently, and also from

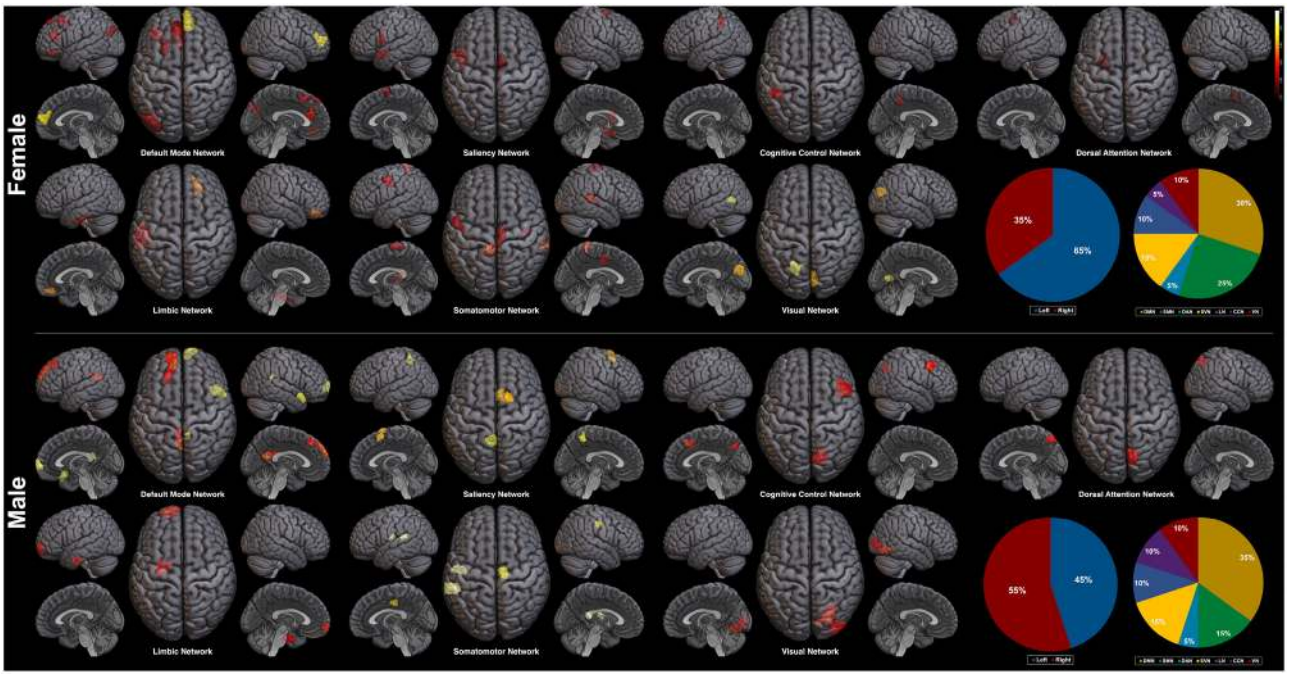

FIGURE 6 | Saliency mapping result of the proposed method. Top 20 salient regions are plotted with respect to the Yeo 7 networks (Thomas Yeo et al., 2011). The pie charts indicate the ratio of the two hemispheres and the ratio of each networks across the salient regions.

the meta-analysis, that the female individuals show stronger functional connectivity of the DMN compared to the males (Bluhm et al., 2008; Biswal et al., 2010; Allen et al., 2011; Mak et al., 2017; Zhang et al., 2018). CAM-based saliency mapping method also reflected this difference in the DMN between the two sexes and has shown predominance of the DMN in the saliency map, which is replicative of the original CAM-based saliency mapping study by Arslan et al. (2018). These findings suggest the validity of our saliency mapping method that it corresponds to the previous neuroimaging evidences regarding the importance of the DMN in sex classification.

