Check for updates

*CORRESPONDENCE
Nojod M. Alotaibi
✉ nalotaibi0351@stu.kau.edu.sa
Areej M. Alhothali
✉ aalhothali@kau.edu.sa

Multi-atlas ensemble graph neural network model for major depressive disorder detection using functional MRI data

Nojod M. Alotaibi*, Areej M. Alhothali* and Manar S. Ali

Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Major depressive disorder (MDD) is one of the most common mental disorders, with significant impacts on many daily activities and quality of life. It stands as one of the most common mental disorders globally and ranks as the second leading cause of disability. The current diagnostic approach for MDD primarily relies on clinical observations and patient-reported symptoms, overlooking the diverse underlying causes and pathophysiological factors contributing to depression. Therefore, scientific researchers and clinicians must gain a deeper understanding of the pathophysiological mechanisms involved in MDD. There is growing evidence in neuroscience that depression is a brain network disorder, and the use of neuroimaging, such as magnetic resonance imaging (MRI), plays a significant role in identifying and treating MDD. Rest-state functional MRI (rs-fMRI) is among the most popular neuroimaging techniques used to study MDD. Deep learning techniques have been widely applied to neuroimaging data to help with early mental health disorder detection. Recent years have seen a rise in interest in graph neural networks (GNNs), which are deep neural architectures specifically designed to handle graph-structured data like rs-fMRI. This research aimed to develop an ensemble-based GNN model capable of detecting discriminative features from rs-fMRI images for the purpose of diagnosing MDD. Specifically, we constructed an ensemble model by combining functional connectivity features from multiple brain region segmentation atlases to capture brain complexity and detect distinct features more accurately than single atlas-based models. Further, the effectiveness of our model is demonstrated by assessing its performance on a large multi-site MDD dataset. We applied the synthetic minority over-sampling technique (SMOTE) to handle class imbalance across sites. Using stratified 10-fold cross-validation, the best performing model achieved an accuracy of 75.80%, a sensitivity of 88.89%, a specificity of 61.84%, a precision of 71.29%, and an F1-score of 79.12%. The results indicate that the proposed multi-atlas ensemble GNN model provides a reliable and generalizable solution for accurately detecting MDD.

KEYWORDS

major depressive disorder, RS-fMRI, brain functional connectivity network, deep learning, graph neural network, data oversampling, ensemble model

1 Introduction

Mental disorders are considered a major cause of disability and are correlated with a higher risk of early death (Freeman, 2022). There is a strong correlation between mental disorders, suicide, and other chronic diseases, such as cardiovascular disease, diabetes, and HIV/AIDS (Freeman, 2022). Major depressive disorder (MDD) is a serious mental disorder

that leads to severe disruptions in the ways in which an individual expresses emotions, rationalizes, and engages in social interactions (Zhuo et al., 2019). It is among the most prevalent mental disorders. According to the Global Burden of Diseases, Injuries and Risk Factors Study 2020 (GBD 2020), MDD prevalence increased by 27.6%, since the COVID-19 outbreak (Santomauro et al., 2021). Before the pandemic, an estimated 193 million people suffered from MDD. After the pandemic, the number of individuals experiencing MDD jumped to 246 million (Santomauro et al., 2021). Furthermore, MDD was the second primary cause of disability in the world, contributing significantly to the global disease burden in 2019 (GBD 2019 Mental Disorders Collaborators, 2022).

At present, diagnosing MDD is primarily dependent on criteria provided by the Diagnostic and Statistical Manual of Mental Disorders (DSM) along with self-reported symptoms obtained through clinical questionnaires and clinical interviews (Zhuo et al., 2019). As MDD shares many symptoms with other mental disorders, distinguishing them can be challenging, requiring the expertise of highly trained psychiatrists. Thus, such methods of diagnosis are time consuming and unable to detect MDD in its earliest stages. Moreover, the lack of recognized neurological biomarkers for MDD makes the diagnostic and prognostic process difficult (Zhuo et al., 2019). Actually, there is evidence from neuroimaging studies that it is closely related to abnormality functionally and structurally in certain regions of the brain (Pilmeyer et al., 2022). In the last decade, neuroimaging has gained considerable prominence due to advances in computing technology, which have enabled researchers to obtain a deeper comprehension of the brain's mechanisms (Tulay et al., 2019). Resting-state functional magnetic resonance imaging (rs-fMRI) has become a widely used neuroimaging method for analyzing MDD as it offers insights into the brain's functional connectivities by detecting dynamic changes in blood oxygenation-dependent signals (BOLDs) of patients (Pilmeyer et al., 2022). Several studies have established that functional connectivity networks (FCNs) are reliable data sources for diagnosing MDD as they represent correlations between different brain regions of interest (ROIs) (Yan et al., 2020). The rs-fMRI findings indicated that patients with MDD had abnormal functional connectivity related to the central executive network (CEN), the default mode network (DMN), and the salience network (SN) (Dai et al., 2019). Several studies based on rs-fMRI also found that MDD patients have abnormal brain function in several cortical and subcortical structures, including the prefrontal cortex, amygdala, insula, hippocampus, and precuneus (Dai et al., 2019).

Currently, most neuroimaging-based MDD studies are primarily focused on exploring functional imaging biomarkers, by utilizing different machine learning (ML) and deep learning (DL) approaches to analyze complex medical images and assist physicians in making an accurate diagnosis. Among these approaches are support vector machines (SVM), Random Forest (RF), 3D convolutional neural networks (CNNs), BrainNetCNN, and ResNet50. Graph neural networks (GNNs) are recently emerging as a new type of DL technique capable of analyzing graph-structured data, such as 4D rs-fMRI, which are difficult to model directly using conventional DL methods created for

Euclidean data such as images and texts (Ahmedt-Aristizabal et al., 2021). Several studies have been conducted to detect MDD using rs-fMRI, including, Yao et al. (2020), Qin et al. (2022), and Noman et al. (2024). Even though the studies mentioned above have proven successful in diagnosing MDD, they only used a single brain region segmentation template (atlas). Nevertheless, the heterogeneity and complexity of brain regions cannot be fully captured using a single brain atlas (Liu et al., 2018). To tackle this issue, several studies have been developed that use multi-atlas fusion to detect mental disorders, including Liu et al. (2016), Lei et al. (2020), Chu et al. (2022), Xia et al. (2023), and Lee et al. (2024). However, most existing studies have concatenated or averaged two or three atlases to provide complementary information about the brain. They also trained their models using a relatively small sample size. In fact, rs-fMRI datasets have a smaller sample size than traditional datasets used in ML since they are more challenging to obtain and analyze. As a result, applying DL methods to rs-fMRI datasets may lead to model underfitting or overfitting (Liu et al., 2023). Consequently, large-scale datasets are more suitable for training DL models, which will enhance the models' classification performance and make them more dependable and generalizable (Alzubaidi et al., 2023). To address this issue, many studies have suggested oversampling techniques to generate new samples from the rs-fMRI dataset, including Venkatapathy et al. (2023) and Liu et al. (2023).

This research is designed to develop an ensemble-based GNN model able to categorize rs-fMRI data into MDD and healthy control (HC) subjects. Our model consists of four GNNs, each of which is trained using FCNs derived from a different brain atlas. Models based on multiple atlases can capture the complexity of the human brain structure and detect more discriminatory features than models based solely on an atlas. Moreover, the Synthetic Minority Over-sampling (SMOTE) technique is applied to produce a wide variety of data suitable for training our ensemble model. To assess the effectiveness of the proposed model, a comprehensive evaluation is conducted on a substantial multi-site MDD dataset, allowing for generalizable and reliable classification performance for MDD.

The rest of the paper is organized as follows. In Section 2, we provide a brief overview of the most relevant studies. Our proposed method is introduced in Section 3 by providing an overview of the materials and methods used to develop it. Section 4 presents experimental settings and evaluation metrics, followed by a discussion of the classification results. Finally, the conclusion of this paper is summarized in Section 5.

2 Related works

Recently, a number of studies have been conducted on the detection of MDD based on rs-fMRI features. In this section, we have grouped prior studies into three main groups according to their methodological focus: (1) GNN-based studies, (2) studies that utilize oversampling to overcome class imbalance, and (3) studies that integrate multi-atlas information using ensemble methods.

Among the GNN-based studies, Yao et al. (2020) introduced a temporal-adaptive graph convolutional network (TAGCN) model for capturing both dynamic and static information about functional

connectivity (FC) patterns toward the diagnosis of MDD. The rs-fMRI time-series signals were first extracted from each ROI and divided into multiple overlapped blocks using fixed-size sliding windows. An adaptive graph convolutional layer was then applied to produce dynamic connectivity matrices for each block. Afterwards, convolution operations were performed on each ROI along various blocks to obtain the temporal dynamics of the entire time series. Finally, the classification of MDD was accomplished through a fully connected layer and a softmax function. The TAGCN method performed better than the other methods and achieved accuracy of 73.8%. Noman et al. (2024) designed a novel graph autoencoder (GAE) architecture to encode the topological structure and node content into low-dimensional latent representations and train a decoder for graph reconstruction. In addition, the authors developed a framework to learn graph embeddings in brain networks in an unsupervised manner. A deep fully connected neural network (FCNN) was then trained to recognize MDD from HC using the learned embeddings. It was found that GAE-FCNN significantly outperformed other existing methods, attaining the highest accuracy of 72.5%.

