Graph Autoencoders for Embedding Learning in Brain Networks and Major Depressive Disorder Identification

Fuad Noman‡, Senior Member, IEEE, Chee-Ming Ting ‡, Senior Member, IEEE, Hakmook Kang ‡, Raphaël C.-W. Phan ‡, Senior Member, IEEE, and Hernando Ombao ‡

Abstract—Brain functional connectivity (FC) networks inferred from functional magnetic resonance imaging (fMRI) have shown altered or aberrant brain functional connectivity in various neuropsychiatric disorders. Recent application of deep neural networks to connectome-based classification mostly relies on traditional convolutional neural networks (CNNs) using input FCs on a regular Euclidean grid to learn spatial maps of brain networks neglecting the topological information of the brain networks, leading to potentially sub-optimal performance in brain disorder identification. We propose a novel graph deep learning framework that leverages non-Euclidean information inherent in the graph structure for classifying brain networks in major depressive disorder (MDD). We introduce a novel graph autoencoder (GAE) architecture, built upon graph convolutional networks (GCNs), to embed the topological structure and node content of large fMRI networks into low-dimensional representations. For constructing the brain networks, we employ the Ledoit-Wolf (LDW) shrinkage approach to efficiently estimate high-dimensional FC metrics from fMRI data. We explore both supervised and unsupervised techniques for graph embedding learning. The resulting embeddings serve as feature inputs for a deep fully-connected neural network (FCNN) to distinguish MDD from healthy controls (HCs). Evaluating our model on resting-state fMRI MDD dataset, we observe that the GAE-FCNN outperforms several state-of-the-art methods for brain connectome classification, achieving the highest accuracy when using LDW-FC edges as node features. The graph embeddings of fMRI FC networks also reveal significant group differences between MDD and HCs. Our framework demonstrates the feasibility of learning graph embeddings from brain networks, providing valuable discriminative information for diagnosing brain disorders.

Index Terms—Brain connectivity networks, graph autoencoder, graph convolutional network, major depressive disorder, resting-state fMRI.

I. INTRODUCTION

Analysis of brain functional connectivity (FC) networks inferred from functional magnetic resonance imaging (fMRI) data has become an important method to probe large-scale functional organization of the human brain in health and disease [1]. Considerable evidence from rs-fMRI studies have shown altered or aberrant brain functional connectome in various neuropsychiatric and neurodegenerative disorders [2], e.g., schizophrenia [3], autism spectrum disorder (ASD) [4], Alzheimer’s disease (AD) [5], suggesting potential use of network-based biomarkers for clinical diagnostics [6]. Functional abnormalities are detected not only in the strengths of individual connections but also topological structure of resting-state FC networks [1]. The brain function in major depressive disorder (MDD) — the most prevalent psychiatric disorder with pervasive depressed mood, cognitive inability and suicidal tendency, has been a subject of intensive studies recently. It is increasingly understood as a network-based disorder with consistent alternations in FC patterns [7]. Disrupted resting-state FC from fMRI has been found in MDD core networks, such as the default mode network (DMN) related to self-referential processing and emotion regulation, central executive network (CEN) for attention and working memory, and other subcortical circuits [8]. Increased connectivity within DMN [9] and decreased connectivity between DMN and CEN have been observed in MDD patients compared to healthy controls (HCs) [7]. Graph theoretical analyses of rs-fMRI also revealed altered network topological properties in MDD, e.g., enhanced global efficiency [10] and high local efficiency and modularity [11].

Machine learning techniques have been increasingly used in turning altered brain FC into biomarkers for fast and automated classification of brain disorders [12]. A vast majority of studies use traditional machine learning algorithms for classification, such as support vector machine (SVM), logistic regression and linear discriminant analysis (For review, see [13], [14], [15]). Compared to other disorders, functional connectome-based classification of MDD is relatively unexplored. Several recent studies [16], [17], [18] have employed SVMs combined with some
ad-hoc feature selection methods to differentiate MDD from HCs using rs-fMRI FC, and obtained reasonable classification accuracies on leave-one-subject-out cross-validation.

Deep learning methods have received significant interest in fMRI-based classification of brain disorders [19]. In recent applications to connectome-based classification, it has shown great potential providing substantial gain in performance over traditional classifiers. Deep neural networks (DNNs) can automatically learn a hierarchy of representations directly from the connectome data, without relying on preliminary feature handcrafting and selection. Fully-connected DNNs have been used as autoencoders (AE) to map high-dimensional input vectors of FC metrics to latent compact representations for rs-fMRI classification of ASD [20] and schizophrenia [21]. Inspired by remarkable success in image and object classification, deep convolutional neural networks (CNNs) have also been used to learn spatial maps of brain functional networks. A CNN architecture (BrainNetCNN) with specially-designed convolutional filters for modeling connectome data was introduced by [22] for predicting neurodevelopment in infants. Various variants of connectome CNNs were subsequently proposed for FC classification. These include one-dimensional (1D) spatial convolutional filters on rs-fMRI FC data for mild cognitive impairment (MCI) identification [23], 2D-CNNs for FC matrices for ASD classification [24], 3D CNNs to combine static and dynamic FC for early MCI detection [25], and multi-domain connectome CNN to integrate different brain network measures [26]. The deep learning models discussed above often neglect the intricate topological information inherent in brain networks which may potentially lead to sub-optimal performance in the identification and analysis of brain disorders. Notably, the process of flattening input FC maps in fully-connected DNNs destroys their spatial structure. Similarly, the use of fixed 1D or 2D regular grid convolution operators in CNNs proves inadequate for effectively capturing the inherent graph-based structure present in connectome data. Brain networks typically exhibit irregular structure with nodes being unordered and connected to a different number of neighbors, which renders convolution operations for regular grid inappropriate for modeling graphs. Thus, preserving the topological structure of brain functional networks remains critical for the analysis of functional irregularities.

Extending deep learning approaches to non-Euclidean data domains, particularly graph structures, is a rapidly evolving field of research [27]. Within this domain, a well-established graph-based neural network (GNN) architecture extends the operations commonly found in CNNs to facilitate the learning of local and global structural patterns within irregular graphs. A spectral-based graph convolutional network (GCN) has been introduced to perform convolutions in the spatial domain of graphs as multiplications in the graph spectral domain [28], [29]. The application of spectral GCNs to the realm of brain disorder detection using brain functional networks is a recent development that remains in its nascent stages, e.g., for predicting ASD and conversion from MCI to AD [30], [31], [32], [33]. These studies used the population graph as input to the GCN. In this configuration, individual nodes represent subjects, with associated resting-state FC feature vectors, while phenotype information is encoded as edge weights within the graph structure. This approach inherently relies on non-imaging data for graph construction and necessitates prior knowledge of relevant phenotype information for specific disorders. Furthermore, it is semi-supervised learning, using all subjects (both training and testing sets), thereby limiting its generalization capability to unseen subjects. A recent benchmarking study [34] also showed that population-based spectral GCN is less effective than the BrainNetCNN in resting-state FC-based behavioral prediction.

