Automated Diagnosis of Major Depressive Disorder With Multi-Modal MRIs Based on Contrastive Learning: A Few-Shot Study

Tongtong Li, Yuhui Guo, Ziyang Zhao, Miao Chen, Qiang Lin, Xiping Hu, DIRECT Consortium, Zhijun Yao, and Bin Hu, *Fellow, IEEE*

***Abstract*— Depression ranks among the most prevalent mood-related psychiatric disorders. Existing clinical diagnostic approaches relying on scale interviews are susceptible to individual and environmental variations. In contrast, the integration of neuroimaging techniques and computer science has provided compelling evidence for the quantitative assessment of major depressive disorder (MDD). However, one of the major challenges in computer-aided diagnosis of MDD is to automatically and effectively mine the complementary cross-modal information from limited datasets. In this study, we proposed a few-shot learning framework that integrates multi-modal MRI data based on contrastive learning. In the upstream task, it is designed to extract knowledge from heterogeneous data. Subsequently, the downstream task is dedicated to transferring the acquired knowledge to the target dataset, where a hierarchical fusion paradigm is also designed to integrate features across inter- and intra-modalities. Lastly, the proposed model was evaluated on a set of multi-modal clinical data, achieving average scores of 73.52% and 73.09% for accuracy and AUC, respectively. Our findings also reveal that the brain regions within the default mode network and cerebellum play a crucial role in the diagnosis, which provides further direction in exploring reproducible biomarkers for MDD diagnosis.**

***Index Terms*— Depression recognition, multi-modal, few-shot, contrastive learning, biomarkers.**

I. Introduction

MAJOR Depressive Disorder (MDD) is considered a prevalent and chronic mental disorder, characterized by typical symptoms such as sadness, anxiety, emotional instability, and even suicidal tendencies, imposing a heavy burden on patients themselves and their families [1]. Conventional clinical diagnostic approaches, relying on subjective interviews and scale-based assessments, are susceptible to environmental and individual variations. Moreover, depression can be challenging to diagnose in clinics due to its symptoms overlap with those of other mental disorders, resulting in a lower clinical detection rate [2]. Under these circumstances, rapidly and accurately detecting depression remains a challenging issue, especially given the limitations of existing diagnostic approaches. This is crucial for alleviating the growing mental health crisis [3], [4], [5].

In recent decades, rapid advances in non-invasive neuroimaging technologies have made it possible to study the structure and function of the human brain. For instance, structural magnetic resonance imaging (sMRI) can measure macro-structural changes in the brain, stemming from normal brain development, aging, and even diseases [6], [7], [8]. Functional magnetic resonance imaging (fMRI) is commonly utilized to track brain functional activities by recording fluctuations of blood oxygen level-dependent (BOLD) [9]. Diffusion tensor imaging (DTI) can be used to explore micro-structural connections and communication pathways in the brain by depicting the trajectories of white matter fiber bundles [10].

Manuscript received 27 November 2023; revised 20 February 2024; accepted 18 March 2024. Date of publication 21 March 2024; date of current version 17 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U21A20520 and Grant 62227807, in part by the National Key Research and Development Program of China under Grant 2019YFA0706200, in part by the Science and Technology Program of Gansu Province under Grant 23YFGA0004, and in part by the Fundamental Research Funds for the Central Universities under Grant lzujbky-2022-it24. *(Corresponding authors: Zhijun Yao; Bin Hu.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Gansu Provincial Hospital under Approval No. 2017-071.

Tongtong Li, Ziyang Zhao, Miao Chen, and Zhijun Yao are with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: yaozj@lzu.edu.cn).

Yuhui Guo is with the School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, China.

Qiang Lin is with the School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730030, China.

Xiping Hu is with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China.

Bin Hu is with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China, also with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China, also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, and also with the Joint Research Center for Cognitive Neurosensor Technology of Lanzhou University and Institute of Semiconductors, Chinese Academy of Sciences, Lanzhou 730000, China (e-mail: bh@lzu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3380357

Compared to single-modality imaging-based studies [11], [12], the application of multiple neuroimaging modalities provides a more comprehensive view, where complementary information could help to improve the classification performance for the clinical diagnosis of MDD [13], [14]. However, a challenge that needs to be addressed is the automatic and efficient extraction of complementary cross-modal information. This is because each imaging modality has its own unique characteristics and practical limitations.

Recent studies have shown that integrating multiple neuroimaging modalities can improve the performance in the diagnosis of mental disorders. For example, Zheng et al. [15] designed a Functional and Structural Co-attention Fusion (FSCF) module to explore potential associations between deep features from different modalities for MDD diagnosis, achieving an accuracy of 75.2%. Yuan et al. [16] developed a Brain Dynamic Attention Network (BDANet) to dynamically generate sample-specific brain graphs using fMRI and sMRI images for identifying depression. Wei et al. [17] proposed a sub-attention mechanism to explore multimodal information for automatic depression estimation. They hypothesized that the clues to depression can be obtained from diverse heterogeneous resources and demonstrated the effectiveness of their multimodal fusion strategies in the classification tasks.

Although existing works above using multi-modal fusion methods have achieved acceptable performance in diagnosing depression, all of them were implemented relying on large-scale multi-modal datasets and supervised learning paradigms, which require massive amounts of data to be incorporated and the huge participation of experienced physicians during the initial data preparation phase. This is impractical in real-world settings due to the limited scale in local hospitals and the heavy burdens on physicians. In addition, the limited sensitivity of neuroimaging in the detection of depression frequently raises issues of inconsistent or mismatched annotations between different modalities. It is therefore desired to develop an effective framework that can automatically learn multi-modal data representations and complementarities from a scale-limited multi-modal neuroimaging dataset for MDD diagnosis.

Knowledge transfer is regarded as an effective strategy in few-shot learning. Unsupervised/self-supervised learning, an important branch of machine learning, can explore the relationships in data by maximizing intra-sample similarity [18], [19]. Contrastive learning [20], [21] is a promising self-supervised learning approach, which learns common features between similar samples from the unlabeled data in the upstream tasks. The learned feature representations are then transferred to the downstream scenario tasks. In other words, contrastive learning enables knowledge transfer from heterogeneous data to the target dataset through upstream training and downstream knowledge transfer, playing an excellent performance in the few-shot studies [22], [23], [24]. In theory, compared to limited labels, images themselves should contain richer and more diverse information, making self-supervised learning easier to implement and more promising.

In this study, we developed a computer-aided diagnosis model to exploit the multi-modal semantic information within a limited sample set for the automatic diagnosis of MDD. Concretely, a novel self-supervised contrastive learning framework is used to learn multi-modal representations for MDD diagnosis. Firstly, as a multi-source, homogeneous dataset and the largest publicly available MDD dataset, the REST-meta-MDD dataset [25] was used to learn feature knowledge in upstream tasks, referred to as stage I: pre-training. Subsequently, the knowledge is transferred to the target dataset through fine-tuning for downstream tasks, called stage II: meta-training. Meanwhile, a 2-stage integration strategy was designed to fuse multi-modal features and classify them. Lastly, the performance of the proposed model was comprehensively evaluated using a series of evaluation metrics with some statistical tests. An occlusion analysis was used to explore the key biomarkers of MDD. The main contributions of this work can be summarized as follows:

- We propose a novel multi-modal contrastive learning framework for the automatic diagnosis of MDD.
- We design a 2-stage hierarchical feature integration paradigm to fuse multi-modal information.
- We demonstrate that feature knowledge transfer strategies can be used to address the challenges of insufficient and imbalanced datasets.
- The experimental findings reveal that the default mode network (DMN)- and cerebellum-related regions play a pivotal role in the diagnosis of MDD.

