

MT-DBDAE: a multi-task dual-branch denoising autoencoder for major depressive disorder diagnosis

Rui Liu, Xiaoyan Sun, Zhenming Yuan and Jian Zhang
*School of Information Science and Technology, Hangzhou Normal University,
Hangzhou, China*

Abstract

Major depressive disorder (MDD) is a globally prevalent mental illness. Functional connectivity (FC) analysis based on resting-state functional magnetic resonance imaging resting-state fMRI (rs-fMRI) provides new opportunities for neurobiological mechanism research and objective diagnosis of MDD. However, existing methods are prone to overfitting when handling high-dimensional FC data and predominantly rely on single-feature learning strategies, which limits their ability to fully leverage synergistic information between local connectivity patterns and global network features. To address these challenges, this study proposes a multi-task dual-branch denoising autoencoder (MT-DBDAE) framework for automatic MDD diagnosis. The framework uses a heterogeneous dual-pathway module (HDM) that integrates FC transformer and graph attention networks (GATv2) to achieve multi-scale feature extraction, applies hidden space mixup technology for data augmentation in latent space and finally jointly optimizes three objectives – classification, reconstruction and contrastive learning – through a multi-task learning strategy. Nested cross-validation experiments were conducted on the large-scale multi-center resting-state (REST)-meta-MDD data set, achieving favorable classification performance on both the CC200 and anatomical automatic labeling atlases. This effectively alleviates issues of high-dimensional FC data and medical data scarcity, providing an effective tool for objective MDD diagnosis.

Keywords Major depressive disorder, Functional connectivity, Graph neural networks, Transformer, Multi-task learning, Mixup

Paper type Research paper

1. Introduction

Major depressive disorder (MDD) is a globally prevalent psychiatric disorder characterized by persistent depressed mood, cognitive dysfunction and significant suicidal tendencies as core symptoms (Mulders *et al.*, 2015). According to the World Health Organization, MDD affects about 300 million people each year and is a leading cause of disability (Lim *et al.*, 2018). Its mortality rate is 60–80% higher than the general population (Cuijpers *et al.*, 2014; Walker *et al.*, 2015). In recent years, the rapid development of neuroimaging technology has made significant contributions to understanding the pathological mechanisms of brain diseases while also providing new avenues for exploring the neurobiological mechanisms of MDD and developing objective biomarkers. In this context, artificial intelligence (AI)-based automated classification is considered a promising tool for the diagnosis of MDD.

Resting-state functional magnetic resonance imaging resting-state fMRI (rs-fMRI) can capture spontaneous brain activity patterns in individuals under non-task-specific conditions,



revealing intrinsic functional connectivity (FC) networks that reflect baseline brain organization. In recent years, FC analysis based on rs-fMRI has been recognized as a powerful tool for identifying potential biomarkers of various brain diseases [such as MDD and Alzheimer's disease (AD)] ([Drysdale et al., 2017](#); [Li et al., 2018](#); [Liu et al., 2023, 2024](#); [Zhang et al., 2023](#)). In MDD research, FC analysis has revealed FC abnormalities in key brain networks [such as the default mode network (DMN) and central executive network (CEN)] ([Mulders et al., 2015](#)), indicating that these abnormal patterns have potential as clinical diagnostic biomarkers ([Yang et al., 2021](#)).

In recent years, machine learning (ML) and deep learning (DL) technologies have opened new pathways for FC-based automatic MDD classification. Traditional ML methods, such as support vector machine, logistic regression typically require feature selection and dimensionality reduction to handle high-dimensional FC data and achieve reasonable classification performance ([Dadi et al., 2019](#); [Zhang et al., 2022a](#); [Zhu et al., 2021](#)). DL methods can automatically learn complex FC representations through multi-layer nonlinear transformations and have attracted widespread attention in fMRI-based brain disease classification ([Deng et al., 2022](#); [Jinji, 2022](#); [Shan et al., 2022](#)). Early CNN applications like BrainNetCNN ([Kawahara et al., 2017](#)) and its improvements such as BrainNet-GA CNN ([Oh et al., 2022](#)) have shown promising results in brain network analysis. However, most CNN-based methods ignore the spatial structure and local connectivity patterns between region of interests (ROIs), overlooking the non-Euclidean topological relationships in brain networks ([Noman et al., 2024](#)).

The emergence of graph neural networks (GNNs), transformers and diffusion architectures provides new perspectives for processing brain FC data. GNNs can effectively model non-Euclidean topological relationships between brain regions and capture complex patterns in graph-structured data ([Dosenbach et al., 2010](#); [Scarselli et al., 2008](#); [Sun et al., 2024](#)). For example, X. [Li et al. \(2021\)](#) proposed BrainGNN based on graph convolutional networks (GCNs) to learn brain functional network representations, demonstrating superior performance on various neural and cognitive data sets. Transformer architecture can naturally model dependencies between ROIs through self-attention mechanisms and has been used to analyze FC variations ([Bullmore and Sporns, 2009](#); [Sporns, 2018](#)). For instance, [Kan et al. \(2022\)](#) proposed BrainNetTF, attempting to use transformers for joint learning and classification within a unified framework. However, pure transformers require sufficient data training, which is a significant challenge in medical scenarios ([Zhang et al., 2022b](#)). Diffusion models can participate in medical image processing by generating synthetic data to augment limited training data sets. Hierarchical diffusion models (HCDM), such as the HCDM proposed by [Tang et al. \(2025\)](#), demonstrated strong generative capabilities through multi-scale conditional generation and hierarchical control. Brain-net-diffusion adapts transformer-based latent diffusion models specifically for fMRI data, achieving accuracy improvement for autism spectrum disorder classification ([Zhao et al., 2025](#)). Nonetheless, such approaches require substantial computational resources and careful parameter tuning.

Although these methods have achieved certain progress in brain disease classification, existing research still faces several key challenges:

- High-dimensional FC data easily leads to model overfitting, while existing methods lack effective regularization and data augmentation strategies to adequately mitigate this problem; and
- Most methods adopt single-feature learning strategies and cannot simultaneously account for collaborative modeling of local connectivity patterns and global network features.

To address these challenges, we propose a multi-task dual-branch denoising autoencoder (MT-DBDAE) framework for MDD classification. Our framework introduces three innovations: First, we propose heterogeneous Fisher-score masking combined with learnable dual-pathway fusion. Heterogeneous masking preprocesses inputs according to each pathway's requirements and then by learnable gating dynamically fuses both pathways to balance global and local information for MDD classification. Second, we adopt an ROI-as-token formulation inspired by vision transformers (Dosovitskiy *et al.*, 2021). For each ROI, we linearly project its ROI-wise FC vector, the corresponding row of the FC matrix, into a patch-like token embedding. Finally, we apply hidden-space mixup combined with latent-space reconstruction and adaptive multi-task learning, which provides synergistic regularization that outperforms traditional data augmentation by operating at the semantic level rather than the input level, reducing overfitting in the limited-data medical imaging regime.

The main innovations of this study include:

- We propose heterogeneous Fisher-score masking and learnable dual-pathway fusion. The Transformer branch uses a higher threshold (70%–80%) to retain feature diversity for global attention. The GNN branch uses a lower threshold (30%–50%) to ensure edge quality for local aggregation. On the resting-state (REST)-meta-MDD data set with 5-fold cross-validation, our model attains 69.06% accuracy and an area under the curve (AUC) of 74.35%, surpassing all baselines.
- We introduce hidden-space mixup in the latent space and train with a multi-task learning to strengthen regularization. Ablation results show that removing either mixup and the multi-task loss reduces accuracy to 64.37% and AUC to 70.82%, confirming effective overfitting mitigation and improved MDD diagnosis.

2. Related work

2.1 Resting-state functional magnetic resonance imaging in major depressive disorder research

Resting-state fMRI (rs-fMRI) is a noninvasive technique that captures spontaneous brain activity without task stimulation, offering a window into the neural basis of MDD. FC matrices quantify temporal synchrony among brain regions and provide candidate biomarkers for MDD neurobiology (Yang *et al.*, 2021). However, conventional statistical analyses rely on strong priors and often fail to exploit the complex structure of high-dimensional FC data. Traditional ML also depends on hand-crafted feature engineering, including feature selection and dimensionality reduction (Zhu *et al.*, 2021). These limitations motivate DL approaches with stronger representation-learning capacity.

2.2 Graph Neural Networks-based functional connectivity analysis

In recent years, GNNs have been extensively studied and have demonstrated strong modeling capabilities across a wide range of domains (Carlo *et al.*, 2025; Ju *et al.*, 2025). Given the inherent graph-structured nature of brain functional networks, GNNs provide a more natural modeling approach for processing brain functional network data. GNNs can directly perform information propagation and aggregation operations on brain network graphs, effectively capturing complex dependencies between ROIs (Bessadok *et al.*, 2022).