Hemisphere related sex differences are also previously reported (Tian et al., 2011; Hjelmervik et al., 2014). The studies indicate that female subjects show higher functional connectivity in the left hemisphere, and male subjects in the right hemisphere (Tian et al., 2011). This difference in hemisphere dominance has shown the same trend in our experiment. In the female class, the salient regions in the left hemisphere outnumbered the salient regions in the right hemisphere (left 65% vs. right 35%), whereas the male class resulted in the right hemisphere lateralized saliency mapping result (left 45% vs. right 55%). The left hemisphere dominance of the female class was also found from the CAM-based saliency mapping results (left 75% vs. right 25%), but was not apparent in the male class (left 50% vs. right 50%). We interpret that the hemisphere related sex differences found in our saliency mapping result further supports the validity of our method.

Given the validity of the proposed saliency mapping method, the novel advantages of our method is highlighted by comparing it with the results from the CAM-based method. We find that the two major advantages over the CAM-based method are the robustness and the mapping sensitivity. The advantage in robustness is suggested from the experiment result that our proposed method captures more consistent top salient regions than the CAM-based method even with small number of subjects (**Table 4**). The other advantage, mapping sensitivity, is implied in the saliency mapping results. Mapping results from our method revealed the involvement of the regions within the SMN and the VN, while the CAM-based method was not able to identify them (**Figures 6**, **7**). There are some previous studies noting that there exist difference between the two sexes in terms of the functional connectivity within the SMN and the VN (Allen et al., 2011; Xu et al., 2015; Zhang et al., 2018). However, the evidences supporting this difference in the SMN and the VN are not as prominent and well-established as the difference in the DMN between the two sexes. It can be said that another supportive evidence of the difference of the SMN and the VN between the two sexes is added to the functional neuroimaging field by the proposed saliency mapping method, which would had not been identified by the CAM-based method. Based on this mapping sensitivity, applying the proposed method other types of classification tasks or to other subject groups is expected to provide new interesting findings to the neuroscientific field. To sum up, the proposed GIN based rs-fMRI analysis framework achieves state-of-the-art classification performance while providing a robust and sensitive saliency map which can be interpreted to add new insights to the field of functional neuroimaging.

TABLE 5 | Top 5-percentile salient regions identified by the proposed method for the female and the male class.

Female							*Male*						
Side	**Region**	**Network**	**R**	**A**	**S**	**Value**	**Side**	**Region**	**Network**	**R**	**A**	**S**	**Value**
L.	Somatomotor area	SMN	−8	−42	70	1.000	R.	Medial PFC	DMN	10	66	0	1.000
R.	Somatomotor area	SMN	64	−34	10	0.968	L.	Somatomotor area	SMN	−48	−12	14	0.986
L.	Visual cortex	VN	−18	−64	6	0.951	R.	PCC	DMN	8	−44	20	0.985
R.	Medial PFC	DMN	8	54	12	0.931	L.	Somatomotor area	SMN	−58	−36	16	0.976
R.	Visual cortex	VN	4	−80	24	0.909	R.	Cingulate cortex	SVN	6	10	58	0.973
R.	Orbitofrontal cortex	LN	20	42	−18	0.887	L.	PCC	DMN	−4	−54	20	0.960
L.	PFC	DMN	−6	34	20	0.863	R.	Temporal lobe	DMN	48	16	−20	0.951
L.	PFC	DMN	−22	20	52	0.835	L.	Cingulate cortex	SVN	−6	−48	56	0.949
L.	PFC	DMN	−36	36	−12	0.835	R.	Somatomotor area	SMN	12	−18	42	0.935
R.	Somatomotor area	SMN	6	−22	72	0.832	L.	PFC	DMN	−14	58	30	0.932
L.	Temporal lobe	DMN	−40	−78	30	0.821	R.	Cingulate cortex	SVN	16	6	70	0.908
L.	Parietal lobe	CCN	−44	−42	46	0.816	R.	Visual cortex	VN	24	−74	−10	0.899
L.	Somatomotor area	SMN	−52	−6	44	0.814	R.	Lateral PFC	CCN	44	18	44	0.881
L.	Frontal operculum	SVN	−52	8	14	0.806	L.	Orbitofrontal cortex	LN	−16	64	−8	0.881
L.	Frontal operculum	SVN	−44	6	−16	0.806	L.	PFC	DMN	−18	36	48	0.854
L.	Temporal pole	LN	−54	−22	−30	0.804	R.	Intraparietal sulcus	DAN	8	−72	52	0.843
L.	PFC	DMN	−8	42	52	0.799	L.	Temporal pole	LN	−26	−10	−32	0.835
R.	Somatomotor area	SMN	40	−20	4	0.784	R.	Visual cortex	VN	36	−88	2	0.835
R.	Cingulate cortex	SVN	6	−2	66	0.781	L.	PCC	DMN	−6	−40	24	0.834
L.	Frontal eye field	DAN	−26	0	56	0.777	R.	Precuneus	CCN	14	−72	40	0.834