Despite the strong classification results reported in these studies, they were based on single-site datasets with small sample sizes, raising concerns regarding generalizability across diverse populations and imaging conditions. To address this limitation, several studies have employed GNN on multi-site datasets to improve robustness. Kong et al. (2021) proposed a spatiotemporal graph convolutional network (STGCN) framework trained on rs-FMRI data from two sites to diagnose MDD patients and predict their response to treatment. First, dynamic FCN were derived using a sliding temporal window method. Then, the spatial graph attention convolution module (SGAC) was created to enhance feature learning, while prior pooling was implemented to reduce feature dimensions. The temporal fusion module was designed for capturing dynamic FCN features among adjacent sliding windows together with the SGAC module. The STGCN achieved the highest diagnostic accuracy of 84.14 and 83.93% on the two sites, respectively. The STGCN also achieved an accuracy of 89.63% in predicting treatment response. In Qin et al. (2022), a model was developed to support early and precise diagnosis of MDD by applying graph convolutional network (GCN) to a large multi-site MDD dataset (17 sites). A whole-brain functional network was used to train the GCN model to recognize MDD patients from HC. Further, the authors conducted a subgroup analysis to distinguish between patients with first-episode drug-naive (FEDN) and HC, recurrent and HC, and recurrent MDD and FEDN. Results showed that GCN had an accuracy of 81.5%, which was higher than that of other competing classifiers.

A second group of studies examined the issue of class imbalance, which affects classification performance in neuroimaging experiments. These studies have used oversampling techniques to overcome this problem. Venkatapathy et al. (2023) proposed an ensemble model that combines GraphSAGE, Graph attention network (GAT), and GCN models trained on Dosenbach's 160 functional ROIs (Dose) atlas for classification of MDD and HC. A further analysis of subgroups between patients with REC and FEDN was conducted. Moreover, they performed random oversampling by copying data from minority classes, and random

undersampling by selecting data from majority classes in order to achieve a balanced sample size. In the effort to classify MDD from HC, upsampling produced 71.40% accuracy, while downsampling yielded 70.24% accuracy. With upsampled REC and HC data, the model yielded 91.6% accuracy; in contrast, with downsampled data, the model produced 68.78% accuracy. Ultimately, the model produced accuracy of 77.78% after upsampling REC with FEDN, and accuracy of 71.96% after downsampling. Liu et al. (2023) created a spatial-temporal data-augmentation-based classification scheme (STDAC) that can fuse spatial-temporal information, enhance classification performance, and expand the sample size. A spatial data augmentation (SDA) module was built using KNN-like techniques based on the Automated Anatomical Labeling (AAL) atlas, and a temporal data augmentation (TDA) module was constructed using discontinuous time series from the original data time period. Finally, the features from the aforementioned two modules were combined using a tensor fusion technique. Extensive experiments were conducted on the Alzheimer's Disease Neuroimaging (ADNI) and REST-meta-MDD datasets in order to assess the efficacy of the proposed model. Additionally, the STDAC model was trained with some traditional ML classifiers, such as Random Forest (RF) and Neural Network (NN). It turned out that the proposed STDAC model with NN yielded the best accuracy for the two datasets, at 84.170 and 63.406%, respectively.

However, all of the above studies had a common limitation. They relied on only one brain atlas to define brain regions, which may result in incomplete representations of the brain's connectivity patterns. As a result, several studies have investigated the fusion of multiple atlases to provide a more comprehensive understanding of brain connectivity. Among these studies, Xia et al. (2023) introduced a two-channel graph neural network (DepressionGraph) model to efficiently aggregate the graph information from the two channels based on the node number and the node feature number. In addition, a transform-encoder architecture was employed to extract relevant information from time-series FCN. Using a majority voting method, the prediction results of models trained with three different atlases were integrated to improve the MDD prediction model. DepressionGraph demonstrated superior performance when compared to previous studies, with an accuracy of 71.48%. In Lee et al. (2024), an innovative multi-atlas fusion technique using GCN for MDD diagnosis was developed that integrates both early and late fusion approaches. During the early fusion phase, ROI signals are combined from each individual atlas. Next, a holistic FCN is created by calculating FCs between all ROI signals derived from three atlases. In the late fusion phase, a soft voting ensemble method was employed to combine feature vectors derived from FCs extracted from each atlas and the holistic FCN. It was found that the proposed model based on the GCN achieved the highest performance with an accuracy of $68.88 \pm 2.55\%$.

In summary, a majority of depression detection studies used single-site datasets with small sample sizes, which limits generalizability and causes significant variations in model's performance. These approaches are often based upon a single atlas, which present considerable limitations, including an incomplete representation of brain structures and a potential bias imposed by

TABLE 1 Demographic information of the 1,563 study subjects.

Group	Number of subjects	Male	Female	Average ± standard deviation (%)	
				Age	Education
MDD	810	293	517	34.39 ± 11.55	11.96 ± 3.37
HC	753	307	446	34.61 ± 13.17	13.57 ± 3.42

the specific atlas. To address these challenges, we proposed a multi-atlas ensemble model leveraging a graph neural network (GNN) for categorizing rs-fMRI data into MDD and healthy controls (HC). This model improves representational power by incorporating diverse brain structures from multiple atlases, thus enhancing robustness and sensitivity to subtle connectivity changes. Besides that, we used the Synthetic Minority Over-sampling Technique (SMOTE) to produce a variety of training data allowing us to improve generalizability and reduce overfitting.

3 Materials and methods

This section provides a detailed description of the materials and methodologies employed in this study to differentiate MDD patients from HCs using rs-fMRI time series.

3.1 REST-meta-MDD dataset

In this research, we utilize a dataset from a public open-access data repository called the REST-meta-MDD consortium, which is the largest public MDD dataset to our knowledge (Chen et al., 2022). The dataset consists of 2, 428 subjects from 25 sites, including 1, 300 patients with MDD (826 are females and 474 are males) and 1, 128 HC (Chen et al., 2022; Yan et al., 2019). At each site, phenotypic data such as age, sex, episode status (recurrent or first episode), medication status, illness duration, and the 17-item Hamilton Depression Rating Scale (HAMD) were gathered. There are also two types of imaging data available, T1-weighted sMRI and rs-fMRI (Chen et al., 2022). Participants in REST-meta-MDD were required to provide written informed consent before participating, and the local Institutional Review Board approved the collection of data at each site (Chen et al., 2022; Yan et al., 2019).

3.1.1 Data preprocessing

At each local site, resting-state functional MRI and three-dimensional structural T1-weighted MRI images were acquired for all participants. The Data Processing Assistant (DPARSF) toolbox was used to perform a unified image preprocessing protocol. Among the preprocessing steps were discarding the first ten volumes to ensure magnetization equilibrium, correcting slice timing and head motion, normalizing to the Montreal Neurological Institute (MNI) template, smoothing, detrending and band-pass filtering, excluding sites with fewer than ten subjects, and regressing covariates (Yan et al., 2019). Further, our analysis of the time series of brain regions using targeted atlases revealed that some brain regions of some subjects were missing signals, so these subjects were

excluded. Finally, 1, 563 participants from 16 sites were included in our study, of which 810 MDD patients and 753 HC. At the end of this process, the data for each subject was transformed into a matrix of size $T \times N$, where T denotes the number of time points and N denotes the number of brain ROIs. In this case, we used 140 time points between each node. Detailed demographic information about the subjects included in our study is provided in Table 1.

To overcome the challenges related to multi-site data, we performed ComBat harmonization to mitigate site-varying effects (also called batch effects), while keeping critical biological covariates such as sex and age (El-Gazzar et al., 2023). ComBat relies on a multivariate linear mixed-effects regression framework, which was developed for batch effect corrections in genomic studies (Johnson et al., 2007; El-Gazzar et al., 2023). This model utilizes empirical Bayes to estimate both biological and non-biological effects and effectively remove the estimated additive and multiplicative site effects (El-Gazzar et al., 2023). As our data consists of time series features represented as $140 \times N$ matrices, where 140 is the number of time points and N is the number of brain ROIs, our ComBat harmonization model can be expressed mathematically as follows:

$$Y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i\epsilon_{ij} \tag{1}$$

In this case, Y_{ij} denotes the time series value for subject j at site i, α is the overall mean, X_{ij} is a design matrix for the biological covariates (i.e. sex and age), β is the regression coefficients for covariates, γ_i and δ_i are the site-specific effects (additive and multiplicative effects), and ϵ_{ij} represents the residual error.

3.2 Proposed model

This paper presents a multi-atlas ensemble GNN model for categorizing rs-fMRI data into MDD and HC subjects. Figure 1 illustrate the overall pipeline of our framework, including time-series extraction, data oversampling, FCN construction, graph generation, and ensemble-based GNN classification. We first extracted time series for each atlas from the rs-fMRI data using four popular brain segmentation atlases, including Dosenbach's 160 functional ROIs (Dose) (Dosenbach et al., 2010), Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002), Craddock's clustering 200 ROIs (CK) (Craddock et al., 2012), and Harvard–Oxford (HO) atlas (Kennedy et al., 1998). In particular, AAL and HO represent anatomical atlases, while Dose and CK represent functional atlases. The use of a variety of brain atlases contributes to improving the accuracy of diagnosis of brain disorders (Xia et al., 2023). Secondly, the Synthetic Minority Over-sampling (SMOTE) technique is adopted to generate a wide variety of data suitable for training our ensemble model. The SMOTE

method is computationally efficient and easy to implement, making it a practical choice for addressing class imbalances in our study (Mansourifar and Shi, 2020). Thirdly, FCNs are generated for each atlas separately using Pearson correlation coefficients (PC). These FCNs are then represented as graphs. Lastly, four homogeneous GNN models are constructed to complement our ensemble model, each having been trained with graphs derived from a different brain atlas. We used the graph attention network (GAT) models as the base models for our multi-atlas ensemble model.

