Cross-Domain Identification of Multisite Major Depressive Disorder Using End-to-End Brain Dynamic Attention Network

Xue Yuan[ID], Maozhou Chen, Peng Ding, Anan Gan, Anmin Gong, Zhaosong Chu[ID], Wenya Nan, Yunfa Fu[ID], and Yuqi Cheng

Abstract—**Establishing objective and quantitative imaging markers at individual level can assist in accurate diagnosis of Major Depressive Disorder (MDD). However, the clinical heterogeneity of MDD and the shift to multisite data decreased identification accuracy. To address these issues, the Brain Dynamic Attention Network (BDANet) is innovatively proposed, and analyzed bimodal scans from 2055 participants of the Rest-meta-MDD consortium. The end-to-end BDANet contains two crucial components. The Dynamic BrainGraph Generator dynamically focuses and represents topological relationships between Regions of Interest, overcoming limitations of static methods. The Ensemble Classifier is constructed to obfuscate domain sources to achieve inter-domain alignment. Finally, BDANet dynamically generates sample-specific brain graphs by downstream recognition tasks. The proposed BDANet achieved an accuracy of 81.6%. The regions with high attribution for classification were mainly located in the insula, cingulate cortex and auditory cortex. The level of brain connectivity in p24 region was negatively correlated ($p<0.05$) with the severity of MDD. Additionally, sex differences in connectivity strength were observed in specific brain regions and functional subnetworks ($p<0.05$ or $p<0.01$). These findings based on a large multisite dataset support the conclusion that BDANet can better solve the problem of the clinical heterogeneity of MDD and the shift of multisite data. It also illustrates the potential utility of BDANet for personalized accurate identification, treatment and intervention of MDD.**

Index Terms—**Major depressive disorder, brain dynamic attention network, clinical heterogeneity, multisite data shift.**

Xue Yuan, Maozhou Chen, Peng Ding, Anan Gan, and Yunfa Fu are with the Faculty of Information Engineering and Automation and the Brain Cognition and Brain-Computer Intelligence Integration Group, Kunming University of Science and Technology, Kunming 650500, China (e-mail: fyf@ynu.edu.cn).

Anmin Gong is with the School of Information Engineering, Chinese People's Armed Police Force Engineering University, Xi'an 710086, China.

Zhaosong Chu and Yuqi Cheng are with the Department of Psychiatry, The First Affiliated Hospital of Kunming Medical University, Kunming 650032, China (e-mail: yuqicheng@126.com).

Wenya Nan is with the Department of Psychology, College of Education, Shanghai Normal University, Shanghai 200234, China.

I. Introduction

MAJOR Depressive Disorder (MDD) is a prevalent psychiatric disorder with a global impact [1], which severely impairs the social function and quality of life of patients [2]. However, the clinical heterogeneity and neurobiological complexity of MDD [3] pose challenges for its the study of the neurogenesis mechanism and early prediction. To objectively characterize brain activity and thus explore the neurological mechanisms of MDD, many studies have used quantitative measurement techniques provide quantitative assessment of disrupted brain function, such as structural Magnetic Resonance Imaging (sMRI) [4], functional Magnetic Resonance Imaging (fMRI) [5], [6], [7] and Electroencephalogram [8], [9], [10]. And MRI playing a crucial role in uncovering the etiology and pathogenesis of MDD, schizophrenia, and other psychiatric disorders [11]. In addition, extensive evidence from resting-state fMRI (rs-fMRI) indicates that patients with MDD exhibit abnormal brain function in multiple cortical and subcortical areas, such as the prefrontal cortex, insula, amygdala and hippocampus [12].

Due to the clinical heterogeneity of MDD, the classification accuracy of MDD-related studies based on small-dataset Machine Learning (ML) varies tremendously [13], led to the absence of widely accepted imaging biomarkers for clinical diagnosis [11], [14]. Yang et al. proposed obtaining stable markers from large MDD datasets enhance the reliability of study results [11], [15], [16], [17]. Several large-sample methods focus on statically modeling the relationships between Regions of Interest (ROIs) by constructing correlation matrix, and the resulting functional brain graph is fed into Graph Neural Networks (GNNs) to interpret abnormal brain patterns in MDD [18], [19], [20], [21], [22]. GNNs, inspired by time-frequency analysis methods [23], deepen the understanding of structural relationships in brain graphs, are becoming popular in neuroscience for modeling static ROI relationships to discover common brain patterns [18]. However, such methods have two limitations, one is that only functional data

are used to characterize ROIs and construct brain graphs, which lack structural features, and the other is the inability to optimize static brain graph according to the recognition tasks, which is not conducive to addressing the clinical heterogeneity of MDD. In addition, developing MDD impairs the structure and function of brain, so extracting hybrid neurobiomarkers from different feature patterns in order to portray the MDD brain state in a multidimensional way is beneficial to address the clinical heterogeneity of MDD and thus improve the reliability of the results [11], [15].

In addition, large-scale MDD datasets collected from multisite suffer from shifts, including exogenous shift caused by imaging at different resolutions, and endogenous shift caused by internal factors such as equipment hardware parameters and the degree of machine aging. These shifts can cause large differences in the spatial distribution between the training and test sets, which can lead to a sharp drop in model performance [24]. And most studies have proposed ideas to solve the exogenous shift with single site as a domain [13], and few have focused on the endogenous shift.

To address the above issues, we innovatively proposed an end-to-end Brain Dynamic Attention Network (BDANet) to analyze large multisite bimodal MDD data (sMRI and rs-fMRI). BDANet can directly encode the rs-fMRI time series and use the attention-based Dynamic BrainGraph Generator (DBG-G) to resolve the temporal features of ROIs to generate dynamic brain graph, and the Graph Convolution Network (GCN) analyzes it and fuses bimodal features, which differs from the analysis approaches of joint structural and functional abnormalities [25], [26], [27]. BDANet uses the downstream MDD identification task to reverse optimize DBG-G to facilitate the generation of specific brain patterns and thus address MDD clinical heterogeneity.

To solve sample disparities caused by multisite data shift, we construct an Ensemble Classifier (E-Classifier) to achieve inter-domain knowledge migration to align the sample space, improve the domain adaptation capability of the model. Finally, high attribution brain regions contributing to classification were identified, potential connections between their functional abnormalities, clinical symptoms, and sex were explored. Additionally, abnormalities in 7 functional subnetworks were analyzed. Thus, our main contributions are as follows: (1) The BDANet we proposed is a high-precision end-to-end model capable of recognizing multisite MDD across domains, which addresses the clinical heterogeneity of MDD and the multisite data shift. (2) The BDANet characterizes the brain state of MDD in multiple dimensions, the DBG-G module in our proposed dynamically focuses on abnormalities in brain connectivity, and the E-Classifier module eliminates domain noise. (3) We explored the correlation between functional connectivity and MDD severity, and gender differences in MDD functional connectivity strength.

II. Materials

A. Dataset

This study was conducted on 23 datasets from the Rest-meta-MDD consortium, including 1645 MDD and 1335 Healthy Controls (HCs). Demographic and clinical data, including age, gender, and the 17-item Hamilton Depression Scale (HAMD), were collected at each site. The study followed the standard quality control procedures of the consortium to exclude poor quality data, data from subjects older than 65 years and younger than 18 years, and missing data from the HAMD scale. Finally, 2055 participants (802 MDD, 1253 HC) were included in our analysis. The studies involving human participants received approval from the Medical Ethics Committee of Kunming University of Science and Technology (ethical approval no: KMUSTMEC056, date: March 11, 2021). All participants provided written informed consent prior to participation, and data collection at each site was approved by the local Institutional Review Board.

