ORIGINAL RESEARCH
published: 29 April 2022
doi: 10.3389/fninf.2022.856175

Adaptive Multimodal Neuroimage Integration for Major Depression Disorder Detection

Qianqian Wang[1], *Long Li*[2], *Lishan Qiao*[1]* *and Mingxia Liu*[3]*

[1] School of Mathematics Science, Liaocheng University, Liaocheng, China, [2] Taian Tumor Prevention and Treatment Hospital, Taian, China, [3] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Major depressive disorder (MDD) is one of the most common mental health disorders that can affect sleep, mood, appetite, and behavior of people. Multimodal neuroimaging data, such as functional and structural magnetic resonance imaging (MRI) scans, have been widely used in computer-aided detection of MDD. However, previous studies usually treat these two modalities separately, without considering their potentially complementary information. Even though a few studies propose integrating these two modalities, they usually suffer from significant inter-modality data heterogeneity. In this paper, we propose an adaptive multimodal neuroimage integration (AMNI) framework for automated MDD detection based on functional and structural MRIs. The AMNI framework consists of four major components: (1) a graph convolutional network to learn feature representations of functional connectivity networks derived from functional MRIs, (2) a convolutional neural network to learn features of T1-weighted structural MRIs, (3) a feature adaptation module to alleviate inter-modality difference, and (4) a feature fusion module to integrate feature representations extracted from two modalities for classification. To the best of our knowledge, this is among the first attempts to adaptively integrate functional and structural MRIs for neuroimaging-based MDD analysis by explicitly alleviating inter-modality heterogeneity. Extensive evaluations are performed on 533 subjects with resting-state functional MRI and T1-weighted MRI, with results suggesting the efficacy of the proposed method.

Keywords: major depressive disorder, resting-state functional MRI, structural MRI, feature adaptation, multimodal data fusion

OPEN ACCESS

Edited by:
Antonio Fernández-Caballero,
University of Castilla-La Mancha,
Spain

Reviewed by:
Yu Zhang,
Lehigh University, United States
Shu Zhang,
Northwestern Polytechnical
University, China

Correspondence:
Lishan Qiao
qiaolishan@lcu.edu.cn
Mingxia Liu
mxliu1226@gmail.com

Received: *16 January 2022*
Accepted: *05 April 2022*
Published: *29 April 2022*

Citation:
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

1. INTRODUCTION

Major depressive disorder (MDD) is one of the most common mental health disorders, affecting as many as 300 million people annually (Organization et al., 2017). This disease is generally characterized by depressed mood, diminished interests, and impaired cognitive function (Alexopoulos, 2005; Pizzagalli et al., 2008; Otte et al., 2016). Despite decades of research in basic science, clinical neuroscience and psychiatry, the pathological, and biological mechanisms of major depression remain unclear (Holtzheimer III and Nemeroff, 2006). The traditional diagnosis of MDD mainly depends on criteria from the diagnostic and statistical manual of mental disorders (DSM) and treatment response (Papakostas, 2009), which could be subjective and susceptible. As a robust complement to clinical neurobehavior-based detection, computer-aided diagnosis based on neuroimaging data hold the promise of objective diagnosis and prognosis of mental disorders (Foti et al., 2014; Liu and Zhang, 2014; Bron et al., 2015; Shi et al., 2018; Zhang L. et al., 2020; Buch and Liston, 2021).

Multiple neuroimaging modalities, such as resting-state functional magnetic resonance imaging (rs-fMRI) and structural MRI (sMRI), can provide complementary information in discovering objective disease biomarkers, and have been increasingly employed in automated diagnosis of various brain disorders (Hinrichs et al., 2011). Resting-state fMRI helps capture large-scale abnormality or dysfunction on functional connectivity network (FCN) by measuring bold-oxygen-level-dependent (BOLD) signals of subjects (Van Den Heuvel and Pol, 2010; Wang et al., 2019; Zhang Y. et al., 2020; Sun et al., 2021), and thus, can measure hemodynamic response related to neural activity in the brain dynamically. Structural MRI provides relatively high-resolution structural information of the brain, enabling us to study pathological changes in different brain tissues, such as gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) (Cuadra et al., 2005). It is critical to integrate rs-fMRI and sMRI data to facilitate automated diagnosis of MDD and related disorders.

Existing neuroimaging-based MDD studies usually focus on discovering structural or functional imaging biomarkers, by employing various machine learning approaches such as support vector machines (SVM), Gaussian process classifier (GPC), linear discriminant analysis (LDA), and deep neural networks (Sato et al., 2015; Bürger et al., 2017; Rubin-Falcone et al., 2018; Li et al., 2021). However, these methods generally ignore the potentially complementary information conveyed by functional and structural MRIs. Several recent studies propose to employ functional and structural MRIs for MDD analysis, but they usually suffer from significant inter-modality data discrepancy (Fu et al., 2015; Maglanoc et al., 2020; Ge et al., 2021).

In this article, we propose an adaptive multimodal neuroimage integration (**AMNI**) framework for automated MDD detection using functional and structural MRI data. As shown in **Figure 1**, the proposed AMNI consists of four major components: (1) a *graph convolutional network* (GCN) for extracting feature representations of functional connectivity networks derived from rs-fMRI scans; (2) a *convolutional neural network* (CNN) for extracting features representations of T1-weighted sMRI scans; (3) a *feature adaptation module* for alleviating inter-modality difference by minimizing a cross-modal maximum mean discrepancy (MMD) loss; and (4) a *feature fusion module* for integrating features of two modalities for classification (*via* Softmax). Experimental results on 533 subjects from the REST-meta-MDD Consortium (Yan et al., 2019) demonstrate the effectiveness of AMNI in MDD detection.

The major contributions of this work are summarized below:

- An adaptive integration framework is developed to fuse functional and structural MRIs for automated MDD diagnosis by taking advantage of the complementary information of the two modalities. This is different from previous approaches that focus on only discovering structural or functional imaging biomarkers for MDD analysis.
- A feature adaptation strategy is designed to explicitly reduce the inter-modality difference by minimizing a cross-modal maximum mean discrepancy loss to re-calibrate features extracted from two heterogeneous modalities.
- Extensive experiments on 533 subjects with rs-fMRI and sMRI scans have been performed to validate the effectiveness of the proposed method in MDD detection.

The rest of this article is organized as follows. In Section 2, we briefly review the most relevant studies. In Section 3, we first introduce the materials and then present the proposed method as well as implementation details. In Section 4, we introduce the experimental settings and report the experimental results. In Section 5, we investigate the effect of several key components in the proposed method and discuss limitations as well as possible future research directions. We finally conclude this article in Section 6.

2. RELATED WORK

In this section, we briefly introduce the most relevant studies on structural and functional brain MRI analysis, as well as multimodal neuroimaging-based diagnosis of brain disorders.

2.1. Brain Structural MR Imaging Analysis

Currently, MRI is the most sensitive imaging test of the brain in routine clinical practice. Structural MRIs can non-invasively capture the internal brain structure and atrophy, assisting us to understand the brain anatomical changes caused by various mental disorders. Conventional sMRI-based MDD analysis is usually performed manually by human beings *via* visual assessment (Scheltens et al., 1992), which could be subjective and susceptible. To this end, many machine learning methods (Gao et al., 2018), such as support vector machines (SVM), Gaussian process classifier (GPC), and linear discriminant analysis (LDA), have been used for automated MRI-based MDD diagnosis. However, these methods generally rely on handcrafted MRI features and these features may be suboptimal for subsequent analysis, thus significantly limiting their practical utility.

In recent years, deep learning methods such as convolutional neural networks (CNNs) have been widely used in the fields of computer vision and medical image analysis (Yue-Hei Ng et al., 2015; Chen et al., 2016; Zhang L. et al., 2020). As a special type of multi-layer neural network, CNN is capable of automatic feature learning, which eliminates the subjectivity in extracting and selecting informative features for specific tasks (Lee et al., 2017). Based on the LeNet5 network, Sarraf and Tofighi (2016) presented a 2D convolutional neural network that could classify sMRI scan slices for Alzheimer's disease diagnosis. With the development of high-performance computing resources, Hosseini-Asl et al. (2016) developed a deep neural network that used 3D convolution layers to extract features of 3D medical images for Alzheimer's disease diagnosis. Chakraborty et al. (2020) developed a 3D CNN architecture for learning intricate patterns in MRI scans for Parkinson's disease diagnosis.

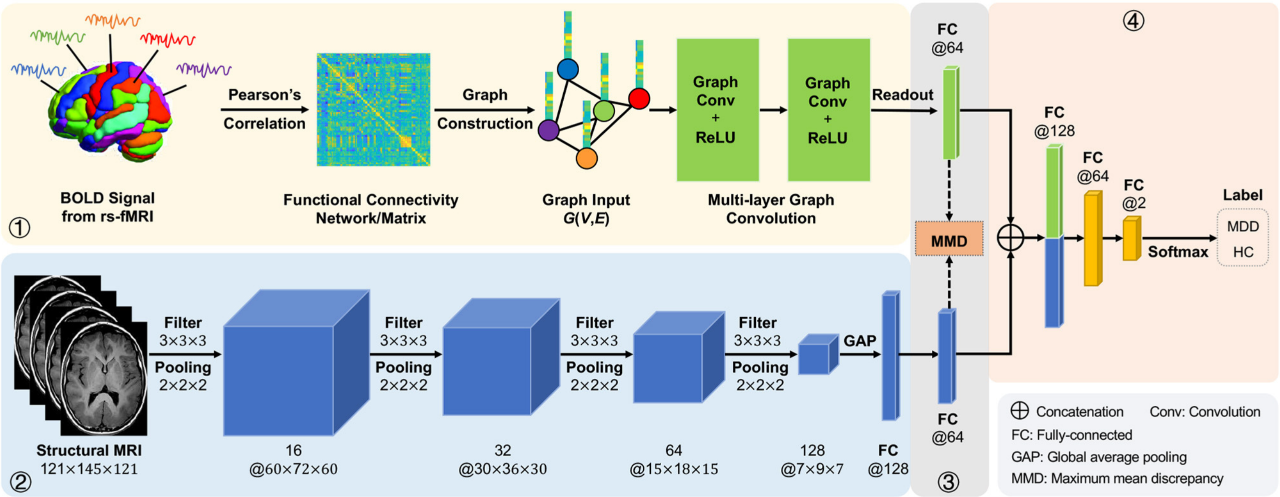

FIGURE 1 | Illustration of the proposed adaptive multimodal neuroimage integration (AMNI) framework, including (1) a graph convolutional network (GCN) for extracting features of functional connectivity networks derived from resting-state functional MRI (rs-fMRI) data, (2) a convolutional neural network (CNN) for extracting features of T1-weighted structural MRI (sMRI) data, (3) a feature adaptation module for alleviating inter-modality difference by minimizing a cross-modal maximum mean discrepancy (MMD) loss, and (4) a feature fusion module for integrating sMRI and fMRI features for classification. MDD, major depressive disorder; HC, healthy control.

