Attention-Like Multimodality Fusion With Data Augmentation for Diagnosis of Mental Disorders Using MRI

Rui Liu, Zhi-An Huang, Yao Hu, Zexuan Zhu, *Senior Member, IEEE*, Ka-Chun Wong, and Kay Chen Tan, *Fellow, IEEE*

***Abstract*— The globally rising prevalence of mental disorders leads to shortfalls in timely diagnosis and therapy to reduce patients' suffering. Facing such an urgent public health problem, professional efforts based on symptom criteria are seriously overstretched. Recently, the successful applications of computer-aided diagnosis approaches have provided timely opportunities to relieve the tension in healthcare services. Particularly, multimodal representation learning gains increasing attention thanks to the high temporal and spatial resolution information extracted from neuroimaging fusion. In this work, we propose an efficient multimodality fusion framework to identify multiple mental disorders based on the combination of functional and structural magnetic resonance imaging. A multioutput conditional generative adversarial network (GAN) is developed to address the scarcity of multimodal data for augmentation. Based on the augmented training data, the multiheaded gating fusion model is proposed for classification by extracting the complementary features across different modalities. The experiments demonstrate that the proposed model can achieve robust accuracies of 75.1 ± 1.5%, 72.9 ± 1.1%, and 87.2 ± 1.5% for autism spectrum disorder (ASD), attention deficit/hyperactivity disorder, and schizophrenia, respectively. In addition, the interpretability of our model is expected to enable the identification of remarkable neuropathology diagnostic biomarkers, leading to well-informed therapeutic decisions.**

***Index Terms*— Attention deficit/hyperactivity disorder, autism spectrum disorder (ASD), computer-aided diagnosis, magnetic resonance imaging, multimodality data generation, multimodality fusion, schizophrenia.**

I. Introduction

MENTAL disorders account for a large portion of the total global burden of disease. Especially for children and adolescents, 10%–20% of them have been affected by mental health problems worldwide [1]. Considering the global prevalence of mental illnesses, effective diagnosis for patients is essential. However, the current diagnosis of mental illnesses by symptom-based criteria such as the Diagnostic and Statistical Manual of Mental Disorders[1] is far from satisfactory. Recent studies have revealed that personal observation and subjective decision-making in the assessment process tend to misdiagnose or overdiagnose mild cases [2]. Meanwhile, the global shortfalls in medical professionals exacerbate the suffering of people in need. Based on noninvasive neuroimaging platforms, a myriad of computer-aided diagnosis (CAD) approaches have been proposed to identify neuropathological biomarkers for intelligent auxiliary diagnosis of mental diseases.

As a versatile imaging technique, magnetic resonance imaging (MRI) has been extensively adopted to capture functional and structural brain neural patterns from two modalities, i.e., functional MRI (fMRI) and structural MRI (sMRI), respectively. fMRI gains widespread attention in neuroscience research by measuring functional alterations in the brain caused by neuronal activities. Specifically, resting-state fMRI (rs-fMRI) allows us to effectively explore the resting-state functional connectivity (FC) of the brain by detecting the variations of blood-oxygen-level-dependent (BOLD) signals. FC refers to the temporal correlations between BOLD signals among spatially separated regions of interest (ROIs) [3] in human brain. The unique neural patterns can be characterized by measuring FC. The significant changes in FC can be identified as a productive indicator to boost the diagnosis performance of mental disorders [4].

Alongside fMRI, sMRI provides information to qualitatively and quantitatively explore brain structures such as shape, size,

[1]https://psychiatry.org/psychiatrists/practice/dsm

Manuscript received 9 October 2021; revised 13 May 2022; accepted 28 October 2022. Date of publication 14 November 2022; date of current version 4 June 2024. This work was supported in part by the National Key Research and Development Project under Grant 2019YFE0109600; in part by the National Natural Science Foundation of China under Grant 62202399, Grant 61871272, Grant U21A20512, and Grant 61876162; in part by the Research Grants Council of the Hong Kong SAR under Grant PolyU11211521; in part by the Open Project of BGIShenzhen under Grant BGIRSZ20200002; and in part by the City University of Hong Kong Dongguan Research Institute. *(Corresponding author: Zhi-An Huang.)*

Rui Liu, Yao Hu, and Ka-Chun Wong are with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong, Shenzhen Research Institute, Shenzhen 518060, China (e-mail: rliu38-c@my.cityu.edu.hk; yaohu4-c@my.cityu.edu.hk; kc.w@cityu.edu.hk).

Zhi-An Huang is with the Center for Computer Science and Information Technology, City University of Hong Kong Dongguan Research Institute, Dongguan 523000, China, and also with the City University of Hong Kong, Shenzhen Research Institute, Shenzhen 518060, China (e-mail: huang.za@cityu.edu.cn).

Zexuan Zhu is with the National Engineering Laboratory for Big Data System Computing Technology, and the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: zhuzx@szu.edu.cn).

Kay Chen Tan is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: kctan@polyu.edu.hk).