B. fMRI and sMRI Images Acquisition and Processing

Each dataset contains rs-fMRI and of the 3-dimensional T1-weighted sMRI images of all subjects, and all images were preprocessed uniformly using the DPARSF toolbox [28]. The preprocessing primarily included slice timing correction, head motion correction, normalization, spatial smoothing. We parceled the whole brain into left and right 360 ROIs according to HCP-MMP1.0 atlas [29]. Time series of BOLD signals $T \in \mathbb{R}^{B \times N \times L}$ from vertexes in each ROI were extracted and averaged. Where B is the number of samples per batch, N is the number of nodes, each node represents ROI, and L is the length of the time series of each node. To depict brain nodes in spatio-temporal context, the 5-dimensional structural features $X^v = \{x_1^v, x_2^v \ldots, x_5^v | x^v \in \mathbb{R}^{B \times N}\}$ for structure-related properties (Sulci, Area, Volume, Thickness, Curve) and 4-dimensional functional features $X^i = \{x_1^i, x_2^i, \ldots, x_4^i | x^i \in \mathbb{R}^{B \times N}\}$ for function-related properties (DC, ALFF, fALFF, ReHo) were calculated and collectively describe the distinct attributes of ROIs in GCN.

III. Methods

The architecture of the end-to-end BDANet is shown in Fig. 1 (A). BDANet first fuses the structural and functional indicators into node features $X = \text{concat}(X^v, X^i)$, $X \in \mathbb{R}^{B \times N \times 9}$. The Temporal Representation Learning component was used to temporally model the ROIs time series T, and the temporal features formed are input to the DBG-G component (Fig. 1 (B)) to generate dynamic brain graphs $A \in \mathbb{R}^{B \times N \times N}$. The generated brain graphs and node features are then parsed using the GCN of the Spatial Represent Learning component to output the sample-level features $\mathcal{H} \in \mathbb{R}^{B \times N \times V}$ of MDD, V is the dimensions of the features. Finally, the sample classifier of E-Classifier (Fig. 1 (C)) analyzes the graph features $\mathcal{H}_{source} \in \mathbb{R}^{B \times N \times V} \subset \mathcal{H}$ of source domain and outputs hybrid neural markers to accurately identify MDDs, and the domain classifier to align the domains.

A. Temporal Representation Learning

Traditional time series models are not effective in characterizing time series data with long short-term dependencies,
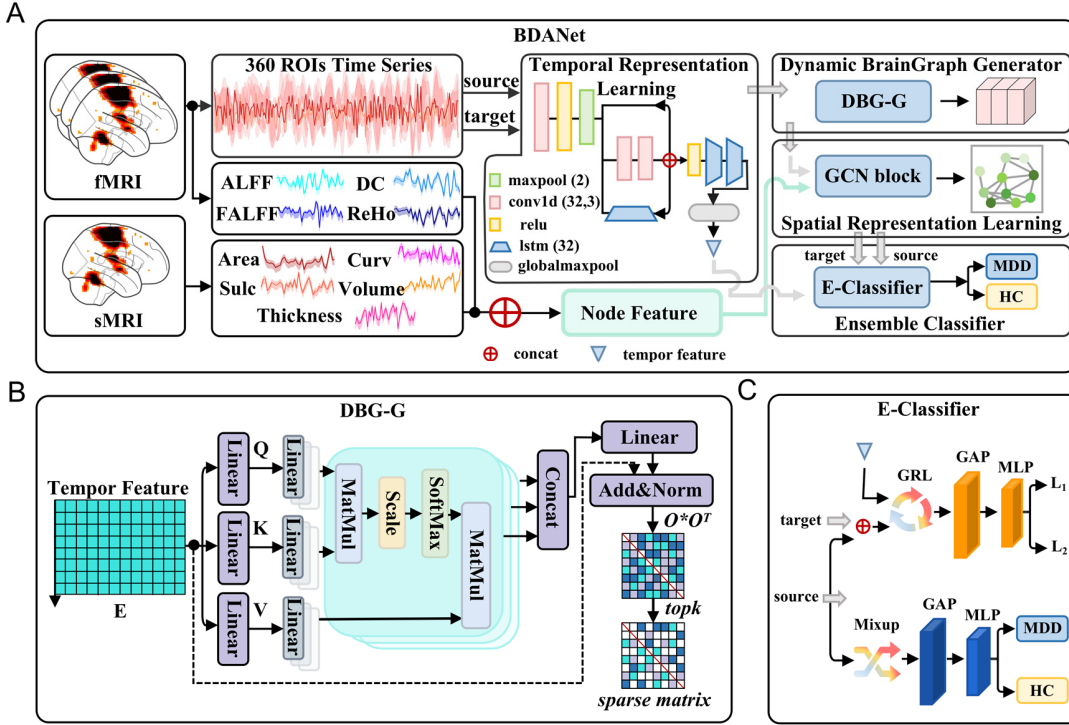
Fig. 1. End-to-end BDANet model architecture and DBG-G and E-Classifier components. (A) Overview of the proposed MDD recognition framework BDANet; (B) DBG-G component; (C) E-Classifier component.

nonlinearity and non-stationarity. To further improve the characterization of ROIs time-series T, our work uses Temporal Representation Learning component to temporally encode T. Subsequently, the temporal features $E \in \mathbb{R}^{B \times N \times S}$ of the extracted ROIs sequences are input to DBG-G, S is the feature dimension of the time.

The Temporal Representation Learning component is shown in Fig. 1 (A), which mainly consists of a parallel stacking structure of one-dimensional residual [30] and Long Short-Term Memory (LSTM) [31]. The temporal features $E_i \in \mathbb{R}^{N \times S}$ of the ROIs sequence $T_i \in \mathbb{R}^{N \times L}$ for the ith sample in each batch is calculated as:

$$S_1 = \text{MaxPool}\left(\text{ReLU}\left(\text{Conv}\left(T_i\right)\right)\right), S_2 = \text{Conv}(\text{Conv}\left(S_1\right))$$
$$S_3 = \text{LSTM}\left(S_2\right), S_4 = \text{ReLU}(\text{LSTM}(\text{LSTM}(S_1 + S_2 + S_3)))$$
$$E_i = \text{GlobalAveragePooling}(S_4) \tag{1}$$

where S_1, S_2, S_3, S_4 denotes intermediate results for output E_i. The 1D convolution layer in the time representation component uses 32 convolution kernels with a size of 3. The maximum pooling layer has a kernel size of 2, and there are 64, 32 units in the LSTM, respectively.

B. DBG-G

DBG-G contains a relevance metric component and a graph sparsification component (Fig. 1 (B)). The Multihead Self Attention (MSA) of the relevance metric component calculates the relevance among ROIs to returns a fully connected graph $G \in \mathbb{R}^{B \times N \times N}$. The graph sparsification component is then used to enhance the sparsity of the learned graph structure.

1) Attention-Based Relevance Metrics Function: Closely related manifestations of brain regions, loops, and functional sub-networks associated with emotion and cognitive abnormalities in MDD need to attract more of our attention. Therefore, we quantify the relevance between nodes based on the MSA [32] metric function to learn the relational representations among ROIs.

MSA, developed from self-attention, and a self-attention function is described as mapping a query and a set of key-value pairs to an output. The output is computed as a weighted sum of values. Specifically, for the given input sequence E, self-attention computes Q, K, V with linear projections as follows:

$$Q = E \times W^Q, K = E \times W^K, V = E \times W^V \tag{2}$$

where W^Q, W^K, W^V are the weight matrix of the linear projections. The output of self-attention is calculated as:

$$Attention\left(Q, K, V\right) = \text{Softmax}(\frac{Q \times K^{\text{T}}}{\sqrt{d_k}}) \times V \tag{3}$$

where T is the matrix transpose operator, d_k is the dimension of Q, K, d_v is the dimension of V.