GCNs were among the first GNN variants applied to brain network analysis. Qin *et al.* (2022) explored the application of GCN in MDD recognition, using spectral graph convolutional filters to convolve graphs and learn low-dimensional representations of networks. Xing *et al.* (2019) proposed dynamic spectral graph convolutional networks for

early mild cognitive impairment diagnosis using fMRI. Kazi *et al.* (2022) proposed combining differentiable graph modules with GCN for AD diagnosis. Li *et al.* (2021) designed novel ROI-aware graph convolutional layers that use topological and functional information from fMRI, along with ROI selection pooling layers (R pooling) to highlight important ROIs.

2.3 Transformer-based functional connectivity analysis

Transformer architectures, due to their powerful sequence modeling capabilities and self-attention mechanisms, have also been introduced into brain network analysis in recent years. They provide a more natural way to model relationships between ROIs based on attention for brain networks, offering new perspectives for understanding global brain network patterns.

Recently, Kan *et al.* proposed BrainNetTF and Dynamic brain Transformer (DRAT), respectively proposing an orthonormal clustering readout operation based on self-supervised soft clustering and orthogonal projection for generating distinguishable cluster-aware node embeddings and informative graph embeddings (Kan *et al.*, 2022) and proposing DART, which combines static and dynamic brain networks for more effective and detailed brain functional analysis (Kan *et al.*, 2023). Bannadabhavi *et al.* (2023) proposed Com-BrainTF, a hierarchical local-global transformer architecture that learns intra-community and inter-community aware node embeddings for autism spectrum disorder prediction tasks.

Although these methods have achieved certain progress in MDD classification, the following key challenges remain:

- Most methods use a single feature-learning strategy and cannot jointly capture local connectivity and global network structure.
- The scarcity of medical data makes DL models prone to overfitting and existing methods lack effective regularization and data augmentation strategies.

We address these gaps with MT-DBDAE that fuses a multi-head self-attention pathway with a GATv2 pathway to model global and local patterns. We further use latent-space mixup and multi-task objectives for strong regularization. This design mitigates overfitting and improves automatic MDD diagnosis.

3. Materials and methods

3.1 Data set

The fMRI data used in this study are from the open-access data repository REST-meta-MDD consortium, which is considered the largest public MDD database (Yan and Zang, 2010). This database contains data from 2,428 participants from 25 hospitals, including 1,300 MDD patients and 1,128 healthy controls (HC). We adopted the same data screening steps mentioned in Yan *et al.*'s paper (Yan and Zang, 2010), selecting 1,642 participants from 17 sites from this data set. Additionally, we added a sample exclusion step to exclude samples with missing ROI signals. Table 1 and Table 2 show detailed statistical information for all participants after using anatomical automatic labeling (AAL) atlas and CC200 atlas and removing samples with missing ROI signals, respectively.

3.2 Data preprocessing

The fMRI data for each subject in the REST-meta-MDD data set underwent standard preprocessing steps provided by the data processing assistant for resting-state fMRI (Yan and Zang, 2010), preprocessing followed the steps described by Yan *et al.*, including removal of

Table 1. Population size (pop), mean age (AVG±SD) and sex distribution (M/F) of both MDD and healthy control (HC) participants from 17 sites in REST-meta-MDD (1,611 subjects with AAL atlas)

Site	Total	Pop	MDD		Pop	HC	
			Age (AVG ± SD)	Sex (M/F)		Age (AVG ± SD)	Sex (M/F)
S1	146	73	31.88 ± 8.07	30/43	73	31.67 ± 8.93	32/41
S2	30	16	41.75 ± 11.13	1/15	14	45.57 ± 11.65	4/10
S4	41	18	31.67 ± 7.37	8/10	23	28.78 ± 4.58	11/12
S7	72	35	41.94 ± 11.56	13/22	37	38.22 ± 11.65	14/23
S8	87	39	31.87 ± 9.67	9/30	48	31.88 ± 10.60	21/27
S9	96	48	28.56 ± 8.62	22/26	48	28.62 ± 7.90	30/18
S10	71	45	32.73 ± 10.70	21/24	26	32.65 ± 7.90	17/9
S11	37	20	30.20 ± 9.03	9/11	17	31.41 ± 9.27	8/9
S13	36	20	32.65 ± 8.39	8/12	16	34.38 ± 10.39	5/11
S14	93	61	30.13 ± 6.91	19/42	32	29.59 ± 4.92	15/17
S15	67	30	46.47 ± 12.40	9/21	37	39.76 ± 14.46	17/20
S17	82	41	21.68 ± 2.97	14/27	41	20.63 ± 1.75	13/28
S19	49	18	34.94 ± 11.09	5/13	31	35.19 ± 10.05	14/17
S20	477	249	38.56 ± 11.82	83/166	228	39.53 ± 15.73	72/156
S21	144	79	34.14 ± 12.05	34/45	65	36.51 ± 12.42	28/37
S22	38	18	33.83 ± 9.55	9/9	20	24.35 ± 6.89	12/8
S23	45	22	26.23 ± 7.27	10/12	23	33.00 ± 11.78	8/15
Total	1611	832	34.39 ± 11.49	304/528	779	34.47 ± 13.01	321/458

Note(s): AVG = average; SD = standard deviation; M/F = male/female; Pop = population size; HC = healthy controls; MDD = major depressive disorder

Table 2. Population size (pop), mean age (AVG±SD) and sex distribution (M/F) of both MDD and healthy control (HC) participants from 17 sites in REST-meta-MDD (1,639 subjects with CC200 atlas)

Site	Total	Pop	MDD		Pop	HC	
			Age (AVG ± SD)	Sex (M/F)		Age (AVG ± SD)	Sex (M/F)
S1	146	73	31.88 ± 8.07	30/43	73	31.67 ± 8.93	32/41
S2	30	16	41.75 ± 11.13	1/15	14	45.57 ± 11.65	4/10
S4	41	18	31.67 ± 7.37	8/10	23	28.78 ± 4.58	11/12
S7	72	35	41.94 ± 11.56	13/22	37	38.22 ± 11.65	14/23
S8	113	51	31.63 ± 9.41	17/34	62	31.13 ± 10.51	26/36
S9	96	48	28.56 ± 8.62	22/26	48	28.62 ± 7.90	30/18
S10	71	45	32.73 ± 10.70	21/24	26	32.65 ± 7.90	17/9
S11	37	20	30.20 ± 9.03	9/11	17	31.41 ± 9.27	8/9
S13	36	20	32.65 ± 8.39	8/12	16	34.38 ± 10.39	5/11
S14	93	61	30.13 ± 6.91	19/42	32	29.59 ± 4.92	15/17
S15	67	30	46.47 ± 12.40	9/21	37	39.76 ± 14.46	17/20
S17	82	41	21.68 ± 2.97	14/27	41	20.63 ± 1.75	13/28
S19	49	18	34.94 ± 11.09	5/13	31	35.19 ± 10.05	14/17
S20	479	250	38.49 ± 11.86	84/166	229	39.57 ± 15.71	73/156
S21	144	79	34.14 ± 12.05	34/45	65	36.51 ± 12.42	28/37
S22	38	18	33.83 ± 9.55	9/9	20	24.35 ± 6.89	12/8
S23	45	22	26.23 ± 7.27	10/12	23	33.00 ± 11.78	8/15
Total	1639	845	34.32 ± 11.46	313/532	794	34.38 ± 12.99	327/467

Note(s): AVG = average; SD = standard deviation; M/F = male/female; Pop = population size; HC = healthy controls; MDD = major depressive disorder

initial time points (to eliminate magnetization instability effects), slice timing correction, realignment, spatial normalization to standard templates, regression of noise covariates (including head motion parameters, white matter signals and cerebrospinal fluid signals) and application of 0.01–0.10 Hz band-pass filtering to preserve low-frequency oscillation signals (Yan *et al.*, 2019).

Considering that using different brain atlases may affect classification results, we adopted two different atlases to extract time series for subsequent comparative analysis of classification performance. These two atlases are the CC200 atlas (Craddock *et al.*, 2012), which divides the human brain into 200 regions based on brain function and the AAL atlas (Tzourio-Mazoyer *et al.*, 2002), which divides the brain into 116 regions based on brain anatomical structure.