Digital Object Identifier 10.1109/TNNLS.2022.3219551

and integrity. It can be used to examine long-term structural alterations of the brain in terms of anatomy and pathology. Thanks to its high spatial resolution, micro and molecular MRI structural markers in the brain, e.g., cortex thickness, gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), can be probed and interpreted. Accumulating evidence suggests that several psychiatric disorders are associated with abnormality in cortex thickness, regional brain volumes, and densities [5]. Microstructural alterations in the frontal and temporal cortex and hippocampus have been found to increase the risk of developing schizophrenia (SCZ) and attention deficit hyperactivity disorder (ADHD) [6]. Based on regional changes in tissue volume, sMRI can provide effective measures to understand the underlying mechanisms of mental disorders [7]. Although MRI-based CAD methods have achieved wide success, most of them are restricted to one single modality (i.e., fMRI or sMRI). The inherent bias in a single modality limits the performance of the existing CAD methods.

Multimodal learning can be a good recipe to build models that process and correlate information patterns from multiple modalities for joint feature representations [8]. Multimodal learning has been an emerging technology of artificial intelligence to achieve successful pattern recognition applications in video segmentation [9], image-text retrieval [10], etc. Since a single modality cannot provide sufficient information for classification in many cases, combining or fusing heterogeneous multimodal data, i.e., multimodality fusion, becomes crucial, particularly in complex tasks. Multimodality fusion in neuroimaging normally combines nonredundant data acquired with the same instrument (in a wide sense) or different instruments (in a narrow sense). It aims to improve our understanding of the functional and structural underpinnings of the brain by exploiting spatiotemporal resolution complementarity. Although multimodality fusion techniques are popular in brain imaging studies, most of the existing methods make little effort to address the practical issue that there is insufficient multimodal imaging data for training at present, and the generic multimodal fusion strategies have become less effective to characterize multimodal variability across subjects. To address the insufficient data problem, recent multimodality fusion methods directly synthesize the neuroimaging data of one modality (e.g., positron emission tomography, PET) based on those of another modality (e.g., MRI), without considering the implicit pattern heterogeneity of the modality gap [11].

In this article, we propose an attention-like multimodality fusion framework with data augmentation (AMFDA) to classify mental disorders by utilizing the data augmentation method and multimodal classification model. Specifically, a multioutput conditional generative adversarial network (MCGAN) is used to simultaneously generate both fMRI and sMRI synthetic features for the training set. Then, a multiheaded gating fusion (MGF) model is proposed to integrate multimodal data in a sophisticated way and jointly extract the fusion representation across different modalities for classification. Comprehensive experiments are conducted for performance evaluation based on the real-world datasets of three mental disorders, i.e., autism spectrum disorder (ASD), ADHD, and SCZ. The main contributions of this article are summarized as follows.

1) An attention-like multimodality fusion model MGF is proposed to diagnose multiple mental disorders based on fMRI and sMRI data. With the flexible fusion strategy, MGF can efficiently extract complementary representations to improve model diagnosis capability.
2) We develop a generative modeling method to synthesize multimodal data for augmentation. Based on multitask learning, the proposed MCGAN learns the joint representation to generate fMRI and sMRI pairwise data simultaneously.
3) The proposed framework AMFDA enables interpretable deep learning for pattern recognition in brain differences between fMRI and sMRI. The identified neural patterns are anticipated to advance our understanding of functional and structural abnormality associated with mental disorders from the data-driven outcomes.

The rest of this article is organized as follows. In Section II, we briefly review the related work regarding data augmentation and multimodality fusion in neuroimaging. Section III introduces the detail of the proposed framework. The experimental results and analysis are discussed in Section IV. Finally, this article is concluded in Section V.

II. Related Works

This section first reviews the related works on neuroimaging data augmentation, including transformation-based methods and generative model-based methods. Then, the multimodality fusion-based models in neuroimaging data for mental disorders diagnosis are introduced.

A. Data Augmentation

The problem of data island and privacy-preserving protocol have led to the shortage of neuroimaging data and the risk of overfitting. More efforts should be devoted to addressing such a challenge for existing CAD methods. Data augmentation techniques serve as a useful solution to improve the generalization capabilities of the learning model by approximating the data probability space with more insightful data points. The existing data augmentation methods in neuroimaging can be roughly divided into two categories: transformation-based data augmentation and generative adversarial network (GAN)-based data augmentation.

1) Transformation-Based Data Augmentation: The central thrust behind the transformation-based method is to increase the quality, quantity, and diversity of datasets while mitigating data bias. The certain transformation is performed on the original neuroimaging data via traditional manipulations, e.g., noise injection, rotation, and cropping. Noise injection refers to adding random variable values drawn from Gaussian distributions into original data. Zhuang et al. [12] conducted experiments to verify the effect of adding Gaussian noise with different variances ranging from 0.01 to 0.3 for fMRI data augmentation. As for the rotation methods, neuroimaging matrices are spun clockwise/counterclockwise at certain degrees. Liu et al. [13] performed MRI data augmentation

through in-plane rotation of ± 1 pixel at seven slices in different orientations. Hao et al. [14] selected the best angle from a predefined range to perform the random clockwise rotation on diffusion-weighted MRI data of prostate cancer. Since some types of neuroimaging data, e.g., fMRI and PET, are in time series, random cropping can be used to segment data with a fixed length along the time dimension for data augmentation. In [15], ten sequences of a fixed length (i.e., 90) are cropped from the fMRI time-series data for each subject. Mao et al. [16] cropped the rs-fMRI scan of each subject into sixteen 3-D frames with a fixed sampling stride (i.e., 2). Nguyen et al. [17] recently adopted a novel T1-based coregistration method to synthesize fMRI images. Each source sMRI scan is registered to both age-matched and gender-matched target one. Then the source fMRI signal is used to synthesize new target fMRI data based on the resulting target sMRI scan.