3.2.1 Graph attention network

A graph attention network (GAT) is an innovative neural network architecture that employs masked self-attentional layers to address the limitations of prior approaches utilizing graph convolutions or their approximations (Veličković et al., 2017). The GAT model implements a self-attention strategy in which the representation of each node is computed by repeatedly attending to its neighbors. Initially, a graph attention layer receives a set of node features as input, $h = \{\vec{h_1}, \vec{h_2}, ..., \vec{h_N}\}$, where N represents the number of nodes. This layer generate new node features, whose dimensions may differ from those of the input features. Next, self-attention is applied to derive attention coefficients e_{ij} between each pair of nodes, as indicated in Equation 2 (Veličković et al., 2017).

$$e_{ij} = a(W\vec{h_i}, W\vec{h_j}) \tag{2}$$

In this formula, e_{ij} reflects the importance of node j's attributes to node i, a refers to the attention mechanism, and W is a linear weight matrix shared by all nodes. Typically, attention coefficients are determined based on first-order neighbors. Further, a softmax function is used to normalize coefficients across all choices of j, enabling them to be compared across nodes. This process is illustrated in Equations 3, 4 (Veličković et al., 2017). In Equation 3, the node $j \in N_i$, where N_i refers to some neighborhood of node i.

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ik})} \tag{3}$$

$$\alpha_{ij} = \frac{exp(LeakyReLU(\vec{a}^T[W\vec{h_i} \| W\vec{h_j}]))}{\sum_{k \in N_i} exp(LeakyReLU(\vec{a}^T[W\vec{h_i} \| W\vec{h_k}]))} \tag{4}$$

where α_{ij} is normalized attention coefficient, $\vec{a}$ is a weight vector, LeakyReLU is a nonlinear function, $.^T$ refers to a transposition, and $\|$ is the concatenation operation. Lastly, these normalized attention coefficients are utilized to calculate a linear combination of nodes' features, resulting in a new embedding for each node (Equation 5) (Veličković et al., 2017).

$$\vec{h_i}' = \sigma(\sum_{j \in N_i} \alpha_{ij} W\vec{h_j}) \tag{5}$$

where σ represents a nonlinear function. Assuming that K distinct attention heads have been used to compute the node embedding, then the results should be concatenated (Equation 6) or averaged (Equation 7) as follows (Veličković et al., 2017):

$$\vec{h_i}' = \|_{k=1}^K \sigma(\sum_{j \in N_i} \alpha_{ij}^k W^k \vec{h_j}) \tag{6}$$

$$\vec{h_i}' = \sigma(\frac{1}{K} \sum_{k=1}^K \sum_{j \in N_i} \alpha_{ij}^k W^k \vec{h_j}) \tag{7}$$

where $\|$ is concatenation, α_{ij}^k are normalized attention coefficients calculated by the k-th attention mechanism, σ is a nonlinear function, and W is a linear weight matrix.

3.2.2 Data oversampling

In general, deep learning models have a large number of parameters that must be optimized during the training process. Therefore, training deep learning models on large datasets increases their reliability, generalizability, and their classification performance (Alzubaidi et al., 2023). As rs-fMRI datasets are difficult to acquire and collect, their sample size is smaller than classic datasets used for machine learning. Thus, there is a risk of underfitting or overfitting the model when analyzing FBN data (Liu et al., 2023). To overcome this issue, many studies have proposed oversampling techniques that generate new data samples from the original dataset. In real-world applications, the synthetic minority oversampling (SMOTE) technique is one of the most widely used oversampling methods due to its simplicity, high performance, and computational efficiency (Mansourifar and Shi, 2020). In SMOTE, the oversampling is carried out by creating one or more synthetic samples for each training point belonging to the minority class. Specifically, these synthetic samples are obtained in feature space by applying linear interpolation between minority class samples and their K nearest neighbors (Chawla et al., 2002). As a first step, the k-nearest neighbors of sample are computed, and one of them is chosen randomly. Next, the difference between the feature vector of the real sample X_0 and its nearest neighbor X is computed. This difference is then multiplied by a random number between 0 and 1, known as the gap, and then added to the real sample's feature vector to produce a synthetic feature vector Z (Chawla et al., 2002). These steps are repeated for each sample based on how many synthetic samples are needed. The Equation 8 used for constructing synthetic samples is as follows (Chawla et al., 2002):

$$Z = X_0 + (X - X_0) \times gap \tag{8}$$

In this study, SMOTE was used to oversample MDD and HC classes to double both training and validation sets. Although the dataset is nearly balanced regarding the number of individuals with MDD and HC, it is not balanced in terms of the samples per site. Consequently, SMOTE was applied to generate synthetic samples that reduce overfitting and balance data across sites. Firstly, we divided the original time series dataset into 80% for training, 10% for validation, and 10% for testing. Next, each of the training and validation sets is divided into two sets, one containing MDD samples and the other containing HC samples. After that, oversampling was performed using SMOTE on each MDD and HC samples independently. Our current implementation uses three nearest neighbors to interpolate new synthetic samples. In our approach, each sample is represented by a matrix of dimensions TXN, where $T = 140$ time points and N is the number of ROIs. For each real sample in a specific class, one of its K-nearest neighbors from the same class is randomly selected. SMOTE was applied across the time-series in a row-wise manner. Specifically, for each
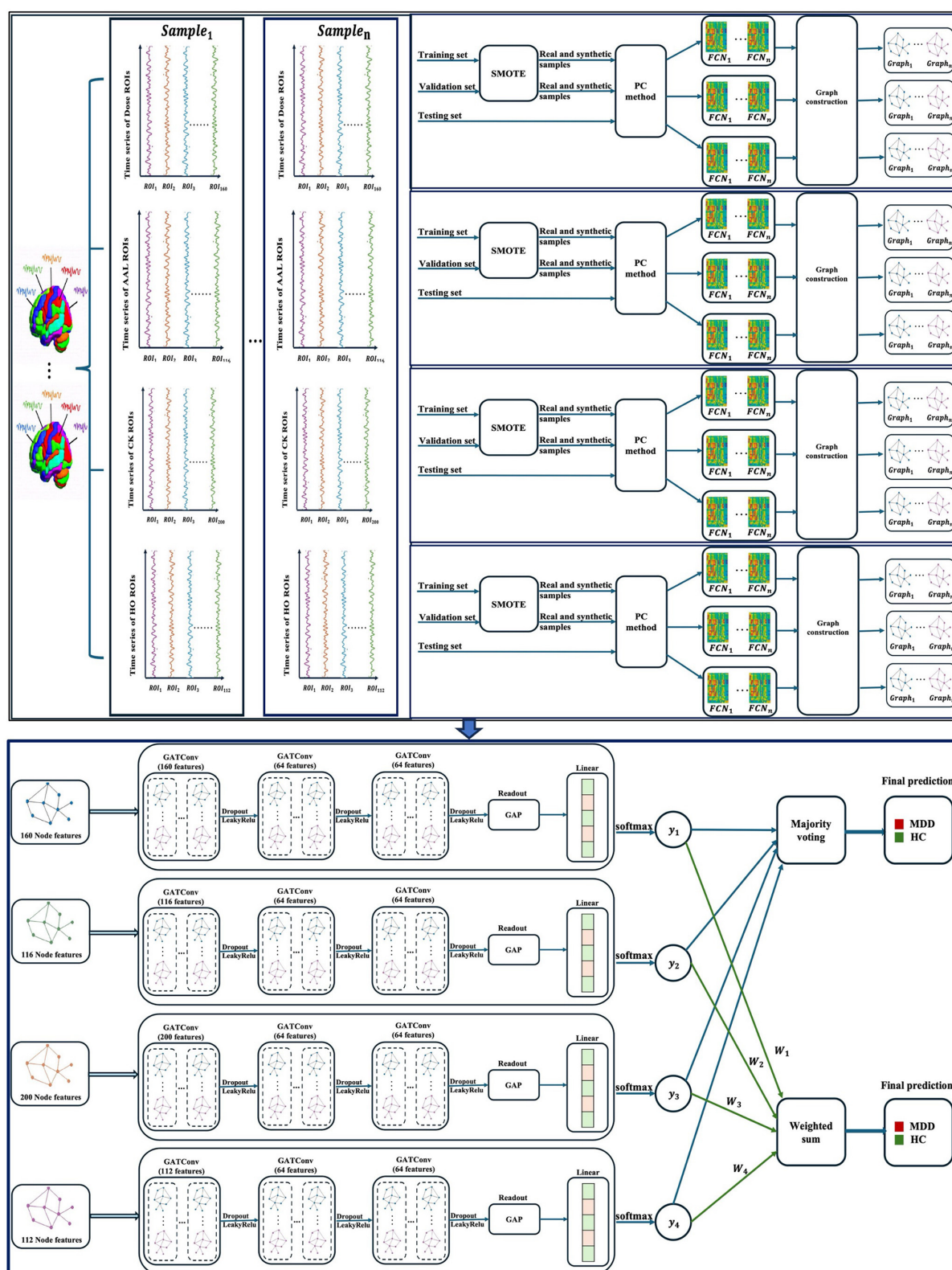
FIGURE 1
An overview of the multi-atlas ensemble GAT model. GATConv, graph attention network; LeakyRelu, leaky rectified linear unit; Dose, Dosenbach's atlas; AAL, automated anatomical labeling atlas; CK, Craddock's clustering atlas; HO, Harvard–Oxford atlas; ROIs, brain regions of interest; SMOTE, synthetic minority oversampling; PC, Pearson correlation coefficients; FCN, functional connectivity network; GAP, global average pooling; MDD, major depressive disorder; HC, healthy control; n, the number of samples.

time point t (from 1 to 140), we interpolated between the t-th row of the real sample and the t-th row of its neighbor to generate the corresponding row in the synthetic sample. This interpolation was performed independently for each time point across all ROIs values using the SMOTE formula (Equation 8). This process ensures that the temporal and spatial characteristics of the original signal are preserved in the generated synthetic samples.