MSA is able to focus on different forms of interaction relationships in different projection spaces to improve the feature representation and generalization ability of the model. And thus, we used MSA to model the inter-ROIs relationships without directly using inner product [33]. MSA performs r different linear projections of Q, K, V. Then, self-attention is computed in parallel to produce d_v/r-dimensional outputs, finally concatenate and project these outputs to produce more effective attention features. The MSA is finally calculated as:

$$MHead\left(Q, K, V\right) = Concat\left(head_1, \ldots, head_r\right) \times W^O$$

$$head_h = Attention\left(Q \times W_h^Q, K \times W_h^K, V \times W_h^V\right) \quad (4)$$

where $h = 1, 2, \ldots, r$, $head_h$ is the hth head self-attention, W_h^Q, W_h^K, W_h^V are the weight matrices of the hth linear projection Q, K, V, with dimensions d_k/r, d_k/r, d_v/r, W^O is the weight matrix of the linear output function.

The DBG-G component of BDANet employs MSA with four parallel heads ($r = 4$) to dynamically adjust the correlation between the ROIs and output a self-attentive feature matrix. In addition, the residual structure is added to prevent the gradient vanishing, and the output O is as in (5). Finally, the fully connected dense graph $G \in \mathbb{R}^{B \times N \times N}$ characterizing the relationship between ROIs is obtained from the inner product of O as in (6), and the elements of G denote the strength of the relationship between ROIs.

$$O = \text{LayerNorm}\left(MultiHead\left(Q, K, V\right) + E\right) \quad (5)$$

$$G = O \times O^{\mathrm{T}} \quad (6)$$

2) Graph Sparsification: The graph relevance metric considers the relationship between each pair of nodes and yields a fully connected graph, which is computationally costly and may introduce noise. For this reason, a K-Nearest Neighbor style graph sparsification method is introduced to sparse the fully connected brain graph, which uses a supervised Deep Learning (DL) approach to parameterize the graph sparsification process and optimizes these parameters based on feedback signals from downstream tasks.

For the fully connected dense graph $G_i \in \mathbb{R}^{N \times N}$ of the ith sample, the topk algorithm [34] is used to retain the connection of each ROI with its K nearest neighbor ROIs and mask the remaining connections to generate a brain graph $A \in \mathbb{R}^{B \times N \times N}$ with the same degree distribution, for the subsequent GCNs, as in (7).

$$A_i^{(n)} = \text{topk}\left(G_i^{(n)} \times \left(\sum_{i=1}^{360} G_i^{(n)}\right)^{-1}\right) \quad (7)$$

where $n = 1, 2, \ldots, N$, $G_i^{(n)}$ is the vector of relationships between the nth node of the graph G_i and the other nodes.

C. Spatial Representation Learning

The graph structure can represent the brain connections and capture the topology of the brain graph. The DBG-G generated learnable brain graph A belongs to non-Euclidean space, cannot be directly used as the input of traditional convolutional neural networks. GCN is introduced to directly model the brain graph A and node features X using two-layer graph convolution in this study. The operation of GCN on A, X at the lth hidden layer is defined as:

$$A' = A + I \quad (8)$$

$$H^{(l+1)} = \sigma\left(DA'DH^{(l)}W^{(l)}\right) \quad (9)$$

where I is the unit matrix, $D \in \mathbb{R}^{B \times N \times N}$ is the degree matrix of the graph A'. $H^{(l)} \in \mathbb{R}^{B \times N \times V}(l > 0)$ is the output feature of the lth layer, V is the dimension of the graph feature vector. When $l = 0$, $H^{(0)}$ is the node feature X. $W^{(l)}$ is the weight

matrix of the lth layer graph convolution operator, $\sigma(\cdot)$ is the activation function ReLU. Finally, the GCN outputs the graph feature $\mathcal{H} = H^{(1)} + H^{(2)}$, $\mathcal{H} \in \mathbb{R}^{B \times N \times V}$. The number of kernels in the graph convolution layer of the spatial representation component is 16.

D. E-Classifier

The E-Classifier employs domain confrontation to mitigate domain-related interference in spatio-temporal features, with the training set with class labels as the source domain and the test set without class labels as the target domain. It accomplishes this by emphasizing accurate domain source discrimination post-Gradient Reversal Layer (GRL) and domain source obfuscation pre-reversal. These training objectives work in opposition to encourage the model to consider both classification accuracy and domain source confusion during supervised training, ultimately achieving domain alignment and addressing exogenous and endogenous shifts in multisite MDD data.

The sample classifier consists of Mixup, Global Average Pooling (GAP), Multilayer Perceptron (MLP), which outputs the classification results of the samples in the source domain. Mixup [35] is used to enhance the representation range of the sample space, thus improving the generalization ability of the model. The sample classifier encodes the GCN graph feature $\mathcal{H}_{source} \in \mathbb{R}^{B \times N \times V}$ for the idx^{th} batch of samples in the source domain and outputs $P_1 \in \mathbb{R}^{B \times 2}$, as in (10).

$$g = \begin{cases} \mathcal{G}\left(\mathcal{H}_{source}\right) if\ idx\ \%5 == 0 \\ \mathcal{H}_{source}\ \ otherwise \end{cases}$$

$$P_1 = \text{MLP}(\text{GlobalAveragePooling}\left(g^{\mathrm{T}}\right)) \quad (10)$$

where $\mathcal{G}$ is the operation of Mixup augmented $\mathcal{H}_{source}$, T is the matrix transpose operator. The domain classifier consists of GRL, GAP and MLP to achieve source and target domain data confusion to solve the data shift. BDANet learns domain invariant features through adversarial training to balance sample classification loss and domain difference loss. The domain classifier filters the temporal features E of the temporal representation component output and the spatial features $\mathcal{H}$ of the GCN output, respectively, to eliminate the domain noise in the spatio-temporal dimension, as in (11).

$$P_2^1, P_2^2 = \text{MLP}(\text{GlobalAveragePooling}(\text{GRL}(E, \mathcal{H}))) \quad (11)$$

where P_2^1, P_2^2 are the outputs of the domain classifier filtering E and $\mathcal{H}$, respectively.