The transformation-based methods are based on the assumption that more information can be captured from the original dataset by following a given transformation operation. Although performance improvement is observed in all of the above-mentioned methods, such inflexible augmentations cannot provide insights into the alternations in neuroimaging data so as to have a better understanding of them. As a result, the massively inflated size of training datasets is not guaranteed to be advantageous.

2) GAN-Based Data Augmentation: As a generative modeling method, GAN-based data augmentation aims to implicitly learn the data distribution from the whole original dataset while generating samples drawn from the learned distribution. This technique is effective in interpreting the latent representation pattern based on the generative network (i.e., generator) and discriminative network (i.e., discriminator). The rationale is to set up a zero-sum "game" between those two "players" (networks): 1) the generator G attempts to fool the discriminator D by forging new samples $G(z)$ based on random noise vector z and 2) D is trained to distinguish $G(z)$ from the real samples. In general, the loss function of GAN can be optimized as a minimax problem as

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{z\sim p_z}\left[\log(1 - D(G(z)))\right] \tag{1}$$

where p_{data} is the real data distribution of x, and p_z is the known prior distribution of noise vector z. The formula (1) aims to train G and D by simultaneously maximizing $G(x)$ and minimizing $1 - D(G(x))$. In an ideal situation, we can observe a Nash equilibrium point of the problem where G learned the same distribution as p_{data}, and $D(x) = (1/2)$ for all x.

Recently, GAN-based data augmentation methods have been used for neuroimaging data generation. For example, Zhuang et al. [12] presented a 3-D convolution-based conditional Wasserstein GAN for fMRI data augmentation. The simulation experiments demonstrated the performance improvement brought by the generated fMRI data. Yao and Lu [18] applied a pretrained deep neural network as a regulator to the normal WGAN to generate single-class FC data of rs-fMRI. To remove the barrier of multimodality, Bi et al. [19] developed a multichannel GAN model to integrate computed tomography images and annotated images as input for augmenting PET images. In [20], the multimodal data of MRI, PET, and diffusion tensor image were combined as input to yield new PET images based on a 3-D conditional GAN. The generated synthetic data are informative to define boundaries between classes. Nevertheless, these GAN-based methods can only estimate the distribution of single modality data for the output, which is restricted to the architecture of single modality classifiers. In this work, a novel MCGAN is used to learn the joint representation and distribution of multimodal data (i.e., fMRI and sMRI) for augmentation. Through MCGAN, a random vector can be used to simultaneously yield pairwise multimodal data of fMRI and sMRI.

B. Multimodal Fusion

Multimodal fusion refers to integrating information extracted from different homologous data and/or heterologous data so as to improve prediction performance. It attracts the increasing focus in various research fields, e.g., audiovisual speech recognition, image captioning, and neuroimaging classification. Based on which stage fusion occurs, the existing multimodal fusion methods in neuroimaging can be divided into three categories: early fusion (at feature level), late fusion (at decision level), and intermediate fusion. The early fusion is performed prior to the classification. For example, by extracting features from fMRI and sMRI data, Guo et al. [21] leveraged the joint feature representation to train a support vector machine (SVM) model for classification. To discriminate ASD patients from control groups, Aghdam et al. [22] directly concatenated the selected features from rs-fMRI data and sMRI data. Unlike early fusion, the late fusion strategy considers the results of multiple classifiers which are trained on separate modalities. Ahmed et al. [23] separately trained an SVM and Bayesian classifier on different modalities. Based on the resulting feature matrix, the second SVM classifier was built for late fusion. A global late fusion method was proposed in Shi's work [24] to integrate local decisions from MRI, PET, and CSF by using a novel boosting classifier. Although the above methods are relatively easy to implement, their fixed fusion strategies cannot effectively model intra- and inter-modality interactions.

Intermediate fusion aims to fuse representations from different modalities at different levels to learn a joint multimodal representation. Intermediate fusion can facilitate intra- and inter-interactions from different modalities by integrating information from both feature level and decision level. Operation-based, tensor-based, and attention-based fusion are three representative categories of intermediate fusion methods. The operation-based fusion methods are based on different types of manipulations, such as concatenation, weighted sums, majority voting, and custom-defined operations. Zhang et al. [25] executed a weighted sum calculation to combine four different modalities (MRI, PET, CSF, and Clinical scores) at different stages involving feature extraction and the final decision making. Based on fMRI and sMRI data, Abdolmaleki and Abadeh [26] concatenated the softmax