As a result, we obtained 2, 502 training samples: 1, 298 MDD samples and 1, 204 HC samples, and 310 validation samples: 160 MDD samples and 150 HC samples. We oversampled only the training and validation sets in order to ensure that only real data were included in the testing set. Consequently, there are 157 real testing samples: 81 MDD samples and 76 HC samples.

3.2.3 The construction of the brain network

In graph theory, the graph structure is expressed as $G = \{V, E, A\}$, where V represents a set of vertices or nodes, E represents a set of edges between these nodes, and A is an adjacency matrix (Lima et al., 2022). An adjacency matrix illustrates the relationships between each pair of nodes in a graph, in which the connection between a given pair of nodes is determined by the entry of A in the i-th row and j-th column of the matrix, and is denoted as A_{ij}. In this matrix, element A_{ij} is 1 if there is an edge from node i to node j else A_{ij} is 0. Furthermore, a graph is considered undirected if none of its edges have a direction, whereas it is considered directed if all its edges have a direction. A graph can also have weighted edges, in which the adjacency matrix entries are arbitrary real-values instead of 0 and 1 (Wu et al., 2022). Feature vectors can also be attached to nodes in a graph. In this case, x_i indicates the feature vector for the node v_i, and $X_V = [x_1, x_2, ..., x_n]$ represents the set of feature vectors for all the nodes in the graph. Similarly, graph edges may be associated with edge feature vectors $x(i, j)$.

Using rs-fMRI time series, functional connectivity networks/matrices (FCNs) were constructed separately for each subject using four brain atlases: Dose (160 ROIs), AAL (116 ROIs), CK (200 ROIs), and HO (112 ROIs). For each atlas, we computed an FCN by calculating the correlation between each pair of ROIs using Pearson correlation coefficients (PC), as shown in Equation 9 (Wang et al., 2022). As a result, four different matrices were created for each subject. Each matrix had a size of ROIs X ROIs depending on the number of ROIs included in each atlas. Fisher's z-transformation was then applied to these matrices to standardize correlations across different sites (Vergara et al., 2018; Dai et al., 2024). Second, an undirected graph was generated from each brain FCN, where graph nodes represent brain regions of interest (ROIs) and node features are determined by the rows in FCN. To establish edges, the k-nearest neighbor (KNN) algorithm was employed to connect each node with its neighbors. Further, the relationships between the nodes are represented by a weighted adjacency matrix A.

$$b_{ij} = \frac{(y_i - \bar{y}_i)(y_j - \bar{y}_j)}{\sqrt{(y_i - \bar{y}_i)(y_i - \bar{y}_i)}\sqrt{(y_j - \bar{y}_j)(y_j - \bar{y}_j)}} \qquad (9)$$

where $b_{ij} \in [-1, 1]$ is the correlation between the i-th and j-th ROIs, y_i and y_j correspond to time series points for ROIs i and j,

while $\bar{y}_i$ and $\bar{y}_j$ represent the means of the regional rs-fMRI singnals in ROIs i and j, respectively.

3.2.4 Ensemble methods

Ensemble learning is the process of combining several baseline models to construct a more powerful and generalizable model than its constituent models (Mohammed and Kora, 2023). A typical ensemble learning system relies on an aggregation function G to combine a set of baseline classifiers $C_1, C_2, ..., C_n$ to predict a single output. The aggregation function is can be derived using different techniques, such as majority voting, weighted voting, averaging, weighted averaging, sum, and weighted sum (Mohammed and Kora, 2023).

In this regard, we developed a multi-atlas ensemble model consisting of four GAT models, each trained on graphs obtained from a different brain atlas. Following this, the GAT models make a set of predictions based on their respective test data, which are then combined to produce a final ensemble prediction. To accomplish this, majority voting and weighted sum methods were applied. In majority voting (or hard voting), each model C_i in the ensemble separately predicts the class label for a given input, which is considered a vote. Afterwards, the ensemble determines the final label prediction $\hat{y}$ by choosing the class label that takes the most votes from the individual models as shown in Equation 10 (Mohammed and Kora, 2023). In this Equation, the mode function refers to the most frequently occurring value within a set of values.

$$\hat{y} = mode[C_1(x), C_2(x), ..., C_n(x)] \qquad (10)$$

Since we have an even number of models, ties may occur if more than one class receives the same score. In this case, we broke the tie by taking the prediction from the model that had the highest accuracy. For weighted sum, the output of each model is taken after applying the softmax function. This produces a vector for each input containing the probability of each class label. The probability values for each class label are then multiplied by the model weight. In this study, each model's weight is determined by its average accuracy based on the validation set, as indicated in Equation 11.

$$w_j = \frac{Acc_j}{\sum_{i=1}^{n} Acc_i} \qquad (11)$$

Assume that w_j is the weight of the model j, n is the number of models, and Acc_i is the accuracy of the ith model. After that, the weighted probabilities for each class are summed, and the class label with the highest sum probability is selected as the final ensemble prediction. A weighted sum method for a binary classification task is demonstrated in the following Equation 12 (Mohammed and Kora, 2023).

$$\hat{y} = argmax_i \sum_{j=1}^{n} w_j \times p_{ij} \qquad (12)$$

In this case, $\hat{y}$ represents the predicted class label, as determined by weighted sum, n represents the number of models, w_j represents the weight of the model j, and p_{ij} represents the probability of class i in the model j.

4 Results and discussion

4.1 Implementation details and evaluation metrics

Throughout this section, we first introduce the implementation details and the evaluation metrics for the classification task. Then, we present and discuss the prediction performance of the proposed model on a large-scale fMRI dataset obtained from the REST-meta-MDD project.

In this study, we developed an ensemble-based GNN model for identifying MDD and HC subjects based on the rs-fMRI data. This model is comprised of four GAT models each trained on a different atlas. For this purpose, several brain segmentation atlases were employed, including Dosenbach's 160 functional ROIs (Dose), Automated Anatomical Labeling (AAL), Craddock's clustering atlas (CK), and Harvard-Oxford (HO). A GAT model is composed of three stacked layers of GATConv with Leaky Rectified Linear Unit (LeakyReLU) activation functions, which are used to learn the representation of each node. Each GATConv layer is composed of 64 hidden units. The number of attention heads used in this study is 4. Before each GATConv layer, the dropout technique is performed to prevent overfitting. Further, global average pooling (GAP) is applied to generate a representation of the whole graph by averaging the final embeddings of each node, as indicated in Equation 13 (Cui et al., 2022). In Equation 13, h_G refers to the representation of a single graph G, N is the number of nodes in G, and h_i is the embedding of the i-th node in G. For graph classification, a fully connected (FC) layer is added and activated by a softmax function to convert output scalars into the predictive probability of each class.

$$h_G = \frac{1}{N} \sum_{i=1}^{N} h_i \tag{13}$$

During the training stage, the cross-entropy loss function is used to optimize the parameters of the model The loss function is regularized using the $L2$ regularization technique known as weight decay. Additionally, Adaptive Moment Estimation (Adam) is utilized as an optimization function to rapidly update all weights based on a constant learning rate alpha. We employed a grid search strategy to identify the optimal values of the hyperparameters to enhance the performance of the model. Specifically, the learning rate was explored in the range $[1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}]$, the weight decay in $[1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}]$, the batch size in $[8, 16, 32, 64]$ and the dropout rate in $[0.2, 0.3, 0.4, 0.5]$. The final configuration included a batch size of 16, learning rate of 1×10^{-3}, a weight decay value of 5×10^{-4}, and dropout rate of 0.5.

Furthermore, we evaluated the proposed model using the REST-meta-MDD dataset based on a stratified 10-fold cross-validation method. The stratified cross-validation method allows us to keep the proportion of samples from each class equal across all folds. Using this method, a dataset is divided into 10 non-overlapping subsets. During each iteration, one subset is selected as

a test set for model evaluation, while the remaining nine subsets are used as a training set. Afterwards, the performance of the proposed model is assessed using five widely used metrics, including accuracy (Acc), sensitivity (Sen), specificity (Spe), precision (Pre), and F1-score (F1). These metrics are calculated as follows (Sharma et al., 2022):

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{14}$$

$$Sen = \frac{TP}{(TP + FN)} \tag{15}$$

$$Spe = \frac{TN}{(FP + TN)} \tag{16}$$

$$Pre = \frac{TP}{(TP + FP)} \tag{17}$$

$$F1 = \frac{(2 \times Pre \times Sen)}{(Pre + Sen)} \tag{18}$$

In this case, TP denotes the correct classification of positive samples (MDD), TN denotes the correct classification of negative samples (HC), FP denotes the incorrect classification of negative samples as positive, and FN denotes the incorrect negative classification.

4.2 Classification results

This section examines the effectiveness of our suggested MDD classification method by analyzing six scenarios: (1) Performance of single-atlas GAT models trained on multi-site data with SMOTE; (2) Performance of the multi-atlas ensemble GAT model trained on multi-site data with SMOTE; (3) An examination of the statistical differences between the multi-atlas GAT model and single-atlas GAT models with SMOTE; (4) Comparison of GAT models with and without SMOTE; (5) Comparison of GAT models using single-site data; (6) Comparison with existing studies.

4.2.1 Results of single-atlas GAT models with SMOTE using multi-site data

This section investigates the performance of four GAT models, each applied to a different brain atlas. We trained these models using the SMOTE algorithm on multi-site data from the REST-meta-MDD dataset. The corresponding performance results are presented in Table 2. The results indicate that the CK-based model has superior performance compared to other models based on the Dose, AAL, and HO atlases in three metrics: an accuracy of $65.41 \pm 2.92\%$, a sensitivity of $75.43 \pm 6.26\%$, and an F1-score of $69.16 \pm 2.90\%$. The HO-based model achieves the second highest performance with $64.33 \pm 1.07\%$ accuracy, $69.63 \pm 9.11\%$ sensitivity, and $66.59 \pm 2.78\%$ F1-score. Nevertheless, the specificity and precision of the HO-based model are $58.68 \pm 10.09\%$ and $64.65 \pm 2.70\%$, which are significantly higher than that of other models based on a single atlas. The AAL-based model

TABLE 2 Average classification results and standard deviation of 10-fold cross-validation for the single-atlas GAT model with SMOTE using multi-site data.