E. Training and Testing BDANet

The advantage of the end-to-end BDANet model is that it avoids errors in manual feature extraction and labeling intermediate features, and trains complex tasks using only the initial class labels and the objective function. For each batch of data, the objective function $\mathcal{L}$ is calculated as

$$\mathcal{L} = 3 * \mathcal{L}_{\text{FocalLoss}} + \mathcal{L}_{\text{CenterLoss}} + \lambda \times \mathcal{L}_{\text{DomainLoss}} \quad (12)$$

where λ is a weight parameter that is used to control the training of the domain classifier. Specifically, λ is set to

TABLE I
PERFORMANCE COMPARISON OF BDANET WITH RELATED REPRESENTATIVE METHODS

		Method	ACC (%)	AUC	F1 Score	SEN	SPE
Deep learning	Gadgil et al.	STGCN	72.33 (±0.54)	0.767 (±1.44%)	0.788 (±0.30%)	0.770 (±0.41%)	0.630(±2.16%)
	Dai et al.	TGCN	71.78 (±1.29)	0.775 (±2.07%)	0.789 (±1.09%)	0.790 (±1.63%)	0.573 (±1.25%)
	Zheng et al.	BFE	75.11 (±0.32)	0.812 (±0.76%)	0.813 (±0.36%)	0.815 (±1.08%)	0.623 (±1.70%)
	Qin et al.	GCN	68.17 (±2.57)	0.727 (±1.58%)	0.782 (±2.40%)	0.743 (±1.18%)	0.563 (±1.70%)
	Kawahara et al.	BrainNetCNN	69.60 (±1.00)	0.760 (±2.24%)	0.773 (±0.30%)	0.713 (±0.47%)	0.650 (±1.63%)
	Li et al.	LSTM	59.33 (±0.82)	0.700 (±1.80%)	0.670 (±0.87%)	0.618 (±1.31%)	0.543 (±1.70%)
Machine learning	-	PCA+RF	54.78 (±1.26)	0.548 (±1.56%)	0.617 (±1.10%)	0.547 (±1.31%)	0.550 (±2.94%)
	-	PCA+RBFSVM	61.33 (±0.72)	0.647 (±0.27%)	0.685 (±0.67%)	0.630 (±0.82%)	0.583 (±0.94%)
Ours	-	**BDANet**	**81.67 (±0.66)**	**0.875 (±0.53%)**	**0.869 (±0.53%)**	**0.878 (±0.62%)**	**0.687(±0.47%)**

0.3 in the training phase and 0 in the testing phase. $\mathcal{L}_{\text{FocalLoss}}$, $\mathcal{L}_{\text{CenterLoss}}$ are used for source domain sample classification, $\mathcal{L}_{\text{DomainLoss}}$ is used for domain classification. $\mathcal{L}_{\text{FocalLoss}}$ is Focal Loss [36], and the focus parameter γ adjusts the sample weights to solve the optimization problem of hard-to-train samples. For notational convenience, we define p_t:

$$p_t = \begin{cases} P_1 \ if \ y_1 = 1 \\ 1 - P_1 \ otherwise \end{cases} \quad (13)$$

where y_1 is the label of source domain samples, P_1 is the prediction probability of sample classifier, the focal loss is computed as $\mathcal{L}_{\text{FocalLoss}}(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$. Center Loss [37] $\mathcal{L}_{\text{CenterLoss}}$, minimizes intra-class distance for more compact features. The center loss is calculated as

$$\mathcal{L}_{\text{CenterLoss}} = \frac{1}{2} \sum_{i=1}^{B} \|C_i - c_{yi}\|_2^2 \quad (14)$$

where B is the number of samples in each batch, C_i is the input feature of MLP in the sample classifier, c_{yi} denotes the feature center of the yi^{th} class. The domain loss $\text{L}_{\text{DomainLoss}}$ consists of two parts: the temporal domain loss, denoted as $\mathcal{L}_1 = \text{CELoss}(P_2^1, y_2)$, and the spatial domain loss, denoted as $\mathcal{L}_2 = \text{CELoss}(P_2^2, y_2)$. CELoss is computed as

$$\text{CELoss}(p, y_2) = \begin{cases} -\log(p) \ if \ y_2 = 1 \\ -\log(1 - p) \ otherwise \end{cases} \quad (15)$$

where y_2 is the domain label, and p is the prediction probability of domain classifier. Finally, the domain loss is calculated as $\mathcal{L}_{\text{DomainLoss}} = \mathcal{L}_1 + \mathcal{L}_2$.

Random sampling divided the entire dataset into 7:3 training and test sets. The test set was then split 1:1 for domain adversarial training and final model evaluation. To prevent BDANet overfitting, GridMask [38] (parameters d1=10, d2=15, ratio=0.4) was applied to augment the training set by a factor of 5. This process was repeated 5 times for result stability and reliability.

F. Interpretations With IG (Integrated Gradients)

IG has higher compatibility and imputation performance than common imputation methods, and it is plug-and-play without modifying the model structure [39]. Thus, we introduce IG to improve the interpretability of end-to-end BDANet by assigning weights to each ROI and evaluating its importance to the model prediction.

Specifically, IG linearly interpolates and gradient sums the paths between the baseline and samples to calculate the contribution of both to the BDANet model separately. The imputation ϕ_j^{IG} for the jth component (ROI) x_j of the ith sample $x = (T_i \in \mathbb{R}^{N*L}, X_i \in \mathbb{R}^{N*D})$ is calculated as

$$\mathcal{F} = x' + \frac{m}{M} \times (x - x') \quad (16)$$

$$\phi_j^{IG}(\Theta, x, x') = (x_j - x_j') \times \sum_{m=1}^{M} \frac{\partial \Theta(\mathcal{F})}{\partial x_j} \times \frac{1}{M} \quad (17)$$

where $m = 1, 2, \ldots, M$, x' is the baseline of x, M denotes the steps in Riemann path integration. $j = 1, 2, \ldots, N$, Θ is the model mapping (BDANet). The attribution of each ROI in each sample is calculated using (16) and (17) to explain their critical role in sample classification.

IV. RESULTS

A. BDANet Classification Performance and Comparison Experiments

The performance metrics for classification include Accuracy (ACC), F1 score, Sensitivity (SEN), Specificity (SPE), and Area Under Receiver Operating Characteristic Curve (AUC) value. The classification performance of the proposed BDANet was compared with traditional ML models (RBFSVM, RF) and the state of the art (BFE [40], STGCN [20], TGCN [41], GCN [13], BrainNetCNN [42], LSTM [43]), as shown in Table I. All results obtained on the same dataset.

DL algorithms outperform ML methods in overall classification. Notably, LSTM can address dynamic temporal patterns, GCN and BrainNetCNN can handle static graph structures, STGCN and TGCN can capture spatio-temporal information, BFE stands out by dynamically attending to ROI relationships, but none match the performance of BDANet. Finally, The ACC of the BDANet model for identifying MDD patients and HC was 81.6%, the AUC was 0.875, the F1 score was 0.869, the SEN was 0.878, the SPE was 0.687, and the confusion matrix, as shown in Fig. 2.

B. BDANet Hyperparameter Settings and Component Contributions

To mimic the actual brain state and enhance brain mapping efficiency, our DBG-G component generates dynamic sparse brain graphs, with sparsity controlled by hyperparameter K. We employ hyperparameter K in a series of parameter comparison experiments to assess the impact of different map

TABLE II
EFFECT OF HYPERPARAMETER K ON CLASSIFICATION METRICS

K	density (%)	ACC (%)	AUC	F1 Score	SEN	SPE
45	12.5	79.33 (±1.65)	0.853 (±1.70%)	0.856 (±1.19%)	0.862 (±1.03%)	0.633 (±2.49%)
54	15	77.17 (±0.83)	0.832 (±1.52%)	0.841 (±0.04%)	0.858 (±0.62%)	0.613 (±0.94%)
90	**25**	**81.67 (±0.66)**	**0.875 (±0.53%)**	**0.869 (±0.53%)**	**0.878 (±0.62%)**	**0.687 (±0.47%)**
126	35	76.84 (±1.16)	0.822 (±1.80%)	0.839 (±0.93%)	0.830 (±1.41%)	0.653 (±1.25%)
171	47.5	77.50 (±1.17)	0.819 (±0.77%)	0.815 (±0.65%)	0.830 (±0.41%)	0.640 (±0.82%)