Compared with 2D convolution, 3D convolution on the entire MR image is able to capture the rich spatial information, which is essential for disease classification.

2.2. Brain Functional MR Imaging Analysis

Existing studies have revealed that fMRI can capture large-scale abnormality or dysfunction on functional connectivity networks by measuring the blood-oxygen-level in the brain (Van Den Heuvel and Pol, 2010; Zhang et al., 2019). With fMRI data, we usually construct a functional connectivity network for representing each subject, where each node represents a specific brain region-of-interest (ROI) and each edge denotes the pairwise relationship between ROIs (Honey et al., 2009; Dvornek et al., 2017). By capturing the dependencies between BOLD signals of paired ROIs, functional connectivity networks (FCNs) have been widely used to identify potential neuroimaging biomarkers for mental disorder analysis. Previous studies often extract handcrafted FCN features (e.g., clustering coefficient and node degree) to build prediction/classification models (Guo et al., 2021; Zhang et al., 2021), but the definition of the optimal FCN features highly relies on expert knowledge, so it is often subjective. Extracting effective feature representations of functional connectivity networks is essential for subsequent analysis.

Recent studies have shown that spectral graph convolutional networks (GCNs) are effective in learning representations of brain functional connectivity networks, where each FCN is treated as a graph (Bruna et al., 2013; Parisot et al., 2018; Bai et al., 2020; Yao et al., 2021). Motivated by breakthroughs of deep learning on grid data, people make efforts to extend CNN to graphs, giving rise to the spectral graph convolutional networks (GCNs) (Bruna et al., 2013). Recent studies have shown that GCNs are effective in learning representations of brain functional connectivity networks compared to traditional machine learning algorithms. For example, Parisot et al. (2018) proposed a GCN-based method for group-level population diagnosis that exploited the concept of spectral graph convolutions. Yao et al. (2021) presented a mutual multi-scale triplet GCN model to extract multi-scale feature representations of brain functional connectivity networks. Bai et al. (2020) developed a backtrackless aligned-spatial GCN model to transitively align vertices between graphs and learn effective features for graph classification. Compared with traditional CNN with Euclidean data, GCN generalizes convolution operations to non-Euclidean data, and helps mine topological information of brain connectivity networks.

2.3. Multimodal Neuroimaging-Based Brain Disease Diagnosis

Previous studies have been shown that multimodal neuroimaging data can provide complementary information of individual subjects to improve the performance of computer-aided disease diagnosis (Sui et al., 2013; Calhoun and Sui, 2016; Maglanoc et al., 2020; Guan and Liu, 2021). For example, Sui et al. (2013) developed a machine learning model to enable fusion of three or more multimodal datasets based on multi-set canonical correlation analysis and joint independent component analysis algorithms. Maglanoc et al. (2020) used linked independent component analysis to fuse structural and functional MRI features for depression diagnosis. Even though previous studies have yielded promising performance, they often extract sMRI and fMRI features manually, which requires domain-specific knowledge (Shen et al., 2017). Several deep learning models of multimodal medical image fusion are proposed to employ multimodal neuroimaging data for brain disease diagnosis (Rajalingam and Priya, 2018). However, existing studies usually focus on combining feature representation of multiple modalities and ignore significant inter-modality heterogeneity (Huang et al., 2019). To this end, we propose an adaptive multimodal neuroimage integration (AMNI) framework for automated MDD diagnosis based on resting-state functional MRI and T1-weighted structural MRI data. The proposed method can not only extract high-level feature representations of structural and functional data *via* CNN and GCN, respectively, but also alleviate the heterogeneity between modalities with the help of a unique feature adaptation module.

3. MATERIALS AND METHODS

In this section, we first introduce the materials and image pre-processing method used in this work, and then present the proposed method and implementation details.

3.1. Materials

3.1.1. Data Acquisition

Resting-state fMRI and T1-weighted structural MRI data were acquired from 282 MDD subjects and 251 healthy controls (HCs) recruited from the Southwest University, an imaging site of the REST-meta-MDD consortium (Yan et al., 2019). Resting-state fMRI were acquired through a Siemens scanner with the following parameters: repetition time (TP) $= 2,000\,ms$, echo time (TE)$=\ 30\,ms$, flip angle$=\ 90^o$, slice thickness$=\ 3.0\,mm$, gap$=\ 1.0\,ms$, time point$=\ 242$, voxel size$=\ 3.44 \times 3.44 \times 4.00\,mm^3$. More detailed information can be found online[1]. The demographic and clinical information of these studied subjects is summarized in **Table 1**.

3.1.2. Image Pre-processing

The resting-state fMRI and structural T1-weighted MRI scans were pre-processed using the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DPARSF) software (Yan and Zang, 2010) with a standardized protocol (Yan et al., 2016). For rs-fMRI data, we first discard the first 10 volumes the initial 10 volumes were discarded, and slice-timing correction was performed. Then, the time series of images for each subject were realigned using a six-parameter (rigid body) linear transformation. After realignment, individual T1-weighted images were co-registered to the mean functional image using a 6 degrees-of-freedom linear transformation without re-sampling and then segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF). Finally, transformations from individual native space to MNI space were computed with the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) tool (Ashburner, 2007). After that, the fMRI data were normalized with an EPI template in the MNI space, and resampled to the resolution of $3 \times 3 \times 3\,mm^3$, followed by spatial smoothing using a $6\,mm$ full width half maximum Gaussian kernel. Note that subjects with poor image quality or excessive head motion (mean framewise-displacement >0.2 mm) were excluded from analysis (Jenkinson et al., 2002). Finally, we extracted the mean rs-fMRI time series with band-pass filtering ($0.01 - 0.1\,Hz$) of a set of 112 pre-defined regions-of-interest (ROIs), including cortical and subcortical areas based on the Harvard-Oxford atlas. Each T1-weighted structural MR image was also segmented into three tissues (i.e., GM, WM, and CSF) and transformed into the MNI space with DARTEL tool (Ashburner, 2007), resulting in a 3D volume (size: 121 × 145 × 121). Here, we employ gray matter volume in the MNI space for representing the original sMRI.

3.2. Proposed Method

As illustrated in **Figure 1**, the proposed AMNI consists of four major components: (1) a GCN module to extract features from rs-fMRI, (2) a CNN module to extract features from T1-weighted sMRI, (3) a feature adaptation module to reduce inter-modality discrepancy, and (4) a feature fusion module for classification, with details introduced below.

3.2.1. GCN for Functional MRI Feature Learning

Based on resting-state fMRI data, one usually constructs a functional connectivity matrix/network (FCN) for representing each subject, with each node representing a specific brain ROI and each edge denoting the pairwise functional connection/relationship between ROIs (Honey et al., 2009; Dvornek et al., 2017). That is, FCNs help capture the dependencies between BOLD signals of paired ROIs. Considering the fact that FCNs are non-Euclidean data, we treat each functional connectivity network as a specific graph and resort to spectral graph convolutional network (GCN) for FCN feature learning by capturing graph topology information. Previous studies have shown that GCN is effective in learning graph-level representations by gradually aggregating feature vectors of all nodes (Yao et al., 2019). In this work, we aim to learn graph-level representations based on node representations of input FCNs.

(**i**) **Graph Construction**. Denote N and M as the numbers of ROIs and time points, respectively, where $N = 112$ and $M = 232$ in this work. We assume that the rs-fMRI time-series data for a subject is $Y = (y_1, \cdots, y_N)^T \in R^{N\times M}$, where each element $y_n \in R^M$ $(n = 1, \cdots, N)$ denotes BOLD measurements of the n-th ROI at M successive time points.

As the simplest and most widely used method, Pearson correlation (PC) is usually used to construct functional connectivity networks from raw rs-fMRI time-series data. Denote $B = (b_{ij}) \in R^{N\times N}$ as the functional connectivity matrix based on the Pearson correlation algorithm. Each element $b_{ij} \in [-1, 1]$ in B represents the Pearson correlation coefficient between the i-th and j-th ROIs, defined as follows:

$$b_{ij} = \frac{(y_i - \bar{y}_i)^T(y_j - \bar{y}_j)}{\sqrt{(y_i - \bar{y}_i)^T(y_i - \bar{y}_i)}\sqrt{(y_j - \bar{y}_j)^T(y_j - \bar{y}_j)}} \tag{1}$$

where $\bar{y}_i$ and $\bar{y}_j$ are the mean vector corresponding to $y_i \in R^M$ and $y_j \in R^M$, respectively, and M represents the length of time points of BOLD signals in each brain region.

[1] http://rfmri.org/REST-meta-MDD

TABLE 1 | Demographic and clinical information of subjects from Southwest University [a part of the REST-meta-MDD consortium (Yan et al., 2019)].