In this paper, we propose a novel framework based on deep GNN for graph embedding on brain functional networks for classifying neuropsychiatric disorders associated with functional dysconnectivity. Precisely, we develop a graph autoencoder (GAE) architecture that leverages GCN to encode the non-Euclidean information about brain connectome into low-dimensional latent representations (or network embeddings), on which a decoder is trained to reconstruct the graph structure. The learned embeddings allow dimensionality reduction of large-sized brain network data, and preserve both the network topological structure and node content information as discriminative features to enhance subsequent connectome-based classification. The extracted patterns by the multiple graph convolutional layers in GCNs can include high-level representations of nodes’ local graph neighborhood. We utilize the GAE in an inductive framework of embedding generation for network-level classification. In contrast to the GCN used in transductive settings in existing GAEs for a single fixed graph [29], [33], [35], our GAE is designed to generate node embeddings for completely unseen graphs. By learning an embedding function that is shared across networks from different subjects, one of its advantages is that it can be readily generalized to multiple brain networks of unseen subjects in the downstream brain network classification. In addition to the unsupervised embedding learning using GAE, we also consider supervised learning where the model makes use of disorder class labels to optimize the embeddings. Finally, a readout layer is added to summarize the node representations of each graph into a graph representation, which is then used as feature inputs to a fully-connected DNN (FCNN) for network classification.

We validate the utility of our proposed GAE-FCNN to rs-fMRI data for classification of MDD and HCs using whole-brain FC networks. The GAE-FCNN is trained on high-dimensional functional networks constructed from rs-fMRI using the Ledoit-Wolf (LDW) covariance estimator [36]. We also explore different types of node features: fMRI time series, associated FC edges and local graph measures. Our primary research questions revolve around the effectiveness and advantages of the proposed approach. We seek to answer how effective the inductive GAE framework is in classifying neuropsychiatric disorders, especially in the context of MDD. Furthermore, we aim to elucidate the unique contributions of the GAE architecture, which integrates GCNs for encoding complex network structures, and its potential for encoding irregular topological features in brain networks.

The main contributions of this work are summarized as follows:

1) We propose, for the first time, a graph deep learning framework for brain FC-based identification of MDD.
2) The proposed GAE-FCNN framework offers a novel approach to directly leverage on the alterations in network structure for brain disorder classification via the learned network embeddings. The GCN-based GAE architecture provides a purely unsupervised way to learn embeddings that encode the irregular topological structure of brain networks, which are inadequately modeled by the connectome CNNs and the vectorized FC features in population.
The proposed GAE-FCNN framework for functional brain network classification consists of two components: an unsupervised and a supervised model. The unsupervised model employs a GCN-based encoder to encode fMRI connectome data (graph structure $A$ & node content $X$) into latent representations $Z$ on which a decoder is used to reconstruct the graph information. A deep FCNN performs network-level classification to discriminate MDD patients and HCs based on the learned representations. The supervised model combines the GCN encoder with a deep DNN to perform graph-level classification and predict class labels for the entire brain graph, rather than node/subject-level classification based on population graphs.

3) We demonstrate that our approach outperforms both the BrainNetCNN and population-based GCN by a large margin in identifying MDD based on resting-state functional brain networks from fMRI.

4) We show that high-order network reconstructed from nodes embeddings learned by the proposed GCN-based GAE can reveal differences in network organization between MDD and HCs related to emotion processing.

II. RS-FMRI DATASET FOR MDD

A. Subjects & Data Acquisition

To demonstrate the proposed GAE-FCNN framework, we analyzed a rs-fMRI data from the open-access REST-meta-MDD Consortium database [37] for evaluation. We considered the largest dataset from site 20, consisting of a total 477 subjects (250 MDD and 227 HC) recruited from Southwest University, China. The rs-fMRI scans were acquired using Siemens scanner with an echo-planar imaging sequence (TR/TE = 2000/30 ms; flip angle = 90°; thickness/gap = 3.0/1.0 mm; time points = 242; field of view = 220 mm; voxel size = $3.44 \times 3.44 \times 4.00$; matrix size = $61 \times 73 \times 61$).

B. Preprocessing

The data were preprocessed using the Data Processing Assistant for Resting-State fMRI (DPARSF) [38], following steps in [37]. Standard preprocessing steps were applied, including, removal of the first ten volumes, followed by slice timing correction, head motion correction, regression of nuisance covariates of head motion parameters, white matter, and cerebrospinal fluid (CSF). The fMRI data are then normalized with an EPI template in the MNI space, and resampled to $3 \times 3 \times 3$ mm$^3$ resolution, followed by spatial smoothing using a 6 mm full-width half-maximum Gaussian kernel, and temporal bandpass filtering (0.01-0.1 Hz).

III. METHODS

An overview of the proposed GAE-FCNN framework for identifying brain disorders using fMRI-based FC networks is shown in Fig. 1 with the three stages: (1) Network construction.
High-dimensional FC networks are constructed from fMRI data using LDW shrinkage covariance estimator, and associated node features are extracted. (2) Network embedding via a GCN-based GAE. The GAE learns network embeddings by using an encoder of stacked GCNs to map the input graph structure and node content of FC networks into latent representation (or embeddings), and using an inner-product decoder to enforce embeddings to preserve graph topological information. (3) Network classification. The learned network embeddings are then used as inputs to a fully-connected DNN to discriminate between MDD patients and HCs. We develop an unsupervised (Fig. 1(a)) and a supervised (Fig. 1(b)) framework for learning graph embeddings in brain networks.

A. Connectivity Network Construction

We consider an undirected graph of brain functional network for each subject, represented by $G \equiv \{V, E\}$ where $V \equiv \{v_1, \ldots , v_N\}$ is a set of $N$ nodes (voxels or ROIs) and $e_{ij} \in E$ denotes the connectivity edge $(i,j)$ between nodes $v_i$ and $v_j$. The topological structure of the graph $G$ can be represented by an adjacency matrix $A = [a_{ij}] \in \{0, 1\}^{N \times N}$, where $a_{ij} = 1$ if nodes $v_i$ and $v_j$ are connected, otherwise $a_{ij} = 0$. We denote by $X = [x_1, \ldots , x_N] \in \mathbb{R}^{N \times d}$ the node feature matrix for $G$, with $x_i \in \mathbb{R}^d$ representing the content feature vector associated with each node $v_i$.

1) Network Connectivity: In constructing FC networks, we compute the graph matrix based on the temporal correlations of fMRI time series between pairs of ROIs. Let $y_t \in \mathbb{R}^N$, $t=1, \ldots , T$ be the fMRI time series of length $T$ measured from the $N$ ROIs. For large-sized fMRI-derived networks in which the number of nodes $N$ is large or comparable to the number of scans $T$, traditional sample correlation matrix is no longer a reliable and accurate estimator of FC. This is due to large number of correlation coefficients (i.e., $N(N-1)/2$) to be estimated relative to the sample size. This condition applies to the MDD fMRI data considered here ($T=232$ and $N=116$ ROIs). To estimate functional connectome efficiently, we use the Ledoit-Wolf (LDW) regularized shrinkage estimator [36], [39], [40] that can yield well-conditioned FC estimates in high-dimensional settings when the ratio of $N/T$ is large. The LDW covariance estimator is defined by $\Sigma = (1-\alpha) \Sigma + \alpha \Delta$ with $\Delta = (Tr(\Sigma)/N)I_N$, where $\alpha$ is a shrinkage parameter, $I_N$ is a $N \times N$ identity matrix, $Tr(\cdot)$ is the trace and $\Sigma = \frac{1}{T} \sum_{t=1}^{T} (y_t - \bar{y})(y_t - \bar{y})^T$ is the $N \times N$ sample covariance matrix with sample mean $\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$. The shrinkage coefficient $\alpha$ can be estimated data-adaptively [41]. The correlation matrix is then computed as $D = \text{Diag}(\Sigma)$.