3.3 Functional connectivity

Based on the time series extracted from the above atlases after preprocessing, a whole-brain FC matrix was constructed for each subject. FC is defined as the temporal correlation of blood-oxygen-level-dependent (BOLD) signal fluctuations between different ROIs (temporal correlation). The Pearson correlation coefficient (PCC) is the most commonly used metric for defining this correlation (Rubinov and Sporns, 2010). The PCC formula used in this study is defined as follows:

$$r_{ij} = \frac{\sum_{t=1}^T (z_{it} - \bar{z}_i)(z_{jt} - \bar{z}_j)}{\sqrt{\sum_{t=1}^T (z_{it} - \bar{z}_i)^2 \sum_{t=1}^T (z_{jt} - \bar{z}_j)^2}} \quad (1)$$

where r_{ij} represents the PCC between the i -th and j -th ROIs; z_{it} and z_{jt} are the BOLD signal values of the i -th ROI and j -th ROI at time point t , respectively; $\bar{z}_i$ and $\bar{z}_j$ are the mean values of the BOLD signals of the i -th ROI and j -th ROI over the entire time series, respectively; T is the length of the time series.

3.4 Feature selection method based on fisher score

High-dimensional FC data ($d = N \times N$, where N is 116 or 200) contains substantial noise and redundant connections, directly training tends to result in model overfitting. We use Fisher discriminant analysis to quantify the inter-class discriminative ability of each FC connection (Supekar *et al.*, 2022) and use its discriminative strength as the basis for feature selection. Due to the symmetry of the FC matrix, we extract its upper triangular part and vectorize it to form a one-dimensional feature vector for subsequent analysis. The Fisher score for each feature can be calculated using the following formula:

$$F_j = \frac{n_{\text{pos}}(r_{\text{pos}}^j - r^j)^2 + n_{\text{neg}}(r_{\text{neg}}^j - r^j)^2}{n_{\text{pos}}(\sigma_{\text{pos}}^j)^2 + n_{\text{neg}}(\sigma_{\text{neg}}^j)^2}, \quad (2)$$

where n_{pos} and n_{neg} represent the sample sizes of MDD patients and HCs, respectively; r_{pos}^j and r_{neg}^j are the mean values of the j -th feature in both classes of samples; $(\sigma_{\text{pos}}^j)^2$ and $(\sigma_{\text{neg}}^j)^2$ are the variances of the j -th feature in both classes of samples; r^j is the overall mean of the j -th feature across all samples. The larger the Fisher score, the stronger the

discriminative ability of that feature (Gu *et al.*, 2012). However, feature selection requires finding a balance between reducing overfitting risk and retaining effective information: excessive removal of low-scoring features may reduce noise but may also lose potentially useful discriminative information.

This study designs a task-specific feature selection strategy. Specifically, we rank features by Fisher score and retain the top $p_{\text{main}}\%$ and $p_{\text{graph}}\%$ to create two binary masks. Then, each mask is mapped back to $N \times N$ by filling the strictly upper triangle, mirroring across the diagonal and zeroing the diagonal, yielding M_{main} for the Transformer branch and M_{graph} for the GNN branch. The key innovation lies in the asymmetric threshold design: M_{main} retains a larger proportion ($p_{\text{main}}\% = 70\text{--}80\%$) of connections because the Transformer pathway operates via multi-head self-attention, which computes global dependencies across all ROI pairs without locality constraints. This global, sequence-agnostic mechanism requires feature diversity to learn rich long-range dependency patterns. A permissive threshold preserves the information density necessary for effective attention-weight computation across varied ROI combinations. Conversely, M_{graph} retains a smaller proportion ($p_{\text{graph}}\% = 30\text{--}50\%$) because the GNN pathway operates via local graph aggregation, each node's feature is updated by weighted aggregation from its neighborhood. In high-dimensional spaces, low-magnitude edges introduce noise that disrupts aggregation quality. A strict threshold ensures the graph comprises exclusively high-confidence, class-discriminative edges, improving representation learning.

3.5 Multi-task dual-branch denoising autoencoder

The framework architecture of the MT-DBDAE model is shown in Figure 1. The main input to the model is the FC matrix $X \in \mathbb{R}^{N \times N}$. By applying two mask matrices to X , we obtain two different versions of FC matrices $H_{\text{main}} \in \mathbb{R}^{N \times N}$ and $H_{\text{graph}} \in \mathbb{R}^{N \times N}$. Meanwhile, to enhance the model's robustness, adaptive Gaussian noise is added to X during the encoding stage. The noise intensity is dynamically adjusted based on the statistical characteristics of the input data, effectively improving the model's resistance to data perturbations.

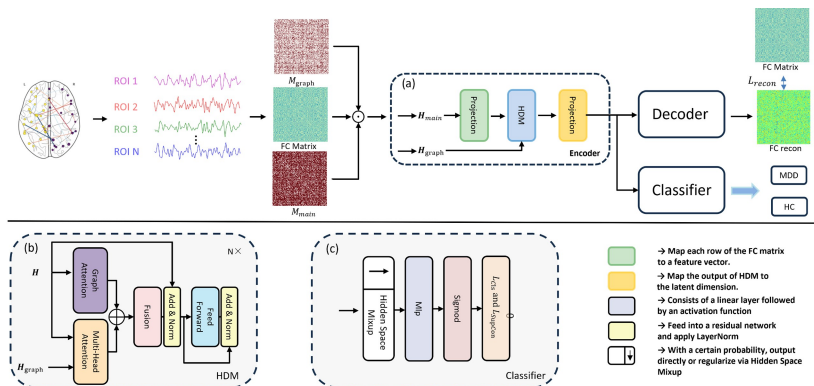


Figure 1. Overall model architecture: (a) encoder receives preprocessed inputs H_{main} and H_{graph} , extracts features through HDM modules and projects to latent space. (b) HDM fuses multi-head attention and graph attention mechanisms with feedforward networks. (c) classifier applies hidden space mixup in latent space, followed by MLP and joint optimization with classification and contrastive losses

3.5.1 Heterogeneous dual-pathway module. Heterogeneous dual-pathway module (HDM) is the core component of the encoder, which achieves multi-scale feature extraction of FC data by fusing multi-head self-attention mechanisms and graph attention networks (GAT). This module contains two parallel feature learning pathways that separately capture global dependency relationships and local topological structural information:

Multi-head self-attention pathway: The self-attention mechanism can effectively capture dependency relationships between arbitrary positions in sequences without distance limitations (Vaswani et al., 2017). First, each row of H_{main} is projected using projection matrix $W^{main} \in \mathbb{R}^{N \times d_{model}}$ for feature projection to generate $H \in \mathbb{R}^{N \times d_{model}}$, where N is the number of ROIs and d_{model} is the feature dimension. The self-attention mechanism first generates Query, Key and Value matrices through linear transformations:

$$Q = HW^Q, K = HW^K, V = HW^V, \quad (3)$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d_{model} \times d_k}$ are learnable projection matrices. The attention weights are computed through the dot product between queries and keys and softmax normalization is applied to each row of the matrices. Finally, the obtained attention weights are used to perform weighted aggregation on the value matrices, yielding the self-attention output:

$$\text{Attention}(Q, K, V) = \text{softmax}_r \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (4)$$

To enhance the model's expressive capability, multi-head self-attention is used for parallel attention computation:

$$\text{head}_i = \text{Attention}(HW_i^Q, HW_i^K, HW_i^V) \quad (5)$$

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \quad (6)$$

where $W^O \in \mathbb{R}^{hd_k \times d_{model}}$ is the output projection matrix. Finally, the multi-head output is mapped back to the original feature dimension to obtain $H^{MHA} \in \mathbb{R}^{N \times d_{model}}$.

GATv2 attention pathway: GAT is a neural network architecture specifically designed for graph-structured data (Velikovi et al., 2017), capable of effectively handling irregular graph topological structures. GATv2 improved upon the original GAT by enhancing the model's expressive capability and dynamic properties through the nonlinear activation transfer of attention mechanisms (Brody et al., 2021). Furthermore, to fully use neighborhood information, GATv2 introduces edge features (such as edge weights) into the attention mechanism, achieving node-wise adaptive aggregation of neighbor information.

Specifically, first, the FC strength is calculated for each ROI feature vector h_i to construct its neighborhood set $\mathcal{N}(h_i)$:

$$\mathcal{N}(h_i) = \{h_j | j \in N, j \neq i\} \quad (7)$$

After inputting H_{graph} into GATv2, we concatenate h_i, h_j and $H_{graph}(i, j)$ and multiply with a transformation matrix W . Then we learn an attention weight vector to transform the activation value of the product to attention coefficients:

$$e_{ij} = a^T \sigma(\text{Concat}(h_i, h_j, H_{graph}(i, j))W), \quad (8)$$

where $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^{d_{model}}$ are the i -th and j -th rows of $\mathbf{H}$, representing the feature vectors of nodes i and j , respectively, $\mathbf{W}$ is the linear transformation matrix, σ is the activation function and $\mathbf{a}$ is the learnable attention weight vector.