Category	Gender	Age	Education	First period	On medication	Duration of illness
MDD	99 M 183 F	38.7 ± 13.6	10.8 ± 3.6	209 (Y)/49 (N) 24 (D)	124 (Y)/125 (N) 33 (D)	50.0 ± 65.9 35 (D)
HC	87 M 164 F	39.6 ± 15.8	13.0 ± 3.9	–	–	–

Values are reported as Mean ± Standard deviation. M, Male; F, Female; Y, Yes; N, No; D, Lack of record.

For each subject, we regard each brain FCN as an undirected graph $G = \{V, E\}$, where $V = \{v_1, \cdots, v_N\}$ is a set of N nodes/ROIs and $b_{ij} \in B$ denotes the functional connectivity between a paired nodes v_i and v_j. Since spectral GCNs work on adjacency matrices by updating and aggregating node features (Bruna et al., 2013), it is essential to generate such an adjacency matrix A and a node feature matrix X from each graph G.

To reduce the influence of noisy/redundant information, we propose to construct a K-Nearest Neighbor (KNN) graph based on each densely-connected functional connectivity matrix. Specifically, a KNN graph is generated by only keep the top k important edges according to their functional connectivity strength (i.e., PC coefficient) for each node. Then, the topology structure of the graph G can be described by adjacency matrix $A = (a_{ij}) \in \{0,1\}^{N\times N}$, where $a_{ij} = 1$ if there exists an edge between the i-th and the j-th ROIs, and $a_{ij} = 0$, otherwise. In addition, the node features are defined by the functional connection weights of edges connected to each node, i.e., corresponding to a specific row in the functional connectivity matrix. Thus, the node features of the graph G can be represented by the node feature matrix $X = B$.

(**ii**) **Graph Feature Learning**. In GCN models, the convolution operation on the graph is defined as the multiplication of filters and signals in the Fourier domain. Specifically, GCN model learns new node representations by calculating the weighted sum of feature vectors of central nodes and the neighboring nodes. Mathematically, the simplest spectral GCN layer (Kipf and Welling, 2016) can be formulated as:

$$H^{l+1} = f(H^l, A) = \sigma(\widetilde{A} H^l W^l) \tag{2}$$

where H^l is the matrix of activations in the l-th layer, and W^l is a layer-specific trainable weight matrix.

In addition, $\widetilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ is the normalized adjacency matrix with self loops, and $\sigma(\cdot)$ is an activation function, such as the $ReLU(\cdot) = max(0, \cdot)$. In addition, D is the diagonal degree matrix, with the i-th diagonal element defined as $d_i = \sum_{i\neq j} A_{ij}$.

In the GCN module in our AMNI framework, we stack two graph convolutional layers with the adjacency matrix A and node features matrix X as inputs. The output of this two-layer GCN module is calculated as:

$$Z = f(A, X) = ReLU(\widetilde{A} ReLU(\widetilde{A} X W^{(0)}) W^1) \tag{3}$$

Note that the number of neurons in the two graph convolutional layers is set as 64 and 64, respectively.

Given that this is a graph classification task, we employ a simple graph pooling strategy (Lee et al., 2019) to generate graph-level FCN representations. To be specific, we employ both global average pooling and global max pooling that aggregate node features to generate new feature representations. The output feature of the graph pooling layer is as follows:

$$g_F = \frac{1}{N}\sum_{i=1}^{N} z_i || \max_{i=1}^{N} z_i \tag{4}$$

where N is the number of ROIs, z_i is the feature vector of i-th ROI obtained by the graph convolution operation, and $||$ denotes concatenation.

By stacking multiple graph convolution layers and graph pooling layers, GCN can learn higher-order node features from neighboring nodes. In addition, GCN propagates information on a graph structure and gradually aggregates the information of neighboring nodes, which allows us to effectively capture the complex dependencies among ROIs.

3.2.2. CNN for Structural MRI Feature Learning

In recent years, convolutional neural networks (CNNs) have shown much predomination in image recognition and classification (Simonyan and Zisserman, 2014; He et al., 2016). Due to the 3D nature of structural MR images (sMRI), it is important to learn feature representations of all three dimensions from volumetric medical data. Considering that 3D convolutional kernels can encode richer spatial information, we adopt 3D CNN model to extract feature representations of T1-weighted MRIs.

In the AMNI framework, the CNN module consists of four convolution blocks and two fully-connected (FC) layers for local to global sMRI feature extraction. To be specific, each convolution block consists of one convolutional layer, one batch normalization layer, one activation function and one max pooling layer. To capture local patterns, 3D convolution is achieved by convolving a 3D kernel over 3D feature cubes. Formally, the j-th feature map in the i-th layer, denoted as $v_{i,j}$, is given by

$$v_{i,j} = f((W_{i,j} * V_{i-1}) + b_{i,j}) \tag{5}$$

where $W_{i,j}$ and $b_{i,j}$ are the kernel weights and the bias for the j-th feature map, respectively, V_{i-1} are the sets of input feature maps connected to the current layer from the $(i-1)th$ layer, $*$ is the convolution operation, and f is the non-linear activation function. The size of each convolution filter is 3 × 3 × 3, and the numbers of convolution filters are set to 16, 32, 64, 128, respectively. In addition, max pooling is applied for each 2 × 2 × 2 region which reduces the spatial size of the feature maps and the number of parameters, and ReLU is used as the activation function. Meanwhile, batch normalization technique can promote faster convergence and better generalization of trained networks.

For the pooling layer, we use the Global Average Pooling (GAP) operation (Lin et al., 2013), which performs downsampling by computing the mean of the height, width, and depth dimensions of the input. The formula for GAP is as follows:

$$g_j = \frac{\sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{d=1}^{D} v_j^{h,w,d}}{H \times W \times D} \tag{6}$$

where $v_j^{h,w,d}$ is the value at position (h, w, d) of the j-th input feature map, H, W, and D are the height, width, and depth respectively and g_j is getting value of the j-th input feature map through GAP. Thus, the sMRI feature g_S generated by CNN is given by:

$$g_S = [g_1, g_2, \cdots, g_c]^T \tag{7}$$

where c is the number of input feature map. It can be seen that the GAP layer converts a 4D tensor to a 1-dimensional feature vector, thus significantly reducing the number of network parameters.

The two fully-connected layers have 128 and 64 neurons, respectively. To avoid overfitting, we employ the dropout technique (Srivastava et al., 2014), with a probability of 0.5 after each fully-connected layer. More detailed information about the CNN architecture can be found in **Table 2**.

3.2.3. Feature Adaptation Module

Due to the heterogeneous nature of multimodal data, it is necessary to reduce the discrepancy between feature representations of different modalities before feature fusion. Inspired by existing studies on domain adaptation (Tzeng et al., 2014), we employ a cross-modal loss based on maximum mean discrepancy (MMD) (Gretton et al., 2012) to re-calibrate channel-wise features extracted from sMRI and fMRI. Denote G_F and G_S as feature representations of fMRI and sMRI, respectively. The cross-modal MMD loss L_M is formulated as follows:

$$\begin{aligned} L_M &= MMD(G_F, G_S) \\ &= \left\| \frac{1}{|G_F|} \sum_{g_F \in G_F} \phi(g_F) - \frac{1}{|G_S|} \sum_{g_S \in G_S} \phi(g_S) \right\| \end{aligned} \tag{8}$$

where $\phi(\cdot)$ denotes the feature map associated with the kernel map, and g_F and g_S are elements in G_F and G_S, respectively. During model training, the cross-modal MMD loss will be used as a regularization term to penalize heterogeneity of the features between the two modalities.

As shown in **Figure 1**, this cross-modal MMD loss is applied to features from two fully-connected layers in the proposed CNN and GCN modules. This would enable the feature adaptation module to learn shared and aligned information across modalities by minimizing the distribution difference between two feature representations.

3.2.4. Feature Fusion Module

To enable our AMNI method to capture the complementary information provided by functional and structural MRIs, we also design a feature fusion module for classification/prediction.

Assuming that F_1 and F_2 are two feature representations obtained by feature adaptation module, we first concatenate them to obtain a new representation. The new representation F can be described as follows:

$$F = [F_1^T, F_2^T]^T \tag{9}$$

After concatenation, the obtained new representation is fed to two fully-connected layers (with 64 and 2 neurons, respectively), and the learned features are further fed into a Softmax layer for classification.

During the training stage, we use the cross-entropy loss function to optimize the parameters in our AMINI model. The classification loss L_C is defined as:

$$L_C = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i log(p_i) + (1 - y_i) log(1 - p_i) \right) \tag{10}$$

where N is the number of samples, and y_i is the true label of the i-th sample, with 1 representing the sample being a MDD patient and 0 denoting the sample being a healthy control. In addition, p is the predicted probability that the sample belongs to the MDD category.

In our model, we aim to minimize not only the classification loss, but also the cross-modal loss to reduce the inter-modality difference. Hence, the total loss function L of the proposed AMNI is defined as follows:

$$L = L_C + \lambda L_M \tag{11}$$

where λ is a hyperparameter to tune the contributions of two terms in Equation (11).

3.3. Implementation Details

We optimize the proposed AMNI model *via* the Adam (Kingma and Ba, 2014) algorithm, with the learning rate of 0.0001, weight decay rate of 0.0015, training epoch of 100, and mini-batch size of 16. The proposed model is implemented based on Pytorch (Paszke et al., 2017), and the model is trained by using a single GPU (NVIDIA Quadro RTX 6000 with 24 GB memory). The hyperparameter λ in Equation (11) is empirically set as 0.01. And we will experimentally investigate its influence in Section 5.

4. EXPERIMENTS

In this section, we introduce experimental settings and several competing methods, present the experimental results, and visualize feature distributions of different methods.

TABLE 2 | Architecture of the CNN module in the proposed AMINI framework.

Layer	Kernel size	Stride	Output size	Feature volumes
Input	–	–	121×145×121	1
C1	3×3×3	1	121×145×121	16
M1	2×2×2	2	60×72×60	16
C2	3×3×3	1	60×72×60	32
M2	2×2×2	2	30×36×30	32
C3	3×3×3	1	30×36×30	64
M3	2×2×2	2	15×18×15	64
C4	3×3×3	1	15×18×15	128
M4	2×2×2	2	7×9×7	128
GAP	–	–	1×1×1	128
FC	–		1×1×1	64

Cn, the n-th convolutional layer; Mn, the n-th max pooling layer; GAP, global average pooling; FC, fully-connected layer.