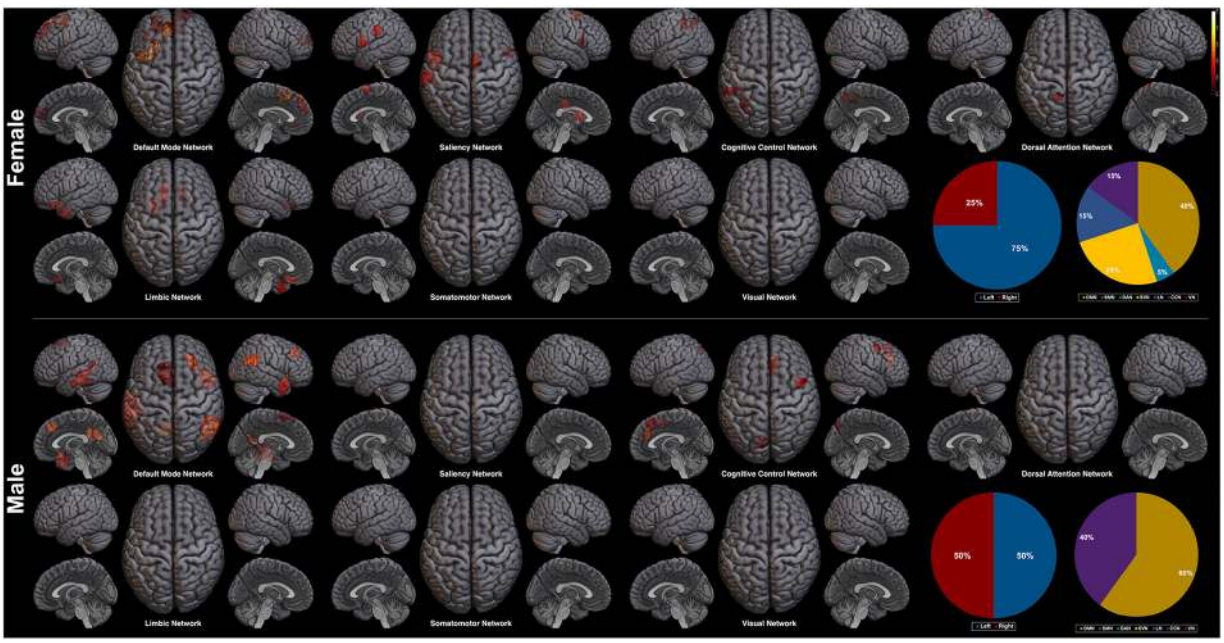

FIGURE 7 | Saliency mapping result of the CAM-based method. The pie charts indicate the ratio of the two hemispheres and the ratio of each networks across the salient regions.

There are some limitations and caveats that needs to be discussed. First, the demographics that can affect the analysis have not been considered or controlled thoroughly. It is well-known that the resting-state network can be affected by the age, handedness, fluid intelligence, and other subject characteristics. The results are expected to have stronger explainability by taking the demographics of the subjects into account in the analysis. Second, the cutoff threshold for

TABLE 6 | Top 5-percentile salient regions identified by the CAM-based method for the female and the male class.