The attention coefficients are normalized using softmax across all neighbors of $\mathbf{h}_i$, i.e. $\mathcal{N}(\mathbf{h}_i)$:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(\mathbf{h}_i)} \exp(e_{ik})} \quad (9)$$

Finally, the node's feature aggregation is obtained:

$$\mathbf{h}_i^{GAT} = \frac{1}{K} \sum_{k=1}^K \sigma \left(\sum_{j \in \mathcal{N}(\mathbf{h}_i)} \alpha_{ij}^k \mathbf{Q}^k \mathbf{h}_j \right), \quad (10)$$

where K is the number of GAT heads and $\mathbf{Q}^k$ is the linear transformation matrix based on multi-head. This finally yields $\mathbf{H}^{GAT} \in \mathbb{R}^{N \times d_{model}}$.

The outputs of the dual pathways are fused via learnable gating weights and connected through a residual network:

$$\mathbf{G} = \sigma \left(\mathbf{W}_g \text{Concat}(\mathbf{H}^{MHA}, \mathbf{H}^{GAT}) + \mathbf{b}_g \right) \quad (11)$$

$$\mathbf{H}_{fused} = \text{LayerNorm}(\mathbf{G} \odot \mathbf{H}^{MHA} \circ (\mathbf{1} - \mathbf{G}) \odot \mathbf{H}^{GAT} \circ \mathbf{H}), \quad (12)$$

where the gate parameter $\mathbf{W}_g \in \mathbb{R}^{d_{model} \times 2d_{model}}$, $\mathbf{b}_g \in \mathbb{R}^{d_{model}}$ and σ is the sigmoid activation function.

Finally, feedforward network propagation is performed to obtain the HDM module output:

$$\mathbf{H}_{att}^{(l+1)} = \text{LayerNorm}(\mathbf{H}^{(l)} + \mathbf{H}_{fused}) \quad (13)$$

$$\mathbf{H}^{(l+1)} = \text{LayerNorm}(\mathbf{H}_{att}^{(l+1)} + \text{FFN}(\mathbf{H}_{att}^{(l+1)})), \quad (14)$$

where l is the module layer number and FFN is the feedforward network layer. Multiple HDM modules are stacked to form the main body of the encoder. After the final output undergoes a Flatten operation, it passes through a multi-layer perceptron to map to hidden space.

The latent vector $\mathbf{c} \in \mathbb{R}^{(N \times d_{latent}) \times 1}$ is obtained, which serves as the latent representation $\mathbf{c}$ for subsequent classifier and decoder classification.

3.5.2 Decoder. The decoder is responsible for reconstructing the latent representation $\mathbf{c}$ (after hidden space mixup combination transformation) back to the original FC vector $\hat{\mathbf{x}}$. We use multi-layer perceptrons for decoding:

$$\mathbf{h}_{dec} = \sigma(\mathbf{W}_{dec1}\mathbf{c} + \mathbf{b}_{dec1}) \quad (15)$$

$$\hat{\mathbf{x}} = \mathbf{W}_{dec2}\mathbf{h}_{dec} + \mathbf{b}_{dec2}, \quad (16)$$

where $W_{dec1} \in \mathbb{R}^{d_{dec} \times (N \times d_{latent})}$ and $W_{dec2} \in \mathbb{R}^{d \times d_{dec}}$ are the weight matrices of the decoder, d_{dec} is hidden layer dimension of the decoder and the latent vector $\hat{x} \in \mathbb{R}^d$.

3.5.3 Hidden space mixup regularization. Compared to data augmentation performed at the raw input level, mixing in latent space can generate new and diverse training samples at higher semantic levels, effectively improving the model's generalization ability (Verma et al., 2019).

During training, we apply hidden-space mixup with probability p . Mixed samples are generated as follows:

$$c = \lambda c_i + (1 - \lambda) c_j \quad (17)$$

$$y = \lambda y_i + (1 - \lambda) y_j, \quad (18)$$

where $i, j \in \{1, \dots, S\}$, $j \neq i$, S is the number of samples, $\lambda \sim \text{Beta}(\alpha, \alpha)$ is the mixing coefficient, following a Beta distribution to ensure diverse mixing characteristics. α is the shape parameter of the Beta distribution, controlling the mixing degree.

3.5.4 Classifier. The classifier uses the latent layer features obtained after mixup (which can be original or mixed latent representations) for classification:

$$y_{logit} = \sigma(W_{cls}c + b_{cls}), \quad (19)$$

where $W_{cls} \in \mathbb{R}^{1 \times (N \times d_{latent})}$ is the classifier weight matrix, $b_{cls} \in \mathbb{R}^1$ is the bias term, outputting a single logit value. The classifier output passes through the sigmoid function to convert to probability y_{logit} . The final classification decision is obtained through threshold comparison:

$$\hat{y} = \begin{cases} 1 & \text{if } y_{logit} > \theta \\ 0 & \text{otherwise} \end{cases}, \quad (20)$$

where θ is the classification threshold. We select θ via grid search on the validation set. We sweep $\theta \in [0, 1][0, 1]$, compute true positive rate (TPR) and false positive rate (FPR) and choose the threshold that maximizes $\text{TPR}(\theta) - \text{FPR}(\theta)$:

$$\theta^* = \arg \max(\text{TPR}(\theta) - \text{FPR}(\theta)) \quad (21)$$

Due to the existence of hidden space mixup, the loss function needs to consider both mixed and unmixed samples. For unmixed samples, the standard binary cross-entropy loss is used:

$$L_{cls} = \frac{1}{S} \sum_{i=1}^S \text{BCE}(\hat{y}_i, y_i) = -\frac{1}{S} \sum_{i=1}^S [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)]. \quad (22)$$

For mixed samples, the classification loss needs to consider the weighted combination of both original sample labels:

$$L_{cls} = \frac{1}{S} \sum_{i=1}^S [\lambda_i \text{BCE}(\hat{y}_i, y_i^a) + (1 - \lambda_i) \text{BCE}(\hat{y}_i, y_i^b)]. \quad (23)$$

where y_i^a and y_i^b are the true labels of the two samples participating in the mixture, respectively and $\hat{y}_i$ is the predicted label for the i -th sample. S is the number of samples and λ_i is the corresponding mixing coefficient.

3.5.5 Multi-task joint optimization. To alleviate the overfitting problem during training, this study adopts a multi-task joint optimization strategy that combines reconstruction loss, classification loss and supervised contrastive loss.

First, the reconstruction loss L_{recon} is introduced as a regularization term to provide a self-supervised signal for the encoder, encouraging it to preserve the primary structural characteristics of the input in the latent space, thereby improving the generalization ability of the representations. The reconstruction loss is defined as follows:

$$L_{recon} = \frac{1}{S} \sum_{i=1}^S \|\hat{x}_i - x_i\|_2^2, \quad (24)$$

where $x_i \in \mathbb{R}^d$ is the vector obtained by flattening the FC matrix X_i of the i -th sample.

In addition, the model uses the supervised contrastive loss L_{SupCon} to enhance the intra-class compactness and inter-class separability of representations.

$$L_{SupCon} = -\frac{1}{S} \sum_{i=1}^S \frac{1}{|P_{c_i}|} \sum_{c_p \in P_{c_i}} \log \frac{\exp(\text{sim}(c_i, c_p)/\tau)}{\sum_{c_n \in Q_{c_i}} \exp(\text{sim}(c_i, c_n)/\tau)}, \quad (25)$$

For each sample c_i , P_{c_i} denotes the set of samples belonging to the same class as c_i and Q_{c_i} denotes the set of all samples except itself, $\text{sim}(\cdot, \cdot)$ is the cosine similarity function and τ is the temperature hyperparameter.

The final multi-task objective is formulated as follows:

$$L_{total} = \alpha L_{recon} + \beta L_{cls} + \gamma L_{SupCon}, \quad (26)$$

where α, β, γ are the weighting coefficients for each loss term.

We adopt a multi-level joint optimization strategy to better achieve the model's objectives. Hidden space mixup is used to generate diversified training samples, thereby enhancing the model's generalization ability. Reconstruction learning provides self-supervised signals to facilitate the acquisition of effective data representations, while contrastive learning reinforces clustering characteristics and promotes clear class boundaries. This integrated approach effectively mitigates overfitting and ensures high-quality representations for MDD diagnostic tasks.