The rest of the paper is organized as follows: Section II provides an overview describes of data preparation and preprocessing. Section III exhibits the details of the proposed framework. Section IV provides the experimental setup and results. Section V and Section VI present the discussion and conclude the findings of the study, respectively.

II. Materials

The materials section consists of the following main components (a) participants, and (b) data preprocessing.

A. Participants

In this study, we collected the multi-modal MRI data as the target dataset, including sMRI, rs-fMRI and DTI scans, from a total of 128 participants from Gansu Provincial Hospital, consisting of 62 depression patients and 66 age- and sex-matched healthy controls (HCs). All patients with MDD in the study received a clinical diagnosis based on the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID). HCs were assessed using the non-patient edition of the Structured Clinical Interview for DSM-IV. All participants were between the ages of 18 and 65, right-handed, and didn't have any other mental illness, or history of illegal substance abuse such as heroin, etc. Of note, this study was supported by the Ethics Committee of Gansu Provincial Hospital (Approval No. 2017-071). The participants provided informed consent after thoroughly understanding and receiving the nature of the study, potential risks, and benefits.

All of the above participants were scanned using a 3.0T MRI Siemens Trio scanner. The T1-weighted, rs-fMRI and DTI scan-imaging parameter settings and cautions are provided in [26] and [27].

In addition, the heterogeneous dataset is a multi-source, homogeneous one. It was obtained from a publicly accessible REST-meta-MDD dataset. It is currently the largest public MDD dataset, containing 1300 MDD patients and 1128 HC individuals from 25 research groups affiliated with 17 hospitals across China. Theoretically, the transferred knowledge can be obtained using any dataset and it is not the only option in the pre-training stage, as the upstream task is a self-supervised learning paradigm based on contrastive learning [20]. However, due to the black-box nature of deep learning, the other options may introduce a degree of uncertainty and weaken the interpretability of the model.

B. Data Preprocessing

1) sMRI Preprocessing: The Voxel-based morphometry (VBM) standard preprocessing pipeline procedures are provided in reference [28]. CAT12 (https://neuro-jena.github.io/cat) is an extended toolkit for SPM12 and was used to process T1-weighted MRI. The main processes include: (a) segmentation of T1w images into grey matter, white matter, and cerebrospinal fluid; (b) normalization; and (c) smoothing. The default parameters were set during this process, and the grey matter volumes (GMV) were estimated from T1-weighted MRI.

2) Rs-fMRI Preprocessing: We employed the standard preprocessing pipeline procedures in [27], which uses the DPARSF toolbox (http://www.restfmri.net) based on SPM12. The main processes include: (a) removing the first 10 volumes; (b) slice timing and head motion correction; (c) coregistration; (d) anatomical segmentation; (e) regression; (f) spatial normalization; (g) smoothing with 8mm Gaussian kernel; and (h) bandpass filtering of 0.01–0.08 Hz. It is noteworthy that the BOLD signals of 116 brain regions were first extracted using Automated Anatomical Labeling templates (AAL templates) [27], [29], and then the functional connectivity matrix (FCM) was obtained by calculating the Pearson correlation coefficient between each pair of BOLD signals according to eq.1.

$$corr(r_i, r_j) = \frac{cov(r_i, r_j)}{\sigma_{ri}\sigma_{rj}}, \tag{1}$$

where r_i and r_j represent the fMRI signals of region i and region j, respectively. $cov(r_i,\ r_j)$ represents the covariance between r_i and r_j. σ_{ri}, σ_{rj} denote the standard deviation of r_i and r_j, respectively.

3) DTI Preprocessing: Raw DTI data were preprocessed based on the PANDA (http://www.nitrc.org/projects/panda). The detailed standard pipeline is provided in [30], which involves: (a) removing non-brain sections; (b) head movement correction and eddycurrent correction; (c) acquisition of the fractional anisotropy coefficients; and (d) Gaussian smoothing. Finally, the fractional anisotropy mapping (FAM) was generated by mapping the MNI space to the AAL template.

The main objective of this study is to develop a deep learning architecture for recognizing depression with a few-shot multi-modal dataset. Those data with missing modalities and excessive head movement (rotation degree $>$ 2 or translation distances $>$ 2 mm or mean FD (Jenkinson) $>$ 0.2) were excluded. Patients clinically diagnosed with MDD and having HAMD scores $>$ 7 were included. In total, 54 MDD and 62 HC were included for further analysis. The demographic of the participants was reported in Table I.

In addition, the REST-meta-MDD dataset provides the GMV and is used for pre-training in the upstream task.

III. Methods

A. Overview

The contrastive learning model can learn the inter-sample differences using a contrasting paradigm for weight pre-training, which improves the performance of subsequent label prediction tasks. Figure 1 provides an overview of the proposed contrastive learning framework based on multi-modal MRIs for the diagnosis of MDD. The overall process consists of two main steps: a pre-training stage and a meta-training stage. As follows:

- In the pre-training stage: For an input REST-meta-MDD (x), the augmented sample pairs (x_i', x'_j) are generated by the Data augmentation module. These generated sample pairs are then input into the Encoder module ($\mathscr{F}(\odot)$) in parallel for feature extraction. The extracted feature representations (h_j) are sequentially imported into the MLP module for feature projection. Finally, the inter-subject distances are calculated according to the contrastive loss function (Eq.3) in the potential feature space.
- In the meta-training stage: After data preprocessing, GMV and FAM are fed into the feature extractor. The trained weights from the upstream task are transferred to the backbone in the downstream task and fine-tuned. The FCM is fed into an MLP module for classification. Lastly, a hierarchical fusion paradigm integrates multimodal predictions and outputs the final results.

B. Data Preparation

1) Data Augmentation: The data augmentation module plays an important role in contrastive learning and aims to generate augmented sample pairs on a batch of input data to construct positive samples and negative samples for contrast. Three data augmentation methods were employed in this study, including image cropping, color distortion, and Gaussian blur.

a) Image cropping: The image was resized from an original size of 121 × 145 (a slice of GMV) to the target size of 224 × 224. Random cropping was then performed using a 50 × 50 window size, covering 22% of the image.

b) Color distortion: The GMV slices were converted to a greyscale map with probability $(p) = 0.2$ to achieve a greyscale color distortion.

c) Gaussian blur: A 3×3 Gaussian kernel was utilized to compute the weighted average of adjacent pixels to achieve Gaussian blur.