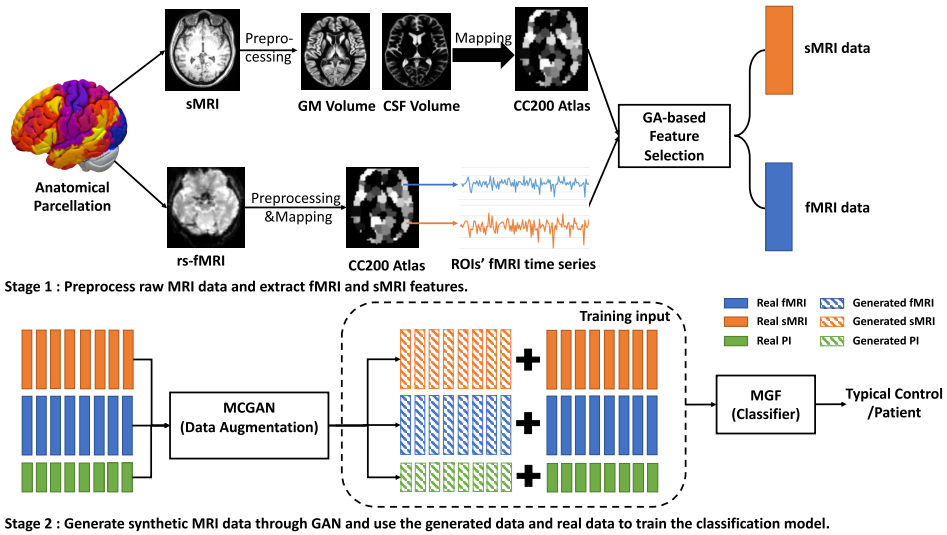


Fig. 1. Flowchart of the proposed model in two stages. GM volume: gray matter volume; CSF volume: cerebrospinal fluid volume; PI: phenotypic information.

output of two convolutional neural networks (CNNs) for early feature fusion. After that, late fusion was performed by linear discriminant analysis to diagnose ADHD.

Most tensor-based fusion methods tend to learn a joint representation space by computing the outer product of all feature tensors from different modalities. To reduce the computational complexity of tensor outer product calculation, Liu et al. [27] proposed a low-rank multimodal fusion method to decompose the weight into a set of low-rank factors. Tensor-based fusion methods struggle to address the curse of dimensionality in neuroimaging. Therefore, they cannot be applied to large sample sizes. On the contrary, attention-based fusion methods pay more focus on the remarkable local features from massive multimodal data, especially achieving success in visual question answering [28]. A dual attention network in [29] is presented to guide the model to probe the expected regions via elementwise multiplication between multimodal data. Currently, few attention-based fusion methods are proposed in this domain. Inspired by Nam et al. [29], we propose an attention-like MGF network to perform flexible fusion patterns on different modalities in multiple stages for auxiliary diagnosis of mental disorders.

III. Model Design

The flowchart of the proposed model is shown in Fig. 1. The primary process can be divided into two stages. First of all, the raw fMRI data and sMRI data are preprocessed by different preprocessing pipelines and then converted into ROIs' mean time-series data and structural volume data, respectively. Pearson correlation coefficient is used to extract the pairwise connection features among N ROIs for fMRI and sMRI data. The total number of connection features adds up to $N \times (N-1)/2$ by retrieving from the upper triangle values of the correlation matrix. A feature selection method based on the genetic algorithm (GA) dubbed GAFS is used to select discriminating connections between different ROIs. To harmonize the feature dimensionality, the sMRI features follow the same selected feature set of fMRI data. In the next stage, the selected fMRI and sMRI features are fed to MCGAN along with the phenotypic information for data augmentation. Based on the augmented multimodality training set, the proposed MGF performs an attention-like weighted fusion to discriminate individuals with mental disorders from healthy controls.

A. GA-Based Feature Selection

Given the dimensionality and complexity of multimodality data, it is essential to select effective features to filter out the irrelevant, redundant information. GAs tend to outperform traditional feature selection methods thanks to their capability of managing datasets with many features. Inspired by Zhao's work [30], a GA-based method is used to select remarkable FC connections. In this work, the solution of feature selection is represented by binary encoding using a bit string of 0 and 1 s. Each bit represents whether the corresponding FC connection is selected. Based on the binary encoding, it is flexible in applying GA algorithms to optimize solutions for feature selection with the guidance of proper fitness function. To efficiently evaluate the possible candidate feature selection solutions, we design a clustering-based similarity measurement as shown in Fig. 2. According to the candidate selected features, the hierarchical clustering algorithm is adopted to iteratively divide candidate solutions into C clusters (ranging from 2 to 5), and the corresponding Calinski-Harabasz indices are calculated, respectively [31]. By maximizing the Calinski-Harabasz index, the optimal number of clusters can be determined for each solution. Meanwhile, we also divide individuals into the patient group and typical controls (TCs) group denoted as L based on their ground-truth labels. Then, the similarity matrix between C and L is constructed as $S \in \mathbb{R}^{|C|\times|L|}$. In this case, Jaccard's Similarity Coefficient (JSC) is used to calculate the similarity between C_i and L_j in S, where C_i and L_j are the ith cluster and jth cluster in

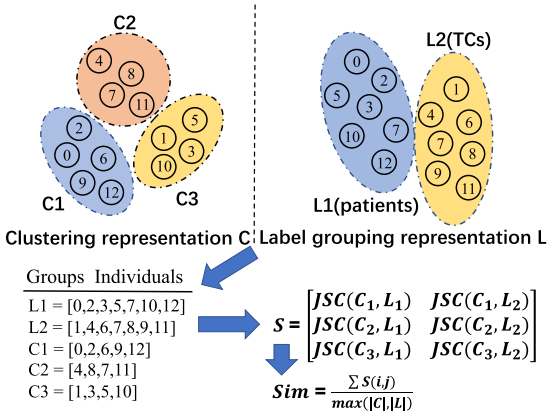


Fig. 2. Illustrative example of similarity measurement steps.