We can generate the adjacency matrix $A$ by thresholding the correlation matrix $R$. We used the k-nearest neighbor graph (k-NNG), which applies a local threshold to FC matrix, selecting the $k$-strongest edges of each node in the network. This approach will result in a fixed density of edges in graphs across all subjects, and thus enabling meaningful comparison of network topology between different groups and conditions. It can also generate more stable network metrics compared to the absolute thresholding [42]. It has been shown that the setting of k-NNG threshold $k$ has a significant impact on the overall performance of the network classification model [32]. Besides, when $k$ decreases, networks become sparser and may lead to the zero-degree nodes (isolated nodes totally disconnected from the rest of the graph).

2) Node Features: We consider three types of node features for $X$: (1) $T \times 1$ raw rs-fMRI time series associated with each node which can capture spontaneous fluctuations in the BOLD signal in individual brain regions. (2) $N \times 1$ FC weights of edges connected to each node, i.e., each column of the LDW-estimated correlation matrix. (3) Graph-theoretic measures to characterize graph topological attributes at local (nodal) level. A list of 18 different nodal graph measures [43] was extracted for each individual node, including degree, eigenvector centrality, modularity, PageRank centrality, nodal eccentricity, community Louvain, module degree z-score, participation coefficient, routing efficiency, clustering coefficient, diversity coefficient, gateway coefficient (node strength), gateway coefficient (betweenness centrality), local assortativity, participation coefficient, node strength, node betweenness, and global efficiency.

B. Graph Convolutional Autoencoder

We propose a new approach that builds on the graph autoencoder (GAE) [35], [44] to learn graph embeddings in brain networks in a purely unsupervised framework. Given the brain network $G$ for each subject, the autoencoder maps the nodes $v_i \in V$ to low-dimensional vectors $z_i \in \mathbb{R}^k$ (or embeddings), using an encoder $f : (A, X) \rightarrow Z$ where $Z = [z_1, \ldots , z_N] \in \mathbb{R}^{N \times k}$ with $k \ll N$ the dimension of embedding, and then reconstruct the graph structure from the embeddings $Z$ using a decoder. The learned latent representations $Z$ should reflect the topological structure of the graph $A$ and the node content information $X$. It contains all the information necessary for downstream graph classification tasks for brain disorders. We consider two variants of GAE: (1) Generic GAE which aims to reconstruct the original input graph adjacency matrix, (2) Variational GAE (VGAE) [35], a variational extension of GAE to learn the distribution of embeddings, which could prevent potential model overfitting. The GAE proposed originally in [35] was applied for transductive problems (e.g., semi-supervised node or link prediction within a single fixed graph). In contrast, we apply the GAE in an inductive setting for multi-graph representation learning for whole-network classification, where our GAE is trained on multi-subject brain networks from the training set, and the trained graph encoder is then used to generate embeddings for completely unseen networks in the test set for subsequent classification. The weight parameters of our graph encoder are shared among networks of different subjects, which allows learning of graph representations across subjects and generalization over unseen graphs.

1) Graph Convolutional Encoder Model: To encode both graph structure $A$ and node content $X$ into $Z$ in a unified way, we employ a variant of graph convolutional network (GCN) [29] as the graph encoder of GAE. The GCN is a first-order approximation of graph convolutions in the spectral domain. The multi-layer GCN learns a layer-wise transformation by a spectral graph convolutional function $f$

$$Z^{(l+1)} = f(Z^{(l)}, A, W^{(l)})$$  \hspace{1cm} (1)$$

where $Z^{(l)}$ is the latent feature matrix after convolution at $l$-th layer of GCN with layer-dependent dimensions, $W^{(l)}$ is a layer-specific trainable weight matrix. Here, $Z^{(0)} = X \in \mathbb{R}^{N \times d}$ is the input node feature matrix. The propagation for each layer of the
The graph vector embeddings \( \sigma \) and \( \mu \) vec learned in the GCN can be attributed to can then be used to make predictions (i.e., \( \tilde{f} \)). For the VGAE, we maximize the variational lower bound w.r.t the parameters \( W \)

\[
L(\mathbf{X}', \mathbf{A}') = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X}', \mathbf{A}')} \left[ \log p(\mathbf{A}' | \mathbf{Z}) \right] - KL[q(\mathbf{Z}|\mathbf{X}', \mathbf{A}') || p(\mathbf{Z})]
\]

where \( KL(\cdot) \) is the Kullback-Leibler divergence function that measures the distance between two distributions. A Gaussian prior \( p(Z) = \prod_i p(z_i) = \prod_i N(z_i|0, I) \) was used along with the mini-batch gradient descent. Moreover, a reparametrization trick [45] was implemented for training.

**C. GAE-FCNN for Network Classification**

We design a GAE-FCNN framework for brain connectome classification by combining the GAE with a fully-connected DNN (FCNN). A readout layer is added to summarize latent node representations \( Z \) learned by the GAE for each graph into graph-level representations, which are then fed into an FCNN to classify individual networks into MDD and HC.

1) **Graph Embeddings Vectorization (Readout):** We apply a readout operation on the network node representations to generate higher graph-level representations. In the readout layer, a vector representation \( z_G \in \mathbb{R}^k \) of the graph \( G \) can be learned by aggregating all individual node embeddings in the graph via some statistical summary measures

\[
z_G = \text{mean/max/sum}(z_i^{(L)}, \ldots, z_N^{(L)})
\]

where \( L \) is the index of the last graph convolutional layer. The graph embedding \( z_G \) can then be used to make predictions about the entire graph. The mean/max/sum-based embeddings can be used individually or concatenated into a single vector to capture different graph-level information. In addition, to retain embedding information for all nodes, we also compute the graph embedding as \( z_G = \text{vec}(Z) \) by flattening of \( Z \).

2) **FCNN Classifier:** The graph vector embeddings \( z_G \) are then used as inputs to a deep FCNN for network-level classification. The FCNN classifier consists of multiple fully-connected/dense layers, plus a final softmax classification layer to output the predictive probabilities of class labels for each network. The dense layer approximates a non-linear mapping function to further capture relational information in the graph embeddings to discriminate between MDD and HC. The weight parameters of the FCNN are trained by minimizing cross-entropy loss function using stochastic gradient descent methods and backpropagation of error. Dropout is also applied to prevent overfitting.

3) **Supervised & Unsupervised Embedding Learning:** We consider two classification schemes using the network embeddings learned in supervised and unsupervised ways. The proposed encoder-decoder framework (Fig. 1(a)) to extract network embeddings described thus far is by default unsupervised, i.e., the GAE is trained to reconstruct the original graph structure. Algorithm 1 summarizes the procedure of Unsupervised GAE-FCNN for MDD classification. The unsupervised learning makes use of only information in \( A \) and \( X \), without knowledge of a particular downstream connectomic classification task. We further develop a supervised framework, as shown in Fig. 1(b), which utilizes the task-specific classification labels in order to learn the network embeddings. The inner-product decoder in the supervised model is replaced with an FCNN to decode the embeddings from the output of GCN encoder to class labels. The parameters of the GCN encoder can be trained based on...
Algorithm 1: GAE-FCNN Training.