Female							*Male*						
Side	**Region**	**Network**	**R**	**A**	**S**	**Value**	**Side**	**Region**	**Network**	**R**	**A**	**S**	**Value**
L.	PFC	DMN	−30	14	58	1.000	L.	PCC	DMN	−8	−52	10	1.000
L.	PFC	DMN	−8	42	52	0.965	R.	Parietal lobe	DMN	54	−46	20	0.983
L.	PFC	DMN	−42	8	48	0.964	L.	PCC	DMN	−14	−60	18	0.970
L.	PFC	DMN	−22	20	52	0.944	R.	Parietal lobe	DMN	48	−64	22	0.961
L.	Frontal operculum	SVN	−52	8	14	0.884	R.	Medial PFC	DMN	26	34	38	0.953
R.	Cingulate cortex	SVN	6	−2	66	0.880	R.	Parietal lobe	DMN	56	−46	32	0.941
L.	Orbitofrontal cortex	LN	−12	24	−20	0.870	R.	Medial PFC	CCN	8	34	24	0.922
L.	PFC	DMN	−22	50	32	0.852	R.	Parietal lobe	DMN	54	−54	26	0.920
L.	Parietal lobe	CCN	−58	−42	46	0.835	R.	Temporal lobe	DMN	48	16	−20	0.914
L.	Parietal lobe	SVN	−62	−24	32	0.819	L.	Temporal lobe	DMN	−60	−36	−18	0.865
L.	Frontal operculum	SVN	−50	2	4	0.812	L.	Temporal lobe	DMN	−62	−18	−20	0.865
L.	Temporal pole	LN	−24	6	−40	0.804	L.	Temporal lobe	DMN	−60	−34	−4	0.858
L.	Intraparietal sulcus	DAN	−14	−50	72	0.798	L.	Temporal lobe	DMN	−52	−22	−6	0.853
L.	Parietal lobe	CCN	−34	−62	48	0.796	R.	Lateral PFC	CCN	42	6	50	0.844
R.	Frontal operculum	SVN	54	12	12	0.784	R.	Temporal lobe	DMN	50	8	−32	0.816
R.	Orbitofrontal cortex	LN	14	24	−20	0.780	L.	Temporal lobe	DMN	−58	−48	16	0.811
L.	Parietal lobe	CCN	−44	−42	46	0.777	L.	PFC	DMN	−6	10	64	0.796
R.	Medial PFC	DMN	18	64	16	0.770	R.	Medial PFC	CCN	4	28	48	0.789
L.	PFC	DMN	−36	36	−12	0.770	L.	Precuneus	CCN	−10	−78	46	0.777
R.	Medial PFC	DMN	8	54	12	0.769	L.	PFC	DMN	−12	24	60	0.772

determining the salient region was heuristically set. We have set the regions with the top 5 percentile values as salient, but the method would have even more validity if the salient regions were determined in a more data-driven way, as in the classical methods perform statistical testing to determine the significance of each voxels. We have not gone through extensive study on the topic of determining the significant regions from the saliency map, but is worth further studies and discussion.

Still, we insist that analyzing the fMRI data based on the GIN has shown its theoretical and experimental validity in this study. We believe that the GIN based analysis method offers a potential advancement in the area, by opening a way to exploit the capability of the GIN to learn highly non-linear mappings. Some interesting topics related to this work can be considered. Theoretically, exploring the operations beyond the two-tab convolution filter by GIN can potentially provide better performance than the existing GIN. Neuroscientifically, extension of the method to clinical data interpretation or to the multi-class graph classification problem can be interesting topics in the future. With enough data assured, the proposed method is expected to help reveal new findings from the functional networks of the brain.

DATA AVAILABILITY STATEMENT

The data analyzed for this study can be found here: https://db.humanconnectome.org/.

AUTHOR CONTRIBUTIONS

B-HK designed and conducted the experiments, interpreted the neuroscientific findings, and wrote the manuscript. JY supervised the experiments, deduced the theoretical findings, and wrote the manuscript.

FUNDING

This work was supported by the National Research Foundation (NRF) of Korea, Grant number NRF-2020R1A2B5B03001980. This work was supported by the Industrial Strategic Technology Development Program (10072064, Development of Novel Artificial Intelligence Technologies to Assist Imaging Diagnosis of Pulmonary, Hepatic, and Cardiac Diseases and their Integration into Commercial Clinical PACS Platforms) funded by the Ministry of Trade Industry and Energy (MI, Korea).