Method	Average ± standard deviation (%)				
	Accuracy	Sensitivity	Specificity	Precision	F1-score
Dose-based model	61.91 ± 1.84	71.23 ± 8.50	51.97 ± 8.87	61.47 ± 2.29	65.66 ± 3.32
AAL-based model	63.95 ± 4.01	70.99 ± 7.27	56.45 ± 11.88	64.04 ± 4.33	66.95 ± 3.03
CK-based model	**65.41±2.92**	**75.43±6.26**	54.74 ± 7.16	64.15 ± 2.91	**69.16±2.90**
HO-based model	64.33 ± 1.07	69.63 ± 9.11	**58.68±10.09**	**64.65±2.70**	66.59 ± 2.78

The bolded values indicate the best performing results.

TABLE 3 Classification results for the fold with the highest accuracy from 10-fold cross-validation for the single-atlas GAT model with SMOTE using multi-site data.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)
Dose-based model	65.61	65.43	65.79	67.09	66.25
AAL-based model	67.52	76.54	57.89	65.96	70.86
CK-based model	**70.06**	**86.42**	52.63	66.04	**74.87**
HO-based model	66.24	61.73	**71.05**	**69.44**	65.36

The bolded values indicate the best performing results.

produces an accuracy of 63.95 ± 4.01%, a sensitivity of 70.99 ± 7.27%, a specificity of 56.45 ± 11.88%, a precision of 64.04 ± 4.33%, and an F1-score of 66.95 ± 3.03%. On the other hand, the Dose-based model achieves the lowest performance with 61.91 ± 1.84% accuracy, 71.23 ± 8.50% sensitivity, 51.97 ± 8.87% specificity, 61.47 ± 2.29% precision, and 65.66 ± 3.32% F1-score.

Furthermore, the evaluation metrics for the fold that has the highest accuracy for each atlas-based model are displayed in Table 3. In the CK-based model, the best model among all 10 folds exhibits an accuracy of 70.06%, a sensitivity of 86.42%, a specificity of 52.63%, a precision of 66.04%, and an F1-score of 74.87%. The best AAL-based model has 67.52% accuracy, 76.54% sensitivity, 57.89% specificity, 65.96% precision, and 70.86% F1-score. While the best HO-based model has 66.24% accuracy, 61.73% sensitivity, 71.05% specificity, 69.44% precision, and 65.36% F1-score. Using the Dose-based model, the best model achieves an accuracy of 65.61%, a sensitivity of 65.43%, a specificity of 65.79%, a precision of 67.09%, and an F1-score of 66.25%.

4.2.2 Results of multi-atlas GAT ensemble models with SMOTE using multi-site data

This section demonstrates the effectiveness of applying different ensemble methods within our proposed multi-atlas GAT framework. The atlas-based models were trained independently, and their predictions were then aggregated using three ensemble methods: majority voting, sum, and weighted sum. Table 4 provides a comparison of the classification performance obtained with each ensemble method. Based on the results, a majority voting ensemble outperforms other ensemble methods across all the metrics, including an accuracy of 69.49 ± 3.54%, a sensitivity of 78.02 ± 6.55%, a specificity of 60.39 ± 5.41%, a precision of 67.79 ± 2.87%, and an F1-score of 72.43 ± 3.71%. The ensemble model based on the sum method achieves the second best performance with 67.83 ± 3.66% accuracy, 77.78 ± 6.51% sensitivity, 57.24 ±

8.71% specificity, 66.24 ± 3.77% precision, and 71.34 ± 3.28% F1-score. Meanwhile, a weighted sum ensemble provides the lowest performance compared to other methods. It has an accuracy of 67.45 ± 3.20%, a sensitivity of 77.65 ± 6.33%, a specificity of 56.58 ± 8.40%, a precision of 65.84 ± 3.45%, and an F1-score of 71.06 ± 2.94%. In addition, the evaluation metrics for the fold with the highest accuracy for each ensemble model are presented in Table 5. In a majority voting ensemble, the best performing model among all 10 folds attains an accuracy of 75.80%, a sensitivity of 88.89%, a specificity of 61.84%, a precision of 71.29%, and an F1-score of 79.12%. Using the sum method, the best ensemble model produces 73.25% accuracy, 85.19% sensitivity, 60.53% specificity, 69.70% precision, and 76.67% F1-score. On the other hand, the best performing model obtained by a weighted sum ensemble has 70.70% accuracy, 82.72% sensitivity, 57.89% specificity, 67.68% precision, and 74.44% F1-score.

Accordingly, we chose a majority voting ensemble since it outperformed both sum-based and weighted sum-based models. Moreover, we compared the performance of the multi-atlas model based on a majority voting with that of other single-atlas models in order to obtain further clarity on its prediction performance. Based on Tables 2, 4 and Figure 2, the multi-atlas model shows better prediction performance than the four models that relied on a single atlas. The multi-atlas model improves all metrics when compared with the Dose-based model: accuracy by 7.58%, sensitivity by 6.79%, specificity by 8.42%, precision by 6.32%, and F1-score by 6.77%. The multi-atlas model also improves accuracy by 5.54%, sensitivity by 7.03%, specificity by 3.94%, precision by 3.75%, and F1-score by 5.48% over the AAL-based model. A comparison with the CK-based model reveal an improvement in accuracy, sensitivity, specificity, precision, and F1-score by 4.08%, 2.59%, 5.65%, 3.64%, and 3.27%, respectively. Additionally, the accuracy, sensitivity, specificity, precision, and F1-score of the multi-atlas model are enhanced over the HO-based model by 5.16%, 8.39%, 1.71%, 3.14%, and 5.84%, respectively.

TABLE 4 Average results and standard deviation from 10-fold stratified cross-validation of the multi-atlas ensemble GAT model with SMOTE across different ensembling approaches using multi-site data.

Method	Average ± standard deviation (%)				
	Accuracy	Sensitivity	Specificity	Precision	F1-score
Majority voting	**69.49±3.54**	**78.02±6.55**	**60.39±5.41**	**67.79±2.87**	**72.43±3.71**
Sum	67.83 ± 3.66	77.78 ± 6.51	57.24 ± 8.71	66.24 ± 3.77	71.34 ± 3.28
Weighted sum	67.45 ± 3.20	77.65 ± 6.33	56.58 ± 8.40	65.84 ± 3.45	71.06 ± 2.94

The bolded values indicate the best performing results.

TABLE 5 Classification results for the fold with the highest accuracy from 10-fold stratified cross-validation of the multi-atlas ensemble GAT model with SMOTE across different ensembling approaches using multi-site data.
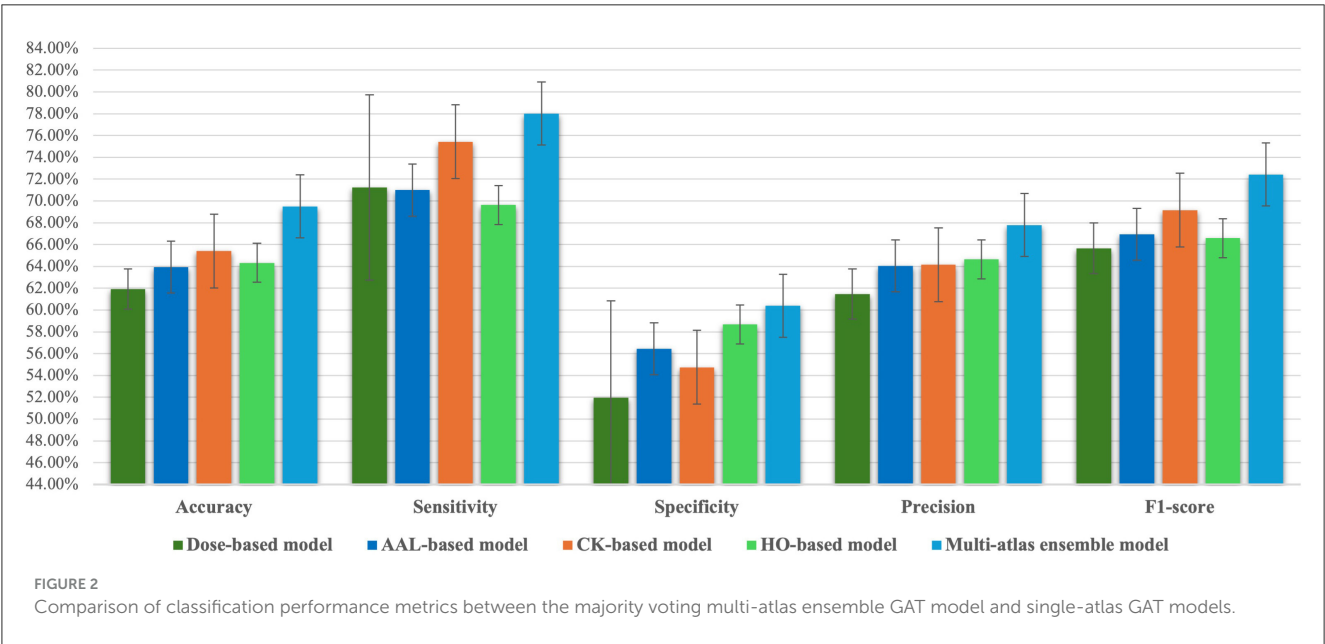
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)
Majority voting	**75.80**	**88.89**	**61.84**	**71.29**	**79.12**
Sum	73.25	85.19	60.53	69.70	76.67
Weighted sum	70.70	82.72	57.89	67.68	74.44

The bolded values indicate the best performing results.