Notably, the data augmentation module generates the augmented view (x_i') by transforming the given data (x_i). The sample pairs x_i' and x_j' are considered as positive sample pairs

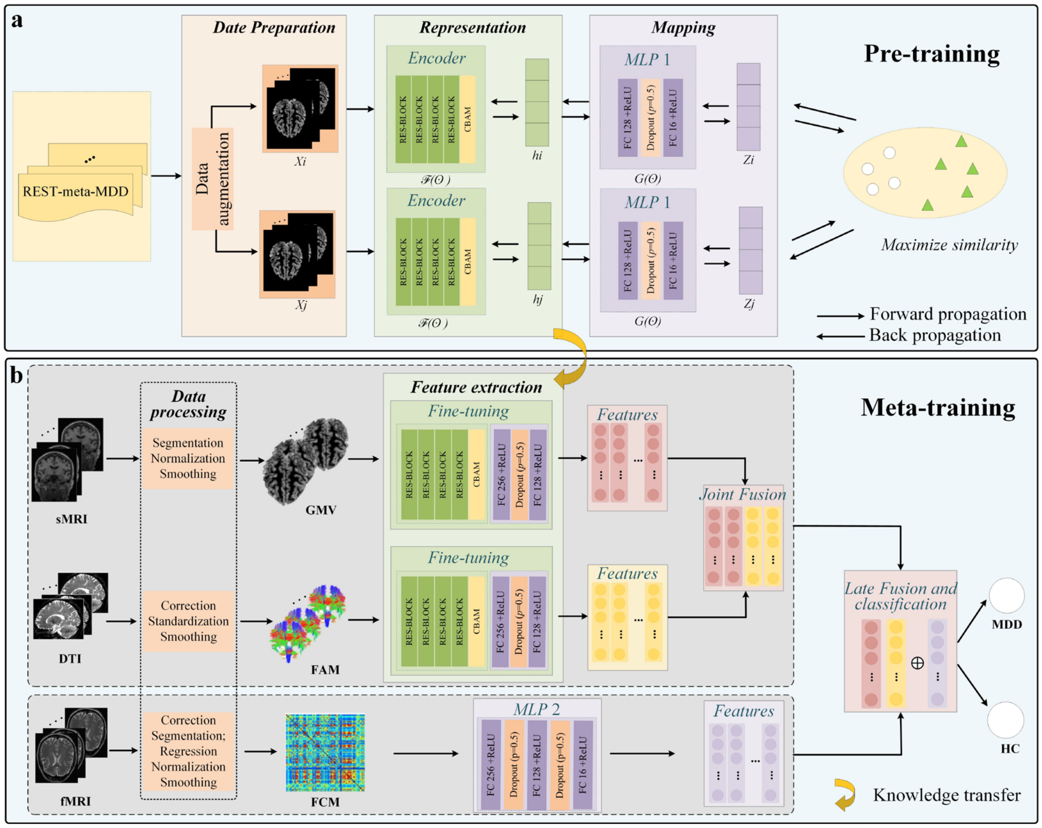


Fig. 1. Overview of the multi-modal contrastive learning-based image classification method. (a) Stage I: pre-training sub-network, and (b) Stage II: meta-training sub-network. *X*i and *X*j denote sample pairs, *h*i and *h*j represent feature representation through encoder *F*(⊙), and *Z*j is the projected feature through the multilayer perceptron.

TABLE I
THE DEMOGRAPHIC OF THE PARTICIPANTS

	MDD	HC	p-valve
Number of participants	54	62	-
Gender (Male/ Female)	30/24	27/35	0.1969 [a]
Age (years)	33±11.62	33.52±12.17	0.8132 [b]
HAMA	17.19±7.58	-	-
HAMD (17-item)	17.62±5.95	-	-

Abbreviations: MDD = Major Depressive Disorder, HC = healthy controls, HAMA = Hamilton anxiety scale, HAMD = Hamilton depression rating scale,
[a] represents two-sided Pearson chi-square test.
[b] represents two-sided two-sample t-test.

if they come from the same subject (x_i). Otherwise, they are regarded as negative sample pairs. It is noteworthy that the construction of negative samples is indispensable. Without them, the network may collapse [31], [32].

2) Data Normalization: Unlike the representations of natural images, where pixel values range in [0, 255], MRI images stored in DICOM files are encoded as 16-bit unsigned integers. However, excessively large values in voxels may bias the model and hinder model convergence during training, ultimately resulting in a decline in model performance [33].

To mitigate this undesirable effect, the max-min normalization was applied to normalize the MRI data to [0, 1] according to eq. (2).

$$V_{nor} = \frac{V}{Vmax - Vmin}, \qquad (2)$$

where V_{max} and V_{min} represent the maximum value and minimum value in MRI, and normalized data is denoted as V_{nor}.

C. Contrastive Learning Network

1) Stage I: Pre-Training Sub-Network: In contrastive learning, the pre-training stage focuses on feature clustering and representation by maximizing differences between negative pairs and minimizing agreement between positive ones via contrastive loss in the potential space. The pre-training sub-network is depicted in Figure 1 (a), consisting of 4 main parts as follows:

- Data augmentation module: The positive and negative sample sets were simultaneously constructed by sequentially applying three augmentations with batches of

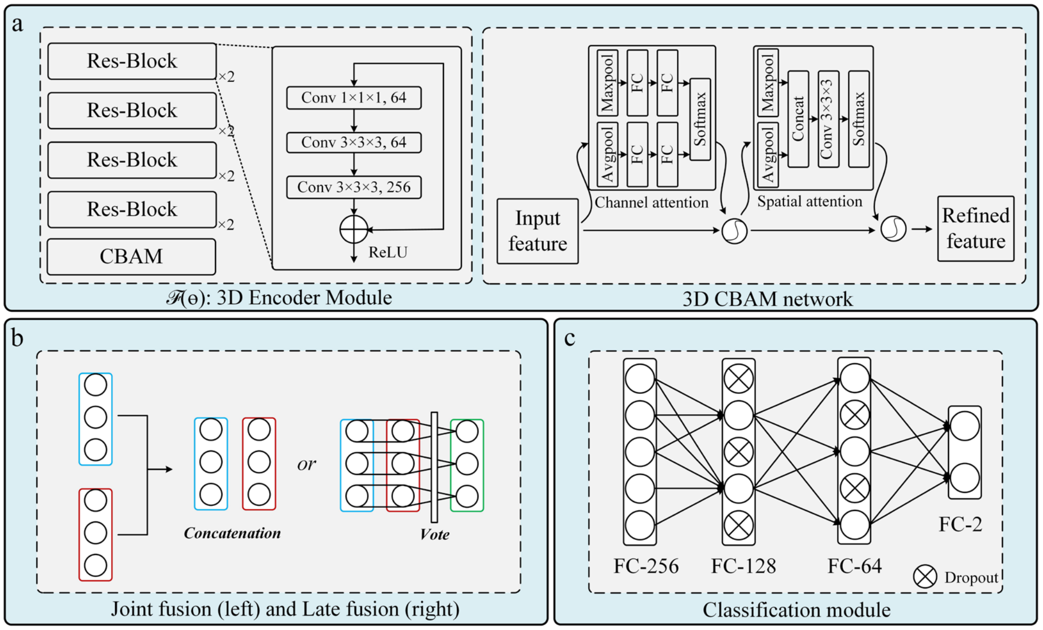


Fig. 2. The downstream task architecture consists mainly: (a) feature extraction module, (b) feature fusion module, and (c) feature classification module.

heterogeneous data: image cropping, color distortion, and Gaussian blur.