ACKNOWLEDGMENTS

This manuscript has been released as a preprint at arXiv (Kim and Ye, 2020). The authors would like to thank Sangmin Lee for his useful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00630/full#supplementary-material

REFERENCES

Allen, E. A., Erhardt, E. B., Damaraju, E., Gruner, W., Segall, J. M., Silva, R. F., et al. (2011). A baseline for the multivariate comparison of resting-state networks. *Front. Syst. Neurosci.* 5:2. doi: 10.3389/fnsys.2011.00002

Arslan, S., Ktena, S. I., Glocker, B., and Rueckert, D. (2018). "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* (Granada: Springer), 3–13. doi: 10.1007/978-3-030-00689-1_1

Bassett, D. S., and Bullmore, E. T. (2009). Human brain networks in health and disease. *Curr. Opin. Neurol.* 22:340. doi: 10.1097/WCO.0b013e32832d93dd

Bassett, D. S., and Sporns, O. (2017). Network neuroscience. *Nat. Neurosci.* 20:353. doi: 10.1038/nn.4502

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107

Bluhm, R. L., Osuch, E. A., Lanius, R. A., Boksman, K., Neufeld, R. W., Théberge, J., et al. (2008). Default mode network connectivity: effects of age, sex, and analytic approach. *Neuroreport* 19, 887–891. doi: 10.1097/WNR.0b013e328300ebbf

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Duffy, B. A., Liu, M., Flynn, T., Toga, A., Barkovich, A. J., Xu, D., et al. (2019). "Regression activation mapping on the cortical surface using graph convolutional networks," in *International Conference on Medical Imaging with Deep Learning-Extended Abstract Track* (London).

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning* (Sydney), Vol. 70, 1263–1272.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., et al. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage* 95, 232–247. doi: 10.1016/j.neuroimage.2014.03.034

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.* 30, 129–150. doi: 10.1016/j.acha.2010.04.005

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., et al. (2019). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* 206:116276. doi: 10.1016/j.neuroimage.2019.116276

He, Y., and Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Curr. Opin. Neurol.* 23, 341–350. doi: 10.1097/WCO.0b013e32833aa567

Hjelmervik, H., Hausmann, M., Osnes, B., Westerhausen, R., and Specht, K. (2014). Resting states are resting traits-an fMRI study of sex differences and menstrual cycle effects in resting state cognitive control networks. *PLoS ONE* 9:e103492. doi: 10.1371/journal.pone.0103492

Huang, W., Bolton, T. A., Medaglia, J. D., Bassett, D. S., Ribeiro, A., and Van De Ville, D. (2018). A graph signal processing perspective on functional brain imaging. *Proc. IEEE* 106, 868–885. doi: 10.1109/JPROC.2018.2798928

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Kashyap, R., Kong, R., Bhattacharjee, S., Li, J., Zhou, J., and Yeo, B. T. (2019). Individual-specific fmri-subspaces improve functional connectivity prediction of behavior. *Neuroimage* 189, 804–812. doi: 10.1016/j.neuroimage.2019.01.069

Kim, B.-H., and Ye, J. C. (2020). Understanding graph isomorphism network for brain mr functional connectivity analysis. *arXiv preprint arXiv:2001.03690*.

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., et al. (2017). "Distance metric learning using graph convolutional networks: application to functional brain networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC: Springer), 469–477. doi: 10.1007/978-3-319-66182-7_54

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., et al. (2018). Metric learning with spectral graph convolutions on brain connectivity networks. *Neuroimage* 169, 431–442. doi: 10.1016/j.neuroimage.2017.12.052

Li, X., Dvornek, N. C., Zhou, Y., Zhuang, J., Ventola, P., and Duncan, J. S. (2019a). "Graph neural network for interpreting task-fMRI biomarkers," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen: Springer), 485–493. doi: 10.1007/978-3-030-32254-0_54

Li, X., Dvornek, N. C., Zhuang, J., Ventola, P., and Duncana, J. (2019b). Graph embedding using infomax for ASD classification and brain functional difference detection. *arXiv preprint arXiv:1908.04769*. doi: 10.1117/12.2549451

Ma, G., Ahmed, N. K., Willke, T., Sengupta, D., Cole, M. W., Turk-Browne, N., et al. (2018). Similarity learning with higher-order proximity for brain network analysis. *arXiv preprint arXiv:1811.02662*.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Mak, L. E., Minuzzi, L., MacQueen, G., Hall, G., Kennedy, S. H., and Milev, R. (2017). The default mode network in healthy individuals: a systematic review and meta-analysis. *Brain Connect.* 7, 25–33. doi: 10.1089/brain.2016.0438