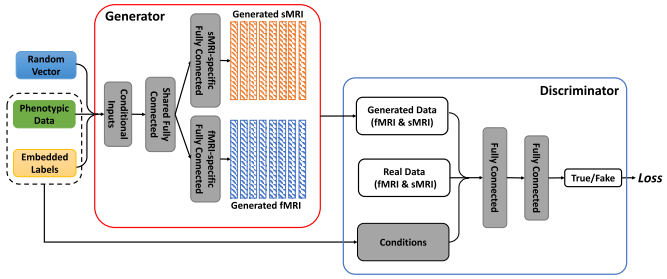


Fig. 3. Proposed architecture of MCGAN in block diagram.

C and L, respectively. The final similarity between C and L can be reached by aggregating each unit in S as follows:

$$S(i, j) = \mathrm{JSC}(C_i, L_j) = \frac{C_i \cap L_j}{C_i \cup L_j} \tag{2}$$

$$\mathrm{Sim} = \sum S(i, j)/\max(|C|, |L|). \tag{3}$$

After the optimization process of GAFS, the optimal feature selection solution with the highest similarity score is selected. Since GA, to a certain extent, is deficient in preventing falling into plateaus (all the surrounding points are very close to the same fitness) for solving complex problems. To overcome the obstacles caused by plateaus, the fitness scaling function [32] is also adopted to scale the fitness values for rewarding the more suitable solutions escaping from the local optimum traps.

B. Multioutput Conditional GAN

Based on the selected fMRI and sMRI features, a novel MCGAN model enables joint learning for multimodality data augmentation. The architecture of MCGAN is shown in Fig. 3. Since MCGAN is a variant of conditional GAN (CGAN) [33], the main principle of CGAN is first described. The purpose of CGAN is to train the generator and discriminator on the condition of some extra information y (e.g., class labels or data from other modalities). The conditioning process is performed by adding y into both discriminator and generator as auxiliary inputs in CGAN. During the training process of CGAN, the generator G combines prior noise vector z and conditional information y into a joint hidden representation to produce synthetic examples for each y in the training dataset. Unlike traditional GANs, the discriminator D of CGAN learns to estimate the probability of synthetic samples matching y. In this case, both G and D are trained simultaneously. The objective function of CGAN can be formulated by rewriting (1) as follows:

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}\left[\log D(\boldsymbol{x}|\boldsymbol{y})\right] + \mathbb{E}_{z\sim p_z}\left[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))\right]. \tag{4}$$

Compared with CGAN, MCGAN is improved to complete two tasks: 1) multitask learning is applied to learn the hidden representation shared between different modalities for generating fMRI and sMRI features simultaneously and 2) the conditional inputs can also be multimodal. In this work, both class labels and phenotypic information (e.g., age, sex, and handedness) are included.

The generator architecture of MCGAN is developed from the shared-bottom model, which is widely used in multitask learning. Two task-specific layers are placed over the bottom network, respectively. It needs to note that the leaky rectified linear unit (ReLU) layer and batch normalization are sequentially added between different fully connected layers, and the tanh activation function is used for the last layer. As for the structure of the discriminator, the model is composed of two fully connected layers. Leaky ReLU and dropout regularization is used for reducing overfitting after each layer. As inputs of the discriminator, the phenotypic information is first concatenated with the real data or synthetic data. In order to solve the instability problem in [33], the Wasserstein distance is applied in the training process of GAN in this work. The distance formula for the probability domain is

$$W(p_{\mathrm{data}}, p_z) = \inf_{\gamma\sim\prod(p_{\mathrm{data}}, p_z)} \mathbb{E}_{(x_r, x_g)\sim\gamma}\left[||x_r - x_g||\right] \tag{5}$$

where x_r and x_g represent a real sample and a generated sample, respectively. $\prod(p_{\mathrm{data}}, p_z)$ is the set of all possible joint probability distributions between p_{data} and p_z. However, the infimum (inf) of (5) is difficult to implement. By taking advantages of Kantorovich-Rubinstein duality and Lipschitz functions, Arjovsky et al. [34] proposed an alternative method to calculate the Wasserstein distance in GAN as

$$\mathcal{L} = \mathbb{E}_{x\sim p_{\mathrm{data}}}[f_w(x)] - \mathbb{E}_{x\sim p_z}[f_w(x)] \tag{6}$$

where f_w denotes a discriminator network with parameters w. The distance $\mathcal{L}$ approximates closer to the Wasserstein distance between actual data distribution and generated data distribution as $\mathcal{L}$ becomes larger. Thus, in this method, the learning loss of MCGAN can be reached as follows:

$$\mathcal{L} = \min_G \max_D -\mathbb{E}_{\boldsymbol{x}'\sim p_{\mathrm{data}}}\left[D(\boldsymbol{x}'|\boldsymbol{y}_l, \boldsymbol{y}_p)\right] + \mathbb{E}_{z\sim p_z}\left[D(G(\boldsymbol{z}|\boldsymbol{y}_l, \boldsymbol{y}_p))\right] \tag{7}$$

where y_l and y_p represent the class labels and phenotypic information, respectively. x' means the real samples integrated with fMRI and sMRI features. Moreover, the generator and discriminator are trained to optimize the model in turns. For each iteration, we train the generator once and the discriminator twice. Then, the SVM is trained to select qualified synthetic