Input: $Y$, is the fMRI time series of length $T$ measured from $N$ ROIs for each subject. $labels$, are the associated class labels for each subject.

Output: The aggregated node embeddings $Z_G$ and predictions of each $Y$ as HC or MDD.

1: for each subject in the dataset do
2:   Construct FC matrix: $FC = LDW$-estimator($Y$)
3:   Construct node feature matrix $X = FC$
4:   Compute adjacency matrix $A = k$-NNG($FC$)
5: end for
6: Initialize model parameters $W = \{W^{(0)}, \ldots, W^{(l)}\}$ and $Z^{(0)} = X$
7: for epoch $\leftarrow 1$ to $P$ do $\triangleright$ GAE training
8:   for batch $\leftarrow 1$ to $B$ do $\triangleright$ Multi-subject training
9:     for $l = 0, 1, \ldots, L$ do
10:       $Z^{(l+1)} = \sigma(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}Z^{(l)}W^{(l)})$
11:     end for
12:     Reconstruct adjacency matrix $\hat{A} = \sigma(ZZ^T)$
13:     Compute GAE loss $\mathcal{L}$ according to (9)
14:   end for
15: end for
16: $Z_G = \text{Readout}(Z^{(l+1)})$ $\triangleright$ Extract embedding
17: $\text{FCNN} = \text{Train}(Z_G, labels)$ $\triangleright$ FCNN Training

cross-entropy loss between the predicted and true class labels using the backpropagation algorithm. By incorporating task-specific supervision, the encoder model is optimized to generate embeddings that may be more discriminative of the MDD and HC classes. This model provides an end-to-end framework for the brain network classification.

IV. EXPERIMENTS

In this section, we present experimental evaluation of the proposed GAE-FCNN models for connectome classification on the rs-fMRI MDD dataset described in Section II.

A. Experimental Setup

1) Data Partitioning: We applied a nested-stratified 5-fold cross-validation (CV) data partitioning scheme [46] to evaluate the performance of different models in classifying MDD and HC. Specifically, a two-level 5-fold CV was used comprising an outer-loop for testing and an inner-loop for model hyperparameter optimization. For each iteration in the outer-loop, a test set was assigned, and the rest of the data were split into five train-validation partitions to tune the model hyper-parameters. This process was repeated for all outer-loop 5-fold partitions. The best-performing model (on the validation set) of the five candidate models was then selected to evaluate the performance on the unseen test sets. The classification performance was evaluated using the following metrics: classification accuracy ($Acc$), sensitivity ($Sen$), specificity ($Spe$), precision ($Pre$), and F-score ($F_1$).

2) Model Architecture and Training: We implement the proposed GAE-FCNN based on PyTorch [47] using the GraphConv module from DGL library [48] for GCN. For the unsupervised model, the architecture and hyper-parameters of GAE and FCNN were determined separately. We computed the reconstruction error of graph over a range of hyper-parameters for the GAE, and a two-layered GCN with respective embedding dimensions of 256 and 32 was identified as the optimal architecture for GAE with the minimum reconstruction error. Further increase in the number of GCN layers gave no further improvement. Using the extracted $116 \times 116$ network adjacency matrices and $116 \times d$ node feature matrices (dimension $d$ depends on type of features used) as inputs, the GAES were trained using Adam optimizer [49] to minimize graph reconstruction loss, with learning rate of 0.00214, reduce-factor of 0.3, 200 training epochs and a batch size of 26. Fig. 2 illustrates a training curve of the GAE model with decreasing reconstruction error over epochs. The trained GAE decoder was then used to generate $116 \times 32$ node embedding matrices $Z$ as inputs to the FCNN. Bayesian optimization [50] with Expected Improvement (EI) acquisition function was used to optimize the hyper-parameters of FCNN, which suggested an architecture of 3 dense layers (with respective 256, 128, 64 hidden nodes), learning rate of 0.00082, reduce factor of 0.3 and a batch size of 8. The FCNN was also trained on the extracted graph embeddings $Z_G$ using Adam algorithm.

For the supervised model, the hyper-parameters of the GCN and FCNN were optimized simultaneously using the Bayesian optimizer. The selected hyper-parameters are: two layers with dimensions of 128 and 128 for GCN, 1 dense layer (with 2 hidden nodes) for FCNN with learning rate of 0.00027, reduce factor of 0.8 and a batch size of 7. The model was trained on the fMRI network data with target class labels, using the Adam algorithm to minimize cross-entropy loss.

3) Methods for Comparison: We benchmark the performance of the proposed methods with SVM and state-of-the-art connectome-specific DNN models: BrainNetCNN and four GCN-based methods. These competing models were evaluated with the same 5-fold CV as the proposed methods.

1) SVM-RBF: We trained SVM with radial basis function (RBF) on the vectorized LDW-correlation coefficients.

2) BrainNetCNN: The BrainNetCNN [22] is a specially designed deep CNN model which can preserve spatial information in brain connectivity data. Here, the $N \times N$ LDW correlation matrices were used directly as inputs to the BrainNetCNN to predict the class labels of MDD.
and HC as output. It consists of three types of layers: edge-to-edge (E2E) layers, edge-to-node (E2N) layers, and node-to-graph (N2G) layers. The E2E layer applies a cross-shaped convolutional filter to each element of the FC input matrix, and combines the edge weights of neighbor nodes to output an $N \times N$ matrix. The E2N layer is equivalent to the 1D-CNN filter designed for dimensionality reduction. The N2G layer is a dense layer taking the $N \times 1$ E2N output to produce a single scalar. Finally, the output of N2G is fed to classification layer for prediction.

3) Population-based GCN: This method exploits GCN to model a population graph, where each node represents a subject and edges encode similarity between subjects [29]. It performs node/subject level-classification in a semi-supervised manner to predict brain disorders. Similar to [29], we used the vectorized upper triangular part of LDW correlation matrices as inputs to the population-based GCN. We set the model hyper-parameters with Chebyshev polynomial basis filters for spectral convolutions as in [29]. The model was trained using 500 epochs with early stopping patience of 10 epochs.

4) GroupINN [51]: The group-based GCN (GroupINN) uses an ensemble of GCNs to learn graph-level latent embedding representations. The unified framework uses multi-graph clustering and embedding learning to jointly optimize the training process of graph convolutions.

5) Hi-GCN [32]: Hierarchical GCN (Hi-GCN) is a two-level GCN. The first level learns topological embeddings from brain connectivity networks of individual subjects. The second level is a population-based GCN using individual network embedding as node features to incorporate contextual associations between subjects for classification. It can jointly learn the graph embeddings from the brain FC and population networks at the same time.

6) E-Hi-GCN [52]: An ensemble of Hi-GCN (E-Hi-GCN) is an ensemble framework combining a set of Hi-GCNs each of which is trained on different sparsity level brain networks. It is capable of handling high-dimensional noisy correlations in brain networks.