FIGURE 2
Comparison of classification performance metrics between the majority voting multi-atlas ensemble GAT model and single-atlas GAT models.

4.2.3 Statistical analysis of multi-atlas and single-atlas GAT models with SMOTE

In this section, we analyze whether the proposed multi-atlas ensemble GAT model differs significantly from each single-atlas model, all of which were trained with SMOTE using multi-site data. Independent two-sample t-test was conducted for each performance metric, as illustrated in Table 6. In this study, we assumed that differences between models were statistically significant when $p < 0.05$. The results show that there is a significant difference between the multi-atlas model and the Dose-based model on four metrics, including accuracy, specificity, precision, and F1-score. Meanwhile, the accuracy, sensitivity, precision, and F1-score of our multi-atlas model differ significantly from those of the AAL-based model and the HO-based model. Furthermore, the CK-based model and the multi-atlas model show significant differences on both accuracy and precision.

4.2.4 Performance comparison of GAT models with and without SMOTE using multi-site data

SMOTE is a powerful technique for dealing with class imbalance, which has achieved robust results in a variety of applications. In SMOTE, synthetic samples are added to the minority class to create a balanced dataset (Chawla et al., 2002). This study utilized SMOTE to oversample MDD and HC classes in both the training and validation sets, so that only real data was included in the testing set. Table 7 presents an analysis of SMOTE's effectiveness on MDD prediction using the proposed multi-atlas model and single-atlas models trained on a multi-site dataset. Based on the results, the SMOTE technique has effectively improved the performance of classification models. Specifically, the Dose-based model shows improvements of 0.06%, 2.09%, and 1.0% in accuracy, sensitivity, and F1-score. In the AAL-based model, sensitivity and F1-score are improved by 4.45 and 1.39%, respectively. Meanwhile,

TABLE 6 Comparison of *p*-values between the majority voting multi-atlas ensemble GAT model and single-atlas GAT models with SMOTE using multi-site data.

Compared models	Accuracy	Sensitivity	Specificity	Precision	F1-score
Multi-atlas model and dose-based model	**0.0000**	0.0738	**0.0257**	**0.0001**	**0.0007**
Multi-atlas model and AAL-based model	**0.0061**	**0.0447**	0.3762	**0.0439**	**0.030**
Multi-atlas model and CK-based model	**0.0158**	0.4019	0.0748	**0.0156**	0.0519
Multi-atlas model and HO-based model	**0.0006**	**0.0376**	0.6594	**0.0282**	**0.0014**

The bolded values indicate statistically significant differences between models at $p < 0.05$.

TABLE 7 Performance comparison between the multi-atlas ensemble GAT model and single-atlas GAT models with and without SMOTE using multi-site data.

	Method	Average ± standard deviation (%)				
		Accuracy	Sensitivity	Specificity	Precision	F1-score
Without SMOTE	Dose-based model	61.85 ± 1.72	69.14 ± 12.28	54.08 ± 15.26	62.72 ± 4.84	64.67 ± 4.12
	AAL-based model	64.46 ± 2.38	66.54 ± 9.59	62.24 ± 9.23	65.69 ± 3.25	65.56 ± 4.53
	CK-based model	65.54 ± 2.71	72.72 ± 8.51	57.89 ± 11.33	65.37 ± 4.15	68.38 ± 2.79
	HO-based model	66.43 ± 1.14	76.91 ± 7.20	55.26 ± 8.63	65.00 ± 2.58	70.16 ± 1.93
	Multi-atlas model	68.92 ± 1.86	74.32 ± 3.86	**63.16±4.52**	**68.35±2.16**	71.13 ± 1.90
With SMOTE	Dose-based model	61.91 ± 1.84	71.23 ± 8.50	51.97 ± 8.87	61.47 ± 2.29	65.66 ± 3.32
	AAL-based model	63.95 ± 4.01	70.99 ± 7.27	56.45 ± 11.88	64.04 ± 4.33	66.95 ± 3.03
	CK-based model	65.41 ± 2.92	75.43 ± 6.26	54.74 ± 7.16	64.15 ± 2.91	69.16 ± 2.90
	HO-based model	64.33 ± 1.07	69.63 ± 9.11	58,68 ± 10.09	64.65 ± 2.70	66.59 ± 2.78
	Multi-atlas model	**69.49±3.54**	**78.02±6.55**	60.39 ± 5.41	67.79 ± 2.87	**72.43±3.71**

The bolded values indicate the best performing results.

the sensitivity and F1-score of the CK-based model are increased by 2.71% and 0.78%, respectively. In contrast, the HO-based model is only improved in terms of specificity by 3.42%. Moreover, the multi-atlas model achieves improvement in accuracy, sensitivity, and F1-score by 0.57%, 3.7%, and 1.3%, respectively. Clearly, the proposed multi-atlas model based on SMOTE has superior performance compared to the other models. This model gives an accuracy of 69.49 ± 3.54%, a sensitivity of 78.02 ± 6.55%, a specificity of 60.39 ± 5.41%, a precision of 67.79 ± 2.87%, and an F1-score of 72.43 ± 3.71%. However, the statistical significance analysis indicates no statistically significant differences between using SMOTE and without it with $p < 0.05$.

4.2.5 Performance comparison of GAT models with and without SMOTE using single site data

In this section, we evaluated our proposed approach on data collected from a single site (site 20), which has the largest number of subjects among the REST-meta-MDD sites. It consisted of 470 subjects, of which 245 had MDD and 225 had HC (Chen et al., 2022; Yan et al., 2019). We trained both single-atlas and multi-atlas GAT models on this single-site data, with and without SMOTE. A comparison of the performance of these models can be found in Table 8. Based on the results, the multi-atlas model without SMOTE has superiority in four metrics: an accuracy of 71.06 ± 5.40%, a sensitivity of 86.00 ± 5.73%, a precision of 68.37 ± 4.98%, and an F1-score of 76.01 ± 4.05%. In contract, the multi-atlas

model based on SMOTE achieves the second highest performance, with 69.79 ± 4.74% accuracy, 84.80 ± 6.14% sensitivity, 52.73 ± 11.89% specificity, 67.59 ± 5.24% precision, and 74.95 ± 3.26% F1-score. Compared to single-atlas models, the CK-based model with SMOTE achieves the best performance, with an accuracy of 68.09 ± 3.81%, a sensitivity of 84.00 ± 7.16%, a specificity of 50.00 ± 11.13%, a precision of 66.02 ± 4.22%, and an F1-score of 73.64 ± 3.00%. Similarly, the Dose-based model shows good performance with and without SMOTE. It delivers 67.45 ± 4.37% accuracy, 82.40 ± 8.04% sensitivity, 50.45 ± 16.3% specificity, 66.33 ± 5.58% precision, and 72.93 ± 2.22% F1-score. The AAL-based model is improved with SMOTE to produce 61.70 ± 5.21% accuracy, 70.40 ± 8.98% Sensitivity, 51.82 ± 17.74% specificity, 63.59 ± 6.47% precision, and 66.12 ± 3.18% F1-score. However, the HO-based model with SMOTE achieves the lowest performance with 57.23 ± 2.2% accuracy, 66.40 ± 12.67% sensitivity, 46.82 ± 13.9% specificity, 59.05 ± 2.44% precision, and 61.72 ± 5.37% F1-score.

As a matter of fact, it could be argued that the differences in SMOTE's effectiveness between single-site and multi-site data can be attributed to their distinctive characteristics. When using SMOTE on single-site data, where imaging protocols and populations are homogeneous, synthetic samples brought into the model may not add significant diversity, resulting in an overall reduced performance of 0.78%–1.27% in all metrics. Conversely, multi-site data have a higher level of variability due to differences in the acquisition protocols, scanner types, and demographic characteristics of the population, resulting in more complex and

TABLE 8 Performance comparison between the multi-atlas ensemble GAT model and single-atlas GAT models with and without SMOTE using single-site data.

Method		Average ± standard deviation (%)				
		Accuracy	Sensitivity	Specificity	Precision	F1-score
Without SMOTE	Dose-based model	67.45 ± 4.37	82.40 ± 8.04	50.45 ± 16.32	66.33 ± 5.58	72.93 ± 2.22
	AAL-based model	58.94 ± 3.69	79.60 ± 17.39	35.45 ± 23.95	59.77 ± 5.39	66.59 ± 5.74
	CK-based model	67.23 ± 5.05	80.00 ± 6.45	52.73 ± 9.79	66.06 ± 4.52	72.19 ± 3.97
	HO-based model	62.55 ± 4.06	64.80 ± 15.68	**60.00±19.58**	67.11 ± 7.87	63.81 ± 8.03
	Multi-atlas model	**71.06±5.40**	**86.00±5.73**	54.00 ± 10.84	**68.37±4.98**	**76.01±4.05**
With SMOTE	Dose-based model	67.45 ± 3.81	76.00 ± 10.28	57.73 ± 11.33	67.60 ± 4.09	71.02 ± 4.74
	AAL-based model	61.70 ± 5.21	70.40 ± 8.98	51.82 ± 17.74	63.59 ± 6.47	66.12 ± 3.18
	CK-based model	68.09 ± 3.81	84.00 ± 7.16	50.00 ± 11.13	66.02 ± 4.22	73.64 ± 3.00
	HO-based model	57.23 ± 2.22	66.40 ± 12.67	46.82 ± 13.94	59.05 ± 2.44	61.72 ± 5.37
	Multi-atlas model	69.79 ± 4.74	84.80 ± 6.14	52.73 ± 11.89	67.59 ± 5.24	74.95 ± 3.26

The bolded values indicate the best performing results.