4.1. Experimental Settings

We randomly select 80% samples as training data, and the remaining 20% samples are used as test data. To avoid bias introduced by random partition, we repeat the random partition procedure 10 times independently, and record the mean and standard deviation results. Eight metrics are used to evaluate the performance of different methods in the task of MDD detection (i.e., MDD vs. HC classification), including accuracy (ACC), sensitivity (SEN), specificity (SPE), balanced accuracy (BAC), positive predicted value (PPV), negative predictive value (NPV), F1-Score (F1), and area under the receiver operating characteristic curve (AUC).

4.2. Methods for Comparison

In this work, we compare the proposed AMNI method with six traditional machine learning methods and three popular deep learning methods. More details can be found below.

(1) **PCA+SVM-s**: The PCA+SVM-s method only uses sMRI data. The 3D image of the whole brain is down-sampled from 121 × 145 × 121 to 61 × 73 × 61, and further flattened into a vectorized feature representation for each subject. We use principal component analysis (PCA) (Wold et al., 1987) by keeping the top 32 principal components to reduce feature dimension based on the above feature representations of all subjects. Finally, the support vector machine (SVM) with Radial Basis Function (RBF) kernel is employed for classification.

(2) **EC+SVM**: The EC+SVM method uses rs-fMRI data. Similar to our AMNI, we first construct a functional connectivity matrix based on Pearson correlation coefficient for each subject. We then extract eigenvector centralities (EC) (Bonacich, 2007), which measure a node's importance while giving consideration to the importance of its neighbors in the FC network, as features of the FCN and feed these 112-dimensional features into an SVM classifier with RBF kernel for disease detection.

(3) **DC+SVM**: Similar to EC+SVM, the DC+SVM method first constructs a FCN based on Pearson correlation coefficient for each subject, and then extracts degree centrality (DC) (Nieminen, 1974) as FCN features by measuring node importance based on the number of links incident upon a node. The 112-dimensional DC features are finally feed into an SVM for classification.

(4) **CC+SVM**: Similar to EC/CC+SVM, this method extracts the local clustering coefficient (CC) (Wee et al., 2012) to measure clustering degree of each node in each FCN. The 112-dimensional CC features are fed into an SVM for classification.

(5) **PCA+SVM-f**: In the PCA+SVM-f method, the upper triangle of a FC matrix is flattened into a vector for each subject after the FC matrix is constructed. Then, we use PCA by keeping the top 32 principal components to reduce feature dimension based on the above feature representations of all subjects. Finally, an SVM is used for classification.

(6) **PP+SVM**: In this method, we integrate rs-fMRI and sMRI features for classification based on SVM. Specifically, we first employ PCA+SVM-s and PCA+SVM-f to extract features from structural and functional MRIs, respectively. Then, we concatenate features of these two modalities for the same subject, followed by an SVM for classification.

(7) **2DCNN**: In this method, we employ the original FC matrix of each subject as input of a CNN model (LeCun et al., 1989). Specifically, this CNN contains three convolutional layers and two fully-connected layers. Each convolutional layer is followed by batch normalization and ReLU activation. The channel numbers for the three convolutional layers are 4, 8, and 8, respectively, and the corresponding size of the convolution kernel is 3 × 3, 5 × 5, 7 × ×7, respectively. The two fully-connected (FC) layers contain 4, 096 and 2 neurons, respectively.

(8) **ST-GCN**: We also compare our method with the spatio-temporal graph convolutional network (ST-GCN), a state-of-the-art method for modeling spatio-temporal dependency of fMRI data (Gadgil et al., 2020). Specifically, the ST-GCN comprises two layers of spatio-temporal graph convolution (ST-GC) units, global average pooling and a fully connected layer. Note that each ST-GC layer produces 64-channel outputs with the temporal kernel size of 11, a stride of 1, and a dropout rate of 0.5.

(9) **3DCNN+2DCNN**: In this method, we employ 3DCNN and 2DCNN to extract features from sMRI and fMRI, respectively. We then concatenate features learned from 3DCNN and 2DCNN, and feed the concatenated features to a fully-connected layer and the softmax layer for classification.

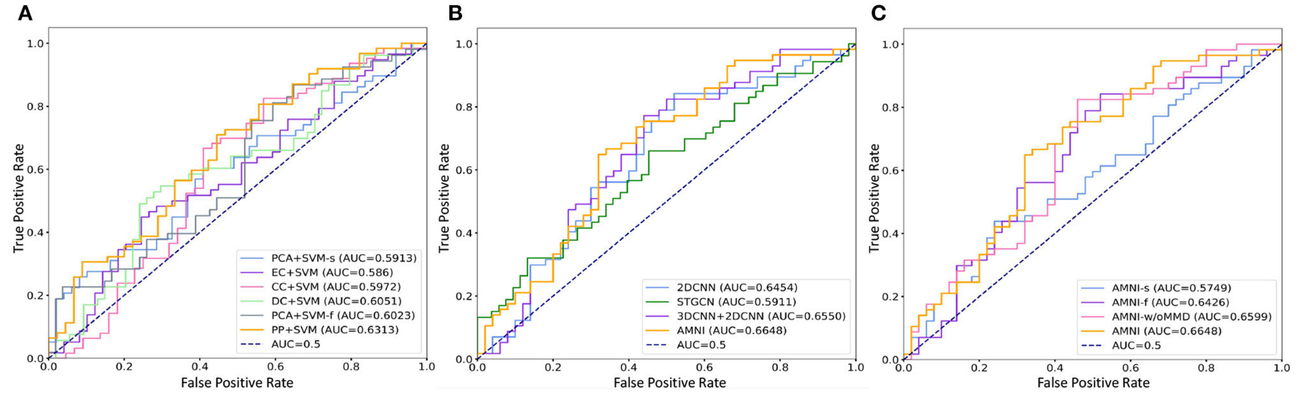

FIGURE 2 | ROC curves and related AUC values achieved by different methods in MDD vs. HC classification. **(A)** AMNI vs. six conventional methods. **(B)** AMNI vs. three deep learning methods. **(C)** AMNI vs. its three variants.

TABLE 3 | Classification results in terms of "mean (standard deviation)" achieved by ten methods in MDD vs. HC classification, with best results shown in bold.

Method	Data	ACC	SEN	SPE	BAC	PPV	NPV	F1	AUC
PCA+SVM-s	S	0.566 (0.011)	0.669 (0.021)	0.456 (0.007)	0.563 (0.010)	0.580 (0.006)	0.553 (0.017)	0.618 (0.013)	0.591 (0.008)
EC+SVM	F	0.560 (0.014)	0.651 (0.009)	0.462 (0.029)	0.557 (0.015)	0.577 (0.013)	0.539 (0.018)	0.609 (0.009)	0.586 (0.019)
CC+SVM	F	0.574 (0.007)	0.674 (0.018)	0.470 (0.014)	0.572 (0.006)	0.589 (0.005)	0.562 (0.011)	0.625(0.009)	0.597(0.014)
DC+SVM	F	0.578 (0.014)	0.676 (0.019)	0.477 (0.016)	0.577 (0.017)	0.593 (0.015)	0.568 (0.021)	0.627 (0.014)	0.605 (0.015)
PCA+SVM-f	F	0.570 (0.011)	0.653 (0.014)	0.483 (0.019)	0.568 (0.012)	0.588 (0.010)	0.554 (0.016)	0.614 (0.009)	0.602 (0.013)
PP+SVM	SF	0.593 (0.026)	0.675 (0.022)	0.502 (0.036)	0.588 (0.027)	0.605 (0.026)	0.578 (0.030)	0.636 (0.022)	0.631 (0.027)
2DCNN	F	0.613 (0.013)	0.670 (0.022)	0.551 (0.024)	0.611 (0.013)	0.628 (0.013)	0.599 (0.016)	0.643 (0.014)	0.645 (0.013)
STGCN	F	0.583(0.022)	0.616 (0.027)	0.544 (0.026)	0.580 (0.022)	0.612 (0.015)	0.548 (0.037)	0.614 (0.018)	0.591 (0.008)
3D+2DCNN	SF	0.632 (0.028)	0.667 (0.022)	0.593 (0.043)	0.630 (0.029)	**0.649 (0.034)**	0.617(0.041)	0.656 (0.026)	0.655 (0.013)
AMNI (Ours)	SF	**0.650 (0.016)**	**0.694 (0.068)**	**0.609 (0.056)**	**0.651 (0.016)**	0.640 (0.031)	**0.667 (0.055)**	**0.663 (0.021)**	**0.665 (0.017)**

S, sMRI; F, fMRI; SF, sMRI+fMRI.

4.3. Experimental Results

The quantitative results of the proposed AMNI and nine competing methods in the task of MDD vs. HC classification are reported in **Table 3**. In **Figures 2A,B**, we also show ROC curves of different methods. From **Table 3** and **Figures 2A,B**, we have the following interesting observations.

First, our AMNI and two deep learning methods (i.e., 2DCNN and 3DCNN+2DCNN) generally achieve better performance in terms of eight metrics, compared with six traditional machine learning methods. For example, in terms of ACC values, the AMNI yields the performance improvement of 5.7%, compared with the best traditional machine learning method (e.g., PP+SVM) in MDD detection. These results demonstrate that, deep learning methods that can learn diagnosis-oriented neuroimage features is more effective in MDD detection, compared with traditional machine learning methods that rely on handcrafted features. *Second*, three multimodal methods (i.e., PP+SVM, 3DCNN+2DCNN, and AMNI) generally outperform their single-modality counterparts (i.e., PCA+SVM-s, PCA+SVM-f, and 2DCNN). For instance, both our AMNI and 3DCNN+2DCNN methods that integrate sMRI and fMRI data are superior to 2DCNN which only use functional data. This implies that taking advantage of multimodal MRIs (as we do in this work) helps promote the diagnosis performance, thanks to the complementary information provided by functional and structural MRIs. Furthermore, our proposed AMNI achieves better performance in terms of most metrics, compared with eight competing methods. These results imply that adaptive integration of multimodal neuroimages helps boost the performance of MDD identification.