7) EV_GCN [33]: The EV_GCN method introduces an adaptive population graph model with variational edges that builds on spectral graph convolutional networks and Monte-Carlo edge dropout. The EV_GCN integrates imaging with non-imaging data in populations for uncertainty-aware brain disease prediction.

8) BrainGINN [53]: BrainGINN utilizes ROI-aware graph convolutional (Ra-GConv) layers to capture the topological and functional information of fMRI. It employs salient ROI-selection pooling layers and regularization terms for pooling results. This method is able to perform multi-graph connectome-based classification.

We applied hyper-parameter tuning using the Bayesian optimization on both the proposed and competing methods based on the same cross-validation setting to obtain the optimal set of hyper-parameters for each method.

B. Results

1) Ablation Study: Network Construction Strategies: Table I shows the classification performance (average and standard deviation over 5 folds) of the unsupervised GAE/VGAE-FCNN and supervised GCN-FCNN classifiers. To investigate the impact of choices of network construction strategies on classification, we also evaluated two FC metrics to construct the graph adjacency matrix A: Pearson’s correlation matrix and LDW shrinkage correlation matrix; three types of input node features for $X$: raw rs-fMRI time series, FC weights (LDW correlation coefficients) and nodal graph-theoretic measures. The selected readout schemes are also given, and details will be discussed in the next section. As expected, using input graph data based on the LDW correlations shows superior performance over the traditional Pearson’s correlations in classifying MDD and HC for all classification models, as the LDW shrinkage method can provide more reliable estimate of the high-dimensional network structure. For node features, the use of LDW-FC generally provided better classification than the raw fMRI time series and local graph measures. This indicates more discriminative information in the connection weights compared to the low-level BOLD fluctuations, and learning of higher-level meta representations from local graph features also fails to offer additional advantages for classification.

We can see that the unsupervised GAE-FCNNs performed better than the supervised GCN-FCNN model, with GAE-FCNN achieving the highest classification accuracy when using LDW-FC for both the graph construction and node features. This suggests that embeddings learned in an unsupervised manner to preserve faithfully the brain network topology can be more predictive of MDD and HC than that optimized to discriminate the class labels directly. Among the unsupervised models, however, use of the probabilistic encoding framework in VGAE does not improve classification performance, probably limited by the strong assumption of an i.i.d. Gaussian prior on latent embeddings, and the approximated model parameter inference of the variational method. Future work will investigate better-suited prior distribution in the VGAE for brain network data.

2) Ablation Study: Readout Strategies: Table II shows the classification results for different readout strategies. We compared different readout/transformation methods to obtain graph-level representation $z_{CL}$ as inputs to FCNN classifier, i.e., flattening of $Z$ and mean/max/sum aggregation of node embeddings $\{z_i\}$. It can be seen that the flattening method by concatenating learned embeddings of all nodes as input yields better classification performance for different classifiers generally, compared to the aggregation method which may induce loss of information about individual nodes.

3) Ablation Study: Different Brain Parcellations: To examine the effect of different brain parcellations on the FC classification performance, we evaluated our method on the ROI-wise fMRI time series data extracted based on three parcellation atlases (both anatomical and functional): AAL atlas, Harvard-Oxford (HO) atlas (derived from anatomical landmarks: sulci and gyral) [54], and Power atlas (comprising functional areas associated with 13 large-scale functional networks and a group of unlabeled regions) [55], with respective number of ROIs of 116, 112 and 264. Fig. 4 shows the MDD classification accuracies of different methods on the various brain atlases. It is apparent that the proposed supervised GCN-FCNN performs the best compared to other competing methods on all atlases. Among
TABLE I
CLASSIFICATION PERFORMANCE OF THE PROPOSED GAE/VEGAN-FCNN AND SUPERVISED GCN-FCNN MODELS USING DIFFERENT NETWORK CONSTRUCTION STRATEGIES FOR CLASSIFYING MDD AND HC SUBJECTS BASED ON RS-FMRI FUNCTIONAL NETWORKS

| Classifier | Adjacency A | Node Feature X | Readout | Acc | Sen | Spe | Pre | F1 |
|------------|-------------|----------------|---------|-----|-----|-----|-----|----|
| Unsupervised GAE-FCNN | Pearson | Raw-fMRI | sum | 57.07 ± 5.29 | 63.08 ± 6.80 | 50.56 ± 5.67 | 57.97 ± 4.74 | 60.36 ± 5.41 |
| | | Graph-measures | max | 51.47 ± 7.76 | 63.59 ± 19.95 | 38.33 ± 13.43 | 52.17 ± 6.52 | 56.57 ± 10.37 |
| | Pearson-FC | | flattern | 60.27 ± 5.29 | 63.59 ± 7.84 | 56.67 ± 4.16 | 61.22 ± 4.63 | 62.31 ± 5.99 |
| | LDW | Raw-fMRI | flattern | 52.00 ± 1.89 | 53.33 ± 7.68 | 50.56 ± 7.54 | 53.93 ± 1.80 | 53.34 ± 4.08 |
| | | Graph-measures | mean | 51.20 ± 2.47 | 62.56 ± 10.58 | 38.89 ± 10.39 | 52.62 ± 2.10 | 56.73 ± 5.03 |
| | | LDW-FC | flattern | 65.07 ± 5.56 | 69.74 ± 9.09 | 60.00 ± 7.16 | 65.38 ± 5.04 | 67.29 ± 6.22 |
| Unsupervised VEGAN-FCNN | Pearson | Raw-fMRI | sum | 54.13 ± 3.83 | 60.00 ± 7.71 | 47.78 ± 2.08 | 55.25 ± 2.98 | 57.44 ± 5.10 |
| | | Graph-measures | max | 51.20 ± 5.76 | 68.72 ± 14.63 | 32.22 ± 19.05 | 53.01 ± 5.55 | 58.91 ± 5.87 |
| | Pearson-FC | | flattern | 56.53 ± 9.03 | 57.95 ± 13.82 | 55.00 ± 9.84 | 57.99 ± 8.21 | 57.96 ± 10.33 |
| | LDW | Raw-fMRI | max | 56.00 ± 6.25 | 66.15 ± 12.29 | 45.00 ± 13.19 | 56.79 ± 6.17 | 60.62 ± 7.11 |
| | | Graph-measures | flattern | 49.33 ± 4.99 | 60.00 ± 12.31 | 37.78 ± 17.44 | 51.96 ± 5.43 | 54.69 ± 5.26 |
| | | LDW-FC | flattern | 60.79 ± 4.84 | 62.64 ± 5.34 | 58.78 ± 7.22 | 62.55 ± 5.00 | 62.50 ± 5.42 |
| Supervised GCN-FCNN | Pearson | Raw-fMRI | max | 55.20 ± 2.47 | 60.51 ± 8.04 | 49.44 ± 11.03 | 56.87 ± 3.18 | 58.20 ± 3.18 |
| | | Graph-measures | max | 55.47 ± 2.61 | 79.49 ± 15.21 | 29.44 ± 12.98 | 54.95 ± 1.53 | 64.38 ± 5.75 |
| | Pearson-FC | | flattern | 62.93 ± 5.68 | 62.05 ± 10.18 | 63.89 ± 9.13 | 65.39 ± 6.36 | 65.22 ± 6.53 |
| | LDW | Raw-fMRI | [mean,max,sum] | 58.67 ± 3.04 | 61.03 ± 8.33 | 56.11 ± 3.69 | 59.95 ± 2.02 | 60.30 ± 4.85 |
| | | Graph-measures | mean | 57.87 ± 4.27 | 66.67 ± 13.66 | 48.33 ± 17.44 | 59.15 ± 5.55 | 61.76 ± 4.87 |
| | | LDW-FC | flattern | 59.47 ± 7.09 | 54.87 ± 6.20 | 64.44 ± 12.35 | 63.66 ± 9.92 | 58.37 ± 6.02 |

Networks are constructed based on pearson’s and LDW correlation matrices using k-NNG thresholds of k. Results are averages (standard deviations) of performance measures over 5-fold cross-validation. The bold values indicate the best-performing method in each category.