TABLE III
RESULTS OF THE ABLATION EXPERIMENTS TO ASSESS THE CONTRIBUTION OF COMPONENTS

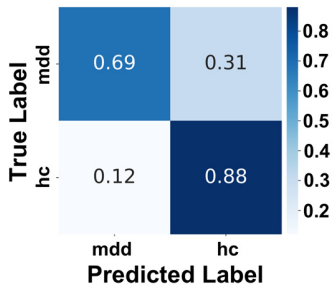|	Paras	Domain-Classifier	DBG-G	GCN-sMRI	ACC (%)	AUC	F1 Score	SEN	SPE
M1	0.054M	NO	YES	YES	75.00 (±0.33)	0.802 (±1.80%)	0.835 (±0.07%)	0.827 (±1.03%)	0.617 (±0.94%)
M2	0.051M	YES	NO	YES	72.50 (±1.50)	0.770 (±1.49%)	0.818 (±0.72%)	0.800 (±1.09%)	0.593 (±1.25%)
M3	0.055M	YES	YES	NO	78.83 (±2.50)	0.857 (±1.86%)	0.854 (±1.59%)	0.862 (±0.24%)	0.663 (±2.49%)
M4	**0.055M**	**YES**	**YES**	**YES**	**81.67 (±0.66)**	**0.875 (±0.53%)**	**0.869 (±0.53%)**	**0.878 (±0.62%)**	**0.687 (±0.47%)**

Fig. 2. The confusion matrix of the proposed BDANet.

sparsity levels on classification, as shown in Table II. The model achieves better recognition performance at K of 90 with a graph density of 25%.

To assess the contribution of each component to MDD identification, model ablation experiments were set up for four different combinations of components:

1) M1 ablation experiment: no (NO) domain classifier, with (YES) DBG-G and GCN-sMRI input.
2) M2 ablation experiment: no DBG-G, with domain classifier and GCN-sMRI input.
3) M3 ablation experiment: without GCN-sMRI input, with domain classifier and DBG-G.
4) M4: BDANet with domain classifier, DBG-G, GCN-sMRI inputs.

The results of the ablation experiments are shown in Table III. Compared with BDANet, the model recognition accuracy decreases by 6.67% when the domain classifier component in E-Classifier is removed and by 9.17% when the DBG-G component is removed, which indicates the effective contribution of the domain classifier and DBG-G components. Only unimodal data for BDANet input and functional features (rs-fMRI) for describe GCN nodes, the model's recognition accuracy also decreases by 2.84%.

C. Results of Hybrid Neural Markers to Identify MDD

The raw rs-fMRI and sMRI data of MDD and HC, high-dimensional and not directly differentiable, were reduced to 2 dimensions using Principal Component Analysis (PCA) for visualization (Fig. 3 (A)), where MDD and HC overlap along both principal component directions. BDANet extract 16-dimensional mixed neural markers to identify MDD, and the 16-dimensional markers were reduced to 2 dimensions by PCA, where each point represents a subject, Fig. 3 (B) shows that the mixed neural markers of MDD and HC are differentiable in both principal component directions. The linear regression line was used to show the differences more clearly, as in Fig. 3 (C), Fig. 3 (D).

D. ROIs With High Attribution for Classification

To interpret the end-to-end BDANet and obtain ROIs with main contribution to classification (high attribution), we used IG to calculate attributions for 360 ROIs and analyzed the functional connectivity strength of the top 19 ROIs in the attribution ranking. The analysis showed that the PI region located in the insula had the highest attribution, followed by the a32pr and p24 regions located in the anterior cingulate gyrus, as shown in Fig. 4 (A).

To analyze the differences in functional connectivity strength between MDD and HC, brain graph of the top 10 ROIs in the attribution ranking were extracted. The analysis showed that in the TE1m, FOP5, PI, Ig and MBelt regions, the functional connectivity strength was significantly lower in MDD patients than in HCs (t-test, p-values 0.0002, 0.0005, 0.0004, 0.0001 and 0.0009), and the TE1m region was the most severely involved (Fig. 4 (B)). In the MDD group, the functional connectivity strength in the Ig region was lower in females than in males (t-test, p-value of 0.0372), as shown in Fig. 4 (C). In the HC group, there was a significant difference in functional connectivity strength between males and females in the STSva and FOP5 regions (t-test, p-value of 0.0127 and 0.0088), as shown in Fig. 4 (D).

In addition, male patients with MDD had significantly lower functional connectivity strength in the TE1m, PI and MBelt regions than healthy males (t-test, p-values of 0.0102, 0.0019, 0.0109, respectively), with the most significant difference in the TE1m region (Fig. 4 (E)). Female MDD patients had significantly lower functional connectivity strength in TE1m, FOP5, PI, Ig and MBelt regions than healthy females (t-test, p-values of 0.0023, 0.0005e-4, 0.0071, 0.0005e-2, 0.0073,
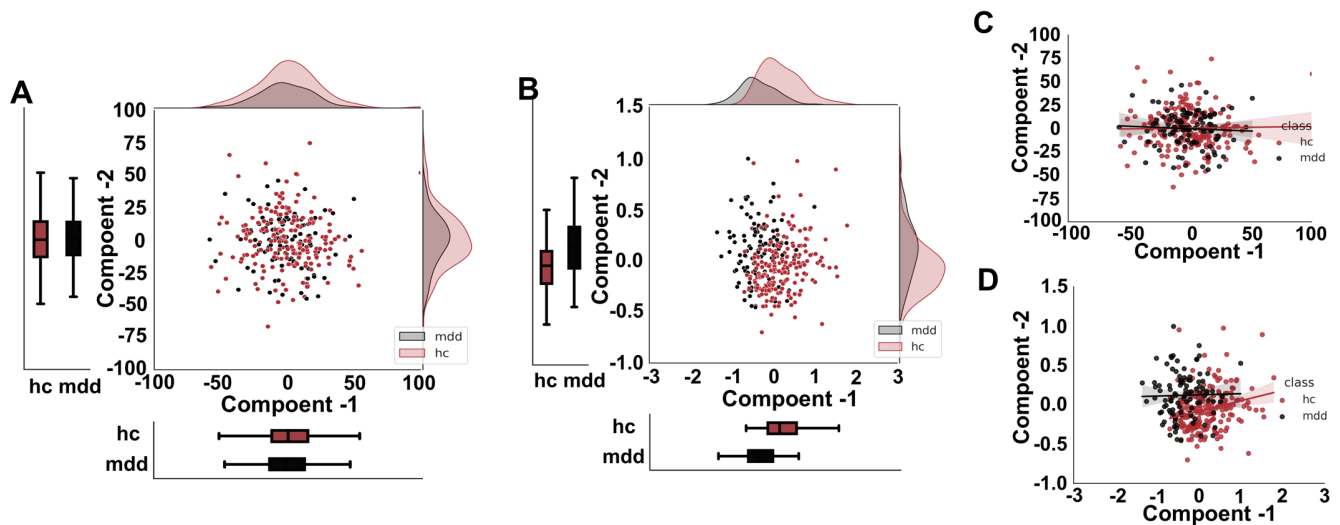
Fig. 3. Raw data and hybrid neural markers of MDD versus HC. (A) Raw data of MDD versus HC; (B) Hybrid neural markers of MDD versus HC; (C) Regression line differences of raw data of MDD versus HC; (D) Regression line differences of hybrid neural markers of MDD versus HC.
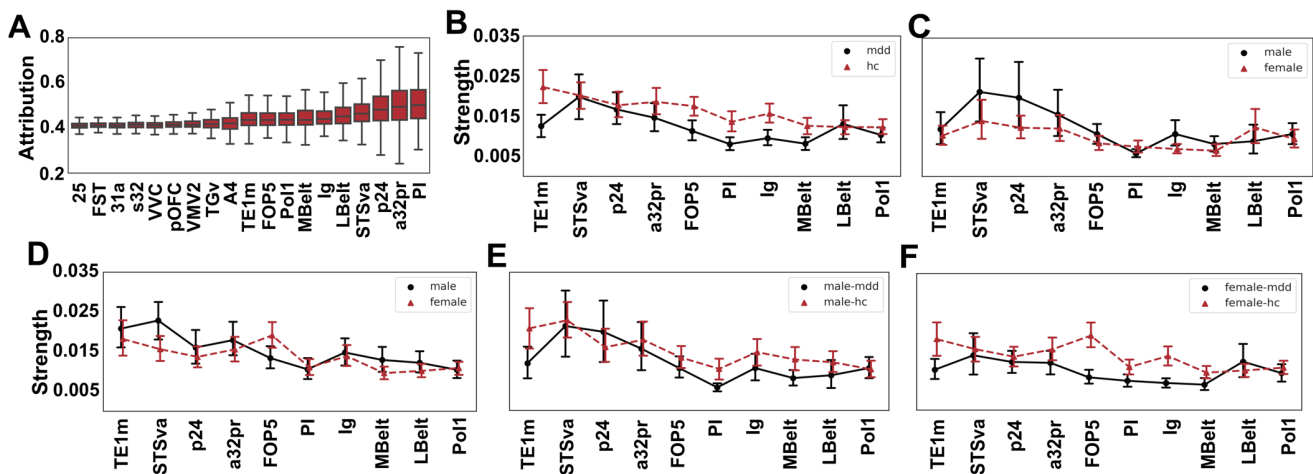