4.4. Statistical Significance Analysis

We further calculate predicted probability distribution difference on test data between our model and each of eight competing methods by paired sample t-test. Denote u_1 and u_2 as the population mean of predicted probability distributions from our AMNI and one competing method, respectively. The hypotheses can be expressed as follows:

$$\begin{aligned} H_0 &: u_1 = u_2 \\ H_1 &: u_1 \neq u_2 \end{aligned} \tag{12}$$

where H_0 is the null hypothesis, meaning that our model and the competing method do not have significant difference. And H_1 is the alternative hypothesis, meaning that our model and the competing method have significance difference. The test statistic for the paired samples t-test is as follows:

$$t = \frac{\bar{x}_{\text{diff}}}{s_{\text{diff}}/\sqrt{n}} \tag{13}$$

where $\bar{x}_{\text{diff}}$ is sample mean of the differences, s_{diff} is sample standard deviation of the differences and n is the sample size (i.e., number of pairs). The p-values that corresponds to the test statistic t are shown in **Table 4**.

As shown in **Table 4**, all obtained p-values are less than our chosen significance level (i.e., 0.05). Therefore, H_0 is rejected, which means that our AMNI method differs significantly from each of the eight competing methods.

TABLE 4 | Results of statistical significance analysis between the proposed AMNI and eight competing methods.

Pairwise comparison	p-value	$p < 0.05$
AMNI vs. PCA+SVM-s	3.40×10^{-4}	Yes
AMNI vs. EC+SVM	3.93×10^{-4}	Yes
AMNI vs. CC+SVM	3.16×10^{-4}	Yes
AMNI vs. DC+SVM	2.43×10^{-4}	Yes
AMNI vs. PCA+SVM-f	1.01×10^{-5}	Yes
AMNI vs. PP+SVM	2.71×10^{-5}	Yes
AMNI vs. 2DCNN	9.48×10^{-3}	Yes
AMNI vs. 3DCNN+2DCNN	1.07×10^{-3}	Yes

4.5. Feature Visualization

In **Figure 3**, we visualize the data distributions of features derived from two multimodal methods (i.e., PP+SVM and AMNI) *via* t-SNE (Van der Maaten and Hinton, 2008). Note that the features of PP+SVM are generated by concatenating handcrafted features from two modalities, while the features of our AMNI are extracted based on an end-to-end deep learning model (see **Figure 1**). As shown in **Figure 3**, the feature distributions of two categories (i.e., MDD and HC) generated from our AMNI method have more significant difference, while their feature distribution gap is not evident for the PP+SVM method. This may indicate that our AMNI can learn more discriminative features for MDD detection by explicitly reducing the inter-modality discrepancy, compared with the traditional PP+SVM method.

5. DISCUSSION

5.1. Ablation Study

To evaluate the effectiveness of each component in the proposed AMNI, we further compare AMNI with its three variants: (1) **AMNI-s** that only uses CNN branch and feature fusion module of AMNI, without considering functional MRI, (2) **AMNI-f** that only uses GCN branch and feature fusion module of AMNI, without considering structural MRI, (3) **AMNI-w/oMMD** that directly feeds concatenated fMRI and sMRI features (*via* GCN and CNN modules, respectively) into the feature fusion module for classification, without using the proposed feature adaption module. The experimental results are reported in **Figures 4**, **2C**.

It can be seen from **Figure 4** that two multimodal methods (i.e., AMNI-w/oMMD and AMNI) generally outperform the single modality methods (i.e., AMNI-s and AMNI-f). This further demonstrates that multimodal data can provide complementary information to help boost the performance of MDD identification. Besides, our AMNI achieves consistently better performance compared with AMNI-w/oMMD that ignores the heterogeneity between the two modalities. These results further validate the effectiveness of the proposed feature adaption module in alleviating the inter-modality discrepancy between different modalities. In addition, **Figure 2C** suggests that our proposed AMNI achieves good ROC performance and the best AUC value compared with its three variants.

5.2. Influence of Hyperparameter

The hyperparameter λ in Equation (11) is used to tune the contribution of the proposed feature adaptation module for re-calibrating feature distributions of two modalities. We now report the classification accuracy of the proposed AMNI with different values of λ in **Figure 5**. As shown in **Figure 5**, with $\lambda = 0.01$, our AMNI can achieve best performance. But using a too large value (e.g., $\lambda = 1$) will yield worse performance. A possible reason is that focusing too much on the reduction of differences between modalities (with a large λ) may lose the specific and unique information of each modality, thereby degrading the learning performance.

5.3. Influence of Graph Construction Strategy

In the main experiment, we build a KNN graph to generate an adjacency matrix for each FCN. To investigate the influence of the use of different graph construction strategies, besides KNN, we also construct a fully-connected graph and a threshold graph to generate the adjacency matrix, respectively. For the fully-connected graph, we directly take $A = (|w_{ij}|)$ as the adjacency matrix, which is an edge-weighted graph. For the threshold graph, we generate the adjacency matrix A by binarizing the FC matrix B to regulate the sparsity of the graph. Thus, the adjacency matrix can be described as $A = (a_{ij}) \in \{0, 1\}^{N \times N}$, where $a_{ij} = 1$ if the connection coefficient between i-th and j-th ROI is greater than a threshold q; and $a_{ij} = 0$, otherwise. The threshold q is set as 0.2 here. The experimental results of our AMNI with three different graph construction strategies are reported in **Figure 6**.

As can be seen from **Figure 6**, our AMNI model based on KNN graph outperforms its two variants that use fully-connected graph and threshold graph. The underlying reason could be that KNN graph can preserve node-centralized local topology information while removing noisy/redundant information in graph (Ktena et al., 2018; Yao et al., 2021).

5.4. Influence of Network Architecture

To explore the influence of different network architectures of AMNI on the experimental results, we adjust the the network depth of two branches of the AMNI model, respectively. *On the one hand*, with the CNN branch fixed, we vary the number of graph convolutional layers for the GCN branch of AMNI

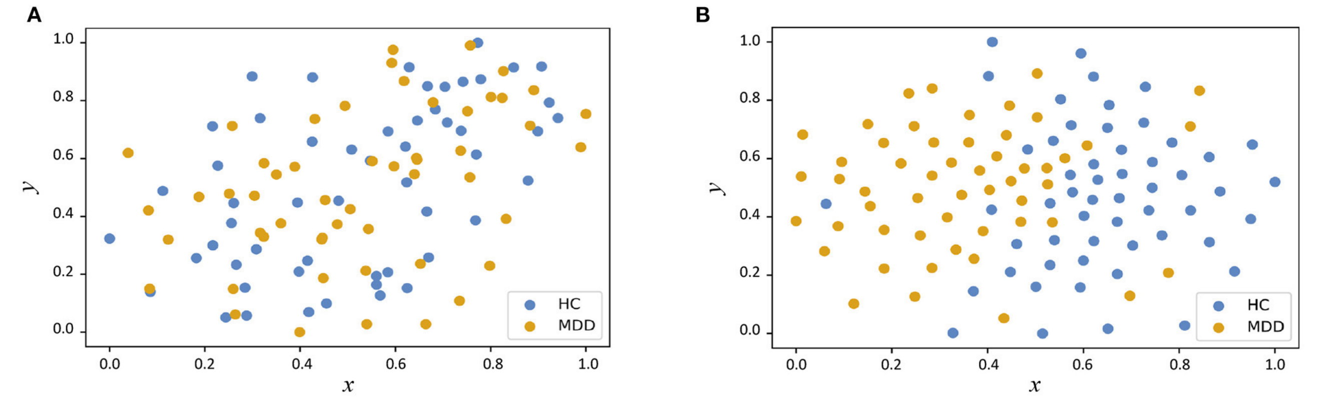

FIGURE 3 | Visualization of feature distributions from the PP+SVM and the proposed AMNI models via t-SNE (Van der Maaten and Hinton, 2008). The horizontal and vertical axes denote two dimensions after feature mapping. **(A)** Distribution of features derived from PP+SVM. **(B)** Distribution of features derived from AMNI.

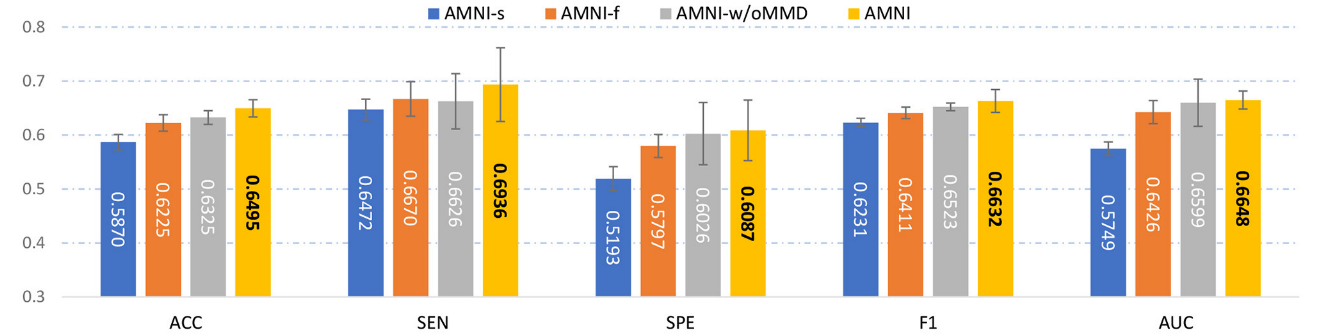

FIGURE 4 | Performance of our AMNI and its three variants in the task of MDD vs. HC classification, with best results shown in bold.