3.6 Implementation details

All experiments in this study were conducted on the following hardware and software environment: operating system Ubuntu 20.04.1 LTS, CPU dual AMD EPYC 7543 32-Core Processor, GPU NVIDIA GeForce RTX 4090, memory supporting NUMA architecture, software environment Python 3.9, PyTorch 2.4.1 (CUDA 11.8). For the model architecture, we stacked 2 HDMs with feature dimension $d_{\text{model}} = 32$ and latent dimension $d_{\text{latent}} = 8$. Hidden-space mixup used Beta distribution parameter $\alpha_m = 0.4$ with application probability $p = 0.7$. For the multi-task loss function in [equation \(26\)](#), the reconstruction loss weight α_r and classification loss weight β_c were set as adaptive learnable parameters initialized at 1.0, while the contrastive loss weight $\gamma_s = 1.0$ was fixed. The temperature parameter in the supervised contrastive loss was $\tau = 0.1$.

Fisher-score thresholds were determined via grid search on the validation set: $p_{\text{main}}\% \in [70, 80\%]$ for the Transformer branch and $p_{\text{graph}}\% \in [30, 50\%]$ for the GNN branch.

4. Results

This study used nested cross-validation to evaluate model performance. Specifically, we used both nested 5-fold cross-validation and nested leave-one-site-out cross-validation schemes. In the nested 5-fold cross-validation, the data set was partitioned into five equally sized subsets. In each iteration, one subset was designated as the external test set, while the remaining four subsets were used for model training and validation. For each external test set, the corresponding training data was further split into five folds, with four folds used for training and one fold for validation, repeated five times. This procedure yielded five external test results, which were then aggregated for performance evaluation. In the nested leave-one-site-out cross-validation, one site's samples are used as the external test set in each iteration and the remaining sites' samples are divided into training and validation sets with a ratio 4:1. We treat each site's samples as an external test set and average the results of them as the final results.

We mainly use accuracy (ACC), sensitivity (SEN), specificity (SPE) and AUC four evaluation metrics to comprehensively evaluate model performance. ACC, SEN, SPE are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (27)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (28)$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad (29)$$

where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively. Sensitivity measures the model's recognition ability for MDD patients, specificity measures the model's recognition ability for HC and AUC comprehensively reflects the model's discriminative ability.

4.1 Comparison with baseline methods

We selected seven traditional ML methods and DL methods as baselines for comparative experiments. Traditional ML methods include support vector machine (SVM), random forest (RF) and logistic regression, which perform classification on original FC matrices. DL methods include ResDAE (Dai *et al.*, 2024), 1DCNN, BrainNetCNN and BrainGNN network architectures. ResDAE uses residual connection-based autoencoders to perform feature learning and reconstruction on the upper triangular part of FC matrices, then uses latent representations for classification; 1DCNN transforms each FC matrix into a one-dimensional vector, which is then used for convolutional processing; BrainNetCNN is a network model specifically designed for processing brain network matrices; and BrainGNN is a graph neural network model specifically designed for brain connectivity network modeling.

Table 3 and Table 4 show the classification experimental results using CC200 and AAL atlas data sets for various methods. From the comparative experimental results, it can be seen that the MT-DBDAE method proposed in this paper achieved the best comprehensive

Table 3. Nested 5-fold cross-validation performance comparison

Method	AAL				CC200			
	ACC (%)	AUC (%)	SEN (%)	SPE (%)	ACC (%)	AUC (%)	SEN (%)	SPE (%)
ResDAE	61.20	68.15	62.47	59.91	64.18	69.87	68.27	60.38
BrainGNN	59.15	61.30	53.83	64.22	59.33	60.89	49.69	68.45
BrainNetCNN	65.18	72.10	67.27	63.22	64.18	70.14	59.18	68.86
1DCNN	54.37	60.00	31.03	76.39	50.76	58.04	67.09	35.41
SVM	64.43	70.83	60.99	67.71	65.71	72.73	62.72	68.50
LogisticReg	62.94	68.59	63.43	62.50	66.93	72.25	66.76	67.10
RandomForest	60.40	64.64	50.57	69.59	57.84	63.23	46.59	68.42
MT-DBDAE	67.17	72.48	67.00	67.17	69.06	74.35	72.53	65.78

Table 4. Nested leave-one-site-out cross-validation performance comparison

Method	AAL				CC200			
	ACC (%)	AUC (%)	SEN (%)	SPE (%)	ACC (%)	AUC (%)	SEN (%)	SPE (%)
ResDAE	62.38	67.36	59.54	65.42	60.62	69.08	60.62	59.27
BrainGNN	58.85	62.94	57.95	61.08	58.17	61.38	55.01	61.40
BrainNetCNN	63.06	68.60	63.33	63.75	62.72	67.97	63.56	62.27
1DCNN	53.39	63.92	38.03	70.35	54.48	56.45	52.50	47.93
SVM	64.66	71.84	59.23	70.74	66.00	72.12	61.51	71.41
LogisticReg	64.39	69.62	64.19	65.32	65.49	72.08	65.64	65.64
RandomForest	60.52	65.32	50.10	71.10	57.52	61.13	42.56	71.54
MT-DBDAE	65.02	69.89	68.83	61.84	66.59	72.36	66.21	69.03

performance on both atlas data sets. On the AAL atlas, MT-DBDAE achieved 67.17% accuracy and a 72.48% AUC value; on the CC200 atlas, MT-DBDAE achieved 69.06% accuracy and a 74.35% AUC value. It is worth noting that MT-DBDAE showed superior performance in sensitivity metrics, obtaining the highest sensitivity scores on both atlases, which is of great significance for early diagnosis of mental disorders.

The leave-one-site-out validation further demonstrates the generalization ability of MT-DBDAE. Leave-one-site-out validation is more stringent than nested 5-fold cross-validation, as it can better reflect true performance capabilities, MT-DBDAE still maintained excellent accuracy and sensitivity performance. On the AAL atlas, MT-DBDAE achieved 65.02% accuracy and 68.83% sensitivity. On the CC200 atlas, accuracy reached 66.59% and sensitivity was 66.21%.

In both types of nested cross-validation. The, ACC and SEN surpass other methods, but the SPE is lower than those of traditional ML models such as SVM. This is because we adopt data-augmentation strategies. These methods improve generalization but may increase false positives, causing some negative cases to be predicted as positive. As shown by the ablation in Table 5, removing these augmentations lowers ACC but raises SPE. In spite of this, MT-DBDAE achieves better overall performance.

Compared to different baseline methods, MT-DBDAE effectively improved model discriminative ability and generalization performance through dual-pathway feature fusion, hidden space mixup regularization and multi-task learning assistance, validating the effectiveness of the proposed method.

Table 5. Ablation study results of MT-DBDAE (CC200 atlas, 5-fold cross-validation)

Model variant	Components										Results		
	Trans (B)	GNN (B)	Mask- <i>T</i> (F)	Mask- <i>G</i> (F)	Mixup (R)	SupCon (R)	Recon (R)	ACC (%)	AUC	SEN (%)			
Transformer only	✓	✗	✓	✓	✓	✓	✓	53.63	55.91	57.46			50.10
GNN only	✗	✓	✓	✓	✓	✓	✓	60.34	65.27	56.19			64.27
w/o mask	✓	✓	✗	✗	✓	✓	✓	67.36	72.96	64.13			70.44
w/o mask- <i>G</i>	✓	✓	✓	✗	✓	✓	✓	68.94	74.34	72.28			65.78
w/o mask- <i>T</i>	✓	✓	✗	✓	✓	✓	✓	67.30	72.96	64.13			70.32
w/o mixup	✓	✓	✓	✓	✗	✓	✓	67.11	73.50	70.65			63.77
w/o supCon	✓	✓	✓	✓	✓	✗	✓	64.37	71.29	60.59			67.92
w/o recon loss	✓	✓	✓	✓	✓	✓	✗	67.97	74.61	68.88			67.07
w/o mixup+recon	✓	✓	✓	✓	✗	✓	✗	64.37	70.82	61.64			67.03
Full model	✓	✓	✓	✓	✓	✓	✓	69.06	74.35	72.53			65.78

Note(s): (B) = branch-wise, (F) = feature selection, (R) = regularization, Mask-*G* = mask for GNN, mask-*T* = mask for transformer

4.2 Ablation experiments

To verify the effectiveness of each module, ablation experiments were conducted, with results shown in [Table 5](#). The ablations are organized into three categories: branch-wise fusion, feature selection and regularization mechanisms. The Transformer branch alone achieves only 53.63% accuracy, while the GNN branch alone reaches 60.34%. This synergistic effect validates that global dependencies and local topology are complementary rather than redundant and their joint modeling is necessary for robust MDD classification.