TABLE II
CLASSIFICATION PERFORMANCE OF PROPOSED MODELS USING DIFFERENT READOUT STRATEGIES FOR TRANSFORMING LEARNED EMBEDDINGS AS INPUTS TO FCNN CLASSIFIERS

| Classifier | Readout | Acc | Sen | Spe | Pre | F1 |
|------------|---------|-----|-----|-----|-----|----|
| Unsupervised GAE-FCNN | flattern | 65.07 ± 5.56 | 69.74 ± 9.09 | 60.00 ± 7.16 | 65.38 ± 5.04 | 67.29 ± 6.22 |
| mean | 53.33 ± 3.77 | 53.33 ± 9.65 | 53.33 ± 4.78 | 55.02 ± 3.46 | 53.91 ± 6.41 |
| max | 56.00 ± 6.95 | 57.95 ± 6.20 | 53.89 ± 9.72 | 57.91 ± 7.31 | 57.84 ± 6.36 |
| sum | 53.33 ± 3.77 | 53.33 ± 9.65 | 53.33 ± 4.78 | 55.02 ± 3.46 | 53.91 ± 6.41 |
| [mean,max,sum] | 54.40 ± 5.02 | 47.18 ± 5.98 | 62.22 ± 11.33 | 58.33 ± 6.67 | 51.76 ± 4.45 |
| Unsupervised VEGAN-FCNN | flattern | 60.79 ± 4.84 | 62.64 ± 5.34 | 58.78 ± 7.32 | 62.55 ± 5.00 | 62.50 ± 4.52 |
| mean | 56.81 ± 3.40 | 60.64 ± 4.28 | 52.62 ± 3.17 | 58.26 ± 2.91 | 59.41 ± 3.48 |
| max | 55.76 ± 1.72 | 60.27 ± 4.60 | 50.85 ± 3.87 | 57.25 ± 1.24 | 58.65 ± 2.41 |
| sum | 58.27 ± 3.93 | 72.29 ± 6.51 | 42.98 ± 9.20 | 58.25 ± 3.15 | 64.34 ± 3.14 |
| [mean,max,sum] | 60.18 ± 4.07 | 69.89 ± 7.64 | 49.59 ± 3.72 | 60.11 ± 2.93 | 64.35 ± 4.72 |
| Supervised GCN-FCNN | flattern | 59.47 ± 7.09 | 54.87 ± 6.20 | 64.44 ± 12.35 | 63.66 ± 9.92 | 58.37 ± 6.02 |
| mean | 61.07 ± 4.80 | 67.69 ± 13.53 | 53.89 ± 16.06 | 62.24 ± 6.58 | 63.96 ± 5.78 |
| max | 60.00 ± 4.99 | 66.15 ± 13.51 | 53.33 ± 15.95 | 61.28 ± 5.47 | 62.70 ± 6.42 |
| sum | 57.33 ± 5.13 | 64.10 ± 14.23 | 50.00 ± 19.33 | 59.30 ± 6.10 | 60.40 ± 5.91 |
| [mean,max,sum] | 58.67 ± 3.96 | 54.36 ± 9.51 | 63.33 ± 4.08 | 61.37 ± 3.10 | 57.37 ± 6.48 |

Network adjacency matrix: LDW. Node feature: LDW-FC. The bold values indicate the best-performing method in each category.

the atlases, AAL-116 generally gives better classification than HO and Power atlases over all methods. This suggests that FC networks based on anatomical ROIs may provide more discriminative information for differentiating between MDD and HC, compared to functional-ROI networks.

4) **Comparison With State-of-the-Art Methods:** Table III shows the performance comparison of different connectome-based classification methods. The proposed methods clearly outperformed the competing models, with the unsupervised GAE-FCNN performing the best. In consistency with recent studies, our results suggest the advantages of DNN methods over traditional SVM classifier with significant improvement in FC classification. The population-based GCNs perform slightly better than the BrainNetCNN. The population-based GCNs, while leveraging on pairwise associations between subjects in a population graph for node/subject-level classification, do not classify brain networks directly as in our proposed models.
TABLE III

PERFORMANCE COMPARISON OF PROPOSED GAE-FCNN WITH VARIOUS STATE-OF-THE-ART METHODS FOR FUNCTIONAL CONNECTOME-BASED CLASSIFICATION OF MDD AND HC ON REST-META-MDD RS-FMRI DATASET

| Classifier       | Acc  | Sen  | Spe  | Pre  | F1   |
|------------------|------|------|------|------|------|
| SVM-RBF          | 62.67 ± 4.22 | 69.74 ± 5.48 | 55.00 ± 4.08 | 62.63 ± 3.39 | 65.97 ± 4.21 |
| BrainNetCNN [22] | 60.53 ± 3.83 | 55.90 ± 5.94 | 65.56 ± 7.97 | 64.14 ± 5.67 | 59.48 ± 4.21 |
| Population-based GCN [29] | 62.93 ± 6.39 | 66.67 ± 8.58 | 58.89 ± 5.39 | 63.54 ± 5.64 | 65.00 ± 6.97 |
| GroupINN [51]    | 60.53 ± 2.47 | 62.00 ± 2.67 | 57.22 ± 9.40 | 63.59 ± 11.05 | 62.15 ± 5.24 |
| Hi-GCN [32]      | 59.73 ± 6.44 | 61.17 ± 5.52 | 61.00 ± 3.77 | 59.49 ± 11.85 | 60.08 ± 8.65 |
| E-Hi-GCN [52]    | 55.47 ± 4.35 | 58.62 ± 5.26 | 59.44 ± 13.45 | 51.79 ± 14.27 | 53.91 ± 7.17 |
| EV_GCN [33]      | 62.55 ± 5.84 | 57.96 ± 15.41 | 67.56 ± 9.80 | 66.00 ± 4.34 | 64.00 ± 4.93 |
| BrainGNN [53]    | 53.19 ± 2.23 | 56.73 ± 12.48 | 49.33 ± 13.29 | 55.14 ± 2.20 | 55.10 ± 6.65 |

Competing

| Classifier       | Acc  | Sen  | Spe  | Pre  | F1   |
|------------------|------|------|------|------|------|
| Supervised GCN-FCNN | 59.47 ± 7.09 | 54.87 ± 6.20 | 64.44 ± 12.35 | 63.66 ± 9.92 | 58.57 ± 6.02 |
| Unsupervised GAE-FCNN | 65.07 ± 5.56 | 69.74 ± 9.09 | 60.00 ± 7.16 | 65.38 ± 5.04 | 67.29 ± 6.22 |
| Unsupervised VGEA-FCNN | 60.79 ± 4.84 | 62.64 ± 5.34 | 58.78 ± 7.22 | 62.55 ± 5.00 | 62.50 ± 4.52 |

All methods used LDW-estimated correlations in rs-FMRI as FC features and network adjacency matrices.