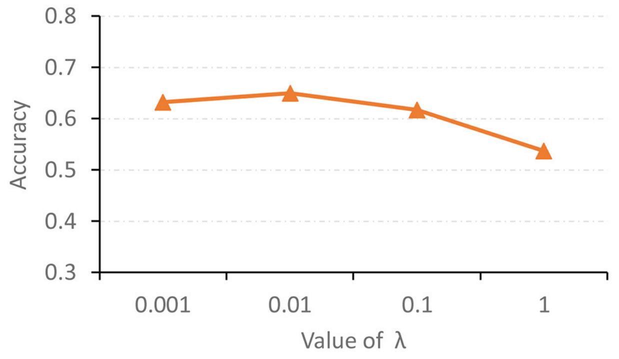

FIGURE 5 | Accuracy achieved by the proposed AMNI method with different values of λ in Equation (11) in the task of MDD vs. HC classification.

and report the corresponding results of AMNI in **Table 5**. This table shows that the AMNI achieves the overall best performances (e.g., ACC=0.6495 and AUC=0.6648) with two graph convolutional layers in the GCN branch. In addition, as the number of graph convolutional layers increases (see AMNI-G3 and AMNI-G4), the performance is not good. This may be due to the over-smoothing problem (that is, Laplacian smoothing makes the node representations more similar as the graph convolutional layer increases; Yang et al., 2020), which may reduce the discriminative compatibility of learned features. *On the other hand*, we fix the GCN branch and vary the architecture of the CNN in AMNI for performance evaluation. Specifically, we vary the number of convoluational layers in CNN within [3, 6] and report the results of AMIN in MDD vs. HC classification in **Table 5**. This table shows that fine-tuning the network architecture of the CNN branch in AMNI achieves comparable results, which implies that our AMNI is robust to different network architectures. Further, AMNI with five convoluational layers in the CNN branch (e.g., AMNI-G5) achieves better performance in terms of accuracy, sensitivity, balanced accuracy, positive predicted value and F1-Score.

Besides, we also further discuss the influence of network width of each branch on the experimental results. *For one thing*, with the CNN branch fixed, we change the number of neurons in the graph convolutional layers and then report the corresponding results of AMNI in **Table 6**. It can be found from **Table 6** that the AMINI model using different numbers of neurons in graph convolutional layers achieves comparable experimental results,

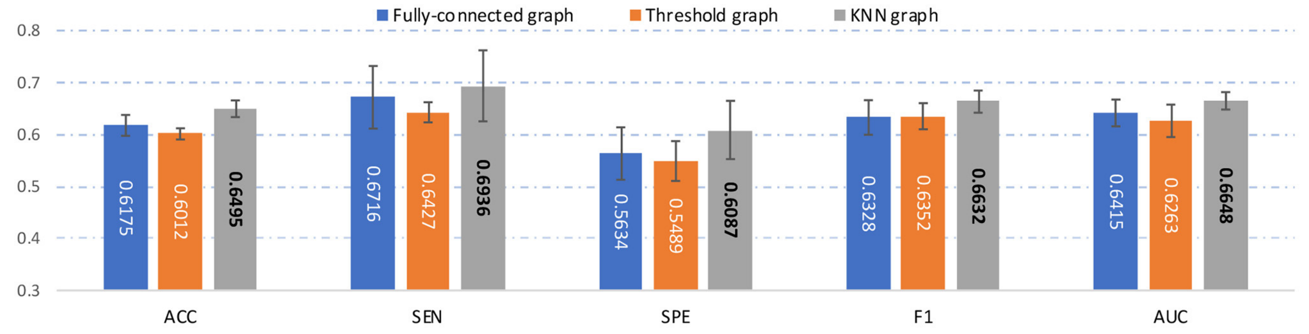

FIGURE 6 | Results of the proposed AMNI based on three different graph construction methods (e.g.fully-connected graph, threshold graph, and KNN graph) in the task of MDD vs. HC classification, with best results shown in bold.

TABLE 5 | Classification results of our AMNI in MDD vs. HC classification with different network depth, with best results shown in bold.

Method	ACC	SEN	SPE	BAC	PPV	NPV	F1	AUC
AMNI-G1	0.634 (0.014)	0.677 (0.065)	0.587 (0.054)	0.632 (0.019)	**0.669 (0.041)**	0.598 (0.077)	**0.669 (0.011)**	0.627 (0.032)
AMNI-G2	**0.650 (0.016)**	**0.694 (0.068)**	**0.609 (0.056)**	**0.651(0.016)**	0.640 (0.031)	**0.667 (0.055)**	0.663 (0.021)	**0.665 (0.017)**
AMNI-G3	0.595 (0.008)	0.629 (0.034)	0.559 (0.041)	0.594 (0.010)	0.600 (0.010)	0.590 (0.019)	0.614 (0.016)	0.605 (0.009)
AMNI-G4	0.587 (0.011)	0.618 (0.023)	0.554 (0.025)	0.586 (0.011)	0.610 (0.042)	0.561 (0.036)	0.613 (0.022)	0.599 (0.022)
AMNI-C3	0.628 (0.005)	0.692 (0.045)	0.551 (0.057)	0.622 (0.007)	0.647 (0.014)	0.603 (0.012)	0.668 (0.013)	0.622 (0.007)
AMNI-C4	0.650 (0.016)	0.694 (0.068)	**0.609 (0.056)**	0.651 (0.016)	0.640 (0.031)	**0.667 (0.055)**	0.663 (0.021)	**0.665 (0.017)**
AMNI-C5	**0.660 (0.022)**	**0.742 (0.042)**	0.565 (0.049)	**0.653 (0.023)**	**0.663 (0.011)**	0.657 (0.040)	**0.700 (0.020)**	0.653 (0.023)
AMNI-C6	0.642 (0.014)	0.701 (0.046)	0.580 (0.041)	0.641 (0.017)	0.651 (0.029)	0.634 (0.053)	0.673 (0.008)	0.628 (0.018)

Note that AMNI-Gn contains n graph convolutional layers in the GCN module of AMNI, and AMNI-Cn contains n convolutional layers in the CNN module of AMNI.

TABLE 6 | Classification results of our AMNI in MDD vs. HC classification with different network width, with best results shown in bold.

Method	ACC	SEN	SPE	BAC	PPV	NPV	F1	AUC
AMNI-g40	0.620 (0.035)	0.626 (0.089)	0.614 (0.097)	0.620 (0.035)	0.652 (0.039)	0.593 (0.040)	0.635 (0.049)	0.650 (0.036)
AMNI-g64	**0.650 (0.016)**	0.694 (0.068)	0.609 (0.056)	**0.651 (0.016)**	0.640 (0.031)	**0.667 (0.055)**	0.663 (0.021)	0.665 (0.017)
AMNI-g88	0.626 (0.015)	**0.697 (0.048)**	0.542 (0.052)	0.620 (0.015)	0.644 (0.016)	0.604 (0.023)	**0.669 (0.021)**	**0.667 (0.011)**
AMNI-g112	0.631 (0.016)	0.647 (0.053)	**0.612 (0.037)**	0.629 (0.015)	**0.659 (0.015)**	0.602 (0.024)	0.651 (0.029)	0.637 (0.037)
AMNI-c1	0.598 (0.017)	0.643 (0.046)	0.535 (0.081)	0.589 (0.028)	**0.643 (0.028)**	0.535 (0.073)	0.642 (0.026)	0.607 (0.0148)
AMNI-c2	0.630 (0.020)	0.693 (0.080)	0.575 (0.096)	0.634 (0.016)	0.593 (0.033)	**0.685 (0.029)**	0.635 (0.023)	0.667 (0.004)
AMNI-c3	**0.650 (0.016)**	**0.694 (0.068)**	0.609 (0.056)	**0.651 (0.016)**	0.640 (0.031)	0.667 (0.055)	**0.663 (0.021)**	0.665 (0.017)
AMNI-c4	0.641 (0.015)	0.654 (0.051)	**0.629 (0.030)**	0.642 (0.015)	0.628 (0.030)	0.658 (0.044)	0.638 (0.028)	**0.689 (0.042)**

Note that AMNI-gn contains n neurons in the graph convolutional layers of the GCN module. And the filter sequences in CNN module of AMNI-c1, AMNI-c2, AMNI-c3 and AMNI-c4 are [4, 8, 16, 32], [8, 16, 32, 64], [16, 32, 64, 128], *and* [32, 64, 128, 256], *respectively.*

which means our model is not very sensitive to the change of network width of the GCN branch. *For another thing*, with the GCN branch fixed, we change the number of filters in each 3D convolutional layer and record the results in **Table 6**. As shown in **Table 6**, with the increase of the number of filters in 3D CNN module of AMNI, the model (i.e., AMNI-c3 and AMNI-c4) generally achieves better performance. This may be due to that using more filters in CNN can capture richer features across global and local information of sMRI.

5.5. Influence of Multimodality Fusion Strategy

We fuse fMRI and sMRI data at the feature-level (see **Figure 1**) in the main experiments. We further investigate the influence of different fusion strategies by comparing our AMNI (using feature-level fusion) with its variant (called **AMNI_lf**) using a decision-level fusion strategy. As shown in **Figure 7**, in the AMNI_lf, the fMRI feature derived from GCN is fed into two fully connected layers and a Softmax layer for feature abstraction and classification. Similarly, the sMRI feature derived from CNN is fed into three fully connected layers and a Softmax layer. The outputs of these two branches are further fused *via* a weighted sum operation. We vary the weighted ratio between fMRI and sMRI branches within $[\frac{0.2}{0.8}, \frac{0.5}{0.5}, \frac{0.8}{0.2}]$ and denote these three methods as AMNI_lf1, AMNI_lf2, and AMNI_lf3, respectively, with the experimental results shown in **Figure 8**.

As shown in **Figure 8**, as the weight of GCN branch increases, the model achieves better performance in terms of most metrics.