Fig. 4. Categorical contribution of the top 19 ROIs in the attribution ranking and brain functional connectivity strength of the top 10 ROIs in the attribution ranking. (A) Top 19 ROIs in the attribution ranking; Brain functional connectivity strength of (B) MDD and HC; (C) males and females in MDD; (D) males and females in HC; (E) MDD and HC in males, and (F) MDD and HC in females; in the top 10 ROIs in the attribution ranking.

respectively), with the most significant difference in FOP5 region, as shown in Fig. 4 (F).

E. Correlation Between the Functional Connectivity of High Attribution ROIs and Clinical Indicators

Correlations between functional connectivity strength in MDD patients at the 10 categorized high attribution ROIs and their clinical HAMD scores were examined. Notably, as shown in Fig. 5, the functional connectivity strength in the p24 region located in the right anterior cingulate cortex showed a negative correlation with HAMD scores (Spearman's correlation coefficient r=-0.2086, p=0.0372).

F. Abnormalities in the Seven Functional Subnetworks of MDD Patients

To describe the functional connectivity abnormalities in MDD from macroscopic perspective, 360 ROIs were divided into 7 functional subnetworks based on the Yeo-7 functional network atlas [44], Visual Network (VN), Somatomotor
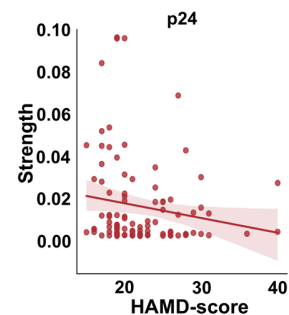

Fig. 5. Correlation between the strength of functional brain connections in the p24 region and HAMD scores in MDD.

Network (SMN), Dorsal Attention Network (DAN), Ventral Attention Network (VAN), Frontoparietal Network (FPN), and Default Mode Network (DMN), Limbic Network (LN).

Compared with HC, all seven functional subnetworks in MDD patients showed varying degrees of reduced functional brain connectivity strength (t-test, p-values of 0.0005e-1,
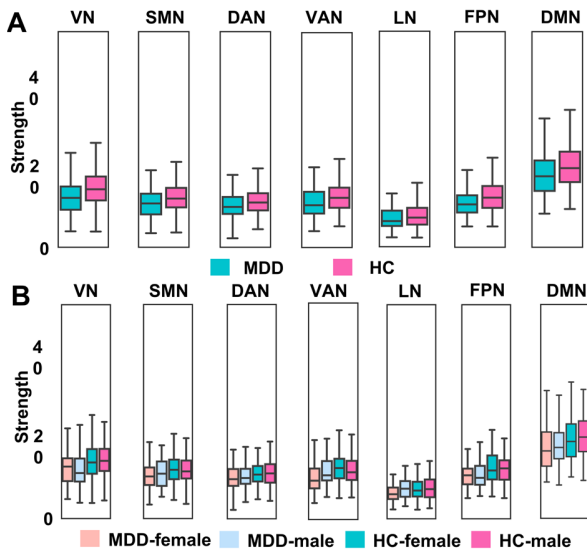
Fig. 6. Connection strength of 7 functional subnetworks of MDD patients and HC. (A) Connection strength of 7 functional subnetworks of MDD and HC; (B) Difference in connection strength of 7 functional subnetworks of MDD and HC between males and females.
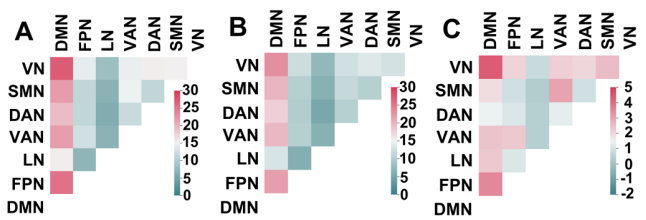

Fig. 7. Connections between the 7 functional subnetworks of MDD and HC. Connections between the 7 functional subnetworks of (A) HC; (B) MDD; (C) Differential functional subnetworks of MDD and HC.

0.0011, 0.0126, 0.0005, 0.0354, 0.0001, and 0.0003, respectively), with more significant abnormalities in the VN, VAN, FPN, and DMN (Fig. 6 (A)). And the functional connectivity intensity of male HC was lower in VAN than that of female HC (t-test, p-value 0.0459), whereas the functional connectivity intensity of male MDD patients was higher in VAN and LN than that of female MDD patients (t-test, p-value 0.0343, 0.0212, respectively), as in Fig. 6 (B).

In addition, the most significant reduction in the connection strength between VN and DMN was found for MDD compared to HC (t-test, p-value 0.0007), followed by the connection strength between DMN and FPN, and between SMN and VAN (t-test, p-value 0.0013, 0.0002, respectively), as shown in Fig. 7 (A), (B), (C).

V. DISCUSSION

The proposed BDANet model was applied to the MDD dataset of the large multisite Rest-meta-MDD consortium, and the hybrid neural markers generated by it were used to characterize MDD, achieving a classification accuracy of more than 81%, outperforming existing DL methods and obtaining more fine-grained MDD-related brain regions than existing studies. Abnormal patterns of MDD that contribute to classification were identified mainly in 10 brain regions, including PI. In further correlation analysis, it was found that the strength of connectivity of p24 with other brain regions was negatively correlated with the severity of depressive symptoms and that functional brain connectivity in MDD had significant gender differences.