The feature selection strategy proves necessary. Without masking, accuracy is 67.36%. Applying the permissive transformer mask improves to 68.94%, while the GNN mask yields 67.30%. The dual-mask configuration reaches the peak at 69.06%, outperforming any single-mask strategy. This result indicates that each mask is most effective when applied jointly with its counterpart, as a single mask alone cannot fully exploit the complementary inductive biases of the two branches. It also supports the principle that different neural architectures have different information requirements: Transformers benefit from permissive feature selection that preserves input diversity, while GNNs benefit from strict selection that ensures aggregation-edge quality.

Among regularization components, supervised contrastive loss has the largest impact: removing it causes a 4.69% accuracy drop to 64.37%. Hidden-space mixup contributes 1.95% (from 69.06% to 67.11%) and reconstruction loss contributes 1.09% (to 67.97%). Removing both mixup and reconstruction simultaneously yields the same result as removing only contrastive loss, indicating that contrastive learning dominates while mixup and reconstruction provide complementary benefits.

These ablation results collectively demonstrate that MT-DBDAE's effectiveness arises from the principled integration of complementary design components. The dual-branch architecture with heterogeneous feature masking captures both global and local connectivity patterns in an architecture-aware manner. The multi-task regularization combining mixup, reconstruction and contrastive objectives provides robust protection against overfitting in the limited-data regime typical of medical imaging. Rather than relying on any single component, the model's strength emerges from their coordinated contribution to representation learning, class separation and generalization.

4.3 Identification of key ROI connectivity patterns for major depressive disorder classification based on Shapley additive explanations analysis

To further analyze the decision mechanism of the MT-DBDAE model and identify key neural biomarkers for MDD classification, this study used the Shapley Additive explanations (SHAP) method ([Lundberg and Lee, 2017](#)) to perform importance quantification analysis on the learned FC features. SHAP is based on the Shapley values in game theory and can decompose the contribution of each input feature to the model prediction, offering an effective approach for interpreting DL models in neuroscience research. We conducted analysis on both the CC200 and AAL atlases. Meanwhile, based on the CC200 atlas FC clustering results, this study further mapped the identified functional regions to standard anatomical structures, validating the neurobiological interpretability of the discovered patterns.

As shown in [Figure 2](#), among the 15 most important connections, the left middle frontal gyrus, right angular gyrus and right temporal pole and other multiple brain regions were identified, where the angular gyrus and middle frontal gyrus are important components of the DMN.

DMN is a brain network that is most active during resting states and is mainly involved in self-referential processing, emotional regulation and introspection ([Philippi and Koenigs, 2014](#)).

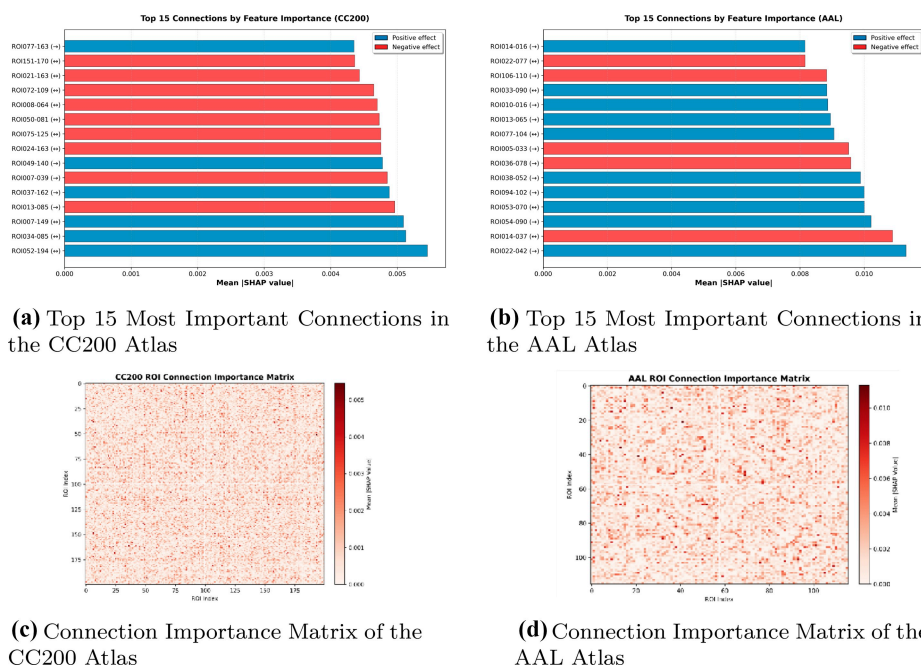


Figure 2. Visualization results of functional connection importance based on SHAP analysis. The top row shows the top 15 functional connections contributing most to MDD classification in the CC200 atlas (a) and the AAL atlas; (b) the bottom row displays the complete connection importance matrices for the respective atlases (c, d), with color intensity indicating the degree of importance

FC disruptions between the precuneus and regions such as the prefrontal cortex and hippocampus are directly associated with excessive self-focus and negative cognitive bias in patients with MDD (Kaiser *et al.*, 2015). Meanwhile, the angular gyrus connection abnormalities reflect MDD patients' deficits in self-reference processing and cognitive integration across multiple functional domains (Gao *et al.*, 2021). At the same time, the precentral gyrus is a key region within the sensorimotor cortex and its abnormal FC suggests impairments in sensorimotor integration among MDD patients. This finding is highly consistent with clinical manifestations of MDD, such as psychomotor retardation, somatic symptoms and altered motor cortical plasticity (Tsujii *et al.*, 2017).

SHAP analysis on the AAL atlas further validated the abnormal neural network patterns associated with MDD. Based on the connection importance ranking, among the 15 most important connections, the connections between the occipital cortex and amygdala, the interhemispheric connection between the triangular part of the inferior frontal gyrus and the left hippocampus, as well as connections between multiple cerebellar subregions, had a substantial impact on classification. The amygdala, a central limbic structure, plays a crucial role in integrating sensory information from the cortex and thalamus (Babaev *et al.*, 2018) and is considered a key structure in emotional response and anxiety circuitry (Duval *et al.*, 2015). Specifically, the amygdala and its connections with other brain regions (such as the hippocampus and olfactory areas) are associated with the emotional dysregulation observed in depression (Chai *et al.*, 2015). Abnormal interhemispheric connections between the

triangular part of the inferior frontal gyrus and the hippocampus reveal dysfunction in the memory-executive control network in MDD patients. The hippocampus is responsible for episodic memory encoding, while the prefrontal cortex is involved in executive control and emotional regulation; abnormalities in their connectivity result in cognitive biases and executive dysfunction in MDD patients (Tahmasian *et al.*, 2013). Recent neuroscience research has demonstrated that the cerebellum is involved in cognitive control and emotional processing and also plays an important role in the pathology of MDD (Ma *et al.*, 2013). Additionally, the spatial distribution of connection importance across both atlases indicated that interhemispheric connections were predominant, suggesting a significant imbalance in brain functional lateralization in MDD patients. This dominant pattern of interhemispheric connectivity may reflect impaired bilateral coordination in networks related to emotional regulation, cognitive control and attention (Li *et al.*, 2025).

This study further used BrainNet Viewer to visualize the top 30 functional connections with classification contribution (Xia *et al.*, 2013), as shown in Figure 3. The analysis results highly align with MDD-related neuroimaging research findings, validating that the MT-DBDAE model learned biologically meaningful neural connectivity patterns, providing potential neuroimaging biomarkers for more precise MDD diagnosis.

5. Discussion

We successfully constructed a MT-DBDAE framework for MDD automatic classification and validated its effectiveness on the large-scale multi-center REST-meta-MDD data set. This framework addresses the overfitting susceptibility issues in high-dimensional FC data analysis through innovative hidden space mixup data augmentation techniques combined with multi-objective joint optimization. The model achieved significant performance improvements in MDD classification tasks.

The HDM module designed in this study achieves dual-pathway integration through multi-head self-attention and GATv2 network architectures, effectively combining global dependency relationships and local topological structural information advantages. The multi-head self-attention mechanism captures long-distance dependencies between ROIs, while GATv2 optimizes the representation and structural properties of graphical networks through attention-enhanced mechanisms.