- Encode module: A modified classical deep learning backbone module was utilized for feature encoding. The 3D-ResNet-18 [34] with the 3D Convolutional Block Attention Module (CBAM) [35] was used as the feature extractor. Note that the residual structure addresses the issue of gradient vanishing, while the CBAM assist in directing the model's attention to the key region. Specifically, the feature extractor is constructed by stacking N Res-Blocks, where N was set to {2, 2, 2, 2}. The detailed architecture setup is illustrated in Figure 2 (a).
- MLP module: The Multilayer Perceptron (MLP) module maps feature representations to a space where the contrastive loss is applied. The MLP is constructed by the stacked Dense layer (i.e., Full Connection layer), Dropout layer and the Rectified Linear Unit activation function.
- Maximize similarity: The contrastive loss function is defined in eq3 and on the right part in Figure. 1(a), respectively.

$$L_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} F_{[k\neq i]} \exp(sim(z_i, z_k)/\tau)}, \quad (3)$$

where $F_{[k\neq i]}$ is an indicator function evaluating to ensure $k \neq i$, $sim(\cdot)$ denotes acquiring similarity scores (i.e., cosine similarities) between example pairs, and τ is a temperature parameter.

It should be noted that the REST-Meta-MDD dataset, as a multi-site dataset, is conducive to the construction of negative samples. Previous studies [31], [36] suggest that introducing more negative samples contributes to strengthening the model's representation capability. This is because additional negative samples can more effectively characterize the underlying distribution of features by increasing the distance between the negative samples and anchors [32]. Therefore, this study does not employ site variance elimination.

TABLE II
PARAMETER SETTINGS OF THE MULTI-MODAL CONTRASTIVE LEARNING

Parameters	Pre-training	Meta-training	Others
Learning rate	3e-4 #	1e-2	cosine annealing decay
Epoch	100	500	early stop
Optimizer	Adam	SGD	-
Batch size	4	32	-
Temperature parameters	0.07	-	-

Abbreviations: # : Default value; - : No settings.

2) Stage II: Meta-Training Sub-Network: The architecture of the meta-training sub-network is shown in Figure 1(b), which consists of three components: feature extraction, fusion, and classification sub-network.

- Feature extraction: The feature extraction network consists of three parallel branches, each corresponding to one input from the multi-modal target data (i.e., GMV, FAM, and FCM). For GMV and FAM branches, we freeze the backbone network from the pre-training stage and train the weights of subsequent structures during the meta-training stage to fine-tune the model. Furthermore, we redesigned the feature classification module (MLP 2 in Figure.1 b) for FCM due to inconsistent input dimensions compared to the other inputs. Simultaneously, we preserved the intrinsic classification capabilities of FCM [37] to some extent by using special fusion methods (see the next section). Detailed training parameters are provided in Table II.
- Feature fusion: The structure of the feature fusion module uses a hierarchical integration strategy as depicted

TABLE III
COMPARISONS OF DIFFERENT INPUTS OF MULTI-MODAL IN CONTRASTIVE LEARNING MODEL

Data modalities	Fusion strategy	*Accuracy*	*Precision*	*Recall*	*Specificity*	*F*-1	p-value$^{\#}$
sMRI	-	0.6500±0.0903	0.6292±0.048	0.6188±0.1299	0.6788±0.0899	0.6214±0.1073	$p < 0.001$
fMRI	-	0.6676±0.0573	0.6621±0.0771	0.6250±0.0412	0.7056±0.1113	0.6404±0.0443	$p < 0.001$
DTI	-	0.6912±0.0470	0.7294±0.0246	0.6062±0.1532	0.7667±0.1278	0.6409±0.0773	$p < 0.001$
sMRI, fMRI	LF	0.6765±0.0820	0.7302±0.1519	0.5625±0.0883	0.7778±0.1870	0.6229±0.0644	$p < 0.001$
sMRI, DTI	JF	0.6941±0.0668	0.7219±0.0718	0.6000±0.2067	0.7778±01309	0.6306±0.1342	$p < 0.001$
fMRI, DTI	LF	0.6970±0.0481	0.6993±0.0581	0.6312±0.0996	0.7556±0.0749	0.6594±0.0674	$p < 0.001$
sMRI, fMRI, DTI	JF+LF	**0.7352±0.0208**	**0.7498±0.0310**	**0.6625±0.0713**	**0.8000±0.0497**	**0.7005±0.0364**	**-**

Abbreviations: LF = Late Fusion; JF = Joint Fusion. $^{\#}$= t-test of the comparison between classification accuracy.

in Figure 1(b) and Figure 2(b). Specifically, the joint fusion strategy (i.e., concatenation fusion, where feature channels are stacked) is employed for the feature fusion of GMV and FAM features, acting as the first integration stage. The late fusion strategy (i.e., score fusion, where maximum probability score for multiple channels serves as prediction results) is employed for integrating the results of the first stage and FCM, serving as the second integration stage.

- Feature classification: As shown in Figure 2(c), feature mapping and classification were performed through the subsequent Dense layers, Dropout layers and soft-max activation functions.

In summary, in this study, the pre-training stage focuses on learning features from heterogeneous data, while the meta-training stage outputs the predictions of MDD and HC by fine-tuning the model and integrating the features.

IV. RESULTS

This section reports the experimental results of comparative and ablation experiments.

A. Experimental Setup

The experimental evaluation of our study metrics includes accuracy, precision, recall, specificity and the F-α score (F-1, $\alpha = 1$), which are defined in eqs. 4-8, respectively

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{7}$$

$$F - 1 = 2 \times \frac{Prec \times Rec}{Prec + Rec}, \tag{8}$$

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively.

A promising classifier is characterized by a high true positive rate (TPR) and a low false positive rate (FPR). The classification performance of the proposed model at various thresholds can be reliably reflected using the receiver operating characteristic (ROC) curve and the Area Under Curve (AUC) value. Generally, closing the ROC curve to the upper left corner indicates better model performance and a higher AUC value. Furthermore, the statistical significance of the accuracy difference between the proposed model and other models was tested using the two-sample t-test.

It is noteworthy that feature weights were learned using all GMV from the REST-meta-MDD dataset in the pre-training stage. The target data (116 subjects) were divided into disjoint training and test subsets using a stratified 4-fold cross-validation in the meta-training stage. In other words, the experimental dataset was divided into a training set (87 subjects) and a test set (29 subjects) with a ratio of 3:1, where one-fold from the training set was used as the validation set (29 subjects) for fine-tuning and assessing the model convergence. Of note, the early stopping strategy was utilized to assess the model convergence by monitoring the loss value of the validation set. The training process was stopped when the loss value of the validation set began to show an upward trend.