The bold values indicate the best-performing method in each category.

TABLE IV

FUNCTIONAL CONNECTOME-BASED MDD CLASSIFICATION RESULTS (IN ACCURACY) OF DIFFERENT METHODS ON REST-META-MDD DATASETS FROM DIFFERENT ACQUISITION SITES

| Classifier       | Site 20 (No. subjects: 477) | Site 1 (No. subjects: 146) | Site 21 (No. subjects: 144) |
|------------------|-----------------------------|----------------------------|----------------------------|
| SVM-RBF          | 62.67 ± 4.22                | 53.64 ± 6.68               | 64.35 ± 10.79              |
| BrainNetCNN [22] | 60.53 ± 3.83                | 53.64 ± 3.40               | 60.00 ± 3.25               |
| Population-based GCN [29] | 62.93 ± 6.39            | 53.55 ± 7.27               | 55.65 ± 6.39               |
| GroupINN [51]    | 60.53 ± 2.47                | 57.27 ± 2.23               | 58.26 ± 6.56               |
| Hi-GCN [32]      | 59.73 ± 6.44                | 62.73 ± 7.27               | 57.39 ± 5.77               |
| E-Hi-GCN [52]    | 55.47 ± 4.35                | 53.64 ± 3.40               | 55.65 ± 6.39               |
| EV_GCN [33]      | 62.55 ± 5.84                | 67.56 ± 5.56               | 63.13 ± 7.49               |
| BrainGNN [53]    | 53.19 ± 2.23                | 54.80 ± 10.13              | 56.23 ± 9.22               |

Competing

| Classifier       | Site 20 (No. subjects: 477) | Site 1 (No. subjects: 146) | Site 21 (No. subjects: 144) |
|------------------|-----------------------------|----------------------------|----------------------------|
| Supervised GCN-FCNN | 59.47 ± 7.09                | 57.27 ± 4.84               | 58.52 ± 7.28               |
| Unsupervised GAE-FCNN | 65.07 ± 5.56              | 60.36 ± 7.39               | 66.61 ± 1.74               |
| Unsupervised VGEA-FCNN | 60.79 ± 4.84              | 62.73 ± 6.68               | 62.61 ± 4.38               |

The bold values indicate the best-performing method in each category.

5) Connectivity Maps Learned by GAE: In Fig. 3, we plot the averaged feature maps of node-level embeddings learned by the GCN-GAE from LDW-based networks for the MDD and HC subjects. Noticeable difference in the learned embedding pattern can be seen between the two groups, with stronger activation for some ROIs in MDD compared to HC. Considerable between-variance is observed, indicating separability of the learned embeddings between two groups. This demonstrates the ability of the proposed GAE-FCNN model to extract latent representations of brain network structure that can clearly distinguish between MDD and controls, which explains the enhanced performance in the downstream classification task.

We further constructed high-order FC by correlating the GAE-learned embeddings $z_i$ between pairs of nodes. Fig. 5(a) shows the difference in connectivity pattern between the MDD and HC groups as quantified by the LDW-estimated raw FC and the embedding-based high-order FC. A group-level t-test was used to contrast the FC between the two groups, and connections with significant difference ($p<0.05$) are shown in Fig. 5(a-right). The embedding-based FC matrices (Fig. 5(a-left & a-middle)) exhibit a visually similar connectivity structure between HC and MDD groups. However, Fig. 5(a-right) shows evidence that the embedding-based FC reveals pronounced difference in connectivity, particularly between specific communities or modules of ROIs. To examine whether these differences are biologically meaningful and related to MDD as a network-based disorder,
we plot the topological maps in Fig. 5(b) and (c) to visualize the increase and decrease in FC between ROIs in MDD relative to HC. The embedding FC identified a spread reduction in intrinsic connectivity of the amygdala with a variety of ROIs involved in emotional processing and regulation in MDD subjects (including caudate, temporal regions, occipital cortex, and cerebellum), as reported in previous rs-fMRI studies [56]. In agreement with previous findings [7], we also found significant increase in FC in the default mode network (DMN). The detected altered rs-FC between cerebellum with the DMN and affective network has also been associated with major depression [57], [58].

V. DISCUSSION

We developed a deep GNN framework for embedding learning in brain functional networks to identify connectome-specific bio-signatures for classifying brain disorders such as MDD. The proposed GAE-FCNN provides a novel approach to incorporating the non-Euclidean information about graph structure into the classification of brain networks. It combines a GCN-based GAE that can learn latent embeddings effectively to encode topological information and node content, and a deep FCNN that leverages on the learned embeddings to reveal disrupted neural connectivity patterns in MDD relative to HC. The embedding FC identified a spread reduction in intrinsic connectivity of the amygdala with a variety of ROIs involved in emotional processing and regulation in MDD subjects (including caudate, temporal regions, occipital cortex, and cerebellum), as reported in previous rs-fMRI studies [56]. In agreement with previous findings [7], we also found significant increase in FC in the default mode network (DMN). The detected altered rs-FC between cerebellum with the DMN and affective network has also been associated with major depression [57], [58].
to HC for classification purpose. On a challenging task of classifying MDD and HC using a small amount of rs-fMRI data, the proposed method substantially outperforms several state-of-the-art brain connectome classifiers, achieving the best accuracy of 65.07% with the unsupervised GAE-FCNN model. Furthermore, high-order networks constructed from the node embeddings generated from the proposed GAE detects altered FC patterns in MDD related to emotional processing, which are not captured by the original FC measures. Our framework is generally applicable to other functional neuroimaging data, e.g., EEG-derived networks, and other neuropsychiatric disorders besides MDD associated with alterations in functional connectivity, showing potential as diagnostic tool in clinical settings.