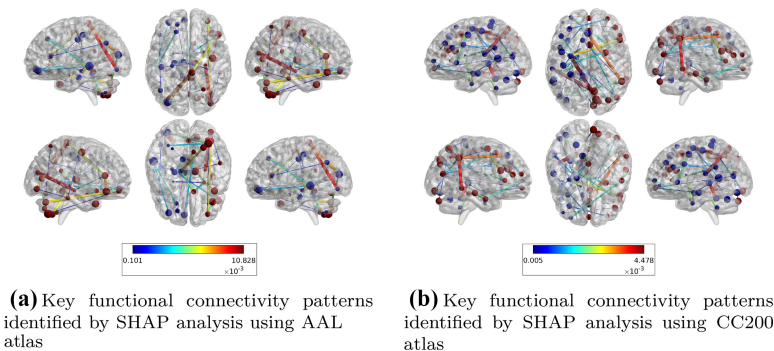


Figure 3. SHAP-Based visualization of key functional connectivity patterns for MDD classification using different brain atlases. (a) Results based on AAL atlas. (b) Results based on CC200 atlas. Node colors indicate left and right hemispheres, while edge colors and thickness represent the importance weights of connections, with warmer colors (red) corresponding to higher SHAP values

This dual-pathway design addresses a key problem by providing more comprehensive feature representations for FC data. Compared to traditional neural data augmentation methods, multi-task learning strategies and hidden space mixup contributed 4.69% accuracy improvement and 3.53% AUC improvement, effectively alleviating the overfitting problem of DL models on limited medical data.

Although our method achieved certain improvements in overfitting problems and classification effectiveness, this study still has some limitations that require further technical improvements in future work. We have not fully used the temporal dynamic connectivity information in large-scale resting-state functional data and dynamic FC can capture richer temporal change characteristics of brain networks, potentially containing more important neural information. Meanwhile, this study primarily focused on single-modal fMRI data and model performance on multi-modal MRI structural imaging diffusion tensor imaging, task-state fMRI and other neuroimaging modalities requires further validation. Multi-modal fusion may further improve diagnostic capabilities. Additionally, although multi-center data was used, differences in site demographics, cultural backgrounds, disease severity and other population factors require more extensive validation studies.

6. Conclusions

This study successfully developed an intelligent diagnostic framework for MDD based on a MT-DBDAE. Through heterogeneous dual-pathway feature fusion, hidden space mixup and multi-objective collaboration, the proposed approach effectively mitigates the challenges of high dimensionality in FC data and the scarcity of medical data. Experimental results on the large-scale, multi-center REST-meta-MDD data set demonstrate that the proposed method achieved optimal classification performance on both brain atlases, with an accuracy of 69.06% and an AUC of 74.35% on the CC200 atlas. SHAP-based interpretability analysis identified key FC patterns that are highly consistent with the neuropathological mechanisms of MDD, including DMN abnormalities, limbic system dysfunction and interhemispheric connectivity imbalance. Ablation experiments validated the effectiveness of each module, with hidden mixup and multi-task learning contributing the most to performance improvement. This study provides an effective tool for the objective diagnosis of MDD based on brain FC, holds significant clinical translational value and offers a new technological pathway for precision medicine in psychiatric disorders.

References

- Babaev, O., Piletti, C., Chatain, Krueger. and D., Burg. (2018), "Inhibition in the amygdala anxiety circuitry", *Experimental and Molecular Medicine*, Vol. 50 No. 4, pp. 1-16.
- Bannadabhavi, A., Lee, S., Deng, W., Ying, R. and Li, X. (2023), "Community-aware transformer for autism prediction in FMRI connectome", pp. 287-297.
- Bessadok, A., Mahjoub, M.A. and Rekik, I. (2022), "Graph neural net-works in network neuroscience", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45 No. 5, pp. 5833-5848.
- Brody, S., Alon, U. and Yahav, E. (2021), "How attentive are graph attention networks?", *arXiv preprint arXiv:2105.14491*.
- Bullmore, E. and Sporns, O. (2009), "Complex brain networks: graph theoretical analysis of structural and functional systems", *Nature Reviews. Neuroscience*, Vol. 10 No. 3, pp. 186-198.
- Carlo, G.D., Mastropietro, A. and Anagnostopoulos, A. (2025), "Kolmogorov- Arnold graph neural networks", *arXiv: 2406.18354 [cs.LG]*.
- Chai, X.J., Hirshfeld-Becker, D., Biederman, J., Uchida, M., Doehrmann, O., Leonard, J.A., Salvatore, J., Kenworthy, T., Brown, A., Kagan, E., *et al.* (2015), "Functional and structural

- brain correlates of risk for major depression in children with familial depression", *NeuroImage. Clinical*, Vol. 8, pp. 398-407.
- Craddock, R.C., James, G., Holtzheimer, P.E. III, Hu, X.P. and Mayberg, H.S. (2012), "A whole brain fMRI atlas generated via spatially constrained spectral clustering", *Human Brain Mapping*, Vol. 33 No. 8, pp. 1914-1928.
- Cuijpers, P., Vogelzangs, N., Twisk, J., Kleiboer, A., Li, J. and Penninx, B.W. (2014), "Comprehensive meta-analysis of excess mortality in depression in the general community versus patients with specific illnesses", *The American Journal of Psychiatry*, Vol. 171 No. 4, pp. 453-462.
- Dadi, K., Rahim, M., Abraham, A., Chyzyk, D., Milham, M., Thirion, B., Varoquaux, G., Initiative, A. D.N., *et al.* (2019), "Benchmarking functional connectome-based predictive models for resting-state fMRI", *NeuroImage*, Vol. 192, pp. 115-134.
- Dai, P., Shi, Y., Lu, D., Zhou, Y., Luo, J., He, Z., Chen, Z., Zou, B., Tang, H., Huang, Z., *et al.* (2024), "Classification of recurrent major depressive disorder using a residual denoising autoencoder framework: insights from large-scale multisite fMRI data", *Computer Meth- Ods and Programs in Biomedicine*, Vol. 247, p. 108114.
- Deng, X., Zhang, J., Liu, R. and Liu, K. (2022), "Classifying ASD based on time-series fMRI using spatial-temporal transformer", *Com- Puters in Biology and Medicine*, Vol. 151, p. 106320.
- Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., *et al.* (2010), "Prediction of individual brain maturity using fMRI", *Science (New York, N.Y.)*, Vol. 329 No. 5997, pp. 1358-1361.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Hounsby, N. (2021), "An image is worth 16x16 words: transformers for image recognition at scale", *arXiv: 2010. 11929 [cs.CV]*.
- Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., *et al.* (2017), "Resting-state connectivity biomarkers define neurophysiological subtypes of depression", *Nature Medicine*, Vol. 23 No. 1, pp. 28-38.
- Duval, E.R., Javanbakht, A. and Liberzon, I. (2015), "Neural circuits in anxiety and stress disorders: a focused review", *Therapeutics and Clinical Risk Management*, Vol. 11, pp. 115-126.
- Gao, J., Li, Y., Wei, Q., Li, X., Wang, K., Tian, Y. and Wang, J. (2021), "Habenula and left angular gyrus circuit contributes to response of electroconvulsive therapy in major depressive disorder", *Brain Imaging and Behavior*, Vol. 15 No. 5, pp. 2246-2253.
- Gu, Q., Li, Z. and Han, J. (2012), "Generalized fisher score for feature selection", *arXiv preprint arXiv:1202.3725*.
- Jinji, G. (2022), "Bionic artificial self-recovery enables autonomous health of machine", *Journal of Bionic Engineering*, Vol. 19 No. 6, pp. 1545-1561.
- Ju, L., Yang, X., Li, Q. and Wang, X. (2025), "GraphBridge: towards arbitrary transfer learning in GNNs", The Thirteenth Inter-national Conference on Learning Representations,
- Kaiser, R.H., Andrews-Hanna, J.R., Wager, T.D. and Pizzagalli, D.A. (2015), "Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity", *JAMA Psychiatry*, Vol. 72 No. 6, pp. 603-611.
- Kan, X., Gu, A.A.C., Cui, H., Guo, Y. and Yang, C. (2023), "Dynamic brain transformer with multi-level attention for functional brain network analysis", pp. 1-4.
- Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y. and Yang, C. (2022), "Brain network transformer", *Advances in Neural Information Process- Ing Systems*, Vol. 35, pp. 25586-25599.
- Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G. and Hamarneh, G. (2017), "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment", *NeuroImage*, Vol. 146, pp. 1038-1049.