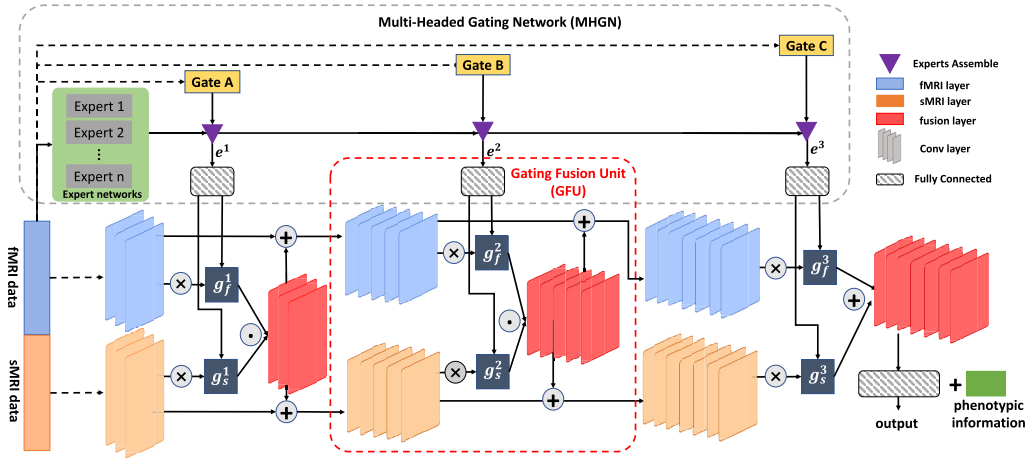


Fig. 4. Proposed architecture of MGF in block diagram. MGF mainly consists of multiheaded gating networks and gating fusion units.

samples with high classification confidence to better clarify decision boundaries. Finally, the resulting synthetic samples are mixed up with the real samples from the training set for training the MGF classification model. The pseudocode is shown in Algorithm 1.

C. Multiheaded Gating Fusion

Aimed at improving the classification performance on multimodal data, an attention-like fusion network called MGF is proposed in this section to identify mental disorders from TCs based on an MGF method. The main idea of MGF is to fuse different modalities in multiple stages of deep neural networks by utilizing the dynamic gating network and the attention-like fusion unit. As we can see in Fig. 4, MGF network is mainly composed of two parts: multiheaded gating network (MHGN) and gating fusion unit (GFU). MHGN dynamically computes dimension-specific scalar weights for different modalities and different levels of feature representation. Thanks to the multiple glimpses (output heads) offered by MHGN, GFU captures the subtle but relevant "attention focus" from different modalities while assuring feature reusability of the same dimensionality from prior layers. Specifically, the intermodality fusion within each GFU is implemented by the multiplication operations including two weighted multiplications and one elementwise product as shown in Fig. 4. Two multiplication operations are applied to assign the scalar weights learned from MHGN to different modalities at different stages. Then, the elementwise product is used for cross-modal fusion of the weighted features of fMRI and sMRI. Based on the weighted multiplication and elementwise product, the joint multimodal representation can be learned by a more flexible mixture pattern. To stabilize training and convergence, elementwise addition operation allows for feature reuse like skip connections in ResNets. Such an addition operation provides an unbroken gradient flow through the whole GFU for addressing the gradient update problem. In this work, the input multimodal data are first concatenated and

Algorithm 1 Pseudocode of MCGAN With Selection of SVM

Require: Preprocessed data X_f, X_s, phenotypic data X_p, and label Y_l

Ensure: Augmented dataset X_f^*, X_s^*, X_p^*, Y^*

$[X'_f, X'_s] \leftarrow \text{GAFS}(X_f, X_s)$

$[X'_{tr}, X'_{te}, Y_{tr}, Y_{te}] \leftarrow \text{Split}(X'_f, X'_s, X_p, Y_l)$ // $[X'_{f\_tr}, X'_{s\_tr}, X'_{p\_tr}]$ as X'_{tr}, $[X'_{f\_te}, X'_{s\_te}, X'_{p\_te}]$ as X'_{te}

Initialize $D(w)\&G(\theta)$. // w: the parameters of D; θ: the parameters of G

for $e = 1, \ldots, epochs$ **do** //e: # of training epoch

  for $m = 1, \ldots, minibatch$ **do** // m: # of minibatch

    // begin training discriminator D

    for $i = 1, \ldots, n$ **do** // train D n times in each epoch.