The experiments were compiled with pytorch-1.13 and executed on Nvidia-A100 GPUs running on Ubuntu 20.04. Due to hardware constraints, we configured a small batch size and used a validation set-based early stopping strategy to ensure convergence of the model in the mate-training stage. A summary of the parameter setting is reported in Table II.

B. Experimental Results

In this section, the performance of the proposed model was evaluated through comparative and ablation experiments. To ensure the reliability of the experimental results and mitigate the impact of the imbalanced dataset, we implemented a stratified cross-validation strategy to partition the target dataset. We comprehensively evaluated the proposed model's performance from different perspectives by employing several evaluation indicators, including accuracy, precision, recall, specificity, the F-1 score, and t-test. In addition, we generated the confusion matrix of the proposed model to analyze the types of errors made by the classifier. It is noteworthy that each of the reported experimental results is an average of 4-fold cross-validation.

Table III and Figure 3 present the comparative results using different combinations of imaging modalities as input. The best performance (Accuracy = 0.7352) was achieved when three modalities were input simultaneously. Besides, the DTI modality demonstrated superior classification

TABLE IV

COMPARISONS OF DIFFERENT MODELS BETWEEN CLASSICAL DEEP LEARNING MODELS AND THE PROPOSED MODEL USING sMRI, fMRI AND DTI AS INPUTS

Models	*Accuracy*	*Precision*	*Recall*	*Specificity*	*F*-1	*p*-value#
AlexNet [34] *	0.6882±0.0206	0.7493±0.0941	0.5526±0.1598	0.8056±0.1367	0.6159±0.0767	$p < 0.001$
VGGNet [35] *	0.7000±0.0387	0.7105±0.0865	0.6625±0.1592	0.7333±0.1500	0.6676±0.0706	$p < 0.001$
ResNet [30]*	0.6676±0.0417	0.6834±0.0692	0.5625±0.0884	0.7611±0.0909	0.6119±0.0597	$p < 0.001$
DensetNet [36]*	0.7118±0.0304	0.7480±0.0547	0.6000±0.1257	0.8111±0.0750	0.6557±0.0690	$p < 0.001$
ViT [37] *	0.6667±0.0509	0.5000±0.0606	0.5208±0.0176	0.7962±0.0962	0.5784±0.1317	$p < 0.001$
Swin-ViT [38] *	0.6911±0.0811	0.7306±0.0957	0.5375±0.1419	**0.8277±0.0665**	0.6138±0.1228	$p < 0.001$
BYOL [39] *	0.7205±0.0465	0.5794±0.0844	**0.7361±0.0806**	0.6562±0.0943	0.6869±0.0539	$p = 0.152$
MoCo v3 [40] *	0.6960±0.0170	0.6078±0.0103	0.7204±0.0767	0.7778±0.1111	0.6495±0.0296	$p < 0.001$
Ours	**0.7352±0.0208**	**0.7498±0.0310**	0.6625±0.0713	0.8000±0.0497	**0.7005±0.0364**	-

Notes: * denotes classic deep learning model. #= *t*-test of the comparison between classification accuracy.

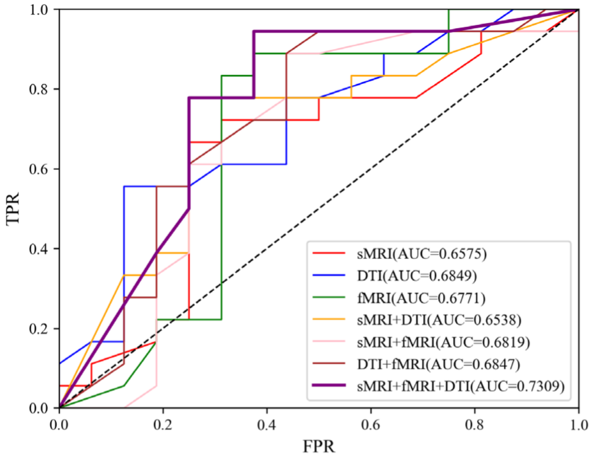


Fig. 3. The ROC curves of the classifiers on test samples with different inputs of multi-modal.

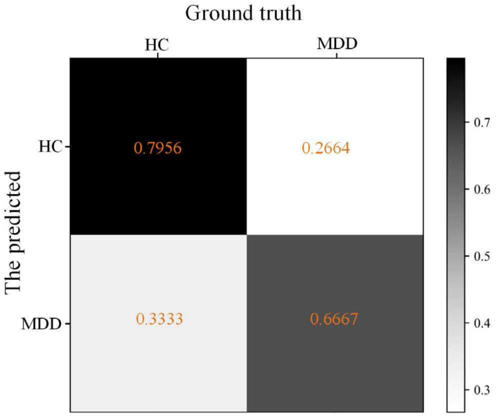


Fig. 4. Confusion matrix of the proposed model using the method of [45]. The rows represent the predicted results and the columns represent the ground truth.

performance (Accuracy =0.6912) in unimodal contrastive learning experiments, while the fusion of fMRI and DTI outperformed (Accuracy = 0.6970) the integration of other modalities in bimodal. This also illustrates that: (a) DTI outperforms the other two modalities in terms of classification performance, and (b) the complementarity of information between inter- and intra-modalities can be exploited through multi-modal fusion techniques.

In Table IV, we conducted a performance comparison between the proposed model and a selection of classical 3D deep learning models (e.g., AlexNet [38], VGGNet [39], ResNet [34], DenseNet [40], Vision Transformer [41], Swin-Transformer [42], BYOL [43] and MoCo v3 [44]), using three modalities as inputs.

In Table V, we evaluated the impact of different encoder modules derived from classical deep learning models on the performance of the model. Table VI presents the results of ablation studies of CBAM modules, which further demonstrates the effectiveness of the proposed model.

Moreover, we provide a confusion matrix of the proposed model using the method of [45] to investigate the predicted labels and the ground truth, as shown in Figure 4. Finally, the optimal parameters were obtained by testing various parameter settings for the Gaussian kernel, cropping window size, temperature coefficient and batch size, as shown in Figure 5.

Detailed explanations are provided in the DISCUSSION section.

V. DISCUSSION

This section presents a discussion of our findings, occlusion analysis, and limitations and future directions.

A. Interpretation of Our Findings

We concentrated on developing a deep learning model to leverage the complementarity of multi-modal neuroimaging data within a limited sample set for the automatic diagnosis of MDD. Here, we hypothesize that deep learning models have the potential to bridge the modality gap by extracting complementary information between functional and structural modalities. In our study, we proposed a self-supervised contrastive learning network as the classifier to learn feature knowledge using the heterogeneous data, and then transferred the acquired knowledge to the target data during the meta-training stage. Ultimately, we employed a two-stage hierarchical integration strategy to output the prediction.