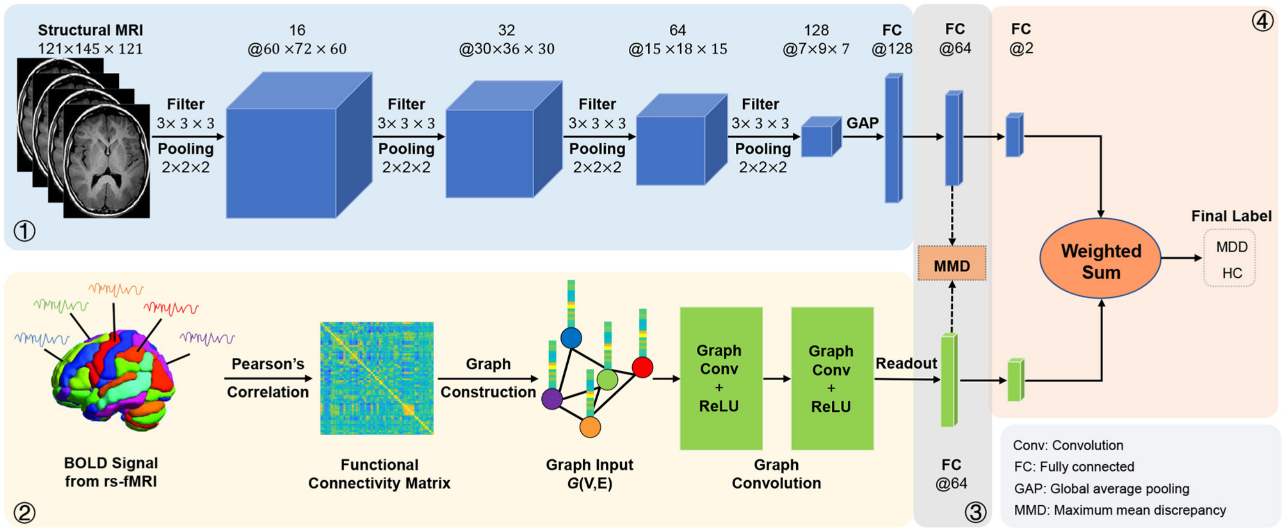

FIGURE 7 | Illustration of the adaptive multimodal neuroimage integration (AMNI) framework based on a decision-level fusion strategy.

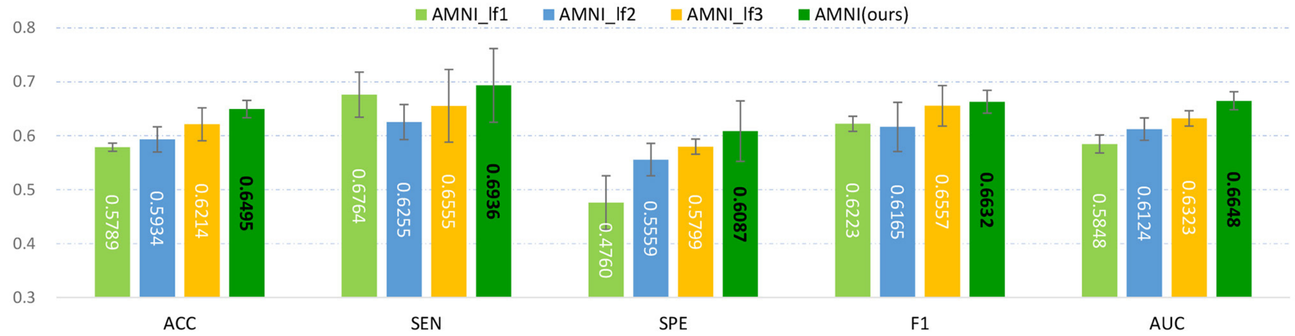

FIGURE 8 | Experimental results of late fusion method and our AMNI method in MDD vs. HC classification. Note that AMNI_lf1, AMNI_lf2, and AMNI_lf3 denote that the weight ratio between fMRI and sMRI branch is $\frac{0.2}{0.8}$, $\frac{0.5}{0.5}$, and $\frac{0.8}{0.2}$, respectively.

However, the results of AMNI using the decision-level fusion method are generally inferior to that of the feature-level fusion method proposed by this article. This implies that feature-level fusion of functional and structural representations could be more effective.

5.6. Limitations and Future Work

Several limitations need to be considered. First, we only integrate T1-weighted MRI and functional MRI data for automated MDD diagnosis. Actually, diffusion tensor imaging (DTI) data can examine and quantify white matter microstructure of the brain, which can further help uncover the neurobiological mechanisms of MDD. Therefore, it is valuable to incorporate DTI data into multimodal research in our future work. Second, we use functional connectivity networks for representing rs-fMRI data and treat them as input of the proposed method. It is interesting to extract diagnosis-oriented fMRI features, as we do for T1-weighed MRIs, which will also be our future work. Besides, a feature adaptation module with a cross-modal MDD loss is designed for reducing inter-modality data heterogeneity. Many other data adaptation methods (Ben-David et al., 2007) can also be incorporated into the proposed AMNI framework for further performance improvement.

6. CONCLUSION

In this article, we propose an adaptive multimodal neuroimage integration (AMNI) framework for automated MDD diagnosis based on functional and structural MRI data. We first employ GCN and CNN to learn feature representations of functional connectivity networks and structural MR images. Then, a feature adaptation module is designed to alleviate inter-modality difference by minimizing the distribution difference between two modalities. Finally, high-level features extracted from functional and structural MRI modalities are integrated and delivered to a classifier for disease detection. Experimental results on 533 subjects with rs-fMRI and T1-weighted sMRI demonstrate the effectiveness of the proposed method in identifying MDD patients from healthy controls.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: REST-meta-MDD Consortium Data Sharing.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by REST-meta-MDD Consortium Data Sharing. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

QW and ML designed the study. QW downloaded and analyzed the data, performed experiments, and drafted the manuscript. QW, LL, LQ, and ML revised the manuscript. All authors read and approved the final manuscript.

FUNDING

QW and LQ were partly supported by National Natural Science Foundation of China (Nos. 62176112, 61976110, and 11931008), Natural Science Foundation of Shandong Province (Nos. ZR2018MF020 and ZR2019YQ27), and Taishan Scholar Program of Shandong Province.

REFERENCES

Alexopoulos, G. S. (2005). Depression in the elderly. *Lancet* 365, 1961–1970. doi: 10.1016/S0140-6736(05)66665-2

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Bai, L., Cui, L., Jiao, Y., Rossi, L., and Hancock, E. (2020). Learning backtrackless aligned-spatial graph convolutional networks for graph classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 783–798. doi: 10.1109/TPAMI.2020.3011866

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. (2007). Analysis of representations for domain adaptation. *Adv. Neural Inform. Process. Syst.* 19:137. Available online at: https://proceedings.neurips.cc/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Soc. Netw.* 29, 555–564. doi: 10.1016/j.socnet.2007.04.002

Bron, E. E., Smits, M., Van Der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., et al. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *Neuroimage* 111, 562–579. doi: 10.1016/j.neuroimage.2015.01.048

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*. Available online at: https://arxiv.org/abs/1312.6203

Buch, A. M., and Liston, C. (2021). Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics. *Neuropsychopharmacology* 46, 156–175. doi: 10.1038/s41386-020-00789-3

Bürger, C., Redlich, R., Grotegerd, D., Meinert, S., Dohm, K., Schneider, I., et al. (2017). Differential abnormal pattern of anterior cingulate gyrus activation in unipolar and bipolar depression: an fMRI and pattern classification approach. *Neuropsychopharmacology* 42, 1399–1408. doi: 10.1038/npp.2017.36

Calhoun, V. D., and Sui, J. (2016). Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness. *Biol. Psychiatry* 1, 230–244. doi: 10.1016/j.bpsc.2015.12.005

Chakraborty, S., Aich, S., and Kim, H.-C. (2020). Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network. *Diagnostics* 10:402. doi: 10.3390/diagnostics10060402

Chen, J., Yang, L., Zhang, Y., Alber, M., and Chen, D. Z. (2016). "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," in *The 30th Conference on Neural Information Processing Systems* (Barcelona), 3036–3044.

Cuadra, M. B., Cammoun, L., Butz, T., Cuisenaire, O., and Thiran, J.-P. (2005). Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. *IEEE Trans. Med. Imaging* 24, 1548–1565. doi: 10.1109/TMI.2005.857652

Dvornek, N. C., Ventola, P., Pelphrey, K. A., and Duncan, J. S. (2017). "Identifying Autism from resting-state fMRI using long short-term memory networks," in *International Workshop on Machine Learning in Medical Imaging* (Quebec City, QC: Springer), 362–370. doi: 10.1007/978-3-319-67389-9_42

Foti, D., Carlson, J. M., Sauder, C. L., and Proudfit, G. H. (2014). Reward dysfunction in major depression: multimodal neuroimaging evidence for refining the melancholic phenotype. *Neuroimage* 101, 50–58. doi: 10.1016/j.neuroimage.2014.06.058

Fu, C. H., Costafreda, S. G., Sankar, A., Adams, T. M., Rasenick, M. M., Liu, P., et al. (2015). Multimodal functional and structural neuroimaging investigation of major depressive disorder following treatment with duloxetine. *BMC Psychiatry* 15:82. doi: 10.1186/s12888-015-0457-2

Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Adeli, E., and Pohl, K. M. (2020). "Spatio-temporal graph convolution for resting-state fMRI analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 528–538. doi: 10.1007/978-3-030-59728-3_52

Gao, S., Calhoun, V. D., and Sui, J. (2018). Machine learning in major depression: from classification to treatment outcome prediction. *CNS Neurosci. Therap.* 24, 1037–1052. doi: 10.1111/cns.13048

Ge, R., Gregory, E., Wang, J., Ainsworth, N., Jian, W., Yang, C., et al. (2021). Magnetic seizure therapy is associated with functional and structural brain changes in MDD: therapeutic versus side effect correlates. *J. Affect. Disord.* 286, 40–48. doi: 10.1016/j.jad.2021.02.051