TABLE 9 Classification results for the fold with the highest accuracy from 10-fold stratified cross-validation of the multi-atlas ensemble GAT models.

Method	Site	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)
Without SMOTE	Single-site	**80.85**	**92.00**	68.18	76.67	**83.64**
	Multi-site	71.97	79.01	64.47	70.33	74.42
With SMOTE	Single-site	76.60	76.00	**77.27**	**79.17**	77.55
	Multi-site	75.80	88.89	61.84	71.26	79.12

The bolded values indicate the best performing results.

TABLE 10 Performance comparison between the proposed multi-atlas ensemble GAT model and other existing models.

Site	Method	Number of subjects	Atlas	Average ± standard deviation (%)				
				Accuracy	Sensitivity	Specificity	Precision	F1-score
Single-site	Xia et al. (2023)	533	AAL, CK, HO	**71.48**	**95.04**	45.02	66.01	**77.91**
	Lee et al. (2024)	470	AAL, CK, HO	68.88±2.55	69.98±5.97	**67.62±6.81**	——	70.02±2.79
	Multi-atlas without SMOTE (Ours)	470	Dose, AAL, CK, HO	71.06±5.40	86.00±5.73	54.09±10.84	**68.37±4.98**	76.01±4.05
	Multi-atlas with SMOTE (Ours)*	940	Dose, AAL, CK, HO	69.79±4.74	84.80±6.14	52.73±11.89	67.59±5.24	74.95±3.26
Multi-site	Liu et al. (2023)*	——	AAL	63.41	MDD = 60.57 HC = 66.13	——	MDD = 62.53 HC = 64.17	61.33
	Multi-atlas without SMOTE (Ours)	1563	Dose, AAL, CK, HO	68.92±1.86	74.32±3.86	**63.16±4.52**	**68.35±2.16**	71.13±1.90
	Multi-atlas with SMOTE (Ours)*	2969	Dose, AAL, CK, HO	**69.49±3.54**	**78.02±6.55**	60.39 ± 5.41	67.79 ± 2.87	**72.43±3.71**

* Methods were conducted using the oversampling technique. The number of subjects shown for these methods reflects the sample size after oversampling. The best performing results are bolded for each site category.

integrated distributions. In this context, SMOTE is beneficial since it generates synthetic samples that reduce overfitting and balance data across sites. However, there is no doubt that the multi-atlas models have a better performance than the single-atlas models. Moreover, the evaluation metrics for the fold that has the highest accuracy for each multi-atlas model with and without SMOTE

are displayed in Table 9. Without SMOTE, the best multi-atlas model among all 10 folds based on single-site data provides an accuracy of 80.85%, a sensitivity of 92.00%, a specificity of 68.18%, a precision of 76.67%, and an F1-score of 83.64%. In contrast, the best performing multi-atlas model based on multi-site data has 71.97% accuracy, 79.01% sensitivity, 64.47% specificity, 70.33% precision, and 74.42% F1-score. Using SMOTE, the best multi-atlas model based on single-site data shows accuracy of 76.60%, sensitivity of 76.00%, specificity of 77.27%, precision of 79.17%, and F1 score of 77.55%. While the best performing model based on multi-site data achieves 75.80% accuracy, 88.89% sensitivity, 61.84% specificity, 71.26% precision, and 79.12% F1-score.

While SMOTE offers an effective solution to class imbalance, it has known limitations, especially when applied to high-dimensional and temporal structured data such as fMRI data. First, it may generate noisy or less realistic patterns due to the complexity of the data. Second, it can cause class overlapping by generating synthetic samples near class boundaries, potentially degrading classifier performance in some cases. To mitigate these issues, we implemented SMOTE in a controlled and structure-aware manner. First, we applied SMOTE separately to each class (MDD and HC) to reduce inter-class blending. In addition, we applied SMOTE on row-wise features, where each row represents a single time point across all ROI values to maintain the spatial and temporal integrity of the original fMRI signals. To further verify that the generated samples remained consistent with real data, we performed a paired T-test for each pair of real and synthetic samples. The average p-values across all samples consistently exceeded 0.05 (ranging from 0.2 to 0.4) for all brain atlases used, indicating no statistically significant differences between the original and synthetic data distributions. This strongly supports the conclusion that our SMOTE application did not generate new or divergent samples, but rather reproduced realistic and distributionally consistent patterns, maintaining the spatiotemporal integrity of the original fMRI data. Notably, our results show that incorporating SMOTE improves the performance of the multi-site model compared to its non-SMOTE counterpart, emphasizing its effectiveness in enhancing robustness without compromising data integrity (see Table 7).

4.2.6 Performance comparison with the existing studies

A performance comparison of the proposed model with other existing models is provided in Table 10. We conducted a comparative experiment with models developed by Liu et al. (2023), Xia et al. (2023), and Lee et al. (2024), which utilized the REST-meta-MDD dataset as this study. We compared these models with our multi-site and single-site models with and without SMOTE. Models by Xia et al. (2023) and Lee et al. (2024) differ from ours in that they have been validated five times, use single-site data (site 20), combine three brain atlases (AAL, CK, and HO), and do not use oversampling. Comparatively, our proposed model and those published in Liu et al. (2023) used a 10-fold validation approach, data from multiple sites, and oversampling techniques. However, the model developed by Liu et al. (2023) employs a single structural atlas (AAL), while our model integrates four atlases:

two structural atlases (AAL and HO) and two functional atlases (Dose and CK). Further, we have highlighted the studies that utilized an oversampling technique with an asterisk (*) next to them in the Table 10. Moreover, a slight variation can be observed in the number of subjects between these models for several reasons, including the number of sites included, the brain atlases used in defining ROIs, and the subject inclusion/exclusion criteria applied during preprocessing.

Among the single-site models, Xia et al. (2023) proposed a two-channel GNN architecture (DepressionGraph) that incorporates a transformer-encoder to extract features from FCNs, and uses majority voting to combine predictions from models trained with three brain atlases: AAL, CK, and HO. Table 10 indicates that their model achieves the best performance in three metrics: an accuracy of 71.48%, a sensitivity of 95.04%, and an F1-score of 77.91%. However, their model exhibits a lower specificity of 45.02% and precision of 66.01% compared to competing single-site models. In contrast, our multi-atlas model without SMOTE and trained on single-site data (site 20) achieved the second best performance, with 71.06 ± 5.40% accuracy, 86.00 ± 5.73% sensitivity, 54.00 ± 10.84% specificity, and 76.01 ± 4.05%. It also exhibits a precision of 68.37 ± 4.98%, which is higher than that of other single-site models. Following this, our multi-atlas model with SMOTE produces an accuracy of 69.79 ± 4.74%, a sensitivity of 84.80 ± 6.14%, a specificity of 52.73 ± 11.89%, a precision of 67.59 ± 5.24%, and an F1-score of 74.95 ± 3.26%. The model proposed by Lee et al. (2024), which integrate a holistic FCN and individual FCNs computed from each atlas (AAL, CK, and HO) using a decision-level voting method, achieved an accuracy of 68.88 ± 2.55%, a sensitivity of 69.98 ± 5.97%, and an F1-score of 70.02 ± 2.79%. However, this model exhibits a specificity of 67.62 ± 6.81%, which is higher than that of all competing models.

Moreover, the proposed multi-atlas model with SMOTE was evaluated using data from 16 sites and demonstrated superior performance over other multi-site models. It achieved 69.49 ± 3.54% accuracy, 78.02 ± 6.55% sensitivity, 60.39 ± 5.41% specificity, 68.35% precision, and 72.43 ± 3.71% F1-score. Particularly, the model showed improvements in accuracy between 0.57 and 6.08%, sensitivity between 3.7 and 17.45%, and F1-score between 1.3 and 11.1% compared to other multi-site models. Meanwhile, our proposed multi-atlas model without SMOTE achieved the second best performance among multi-sites models, with 68.92 ± 1.86% accuracy, 74.32 ± 3.86% sensitivity, and 71.13 ± 1.90% F1-score. The specificity and precision of this model are 63.16 ± 4.52% and 68.35 ± 2.16%, which are higher than those of the multi-site models. However, the specificity may be moderate compared with other metrics due to the inherent heterogeneity within the MDD population and the potential overlap with healthy controls. Several recent studies have reported improved specificity when performing subgroup-based classifications (e.g., recurrent vs. first-episode MDD), suggesting that a finer diagnostic distinction would reduce intera-class ambiguity (Venkatapathy et al., 2023; Qin et al., 2022). Consequently, incorporating subgroup analysis represents a promising future direction for improving specificity and overall diagnostic utility. In Liu et al. (2023), a spatial-temporal data-augmentation-based classification model was developed using data from 24 sites. However, no information was

provided regarding the number of subjects after augmentation. Compared to other models, it attains the lowest performance with an accuracy of 63.41%, a sensitivity of 60.57% for MDD and 66.13% for HC, a precision of 62.53% for MDD and 64.17% for HC, and an F1-score of 61.33%.

5 Conclusion

In this study, we developed an ensemble-based GNN model for the classification of MDD based on rs-fMRI data. Our ensemble model was constructed by combining features derived from four brain segmentation atlases to capture brain complexity and identify distinct features more accurately than single atlas-based models. For this purpose, majority voting and weighted sum methods were applied. The experimental findings clearly indicate that the multi-atlas model with a majority voting ensemble offers superior performance compared to the single-atlas model. Our proposed model achieved improvements in accuracy between 4.08 and 7.58%, sensitivity between 2.59 and 8.39%, precision between 1.71 and 8.42%, and F1-score between 3.27 and 6.77% over other single-atlas models developed within this study.