      Sample$\{X_{tr}^{'(i)}\}_{i=1}^m \sim p_{data}$//real data

      Sample$\{z^{(i)}, z_p^{(i)}, z_l^{(i)}\}_{i=1}^m \sim p_z$// noise

      $[X_{f\_fk}^{(i)}, X_{s\_fk}^{(i)}]$ as $X_{fake}^{(i)} \leftarrow G(z^{(i)}, z_p^{(i)}, z_l^{(i)})$

      $g_w \leftarrow \nabla_w \frac{1}{m} \sum_{i=1}^m \mathcal{L}([X_{tr}^{'(i)}, X_{fake}^{(i)}], [Y_{tr}^{(i)}, z_l^{(i)}])$ //

      $\mathcal{L}$:learning loss via Eq.(7)

      $w \leftarrow w + \alpha \cdot \text{RMSprop}(w, g_w)$ // α: learning rate

      $w \leftarrow \text{clip}(w, [-\epsilon, \epsilon])$ // weights clipping with ϵ

    end for

    // begin training generator G

    Sample$\{z^{(i)}, z_p^{(i)}, z_l^{(i)}\}_{i=1}^m \sim p_z$

    $X_{fake}^{(i)} \leftarrow G(z^{(i)}, z_p^{(i)}, z_l^{(i)})$

    $g_\theta \leftarrow -\nabla_w \frac{1}{m} \sum_{i=1}^m \mathcal{L}(X_{fake}^{(i)}, z_l^{(i)})$ via Eq.(7)

    $\theta \leftarrow \theta - \alpha \cdot \text{RMSprop}(\theta, g_\theta)$

  end for

end for

Sample$\{Z, Z_p, Z_l\} \sim p_z$//synthetic data from MCGAN

$[X_{f\_GAN}, X_{s\_GAN}] \leftarrow$ G.predict (Z, Z_p, Z_l)

SVM.train$(X'_{f_tr}, X'_{s_tr}, Y_{tr})$

$score \leftarrow$ SVM.predict_prob(X_{f_GAN}, X_{s_GAN})

$[X'_g, Z'_l] \leftarrow \text{Sort}(score)$//$[X'_{f\_GAN}, X'_{s\_GAN}, Z'_p]$ as X'_g

$[X_f^*, X_s^*, X_p^*, Y^*] \leftarrow \text{Combine}([X'_{tr}, Y_{tr}], [X'_g, Z'_l])$

RMSprop: Root Mean Square propagation [35]

then feed into MHGN to learn the scalar weights. After that, multimodal data are input into the stacked GFUs network for multimodality fusion with the guidance of the learned weights by MHGN. Finally, the generated joint feature representation is concatenated with phenotypic information for classification.

MHGN can be divided into two steps: input data embedding and dense gating transformation. The data embedding network is developed based on the structure of multigate mixture-of-expert (MMoE) [36], as shown in Fig. 4. The MMoE structure consists of a group of expert networks and several gating networks. The rationale behind MMoE is to model the relationship between different learning tasks where the expert networks are assembled with different learning weights obtained from gating networks. The gating network m^t for stage t first performs a simple linear transformation of the input x with a softmax layer

$$m^t(x) = \text{softmax}(W_t x) \tag{8}$$

$$\text{s.t.} \sum_{i=1}^{n} m_i^t(x) = 1 \tag{9}$$

where $W_t \in \mathbb{R}^{n \times d}$ is a trainable matrix in expert networks. Parameters n and d denote the numbers of experts and feature dimension, respectively. It is obvious that each m_i^t represents the learned weights corresponding to the ith expert. Then, the specific output of stage t denoted as e^t can be integrated by the output of the ith expert network $h_i(x)$ and m_i^t as follows:

$$e^t = \sum_{i=1}^{n} m_i^t(x) h_i(x). \tag{10}$$

Next, a three-layer fully connected network G^t accepts the output of e^t and estimates the weights for different modalities and different complex abstraction provided by convolutional layers. In this way, each gating network is responsible for driving a flexible mixture pattern in the corresponding stage of MGF. Finally, the output of MGF for stage t can be reached with softmax function as

$$g^t = \text{softmax}\left(G^t\left(e^t\right)\right). \tag{11}$$

Through the mediation by MHGN, GFU takes advantage of its effective attention mechanism to summarize the multimodal fusion representation via intra- or inter-modality interactions. The architecture of GFU is indicated by the red dashed line in Fig. 4. GFU serves as a building block in a hierarchical manner to characterize different levels of remarkable features. Three stacked GFUs are used in the proposed MGF. The joint representation r^t integrating fMRI and sMRI features at stage t is formulated as follows:

$$r^t = \left(f^t \cdot g_f^t\right) \odot \left(s^t \cdot g_s^t\right) \tag{12}$$

$$f^t = \text{Conv}\left(f^{t-1} + r^{t-1}\right) \tag{13}$$

$$s^t = \text{Conv}\left(s^{t-1} + r^{t-1}\right) \tag{14}$$

where g_f^t and g_s^t represent the gating outputs in stage t of MGF learned from the input multimodal data. The feature representation of fMRI and sMRI, i.e., f^t and s^t is learned from the previous layer of GFU by convolutional layers (denoted as Conv) with batch normalization. Symbol $\odot$ is elementwise multiplication for sparse joint representation. It is worth noting that for the last GFU, the elementwise multiplication is replaced with an elementwise sum to preserve the scale of the result. In the tail of MGF, a three-layer fully connected network undertakes the flattened output of final gating fusion.

IV. Experiments and Results

In this section, we conduct a series of experiments to evaluate the effectiveness of the proposed model. First, the datasets and preprocessing pipelines used in this work are introduced. The details of model settings and evaluation metrics are provided. Then, several state-of-the-art methods are compared with the proposed model for performance comparison. After that, the contributions of the components within our model, i.e., MGF and MCGAN, are investigated. The distribution of augmented data generated by our MCGAN is further analyzed in this section. Finally, we visualize the learned neural patterns of multiple mental disorders for model interpretation.