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* 13, 723–773. Available online at: https://www.jmlr.org/papers/volume13/gretton12a/gretton12a.pdf?ref=https://githubhelp.com

Guan, H., and Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Trans. Biomed. Eng.* 69, 1173–1185. doi: 10.1109/TBME.2021.3117407

Guo, T., Zhang, Y., Xue, Y., Qiao, L., and Shen, D. (2021). Brain function network: higher order vs. more discrimination. *Front. Neurosci.* 2021:1033. doi: 10.3389/fnins.2021.696639

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Hinrichs, C., Singh, V., Xu, G., and Johnson, S. C. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* 55, 574–589. doi: 10.1016/j.neuroimage.2010.10.081

Holtzheimer, P. E. III, and Nemeroff, C. B. (2006). Future prospects in depression research. *Dial. Clin. Neurosci.* 8:175. doi: 10.31887/DCNS.2006.8.2/pholtzheimer

Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., et al. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2035–2040. doi: 10.1073/pnas.0811168106

Hosseini-Asl, E., Keynton, R., and El-Baz, A. (2016). "Alzheimer's disease diagnostics by adaptation of 3D convolutional network," in *IEEE International Conference on Image Processing (ICIP)* (Phoenix, AZ), 126–130. doi: 10.1109/ICIP.2016.7532332

Huang, Y., Xu, J., Zhou, Y., Tong, T., and Zhuang, X. (2019). Diagnosis of Alzheimer's disease *via* multi-modality 3D convolutional neural network. *Front. Neurosci.* 13:509. doi: 10.3389/fnins.2019.00509

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. Available online at: https://arxiv.org/abs/1412.6980

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*. Available online at: https://arxiv.org/abs/1609.02907

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

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Lee, J., Lee, I., and Kang, J. (2019). "Self-attention graph pooling," in *Proceedings of the 36 th International Conference on Machine Learning* (Long Beach, CA: PMLR), 3734–3743.

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., et al. (2017). Deep learning in medical imaging: general overview. *Korean J. Radiol.* 18, 570–584. doi: 10.3348/kjr.2017.18.4.570

Li, M., Liu, M., Kang, J., Zhang, W., and Lu, S. (2021). "Depression recognition method based on regional homogeneity features from emotional response fMRI using deep convolutional neural network," in 2021 *3rd International Conference on Intelligent Medicine and Image Processing* (Tianjin), 45–49. doi: 10.1145/3468945.3468953

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*. Available online at: https://arxiv.org/abs/1312.4400

Liu, M., and Zhang, D. (2014). Sparsity score: a novel graph-preserving feature selection method. *Int. J. Pattern Recogn. Artif. Intell.* 28:1450009. doi: 10.1142/S0218001414500098

Maglanoc, L. A., Kaufmann, T., Jonassen, R., Hilland, E., Beck, D., Landrø, N. I., et al. (2020). Multimodal fusion of structural and functional brain imaging in depression using linked independent component analysis. *Hum. Brain Mapp.* 41, 241–255. doi: 10.1002/hbm.24802

Nieminen, J. (1974). On the centrality in a graph. *Scand. J. Psychol.* 15, 332–336. doi: 10.1111/j.1467-9450.1974.tb00598.x

Otte, C., Gold, S. M., Penninx, B. W., Pariante, C. M., Etkin, A., Fava, M., et al. (2016). Major depressive disorder. *Nat. Rev. Dis. Rrimers* 2, 1–20. doi: 10.1038/nrdp.2016.65

Papakostas, G. I. (2009). Managing partial response or nonresponse: switching, augmentation, and combination strategies for major depressive disorder. *J. Clin. Psychiatry* 70, 16–25. doi: 10.4088/JCP.8133su1c.03

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

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Devito, Z., et al. (2017). "Automatic differentiation in pytorch," in 31st Conference on Neural Information Processing Systems (Long Beach, CA), 1–4. Available online at: https://openreview.net/pdf?id=BJJsrmfCZ

Pizzagalli, D. A., Iosifescu, D., Hallett, L. A., Ratner, K. G., and Fava, M. (2008). Reduced hedonic capacity in major depressive disorder: evidence from a probabilistic reward task. *J. Psychiatr. Res.* 43, 76–87. doi: 10.1016/j.jpsychires.2008.03.001

Rajalingam, B., and Priya, R. (2018). Multimodal medical image fusion based on deep learning neural network for clinical treatment analysis. *Int. J. ChemTech Res.* 11, 160–176. doi: 10.20902/ijctr.2018.110621

Rubin-Falcone, H., Zanderigo, F., Thapa-Chhetry, B., Lan, M., Miller, J. M., Sublette, M. E., et al. (2018). Pattern recognition of magnetic resonance imaging-based gray matter volume measurements classifies bipolar disorder and major depressive disorder. *J. Affect. Disord.* 227, 498–505. doi: 10.1016/j.jad.2017.11.043

Sarraf, S., and Tofighi, G. (2016). DeepAD: Alzheimer's disease classification *via* deep convolutional neural networks using MRI and fMRI. *BioRxiv* 2016:070441. doi: 10.1101/070441

Sato, J. R., oll, J., Green, S., Deakin, J. F., Thomaz, C. E., and Zahn, R. (2015). Machine learning algorithm accurately detects fMRI signature of vulnerability to major depression. *Psychiatry Res.* 233, 289–291. doi: 10.1016/j.pscychresns.2015.07.001

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H., Vermersch, P., et al. (1992). Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *J. Neurol. Neurosurg. Psychiatry* 55, 967–972. doi: 10.1136/jnnp.55.10.967

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Shi, J., Xue, Z., Dai, Y., Peng, B., Dong, Y., Zhang, Q., et al. (2018). Cascaded multi-column RVFL+ classifier for single-modal neuroimaging-based diagnosis of Parkinson's disease. *IEEE Trans. Biomed. Eng.* 66, 2362–2371. doi: 10.1109/TBME.2018.2889398

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. Available online at: https://arxiv.org/abs/1409.1556

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958. Available online at: https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. A., Yu, Q., et al. (2013). Three-way (N-way) fusion of brain imaging data based on mCCA+ jICA and its application to discriminating schizophrenia. *Neuroimage* 66, 119–132. doi: 10.1016/j.neuroimage.2012.10.051

Sun, L., Xue, Y., Zhang, Y., Qiao, L., Zhang, L., and Liu, M. (2021). Estimating sparse functional connectivity networks via hyperparameter-free learning model. *Artif. Intell. Med.* 111:102004. doi: 10.1016/j.artmed.2020.102004

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*. Available online at: https://arxiv.org/abs/1412.3474

Van Den Heuvel, M. P., and Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fmri functional connectivity. *Eur. Neuropsychopharmacol.* 20, 519–534. doi: 10.1016/j.euroneuro.2010.03.008

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: http://jmlr.org/papers/v9/vandermaaten08a.html

Wang, M., Lian, C., Yao, D., Zhang, D., Liu, M., and Shen, D. (2019). Spatial-temporal dependency modeling and network hub detection for functional MRI analysis via convolutional-recurrent network. *IEEE Trans. Biomed. Eng.* 67, 2241–2252. doi: 10.1109/TBME.2019.2957921

Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., et al. (2012). Identification of MCI individuals using structural and functional connectivity networks. *Neuroimage* 59, 2045–2056. doi: 10.1016/j.neuroimage.2011.10.015

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometr. Intell. Lab. Syst.* 2, 37–52. doi: 10.1016/0169-7439(87)80084-9

World Health Organization (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. Technical report, World Health Organization.

Yan, C., Wang, X., Zuo, X., and Zang, Y. (2016). DPABI: data processing & analysis for (resting-state) brain imaging. *Neuroinformatics* 14, 339–351. doi: 10.1007/s12021-016-9299-4

Yan, C., and Zang, Y. (2010). DPARSF: A matlab toolbox for "pipeline" data analysis of resting-state fMRI. *Front. Syst. Neurosci.* 4:13. doi: 10.3389/fnsys.2010.00013

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 recurrent patients with major depressive disorder: evidence from 25 cohorts. *bioRxiv* 2019:321745. doi: 10.1073/pnas.1900390116

Yang, C., Wang, R., Yao, S., Liu, S., and Abdelzaher, T. (2020). Revisiting oversmoothing in deep GCNs. *arXiv preprint arXiv:2003.13663*. Available online at: https://arxiv.org/abs/2003.13663

Yao, D., Sui, J., Wang, M., Yang, E., Jiaerken, Y., Luo, N., et al. (2021). A mutual multi-scale triplet graph convolutional network for classification of brain disorders using functional or structural connectivity. *IEEE Trans. Med. Imaging* 40, 1279–1289. doi: 10.1109/TMI.2021.3051604

Yao, L., Mao, C., and Luo, Y. (2019). "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33* (Honolulu, HI: AAAI), 7370–7377. doi: 10.1609/aaai.v33i01.33017370

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). "Beyond short snippets: deep networks for video classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 4694–4702. doi: 10.1109/CVPR.2015.7299101

Zhang, L., Wang, M., Liu, M., and Zhang, D. (2020). A survey on deep learning for neuroimaging-based brain disorder analysis. *Front. Neurosci.* 14:779. doi: 10.3389/fnins.2020.00779

Zhang, Y., Jiang, X., Qiao, L., and Liu, M. (2021). Modularity-guided functional brain network analysis for early-stage dementia identification. *Front. Neurosci.* 15:720909. doi: 10.3389/fnins.2021.720909

Zhang, Y., Zhang, H., Adeli, E., Chen, X., Liu, M., and Shen, D. (2020). Multiview feature learning with multiatlas-based functional connectivity networks for MCI diagnosis. *IEEE Trans. Cybern*. doi: 10.1109/TCYB.2020.3016953

Zhang, Y., Zhang, H., Chen, X., Liu, M., Zhu, X., Lee, S.-W., et al. (2019). Strength and similarity guided group-level brain functional network construction for MCI diagnosis. *Pattern Recogn*. 88, 421–430. doi: 10.1016/j.patcog.2018.12.001

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.

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.