For future work, we intend to enhance the interpretability of our GNN models by applying explainability techniques such as Grad-CAM to identify brain regions that significantly contribute to MDD classification. Further, the current approach will be extended by integrating structural MRI with rs-fMRI to construct a multimodal framework that provides complementary information to enrich feature representations, thereby improving the robustness and performance of the classification model. Additionally, we will explore cross-atlas (holistic) FCN modeling strategies to better capture inter-atlas relationships and investigate personalized FCNs to reflect subject-specific patterns. Finally, we propose future subgroup-based analyses to reduce intra-class variability and improve clinical applicability.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

NA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. AA: Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing. MA: Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., and Petersson, L. (2021). Graph-based deep learning for medical diagnosis and analysis: past, present and future. *Sensors* 21:4758. doi: 10.3390/s21144758

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A., Al-dabbagh, B. S. N., Fadhel, M. A., et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J. Big Data* 10:46. doi: 10.1186/s40537-023-00727-2

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, X., Lu, B., Li, H.-X., Li, X.-Y., Wang, Y.-W., Castellanos, F. X., et al. (2022). The direct consortium and the rest-meta-mdd project: towards neuroimaging biomarkers of major depressive disorder. *Psychoradiology* 2, 32–42. doi: 10.1093/psyrad/kkac005

Chu, Y., Wang, G., Cao, L., Qiao, L., and Liu, M. (2022). Multi-scale graph representation learning for autism identification with functional MRI. *Front. Neuroinform.* 15:802305. doi: 10.3389/fninf.2021.802305

Craddock, R. C., James, G. A., Holtzheimer, I. I. I., Hu, P. E., and Mayberg, X. P. H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* 33, 1914–1928. doi: 10.1002/hbm.21333

Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A. A. C., Lukemire, J., et al. (2022). Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE Trans. Med. Imaging* 42, 493–506. doi: 10.1109/TMI.2022.3218745

Dai, L., Zhou, H., Xu, X., and Zuo, Z. (2019). Brain structural and functional changes in patients with major depressive disorder: a literature review. *PeerJ* 7:e8170. doi: 10.7717/peerj.8170

Dai, P., Zhou, Y., Shi, Y., Lu, D., Chen, Z., Zou, B., et al. (2024). Classification of mdd using a transformer classifier with large-scale multisite

resting-state fMRI data. *Hum. Brain Mapp.* 45:e26542. doi: 10.1002/hbm.26542

Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., et al. (2010). Prediction of individual brain maturity using fMRI. *Science* 329, 1358–1361. doi: 10.1126/science.1194144

El-Gazzar, A., Thomas, R. M., and van Wingen, G. (2023). Harmonization techniques for machine learning studies using multi-site functional MRI data. *bioRxiv.* doi: 10.1101/2023.06.14.544758

Freeman, M. (2022). The world mental health report: transforming mental health for all. *World Psychiatry* 21:391. doi: 10.1002/wps.21018

GBD 2019 Mental Disorders Collaborators (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. *Lancet Psychiatry* 9, 137–150. doi: 10.1016/S2215-0366(21)00395-3

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127. doi: 10.1093/biostatistics/kxj037

Kennedy, D. N., Lange, N., Makris, N., Bates, J., Meyer, J., Caviness Jr, V. S., et al. (1998). Gyri of the human neocortex: an MRI-based analysis of volume and variance. *Cereb. Cortex* 8, 372–384. doi: 10.1093/cercor/8.4.372

Kong, Y., Gao, S., Yue, Y., Hou, Z., Shu, H., Xie, C., et al. (2021). Spatio-temporal graph convolutional network for diagnosis and treatment response prediction of major depressive disorder from functional connectivity. *Hum. Brain Mapp.* 42, 3922–3933. doi: 10.1002/hbm.25529

Lee, D.-J., Shin, D.-H., Son, Y.-H., Han, J.-W., Oh, J.-H., Kim, D.-H., et al. (2024). Spectral graph neural network-based multi-atlas brain network fusion for major depressive disorder diagnosis. *IEEE J. Biomed. Health Inform.* 28, 2967–2978. doi: 10.1109/JBHI.2024.3366662

Lei, B., Zhao, Y., Huang, Z., Hao, X., Zhou, F., Elazab, A., et al. (2020). Adaptive sparse learning using multi-template for neurodegenerative disease diagnosis. *Med. Image Anal.* 61:101632. doi: 10.1016/j.media.2019.101632

Lima, A. A., Mridha, M. F., Das, S. C., Kabir, M. M., Islam, M. R., Watanobe, Y., et al. (2022). A comprehensive survey on the detection, classification, and challenges of neurological disorders. *Biology* 11:469. doi: 10.3390/biology11030469

Liu, J., Wang, X., Zhang, X., Pan, Y., Wang, X., Wang, J., et al. (2018). Mmm: classification of schizophrenia using multi-modality multi-atlas feature representation and multi-kernel learning. *Multimed. Tools Appl.* 77, 29651–29667. doi: 10.1007/s11042-017-5470-7

Liu, M., Zhang, D., and Shen, D. (2016). Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. *IEEE Trans. Med. Imaging* 35, 1463–1474. doi: 10.1109/TMI.2016.2515021

Liu, Q., Zhang, Y., Guo, L., and Wang, Z. (2023). Spatial-temporal data-augmentation-based functional brain network analysis for brain disorders identification. *Front. Neurosci.* 17:1194190. doi: 10.3389/fnins.2023.1194190

Mansourifar, H., and Shi, W. (2020). Deep synthetic minority over-sampling technique. *arXiv* [preprint]. arXiv:2003.09788. doi: 10.48550/arXiv.2003.09788

Mohammed, A., and Kora, R. (2023). A comprehensive review on ensemble deep learning: opportunities and challenges. *J. King Saud. Univ. Comput. Inf. Sci.* 35, 757–774. doi: 10.1016/j.jksuci.2023.01.014

Noman, F., Ting, C.-M., Kang, H., Phan, R. C.-W., and Ombao, H. (2024). Graph autoencoders for embedding learning in brain networks and major depressive disorder identification. *IEEE J. Biomed. Health Inform.* 28, 1644–1655. doi: 10.1109/JBHI.2024.3351177

Pilmeyer, J., Huijbers, W., Lamerichs, R., Jansen, J. F., Breeuwer, M., Zinger, S., et al. (2022). Functional MRI in major depressive disorder: a review of findings, limitations, and future prospects. *J. Neuroimaging* 32, 582–595. doi: 10.1111/jon.13011

Qin, K., Lei, D., Pinaya, W. H., Pan, N., Li, W., Zhu, Z., et al. (2022). Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites. *EBioMedicine* 78:103977. doi: 10.1016/j.ebiom.2022.103977

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., et al. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *Lancet* 398, 1700–1712. doi: 10.1016/S0140-6736(21)02143-7

Sharma, D. K., Chatterjee, M., Kaur, G., and Vavilala, S. (2022). "3-deep learning applications for disease diagnosis," in *Deep Learning for Medical Applications with Unique Data*, eds. D. Gupta, U. Kose, A. Khanna, and V. E. Balas (New York, NY: Academic Press), 31–51. doi: 10.1016/B978-0-12-824145-5.00005-8

Tulay, E. E., Metin, B., Tarhan, N., and Arıkan, M. K. (2019). Multimodal neuroimaging: basic concepts and classification of neuropsychiatric diseases. *Clin. EEG Neurosci.* 50, 20–33. doi: 10.1177/1550059418782093

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y. (2017). Graph attention networks. *arXiv* [Preprint]. arXiv:1710.10903. doi: 10.48550/arXiv.1710.10903

Venkatapathy, S., Votinov, M., Wagels, L., Kim, S., Lee, M., Habel, U., et al. (2023). Ensemble graph neural network model for classification of major depressive disorder using whole-brain functional connectivity. *Front. Psychiatry* 14:1125339. doi: 10.3389/fpsyt.2023.1125339

Vergara, V. M., Yu, Q., and Calhoun, V. D. (2018). A method to assess randomness of functional connectivity matrices. *J. Neurosci. Methods* 303, 146–158. doi: 10.1016/j.jneumeth.2018.03.015

Wang, Q., Li, L., Qiao, L., and Liu, M. (2022). Adaptive multimodal neuroimage integration for major depression disorder detection. *Front. Neuroinform.* 16:856175. doi: 10.3389/fninf.2022.856175

Wu, L., Cui, P., Pei, J., Zhao, L., and Guo, X. (2022). "Graph neural networks: foundation, frontiers and applications," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 4840–4841. doi: 10.1145/3534678.3542609

Xia, Z., Fan, Y., Li, K., Wang, Y., Huang, L., Zhou, F., et al. (2023). Depressiongraph: a two-channel graph neural network for the diagnosis of major depressive disorders using rs-fMRI. *Electronics* 12:5040. doi: 10.3390/electronics12245040

Yan, B., Xu, X., Liu, M., Zheng, K., Liu, J., Li, J., et al. (2020). Quantitative identification of major depression based on resting-state dynamic functional connectivity: a machine learning approach. *Front. Neurosci.* 14:191. doi: 10.3389/fnins.2020.00191

Yan, C.-G., Chen, X., Li, L., Castellanos, F. X., Bai, T.-J., Bo, Q.-J., et al. (2019). Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proc. Nat. Acad. Sci.* 116, 9078–9083. doi: 10.1073/pnas.1900390116

Yao, D., Sui, J., Yang, E., Yap, P.-T., Shen, D., Liu, M., et al. (2020). "Temporal-adaptive graph convolutional network for automated identification of major depressive disorder using resting-state fMRI," in *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11* (Cham: Springer), 1–10. doi: 10.1007/978-3-030-59861-7_1

Zhuo, C., Li, G., Lin, X., Jiang, D., Xu, Y., Tian, H., et al. (2019). The rise and fall of MRI studies in major depressive disorder. *Transl. Psychiatry* 9:335. doi: 10.1038/s41398-019-0680-6