We compared the proposed BDANet with RBFSVM and RF, which performed well on small datasets, as well as with the latest relevant methods on large datasets (Table I). The poor performance of these methods on multisite MDD dataset is that they do not comprehensively consider two major issues: (1) The static graph structure is difficult to capture the dynamic

relationships among MDD brain nodes and adapt to dynamic functional states. (2) The collection of MDD large-sample data often relies on multisite collaboration, and there are exogenous and endogenous shifts for multisite data. Both of the issues lead to a sharp decrease in the accuracy of MDD recognition on large datasets. The BDANet solves these resulting in superior multisite MDD recognition performance. Specifically, the DBG-G component of BDANet employs MSA to dynamically focus on the relationship between ROIs that are closely related to the downstream tasks, and its generated homogeneous sparse brain graph enable the GCN to capture higher-order topological features to characterize the heterogeneity of homogeneous networks, which overcomes the limitations of the static brain graph and thus solves MDD heterogeneity. Additionally, our parameter comparison experiments revealed that the BDANet performance is optimized with a 25% density of generated brain graph. The E-Classifier employs domain confrontation to mitigate domain noise in spatio-temporal features, solving the shifts of exogenous and endogenous in multisite MDD data, which can enable BDANet obtain more stable markers to characterize MDD neurobiological features, and improve the overall performance by 6%. And then we reconfirmed the effectiveness of the two components in enhancing MDD recognition performance through ablation experiments.

Most studies have used rs-fMRI data to analyze functional abnormalities in MDD patients. However, the brain has both functional and structural properties, using multimodal data provides a more comprehensive depiction of brain, which is conducive to MDD personalized identification. Therefore, BDANet fuses structural and functional bimodal data and generates hybrid neural markers to comprehensively characterize the brain state of MDD patients. And ablation experiments confirmed improved classification accuracy using bimodal data compared to unimodal data (Table III).

To obtain neurobiological information, IG was used to interpret BDANet, and 10 ROIs with main contributions were extracted, which indicated that MDD may causes functional abnormalities in the insula, cingulate gyrus and auditory cortex, similar to previously reported results [12]. This explains the abnormalities of emotional control [45] and hearing impairment [46] seen in patients with MDD. In addition, analysis of dynamic brain graphs generated by BDANet showed that TE1m had the most significant abnormalities in functional connectivity with other brain regions, possibly related to anhedonia in MDD [47], which are a core symptom of MDD [48]. The degree of abnormal functional connectivity of p24, showed a significant correlation with the severity of clinical symptoms of MDD.

Our study observed greater brain involvement in females with MDD compared to males, in line with the findings of Young et al. [49]. In addition, we found significant gender differences in abnormal functional connectivity between the Ig related to memory function [50] and other brain regions in MDD, which may account for the inconsistent rate of memory loss in male and female MDD. Compared to HCs, male MDD exhibited more significant abnormalities in the TE1m region related to anhedonia while female MDD showed greater abnormalities in FOP5 associated with memory function [50]. This can lead to more severe anhedonia in male MDD and more rapid memory loss in female patients. Female MDD showed significantly weaker connections in VAN and LN with other subnetworks, within the 7 functional subnetworks. These findings indicate that sex-differentiated analysis can enhance our understanding of MDD pathogenesis and that brain regions with significant sex variations may serve as valuable clinical indicators to enhance personalized MDD interventions [51], [52].

This study has the following research gap: (1) we mainly analyzed the abnormalities of functional brain connectivity in MDD, and the structural data were only used for nodal description of the GCN; future work will further fuse structural and functional abnormalities to find the best brain abnormality pattern for identifying MDD. (2) Our subject age range was limited to 18-65 years, and the results may not be applicable to children or elderly individuals with MDD. In addition, to make BDANet more consistent with clinical practical applications, future work will consider comorbidities, medication use, disease duration and number of episodes as factors for identifying MDD to further improve the accuracy of personalized identification of MDD.

VI. CONCLUSION

This study demonstrates the effectiveness of the proposed end-to-end BDANet for cross-domain identification on large multisite MDD datasets, achieving more than 81% classification accuracy, addressing the heterogeneity of MDDs to some extent, and alleviating the data shift. We found that regions with high attribution of classification were mainly located in the insula, cingulate cortex and auditory association cortex, and that there were gender differences in the strength of MDD connectivity in specific brain regions and functional subnetworks. The BDANet model is expected to enhance personalized identification of MDD accuracy and has potential applications for treatment or intervention in MDD-targeted brain regions.

REFERENCES

[1] M. M. Sanchez et al., "A machine learning algorithm to discriminating between bipolar and major depressive disorders based on resting EEG data," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 2635–2638, doi: 10.1109/EMBC48229.2022.9871453.

[2] G. S. Malhi and J. J. Mann, "Course and prognosis," *Lancet*, vol. 392, no. 10161, pp. 2299–2312, Nov. 2018, doi: 10.1016/S0140-6736(18)31948-2.

[3] L. A. Maglanoc et al., "Multimodal fusion of structural and functional brain imaging in depression using linked independent component analysis," *Hum. Brain Mapping*, vol. 41, no. 1, pp. 241–255, Jan. 2020, doi: 10.1002/HBM.24802.

[4] M. Mousavian, J. Chen, and S. Greening, "Depression detection using feature extraction and deep learning from sMRI images," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 1731–1736, doi: 10.1109/ICMLA.2019.00281.

[5] J. Jin and L. Huang, "A region-based feature extraction method for rs-fMRI of depressive disorder classification," in *Proc. Int. Conf. Comput. Vis., Image Deep Learn. (CVIDL)*, Jul. 2020, pp. 707–710, doi: 10.1109/CVIDL51233.2020.00051.

[6] E. Zendehrouh et al., "Aberrant functional network connectivity transition probability in major depressive disorder," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1493–1496, doi: 10.1109/EMBC44109.2020.9175872.

[7] D. Zhi et al., "Abnormal dynamic functional network connectivity and graph theoretical analysis in major depressive disorder," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 558–561, doi: 10.1109/EMBC.2018.8512340.

[8] D. M. Khan, K. Masroor, M. F. M. Jailani, N. Yahya, M. Z. Yusoff, and S. M. Khan, "Development of wavelet coherence EEG as a biomarker for diagnosis of major depressive disorder," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4315–4325, Mar. 2022, doi: 10.1109/JSEN.2022.3143176.

[9] B. Zhang, G. Yan, Z. Yang, Y. Su, J. Wang, and T. Lei, "Brain functional networks based on resting-state EEG data for major depressive disorder analysis and classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 215–229, 2021, doi: 10.1109/TNSRE.2020.3043426.

[10] N. Amirkhan, F. Hasanzadeh, M. Mohebbi, and R. Rostami, "Brain network analysis based on frontal cortical EEG sources in major depressive disorder (MDD)," in *Proc. 28th Nat. 6th Int. Iranian Conf. Biomed. Eng. (ICBME)*, Iran, Nov. 2021, pp. 276–282, doi: 10.1109/ICBME54433.2021.9750285.

[11] C. Zhuo et al., "The rise and fall of MRI studies in major depressive disorder," *Translational Psychiatry*, vol. 9, no. 1, pp. 1–14, Dec. 2019, doi: 10.1038/S41398-019-0680-6.

[12] S. Lui, X. J. Zhou, J. A. Sweeney, and Q. Gong, "Psychoradiology: The frontier of neuroimaging in psychiatry," *Radiology*, vol. 281, no. 2, pp. 357–372, Nov. 2016, doi: 10.1148/RADIOL.2016152149.

[13] K. Qin et al., "Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites," *eBioMedicine*, vol. 78, Apr. 2022, Art. no. 103977, doi: 10.1016/J.EBIOM.2022.103977.

[14] V. I. Müller, E. C. Cieslik, I. Serbanescu, A. R. Laird, P. T. Fox, and S. B. Eickhoff, "Altered brain activity in unipolar depression revisited: Meta-analyses of neuroimaging studies," *JAMA Psychiatry*, vol. 74, no. 1, p. 47, Jan. 2017, doi: 10.1001/JAMAPSYCHIATRY.2016.2783.