TABLE V
COMPARISONS OF DIFFERENT ENCODER MODULES WITH CBAM IN THE PROPOSED MODEL

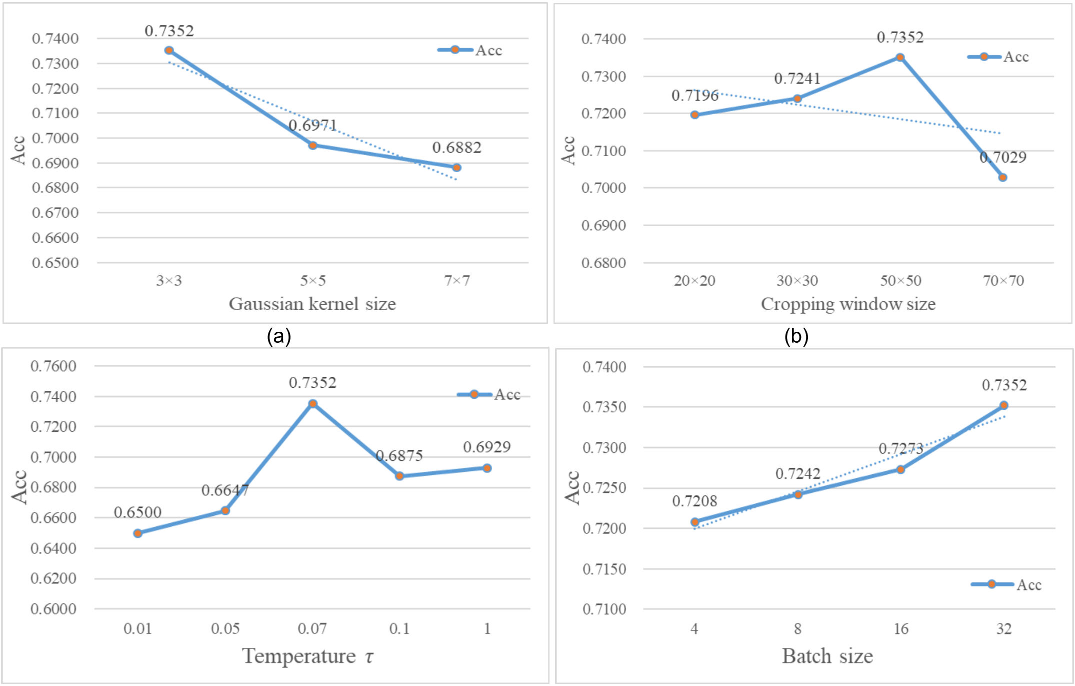
Models	*Accuracy*	*Precision*	*Recall*	*Specificity*	*F*-1	*p*-value#
AlexNet module	0.6352±0.0591	0.6309±0.0882	0.6000±0.1148	0.6667±0.1410	0.6037±0.0742	$p < 0.001$
VGGNet module	0.6588±0.0541	0.6963±0.0744	0.5062±0.1729	0.7944±0.0909	0.5677±0.1296	$p < 0.001$
DensetNet module	0.7294±0.0131	**0.7509±0.0490**	0.6500±0.1045	0.8000±0.0842	0.6903±0.0386	$p = 0.783$
ResNet module **(Ours)**	**0.7352±0.0208**	0.7498±0.0310	**0.6625±0.0713**	**0.8000±0.0497**	**0.7005±0.0364**	-

#= *t*-test of the comparison between classification accuracy.

TABLE VI
ABLATION EXPERIMENTS WITH CBAM MODULE

Models	CBAM module	*Accuracy*	*Precision*	*Recall*	*Specificity*	*F*-1	*p*-value#
Ours		0.7058±0.0208	0.7203±0.0619	0.6375±0.1118	0.7667±0.1069	0.6676±0.0425	$p < 0.001$
Ours	√	**0.7352±0.0208**	**0.7498±0.0310**	**0.6625±0.0713**	**0.8000±0.0497**	**0.7005±0.0364**	-

#= *t*-test of the comparison between classification accuracy.



(a)

(b)

(c)

(d)

Fig. 5. The performance of the model using different parameter settings for Gaussian kernel size, cropping window, temperature τ and batch size. (a) different sizes of the Gaussian kernel; (b) different sizes of the cropping window; (c) different settings of the temperature τ; (d) different settings of the batch size.

The experimental results suggest that the best classification performance is achieved when integrating all three modalities simultaneously as inputs for the contrastive learning model. Of note, in unimodal experiments, DTI has demonstrated the highest classification performance, followed by fMRI and sMRI. It has been suggested that depression can lead to disruptions in the microstructure of the brain, which [46] may explain this phenomenon in our study. The integration of information between fMRI and DTI performs better than other bimodal experiments in the unified model setting. Besides, a factor that has to be taken into account is that changes in structure are considered to underlie functional changes in the brain, and multiple functional connections or even entire functional networks may be affected when one connection in a structure is affected [47]. The functional connection matrix is typically used to describe correlations and connectivity between different brain regions [11], [27]. Although the exact relationship between structural and functional alterations in the brain is still unclear, the multi-modal contrastive learning network can learn implicit relationships among cross-modalities using feature extraction and feature fusion strategies. In particular, late fusion, in contrast to joint fusion, enables the exploration of synergistic effects in multi-modal learning by preserving the predictive capabilities of each modality [48].

The t-test results demonstrate that the proposed model significantly outperforms the other models in the comparison, providing evidence for the robustness of our approach. It is worth noting that the significance level was weak in the

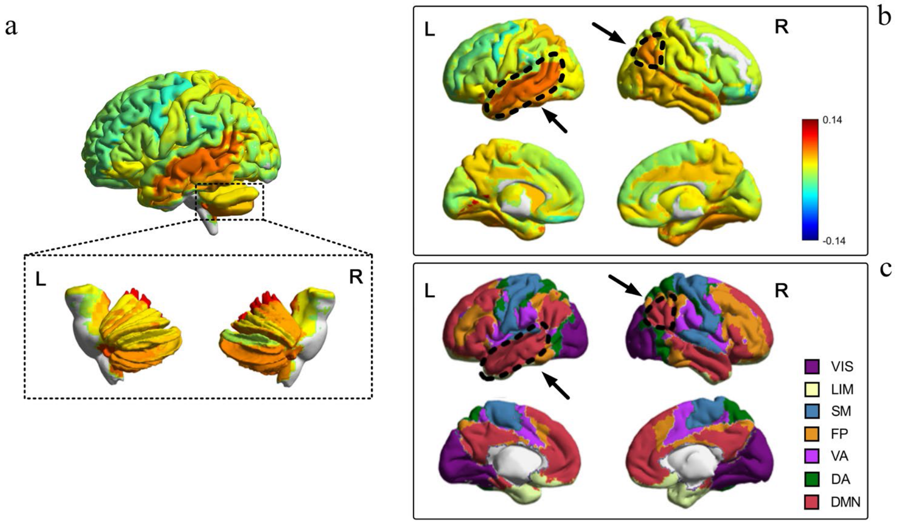


Fig. 6. Effects of different brain regions occluded by the AAL templates on classification performance. (a) effects of different cerebellum regions occluded by the AAL templates on classification performance, (b) effects of different cerebrum regions occluded by the AAL templates on classification performance, (c) Thomas yeo 7-network cortical parcellation. (The black circle areas are the DMN regions).

comparison experiments between the proposed model and the BYOL model (see Table IV), as well as the DenseNet encoder module (see Table V). This may be related to the fact that BYLO used of an asymmetric structure to overcome the reliance on negative samples [43], and the design of densely connected convolutional networks [40], which can guide future research.