Micheloyannis, S., Pachou, E., Stam, C. J., Breakspear, M., Bitsios, P., Vourkas, M., et al. (2006). Small-world networks and disturbed functional connectivity in schizophrenia. *Schizophrenia Res.* 87, 60–66. doi: 10.1016/j.schres.2006.06.028

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., and Vandergheynst, P. (2018). Graph signal processing: overview, challenges, and applications. *Proc. IEEE* 106, 808–828. doi: 10.1109/JPROC.2018.2820126

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., et al. (2018). Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med. Image Anal.* 48, 117–130. doi: 10.1016/j.media.2018.06.001

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Moreno, R. G., Glocker, B., et al. (2017). "Spectral graph convolutions for population-based disease prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC: Springer), 177–185. doi: 10.1007/978-3-319-66179-7_21

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* 90, 449–468. doi: 10.1016/j.neuroimage.2013.11.046

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., et al. (2017). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* 28, 3095–3114. doi: 10.1093/cercor/bhx179

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 618–626. doi: 10.1109/ICCV.2017.74

Shervashidze, N., Schweitzer, P., Leeuwen, E. J. V., Mehlhorn, K., and Borgwardt, K. M. (2011). Weisfeiler-Lehman graph kernels. *J. Mach. Learn. Res.* 12, 2539–2561.

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* 30, 83–98. doi: 10.1109/MSP.2012.2235192

Sporns, O. (2018). Graph theory methods: applications in brain networks. *Dialog. Clin. Neurosci.* 20:111.

Thomas Yeo, B., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106, 1125–1165. doi: 10.1152/jn.00338.2011

Tian, L., Wang, J., Yan, C., and He, Y. (2011). Hemisphere-and gender-related differences in small-world brain networks: a resting-state functional MRI study. *Neuroimage* 54, 191–202. doi: 10.1016/j.neuroimage.2010.07.066

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The Wu-Minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 145, 166–179. doi: 10.1016/j.neuroimage.2016.10.038

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2018). Deep graph infomax. *arXiv preprint arXiv:1809.10341*.

Wang, J., Wang, X., Xia, M., Liao, X., Evans, A., and He, Y. (2015). Gretna: a graph theoretical network analysis toolbox for imaging connectomics. *Front. Hum. Neurosci*. 9:386. doi: 10.3389/fnhum.2015.00458

Wang, J., Zuo, X., and He, Y. (2010). Graph-based network analysis of resting-state functional MRI. *Front. Syst. Neurosci*. 4:16. doi: 10.3389/fnsys.2010.00016

Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T. T., and Eickhoff, S. B. (2019). Sex classification by resting state brain connectivity. *Cereb. Cortex* 30, 824–835. doi: 10.1093/cercor/bhz129

Weisfeiler, B., and Lehman, A. A. (1968). A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno Technicheskaya Informatsia* 2, 12–16.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*. doi: 10.1109/TNNLS.2020.2978386

Xu, C., Li, C., Wu, H., Wu, Y., Hu, S., Zhu, Y., et al. (2015). Gender differences in cerebral regional homogeneity of adult healthy volunteers: a resting-state fMRI study. *BioMed Res. Int*. 2015. doi: 10.1155/2015/183074

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018a). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-I., and Jegelka, S. (2018b). Representation learning on graphs with jumping knowledge networks. *arXiv preprint arXiv:1806.03536*.

Ye, J. C., and Sung, W. K. (2019). "Understanding geometry of encoder-decoder CNNs," in *International Conference on Machine Learning* (Long Beach, CA), 7064–7073.

Zhang, C., Dougherty, C. C., Baum, S. A., White, T., and Michael, A. M. (2018). Functional connectivity predicts gender: evidence for gender differences in resting brain connectivity. *Hum. Brain Mapp*. 39, 1765–1776. doi: 10.1002/hbm.23950

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2921–2929. doi: 10.1109/CVPR.2016.319

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.