- Kazi, A., Cosmo, L., Ahmadi, S.-A., Navab, N. and Bronstein, M.M. (2022), "Differentiable graph module (DGM) for graph convolutional networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45 No. 2, pp. 1606-1617.
- Li, Y., Liu, J., Huang, J., Li, Z. and Liang, P. (2018), "Learning brain connectivity Sub-networks by group-constrained sparse inverse co-variance estimation for Alzheimer's disease classification", *Frontiers in Neuroinformatics*, Vol. 12, p. 58.
- Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P. and Duncan, J.S. (2021), "Braingnn: interpretable brain graph neural network for fMRI analysis", *Medical Image Analysis*, Vol. 74, p. 102233.
- Li, Q., Qi, L., Zhang, G., Hao, J., Ren, Q., Guan, J., Zhan, Y., Yu, Y., Yang, J., Wang, K., *et al.* (2025), "Disrupted interhemispheric functional and structural connectivity in patients with major depressive disorder", *Progress in Neuro-Psychopharmacology and Bio-Logical Psychiatry*, p. 111374.
- Lim, G.Y., Tam, W.W., Lu, Y., Ho, C.S., Zhang, M.W. and Ho, R.C. (2018), "Prevalence of depression in the community from 30 countries between 1994 and 2014", *Scientific Reports*, Vol. 8 No. 1, p. 2861.
- Liu, L., Wen, G., Cao, P., Hong, T., Yang, J., Zhang, X. and Zaiane, O.R. (2023), "BrainTGL: a dynamic graph representation learning model for brain network analysis", *Computers in Biology and Medicine*, Vol. 153, p. 106521.
- Liu, J., Yang, W., Ma, Y., Dong, Q., Li, Y. and Hu, B. and Direct Consortium (2024), "Effective hyper-connectivity network construction and learning: application to major depressive disorder identification", *Computers in Biology and Medicine*, Vol. 171, p. 108069.
- Lundberg, S.M. and Lee, S.-I. (2017), "A unified approach to interpreting model predictions", *in*, Vol. 30.
- Ma, Q., Zeng, L.-L., Shen, H., Liu, L. and Hu, D. (2013), "Altered cerebellar–cerebral resting-state functional connectivity reliably identifies major depressive disorder", *Brain Research*, Vol. 1495, pp. 86-94.
- Mulders, P.C., van, P.F., Eijndhoven, Schene, A.H., Beck-Mann, C.F. and Tendolkar, I. (2015), "Resting-state functional connectivity in major depressive disorder: a review", *Neuroscience and Biobehavioral Reviews*, Vol. 56, pp. 330-344.
- 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 Journal of Biomedical and Health Informatics*, Vol. 28 No. 3, pp. 1644-1655.
- Oh, K.-H., Oh, I.-S., Tsogt, U., Shen, J., Kim, W.-S., Liu, C., Kang, N.-I., Lee, K.-H., Sui, J., Kim, S.-W., *et al.* (2022), "Diagnosis of schizophrenia with functional connectome data: a graph-based convolutional neural network approach", *BMC Neuroscience*, Vol. 23 No. 1, p. 5.
- Philippi, C.L. and Koenigs, M. (2014), "The neuropsychology of self-reflection in psychiatric illness", *Journal of Psychiatric Research*, Vol. 54, pp. 55-63.
- Qin, K., Lei, D., Pinaya, W.H., Pan, N., Li, W., Zhu, Z., Sweeney, J.A., Mechelli, A. and Gong, Q. (2022), "Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites", *EBioMedicine*, Vol. 78.
- Rubinov, M. and Sporns, O. (2010), "Complex network measures of brain connectivity: uses and interpretations", *NeuroImage*, Vol. 52 No. 3, pp. 1059-1069.
- Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M. and Monfardini, G. (2008), "The graph neural network model", *IEEE Transactions on Neural Networks*, Vol. 20 No. 1, pp. 61-80.
- Shan, W., Hu, H., Cai, Z., Chen, H., Liu, H., Wang, M. and Teng, Y. (2022), "Multi-strategies boosted mutative crow search algorithm for global tasks: Cases of continuous and discrete optimization", *Journal of Bionic Engineering*, Vol. 19 No. 6, pp. 1830-1849.
- Sporns, O. (2018), "Graph theory methods: applications in brain networks", *Dialogues in Clinical Neuroscience*, Vol. 20 No. 2, pp. 111-121.

- Sun, C., Jiang, M., Gao, L., Xin, Y. and Dong, Y. (2024), "A novel study for depression detecting using audio signals based on graph neural network", *Biomedical Signal Processing and Control*, Vol. 88, p. 105675.
- Supekar, K., Ryali, S., Yuan, R., Kumar, D., Los Angeles, D C. and Menon, V. (2022), "Robust, generalizable, and interpretable artificial intelligence-derived brain fingerprints of autism and social communication symptom severity", *Biological Psychiatry*, Vol. 92 No. 8, pp. 643-653.
- Tahmasian, M., Knight, D.C., Manoliu, A., Schwerthöffer, D., Scherr, M., Meng, C., Shao, J., Peters, H., Doll, A., Khazaie, H., *et al.* (2013), "Aberrant intrinsic connectivity of hippocampus and amygdala overlap in the frontoinsula and dorsomedial-prefrontal cortex in major depressive disorder", *Frontiers in Human Neuroscience*, Vol. 7, p. 639.
- Tang, S., Xu, X., Qiu, J. and Wang, Z. (2025), "Controllable data generation with hierarchical neural representations", Forty-second International Conference on Machine Learning.
- Tsujii, N., Mikawa, W., Tsujimoto, E., Adachi, T., Niwa, A., Ono, H. and Shirakawa, O. (2017), "Reduced left precentral regional responses in patients with major depressive disorder and history of suicide attempts", *PloS One*, Vol. 12 No. 4, p. e0175249.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. and Joliot, M. (2002), "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI Single-Subject brain", *NeuroImage*, issn: 1053-8119, Vol. 15 No. 1, pp. 273-289.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, N. and Polosukhin, I. (2017), "Attention is all you need", *in*, Vol. 30.
- Velickovi, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y. (2017), "Graph attention networks", *arXiv preprint arXiv:1710.10903*.
- Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D. and Bengio, Y. (2019), "Manifold mixup: Better representations by interpolating hidden states", International Conference on Machine Learning, *PMLR*, pp. 6438-6447.
- Walker, E.R., McGee, R.E. and Druss, B.G. (2015), "Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis", *JAMA Psychiatry*, Vol. 72 No. 4, pp. 334-341.
- Xia, M., Wang, J. and He, Y. (2013), "BrainNet viewer: a network visualization tool for human brain connectomics", *PloS One*, Vol. 8 No. 7, p. e68910.
- Xing, X., Li, Q., Wei, H., Zhang, M., Zhan, Y., Zhou, X.S., Xue, Z. and Shi, F. (2019), "Dynamic spectral graph convolution networks with assistant task training for early MCI diagnosis", International Conference on Medical Image Computing and Computer-Assisted Intervention, *Springer*, pp. 639-646.
- Yan, C. and Zang, Y. (2010), "DPARSF: a MATLAB toolbox for pipeline data analysis of Resting-State fMRI", *Frontiers in Systems Neuroscience*, Vol. 4.
- Yan, C.-G., Chen, X., Li, L., Castellanos, F.X., Bai, T.-J., Bo, Q.-J., Cao, J., Chen, G.-M., Chen, N.-X., Chen, W., *et al.* (2019), "Reduced default mode network functional connectivity in patients with recurrent major depressive disorder", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 116 No. 18, pp. 9078-9083.
- Yang, H., Chen, X., Chen, Z.-B., Li, L., Li, X.-Y., Castellanos, F.X., Bai, T.-J., Bo, Q.-J., Cao, J., Chang, Z.-K., *et al.* (2021), "Disrupted intrinsic functional brain topology in patients with major depressive disorder", *Molecular Psychiatry*, Vol. 26 No. 12, pp. 7363-7371.
- Zhang, J., Zhou, L., Wang, L., Liu, M. and Shen, D. (2022a), "Diffusion kernel attention network for brain disorder classification", *IEEE Transactions on Medical Imaging*, Vol. 41 No. 10, pp. 2814-2827.
- Zhang, B., Liu, S., Liu, X., Chen, S., Ke, Y., Qi, S., Wei, X. and Ming, D. (2022b), "Discriminating subclinical depression from major depression using multi-scale brain functional features: a radiomics analysis", *Journal of Affective Disorders*, Vol. 297, pp. 542-552.

-
- Zhang, M., Long, D., Chen, Z., Fang, C., Li, Y., Huang, P., Chen, F. and Sun, H. (2023), "Multi-view graph network learning framework for identification of major depressive disorder", *Computers in Biology and Medicine*, Vol. 166, p. 107478.
- Zhao, H., Lou, H., Yao, L. and Zhang, Y. (2025), "Diffusion transformer- augmented fMRI functional connectivity for enhanced autism spectrum disorder diagnosis", *Journal of Neural Engineering*, Vol. 22 No. 1, p. 16044.
- Zhu, X., Yuan, F., Zhou, G., Nie, J., Wang, D., Hu, P., Ouyang, L., Kong, L. and Liao, W. (2021), "Cross-network interaction for diagnosis of major depressive disorder based on resting state functional connectivity", *Brain Imaging and Behavior*, Vol. 15 No. 3, pp. 1279-1289.

Corresponding author

Jian Zhang can be contacted at: jeyszhang@outlook.com