The confusion matrix indicates that the model exhibits higher classification reliability for HC compared to MDD. Furthermore, MDD is more likely to be incorrectly predicted as HC, which could be attributed to the absence of organic pathological change in MDD [49], [50]. This is supported by the low recall (sensitivity) and high specificity of the model (see Table I).

As illustrated in Figure 5, the model's performance gradually decreases as the Gaussian kernel size increases (3 × 3, 5 × 5, 7 × 7 [51]), indicating that while larger size kernels enhance the blurring effectiveness, they may also result in the loss of crucial details. The optimal cropping window size was determined by testing various sizes (20 × 20, 30 × 30, 50 × 50, 70 × 70). Excessively small sizes may reduce the cropping effect, while excessively large sizes may result in the loss of critical details. We also experimented with different temperature values ($\tau = 0.01, 0.05, 0.07, 0.1, 1$) to find the optimal parameter. Furthermore, we experimented with different batch sizes (= 4, 8, 16, 32) in the downstream tasks, and the results demonstrated that a larger batch size can improve the model's performance [31].

Interestingly, the transformer-based architecture exhibits lower performance (i.e., ViT, Swin-Transformer in Table IV), which may be attributed to the relatively small sample size in this study. In addition, it can be observed that self-supervised learning exhibits superior performance as compared to supervised learning (i.e., BYOL in Table IV). Possible reasons for this may include: (1) the limited label samples hinder the improvement of supervised learning performance [52], (2) self-supervised learning models have superior feature extraction capabilities compared to supervised learning. At the same time, the upstream architectural design has a significant impacts on a model's performance in the downstream scenario tasks [53], (3) the feature knowledge transfer strategies can be used to address the challenges of insufficient and imbalanced datasets [54], and (4) different fusion strategies have the potential to enhance the complementarity of cross-modalities information [48], [55].

B. Occlusion Analysis

Early recognition of depression can help alleviate or even prevent its progression. Currently, several studies have been devoted to mining biomarkers of depression [11], [27], [47], [56]. In this study, we performed occlusion analysis to find biomarkers for MDD diagnosis. Specifically, after obtaining occlusion masks based on the AAL atlas for each subject, Figure 6 objectively presents the contributions of different brain regions and features to classification performance by occluding brain regions.

In addition, we assessed the impact of local sub-networks in MDD with reference to the Thomas yeo 7-network cortical parcellation [57]. The DMN- and cerebellum-related regions exhibit clearer differences compared to other regions. The DMN, a high-level cognitive network, is widely acknowledged to be closely related to monitoring inter-mental alterations, attentional capture, and cognitive resource allocation [56], [58]. Moreover, it plays a pivotal role in the pathophysiology of depression [59]. Emotional dysregulation and impaired cognitive control are commonly observed symptoms in individuals with MDD and are generally high-level correlated with DMN [27]. Notably, the alterations of functional connectivity dynamics exhibit significant and negative correlations with the severity of symptoms [56]. Furthermore, previous studies also indicate common connectivity dysconnectivity patterns in different cerebellar systems among patients with MDD, implying widespread dysfunction throughout the cerebellum [60]. Structural abnormalities of the cerebellum have also been observed in MDD [61]. Decreased connectivity between the DMN and cerebellum in patients with MDD is primarily observed in the inferior temporal gyrus, precuneus and angular [62]. Previous findings have implicated the temporal lobe in various cognitive processes such as emotion regulation, social cognition and memory processing [63].

In our findings, the results indicated that the DMN- and cerebellum-related regions exhibited significant effects on classification performance. In summary, we conclude that the DMN- and cerebellum-related regions can serve as biomarkers for the diagnosis of MDD.

C. Limitations and Future Directions

Although satisfactory results have been obtained with the proposed model, there are still several limitations. First, we demonstrated that the DMN- and cerebellum-related regions can serve as biomarkers for depression diagnosis, but we encountered limitations in expressing it symbolically. Second, the interpretability of the proposed model still needs to be strengthened due to the black-box nature of deep learning. Third, due to hardware limitations, we are unable to set the optimal parameters, such as a larger batch size. Finally, in this study, we exclusively integrated the most prevalent features from each modality as inputs, without providing evidence for the potential of other features despite this approach being effective and feasible.

In the future, we intend to improve the classification performance of the proposed model in several directions: (1) exploring various data augmentation techniques to improve the generalizability of the proposed framework, (2) designing more powerful encoder modules to extract discriminative features by capturing underlying patterns in multi-modal data, (3) designing a contrastive loss function to minimize the inter-subject differences by maximizing the similarity, and (4) collecting additional multi-modal data and attempting to incorporate domain knowledge to improve model performance.

VI. Conclusion

In conclusion, this study aimed to develop a computer-aided diagnostic model to leverage the complementarity of multi-modal neuroimaging data within a limited sample set for the automatic diagnosis of MDD. We employed a self-supervised contrastive learning framework to extract cross-modalities features in the absence of datasets for MDD diagnosis. Feature knowledge was learned from the heterogeneous data in the pre-training stage and transferred to the target dataset by fine-tuning in the meta-training stage. The prediction was achieved via a 2-stage hierarchical feature integration paradigm. Subsequently, we comprehensively evaluated the robustness of the proposed model through various comparison and ablation experiments. Our findings not only demonstrated improved classification capabilities by exploiting the complementarity between inter- and intra-modalities within a limited dataset but also highlighted critical roles of the DMN- and cerebellum-related regions in MDD recognition, suggesting their potential as biomarkers for MDD diagnosis.

In the future, we aim to improve the performance of the developed classification network in the following directions: First, we plan to augment the existing dataset and collect additional multimodal neuroimaging data with pathology reports to incorporate domain knowledge. Additionally, we will explore the design a more powerful encoder module and a tailored loss function to extract discriminative features.

Acknowledgment

Data Availability Statement: the authors express their utmost gratitude to the REST-meta-MDD consortium for generously providing them with the valuable data, which serves as a robust foundation for aiding in the diagnosis of depression.

Declaration of Competing Interest: The authors contributed effort to this article and have no conflict of interest.

References

[1] T. Vos et al., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019," *Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[2] S. Evans-Lacko et al., "Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the WHO world mental health (WMH) surveys," *Psychol. Med.*, vol. 48, no. 9, pp. 1560–1571, Jul. 2018.

[3] Y. Fang, M. Wang, G. G. Potter, and M. Liu, "Unsupervised cross-domain functional MRI adaptation for automated major depressive disorder identification," *Med. Image Anal.*, vol. 84, Feb. 2023, Art. no. 102707.

[4] A. R. Gerlach et al., "MRI predictors of pharmacotherapy response in major depressive disorder," *NeuroImage, Clin.*, vol. 36, Jan. 2022, Art. no. 103157.

[5] B. Sen, K. R. Cullen, and K. K. Parhi, "Classification of adolescent major depressive disorder via static and dynamic connectivity," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2604–2614, Jul. 2021.

[6] V. Gonuguntla, E. Yang, Y. Guan, B. Koo, and J. Kim, "Brain signatures based on structural MRI: Classification for MCI, PMCI, and AD," *Hum. Brain Mapping*, vol. 43, no. 9, pp. 2845–2860, Jun. 2022.