[15] J. Yang et al., "Development and evaluation of a multimodal marker of major depressive disorder," *Hum. Brain Mapping*, vol. 39, no. 11, pp. 4420–4439, Nov. 2018, doi: 10.1002/HBM.24282.

[16] C.-G. Yan et al., "Reduced default mode network functional connectivity in patients with recurrent major depressive disorder," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 18, pp. 9078–9083, 2019, doi: 10.1073/pnas.1900390116.

[17] M. Xia et al., "Reproducibility of functional brain alterations in major depressive disorder: Evidence from a multisite resting-state functional MRI study with 1,434 individuals," *NeuroImage*, vol. 189, pp. 700–714, Apr. 2019, doi: 10.1016/J.NEUROIMAGE.2019.01.074.

[18] X. Li et al., "BrainGNN: Interpretable brain graph neural network for fMRI analysis," *Med. Image Anal.*, vol. 74, Dec. 2021, Art. no. 102233, doi: 10.1016/J.MEDIA.2021.102233.

[19] B.-H. Kim and J. C. Ye, "Understanding graph isomorphism network for rs-fMRI functional connectivity analysis," *Frontiers Neurosci.*, vol. 14, p. 630, Jun. 2020, doi: 10.3389/FNINS.2020.00630.

[20] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, "Spatio-temporal graph convolution for resting-state fMRI analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* (Lecture Notes in Computer Science), vol. 12267, 2020, pp. 528–538.

[21] A. El-Gazzar, R. M. Thomas, and G. van Wingen, "Dynamic adaptive spatio-temporal graph convolution for fMRI modelling," in *Proc. Int. Workshop Mach. Learn. Clinical Neuroimaging* (Lecture Notes in Computer Science), vol. 13001, 2021, pp. 125–134, doi: 10.1007/978-3-030-87586-2_13.

[22] T. Azevedo, L. Passamonti, P. Lio, and N. Toschi, "Towards a predictive spatio-temporal representation of brain data," 2020, *arXiv:2003.03290*.

[23] R. B. Pachori, "Time-frequency analysis techniques and their applications," in *Time-Frequency Analysis Techniques and Their Applications*. Boca Raton, FL, USA: CRC Press, Jan. 2023, pp. 1–217, doi: 10.1201/9781003367987.

[24] R. Tao et al., "Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2022, pp. 21157-21167, doi: 10.1109/CVPR52688.2022.02051.

[25] S. Zhang et al., "Multimodal MRI reveals alterations of the anterior insula and posterior cingulate cortex in bipolar II disorders: A surface-based approach," *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, vol. 116, Jun. 2022, Art. no. 110533, doi: 10.1016/J.PNPBP.2022.110533.

[26] P. Brenne Kemmer, F. DuBois Bowman, H. Mayberg, and Y. Guo, "Quantifying the strength of structural connectivity underlying functional brain networks," 2017, *arXiv:1703.04056*.

[27] Z. Yao et al., "Structural alterations of the brain preceded functional alterations in major depressive disorder patients: Evidence from multimodal connectivity," *J. Affect. Disorders*, vol. 253, pp. 107–117, Jun. 2019, doi: 10.1016/J.JAD.2019.04.064.

[28] C. Yan and Y. Zang, "DPARSF: A MATLAB toolbox for 'pipeline' data analysis of resting-state fMRI," *Frontiers Syst. Neurosci.*, vol. 4, p. 1377, May 2010, doi: 10.3389/FNSYS.2010.00013.

[29] M. F. Glasser et al., "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, Jul. 2016, doi: 10.1038/NATURE18933.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[33] X. Kan, H. Cui, Y. Guo, and C. Yang, "Effective and interpretable fMRI analysis via functional brain network generation," 2021, *arXiv:2107.11247*.

[34] Y. Chen, L. Wu, and M. J. Zaki, "Reinforcement learning based graph-to-sequence model for natural question generation," 2019, *arXiv:1908.04942*.

[35] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9911, Oct. 2016, pp. 499–515, doi: 10.1007/978-3-319-46478-7_31.

[38] P. Chen, S. Liu, H. Zhao, and J. Jia, "GridMask data augmentation," 2020, *arXiv:2001.04086*.

[39] B. Sanchez-Lengeling et al., "Evaluating attribution for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5898–5910.

[40] G. Zheng et al., "An attention-based multi-modal MRI fusion model for major depressive disorder diagnosis," *J. Neural Eng.*, vol. 20, no. 6, Dec. 2023, Art. no. 066005, doi: 10.1088/1741-2552/AD038C.

[41] P. Dai et al., "Classification of recurrent major depressive disorder using a new time series feature extraction method through multisite rs-fMRI data," *J. Affect. Disorders*, vol. 339, pp. 511–519, Oct. 2023, doi: 10.1016/J.JAD.2023.07.077.

[42] J. Kawahara et al., "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, Feb. 2017, doi: 10.1016/J.NEUROIMAGE.2016.09.046.

[43] H. Li and Y. Fan, "Brain decoding from functional MRI using long short-term memory recurrent neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* (Lecture Notes in Computer Science), vol. 11072, 2018, pp. 320–328, doi: 10.1007/978-3-030-00931-1_37.

[44] B. T. T. Yeo et al., "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *J. Neurophysiol.*, vol. 106, no. 3, pp. 1125–1165, 2011.

[45] L. Q. Uddin, J. S. Nomi, B. Hébert-Seropian, J. Ghaziri, and O. Boucher, "Structure and function of the human insula," *J. Clin. Neurophysiol.*, vol. 34, no. 4, pp. 300–306, Jul. 2017, doi: 10.1097/WNP.0000000000000377.

[46] F. Lu et al., "Anomalous intrinsic connectivity within and between visual and auditory networks in major depressive disorder," *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, vol. 100, Jun. 2020, Art. no. 109889, doi: 10.1016/J.PNPBP.2020.109889.

[47] X.-H. Yang et al., "Anhedonia correlates with abnormal functional connectivity of the superior temporal gyrus and the caudate nucleus in patients with first-episode drug-naive major depressive disorder," *J. Affect. Disorders*, vol. 218, pp. 284–290, Aug. 2017, doi: 10.1016/J.JAD.2017.04.053.

[48] S. Lu et al., "Aberrant interhemispheric functional connectivity in major depressive disorder with and without anhedonia," *BMC Psychiatry*, vol. 22, no. 1, pp. 1–9, Nov. 2022, doi: 10.1186/s12888-022-04343-x.

[49] E. A. Young et al., "Sex differences in response to citalopram: A STAR∗D report," *J. Psychiatric Res.*, vol. 43, no. 5, pp. 503–511, Feb. 2009, doi: 10.1016/J.JPSYCHIRES.2008.07.002.

[50] R. Kinno and K. Ono, "High-density lipoprotein-cholesterol and neuroaging: Memory and gyrification of the insular and frontal opercular cortex," in *Factors Affecting Neurological Aging: Genetics, Neurology, Behavior, and Diet*. New York, NY, USA: Academic, Jan. 2021, pp. 259–268, doi: 10.1016/B978-0-12-817990-1.00023-8.

[51] D. A. Bangasser and A. Cuarenta, "Sex differences in anxiety and depression: Circuits and mechanisms," *Nature Rev. Neurosci.*, vol. 22, no. 11, pp. 674–684, Sep. 2021, doi: 10.1038/S41583-021-00513-0.

[52] J.-Y. Chuang et al., "Adolescent major depressive disorder: Neuroimaging evidence of sex difference during an affective Go/No-go task," *Frontiers Psychiatry*, vol. 8, p. 119, Jul. 2017, doi: 10.3389/FPSYT.2017.00119.