There are potential limitations of our approach. First, our method focuses on embedding learning and classification for static brain networks. However, recent rs-fMRI studies suggest the temporal dynamics of brain FC networks in which connectivity edges between regions evolve over time [59], [60], [61]. Certain neuropsychiatric disorders have also been associated with disruptions in dynamic FC and graph properties such as in MDD [62]. Future work could extend the proposed GAE framework to learn latent representations to embed the time-evolving network structure, by using some recent extensions of GCNs for dynamic graphs in the encoder part, e.g., the EvolveGCN [63] which uses a recurrent neural network (RNN) to evolve the GCN parameters. Second, we analyzed a single type of brain networks from one neuroimaging modality, i.e., functional networks from fMRI. Multimodal fusion by combining different imaging modalities such as fMRI and diffusion imaging [64] could provide multiple views and hence more complete understanding of the brain networks. One possible direction is to characterize the fusion of functional and structural networks as multilayer networks, i.e., networks that can model multiple types of interactions and relations between brain nodes. Our GAE model can be generalized to produce embeddings for these multilayer brain networks, by incorporating the recently proposed multilayer GCN layers [65] in the encoding phase. Moreover, our study uses a single type of node features for classification. One could explore different fusion strategies to learn embeddings for multiple node features in a unified way. Third, our decoder model is designed to reconstruct the network structure only, which is adequate to learn embeddings to capture node relational information in brain networks. It could be extended to reconstruct both the input node features and the adjacency matrix to learn joint embeddings of both network structure and features to improve classification. This could be done by generalizing the decoder function (8) as \( p(A, X, Z) = p(A|Z)p(X|Z) \), and the reconstruction loss (9) to \( L = L_A + L_X \) where \( L_A = \mathbb{E}_{q(Z|X,A)}[\log p(A|Z)] \) and \( L_X = \mathbb{E}_{q(Z|X,A)}[\log p(X|Z)] \). Finally, while this study has devised a novel framework producing network embeddings that differentiate MDD and HC and improve brain connectome classification, the interpretability of the model is important for clinical applications to understand the underlying mechanism behind the predictions and the neurobiological system being classified, instead of being used as a black box. Further studies could explore recent approaches to explaining the predictions in graph neural networks [66], e.g., to identify which input edges and node features of the brain networks are more important in predicting a certain disease class.

VI. CONCLUSION

In this study, we introduced a GNN-based framework, the GAE-FCNN, which seamlessly integrates non-Euclidean graph structure information for classifying brain disorders, specifically MDD. The GAE-FCNN approach effectively encodes topological information and node content using a GCN-based inductive GAE, combined with a deep feedforward FCNN to identify disrupted neural connectivity patterns. The GAE-FCNN model outperforms several state-of-the-art brain connectome classifiers, achieving a peak accuracy of 65.07% on the challenging task of classifying HC and MDD patients using limited resting-state fMRI data. Furthermore, the model’s application extends beyond fMRI data, showcasing potential diagnostic utility in clinical settings. While this approach offers promising results, it is important to consider potential enhancements, including adaptation for dynamic brain networks, exploration of multimodal fusion, joint embedding of network structure and features, and improving model interpretability for clinical applications.

REFERENCES

[1] D. S. Bassett and E. T. Bullmore, “Human brain networks in health and disease,” Curr. Opin. Neurol., vol. 22, no. 4, pp. 340–347, 2009.
[2] N. D. Woodward and C. J. Cascio, “Resting-state functional connectivity in psychiatric disorders,” JAMA Psychiatry, vol. 72, no. 8, pp. 743–744, 2015.
[3] A. Venkataraman, T. J. Whitford, C.-F. Westin, P. Golland, and M. Kubicki, “Whole brain resting state functional connectivity abnormalities in schizophrenia,” Schizophrenia Res., vol. 139, no. 1/3, pp. 7–12, 2012.
[4] R.-A. Müller, P. Shih, B. Keenan, J. R. Deyoe, K. M. Leyden, and D. K. Shakla, “Underconnected, but how? A survey of functional connectivity MRI studies in autism spectrum disorders,” Cereb. Cortex, vol. 21, no. 10, pp. 2233–2243, 2011.
[5] H.-Y. Zhang et al., “Resting brain connectivity: Changes during the progress of Alzheimer disease,” Radiology, vol. 256, no. 2, pp. 598–606, 2010.
[6] M. Hallett et al., “Human brain connectivity: Clinical applications for clinical neurophysiology,” Clin. Neurophysiol., vol. 131, pp. 1621–1651, 2020.
[7] P. C. Mulders, P. F. van Eijndhoven, A. H. Schene, C. F. Beckmann, and I. Tendolkar, “Resting-state functional connectivity in major depressive disorder: A review,” Neurosci. Biobehavioral Rev., vol. 56, pp. 330–344, 2015.
[8] J. Brakowski et al., “Resting state brain network function in major depression–depression symptomatology, antidepressant treatment effects, future research,” J. Psychiatr. Res., vol. 92, pp. 147–159, 2017.
[9] M. D. Greicius et al., “Resting-state functional connectivity in major depression: Abnormally increased contributions from subgenual cingulate cortex and thalamus,” Biol. Psychiatry, vol. 62, no. 5, pp. 429–437, 2007.
[10] J. Zhang et al., “Disrupted brain connectivity networks in drug-naive, first-episode major depressive disorder,” Biol. Psychiatry, vol. 70, no. 4, pp. 334–342, 2011.
[11] M. Ye, T. Yang, P. Qin, X. Lei, J. Qin, and G. Liu, “Changes of functional brain networks in major depressive disorder: A graph theoretical analysis of resting-state fMRI,” PLoS One, vol. 10, no. 9, 2015, Art. no. e0133775.
[12] C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager, “Building better biomarkers: Brain models in translational neuroimaging,” Nature Neurosci., vol. 20, no. 3, pp. 365–377, 2017.
[13] C. J. Brown and G. Hamarneh, “Machine learning on human connectome data from fMRI,” 2016, arXiv:1611.08999.
[14] Y. Du, Z. Fu, and V. D. Calhoun, “Classification and prediction of brain disorders using functional connectivity: Promising but challenging,” Front. Neurosci., vol. 12, 2018, Art. no. 525.
[15] K. Dadi et al., “Benchmarking functional connectome-based predictive models for resting-state fMRI,” Neuroimage, vol. 192, pp. 115–134, 2019.
[16] R. Bhaumik et al., “Multivariate pattern analysis strategies in detection of remitted major depressive disorder using resting state functional connectivity,” Neuroimage: Clin., vol. 16, pp. 390–398, 2017.
[17] X. Geng, J. Xu, B. Liu, and Y. Shi, “Multivariate classification of major depressive disorder using the effective connectivity and functional connectivity,” Front. Neurosci., vol. 12, 2018, Art. no. 38.
[38] C.-G. Yan et al., “Reduced default mode network functional connectivity,” NeuroImage, vol. 142, pp. 124–146, 2016.

[32] H. Jiang, P. Cao, M. Y. Xu, J. Yang, and O. Zaiane, “Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain networks,” in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2019, pp. 2623–2631.

[34] T. He et al., “Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographic,” NeuroImage, vol. 206, 2020, Art. no. 116276.

[47] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 8024–8035.

[35] M. Wang et al., “Deep graph library: A graph-centric, highly-performant package for graph neural networks,” 2019, arXiv:1909.01315.

[36] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” in J. Multivariate Anal., vol. 88, no. 2, pp. 365–411, 2004.

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

[38] C. Yan and Y. Zang, “DPARSF: A MATLAB toolbox for “pipeline” data analysis of resting-state fMRI,” Front. Syst. Neurosci., vol. 4, 2010, Art. no. 13.

[39] M. R. Brier, A. Mitra, J. E. McCarthy, B. M. Ances, and A. Z. Snyder, “Partial covariance based functional connectivity computation using Ledoit–Wolf covariance regularization,” NeuroImage, vol. 121, pp. 29–38, 2015.

[40] M. Fiocas and H. Ombao, “The generalized shrinkage estimator for the analysis of functional connectivity of brain signals,” Ann. Appl. Statist., vol. 5, pp. 1102–1125, 2011.