[7] P. H. Kassani, A. Gossmann, and Y.-P. Wang, "Multimodal sparse classifier for adolescent brain age prediction," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 336–344, Feb. 2020.

[8] E. Yang, L. Wang, D. Steffens, G. Potter, and M. Liu, "Deep factor regression for computer-aided analysis of major depressive disorders with structural MRI data," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 208–211.

[9] M. Mousavian, J. Chen, and S. Greening, "Depression detection using atlas from fMRI images," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2020, pp. 1348–1353.

[10] K. Thiel et al., "Reduced fractional anisotropy in bipolar disorder v. major depressive disorder independent of current symptoms," *Psychol. Med.*, vol. 53, no. 10, pp. 4592–4602, Jul. 2023.

[11] Y. Liang and G. Xu, "Multi-level functional connectivity fusion classification framework for brain disease diagnosis," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 6, pp. 2714–2725, Jun. 2022.

[12] M. Zhao et al., "An attention-based hybrid deep learning framework integrating brain connectivity and activity of resting-state functional MRI data," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102413.

[13] Y.-D. Zhang et al., "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Inf. Fusion*, vol. 64, pp. 149–187, Dec. 2020.

[14] W. Yan et al., "Deep learning in neuroimaging: Promises and challenges," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 87–98, Mar. 2022.

[15] G. Zheng et al., "An attention-based multi-modal MRI fusion model for major depressive disorder diagnosis," *J. Neural Eng.*, vol. 20, no. 6, Dec. 2023, Art. no. 066005.

[16] X. Yuan et al., "Cross-domain identification of multisite major depressive disorder using end-to-end brain dynamic attention network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 33–42, 2024.

[17] P.-C. Wei et al., "Multi-modal depression estimation based on sub-attentional fusion," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 623–639.

[18] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 42–62, May 2022.

[19] X. Tao, X. Gong, X. Zhang, S. Yan, and C. Adak, "Deep learning for unsupervised anomaly localization in industrial images: A survey," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–21, 2022.

[20] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jun. 2021.

[21] T. Shi et al., "A simple and effective self-supervised contrastive learning framework for aspect detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, 2021, pp. 13815–13824.

[22] S. Azizi et al., "Big self-supervised models advance medical image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3458–3468.

[23] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 481–490.

[24] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2228–2237, Sep. 2022.

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

[26] N. Chen, M. Guo, Y. Li, X. Hu, Z. Yao, and B. Hu, "Estimation of discriminative multimodal brain network connectivity using message-passing-based nonlinear network fusion," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 4, pp. 2398–2406, Dec. 2021.

[27] Z. Zhao et al., "Altered temporal reachability highlights the role of sensory perception systems in major depressive disorder," *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, vol. 112, Jan. 2022, Art. no. 110426.

[28] R. A. Bethlehem et al., "Brain charts for the human lifespan," *Nature*, vol. 604, no. 7906, pp. 525–533, 7906.

[29] S. C. Hellewell et al., "Profound and reproducible patterns of reduced regional gray matter characterize major depressive disorder," *Transl. Psychiatry*, vol. 9, no. 1, p. 176, Jul. 2019.

[30] N. Tzourio-Mazoyer et al., "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, Jan. 2002.

[31] T. Chen, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[32] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-D-Swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411415, doi: 10.1109/TGRS.2022.3202036.

[33] T. Li et al., "Automated detection of skeletal metastasis of lung cancer with bone scans using convolutional nuclear network," *Phys. Med. Biol.*, vol. 67, no. 1, Jan. 2022, Art. no. 015004.

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

[35] S. Woo, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[37] H. Zheng et al., "The dynamic characteristics of the anterior cingulate cortex in resting-state fMRI of patients with depression," *J. Affect. Disorders*, vol. 227, pp. 391–397, Feb. 2018.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[42] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[43] J. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 21271–21284.

[44] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," 2021, *arXiv:2104.02057*.

[45] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101759.

[46] K. R. Cullen et al., "Altered white matter microstructure in adolescents with major depression: A preliminary study," *J. Amer. Acad. Child Adolescent Psychiatry*, vol. 49, no. 2, pp. 173–183, Feb. 2010.

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

[48] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, no. 1, p. 136, Oct. 2020.

[49] E. Wee Yun Qing, "Depression and Decay–A case of major depressive disorder with psychotic features in an elderly patient with cancer," *Eur. Psychiatry*, vol. 66, no. S1, pp. S933–S934, Mar. 2023.

[50] N. Ş. Durmuş, B. Can, and A. Tufan, "Unintentional weight loss in adults 65 years or older: A symptom of physical and psychiatric etiologies," *J. Nervous Mental Disease*, vol. 210, no. 8, pp. 640–642, 2022.

[51] Z. M. Ramadan, "Effect of kernel size on Wiener and Gaussian image filtering," *TELKOMNIKA (Telecommun. Comput. Electron. Control)*, vol. 17, no. 3, p. 1455, Jun. 2019.

[52] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2168–2187, Apr. 2022.

[53] K. Ohri and M. Kumar, "Review on self-supervised image recognition using deep neural networks," *Knowl.-Based Syst.*, vol. 224, Jul. 2021, Art. no. 107090.

[54] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jul. 2020.

[55] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[56] Z. Yao et al., "Altered dynamic functional connectivity in weakly-connected state in major depressive disorder," *Clin. Neurophysiol.*, vol. 130, no. 11, pp. 2096–2104, Nov. 2019.

[57] B. T. T. Yeo et al., "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *J. Neurophysiol.*, vol. 106, pp. 1125–1165, Sep. 2011.

[58] P. C. Mulders, P. F. van Eijndhoven, A. H. Schene, C. F. Beckmann, and I. Tendolkar, "Resting-state functional connectivity in major depressive disorder: A review," *Neurosci. Biobehavioral Rev.*, vol. 56, pp. 330–344, Sep. 2015.

[59] J. P. Hamilton, M. C. Chen, and I. H. Gotlib, "Neural systems approaches to understanding major depressive disorder: An intrinsic functional organization perspective," *Neurobiol. Disease*, vol. 52, pp. 4–11, Apr. 2013.

[60] Y. I. Sheline et al., "The default mode network and self-referential processes in depression," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 6, pp. 1942–1947, Feb. 2009.

[61] L. Zhao et al., "Cerebellar microstructural abnormalities in bipolar depression and unipolar depression: A diffusion kurtosis and perfusion imaging study," *J. Affect. Disorders*, vol. 195, pp. 21–31, May 2016.

[62] X. Wang et al., "Disrupted functional connectivity of the cerebellum with default mode and frontoparietal networks in young adults with major depressive disorder," *Psychiatry Res.*, vol. 324, Jun. 2023, Art. no. 115192.

[63] M. Beauregard, V. Paquette, and J. Lévesque, "Dysfunction in the neural circuitry of emotional self-regulation in major depressive disorder," *NeuroReport*, vol. 17, no. 8, pp. 843–846, 2006.