A. Data Acquisition and Processing Pipelines

In this work, three multimodality MRI datasets are downloaded from the Autism Brain Imaging Data Exchange I (ABIDE), the ADHD-200 Consortium (ADHD-200), and the Center for Biomedical Research Excellence (COBRE). Each sample in these datasets contains both functional and structural MRI data along with the key phenotypical information including age, subject gender, and handedness.

ABIDE I[2]: The ABIDE I dataset was created through aggregating independent datasets collected across 17 international brain imaging sites. Here all valid data of 1035 subjects were leveraged including 505 from individuals with ASD and 530 from TCs.

ADHD-200[3]: The ADHD-200 dataset includes 947 samples from 8 international imaging sites, where 362 are children and adolescents with ADHD and 585 are TCs. Due to some errors in preprocessing, eight samples in ADHD-200 cannot be used in this work.

COBRE[4]: The COBRE dataset consists of 146 samples involving 72 SCZ subjects and 74 TC subjects.

fMRI Preprocessing: To facilitate data sharing, the Preprocessed Connectomes Project systematically preprocessed the raw data from ABIDE and ADHD-200 using various preprocessing pipelines. The data through pipelines of the Configurable Pipeline for the Analysis of Connectomes[5] and Athena[6] are downloaded for ABIDE and ADHD-200, respectively. For COBRE, the pipeline of Data Processing Assistant for Resting State fMRI (DPARSF)[7] is used for preprocessing in this work. The preprocessing steps in DPARSF pipeline include: volume removal, slice timing correction and realignment, motion correction, spatial normalization, bandpass filtering, normalization by the MNI template, and smoothing with a

[2]http://fcon_1000.projects.nitrc.org/indi/abide
[3]http://fcon_1000.projects.nitrc.org/indi/adhd200/
[4]http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html
[5]https://preprocessed-connectomes-project.org/abide/cpac
[6]nitrc.org/plugins/mwiki/index.php/neurobureau:Athenapipeline
[7]https://rfmri.org/DPARSF

TABLE I
MODEL PARAMETER SETTINGS FOR GAFS, MCGAN, AND MGF

For GAFS					
Search Range	MaxIteration:[100,200];FeatureSelectionRatio:[0.1,0.5];popSize:[50,200]				
MaxIteration	200	FeatureSelectedRatio	0.2	popSize	100
For MCGAN					
Search Range	Epochs:[100,500];Batch Size:[8,256];Learning rate:[1.0e-05,1.0e-02];Clip value:[0.01,0.1]				
Epochs	500	NoiseVectoerDim	200	Batch Size	200
Optimizer	RMSprop	Learning Rate	1.0e-05	Clip Value	0.01
Configuration of generator			500-1000-Feature Dim		
Configuration of discriminator			1000-500-1		
For MGF					
Search Range	Epochs:[30,100];Batch Size:[8,128];Hidden Units:[32,512];Experts:[4,16]				
Epochs	50	Optimizer(lr=0.001)	Adagrad	Batch_size	16
Settings in MHGN		128(MMOE)-64-2(Dense Gating), experts: 12			
Configuration of convolutional layers			256-512, kernel_size:3		
Configuration of final fully connected layers			256-128-64-1		
Activation function to hidden layers			Relu		
Activation function to output layers			Sigmoid		

6-mm Gaussian kernel of full width at half maximum. The details of these three preprocessing pipelines are available on the websites. In this work, we leveraged Craddock 200 (CC200) functional parcellation atlas to partition the whole brain into 200 ROIs for extracting the mean time-series data.

sMRI Preprocessing: The sMRI data were preprocessed through the DPARSF package. Diffeomorphic Anatomical Registration using Exponentiated Lie algebra toolbox in DPARSF is used to segment the structural images into GM, WM, and CSF. Those segmented images were normalized to the MNI space and then smoothed using 6-mm Gaussian kernel. Then the calibration was performed between the ROI mask of CC200 and the preprocessed GM, WM, and CSF images to extract the ROI-based sMRI features by DPARSF. According to the pathological researches in [37] and [38], GM volumes for ASD and ADHD, and CSF volumes for SCZ become evident as compared to the control group. Based on the simulation experiments, our observations are also consistent with this rule. Namely, this specified fusion mode can achieve the best performance among possible fusion combinations.

B. Model Settings

The model parameter settings for each component of the proposed method are shown in Table I. Since the diagnosis tasks of ASD, ADHD, and SCZ are conducted separately, they are prone to obtain different optimal solutions by GAFS and thereby specify different FC sets for feature selection. The selected feature dimensions are 6400, 3000, and 3600 for the diagnosis of ASD, ADHD, and SCZ, respectively. Considering different feature dimensions selected in the three datasets, the network of MCGAN is different accordingly. To determine the number of expert networks E in MGF, we compared the performance of AMFDA by changing E from 4 to 16 on three datasets. The result suggests that the performance of AMFDA is not sensitive to the setting of E, and a larger E does not necessarily lead to better performance. The best performance is achieved when $E = 12$. Therefore, the number of experts is set to 12 in this work. 10-fold cross-validation (10-fold CV) is used for model evaluation in this work. The classification performance of the proposed method is evaluated by four common metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC). Those metrics are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{16}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{17}$$

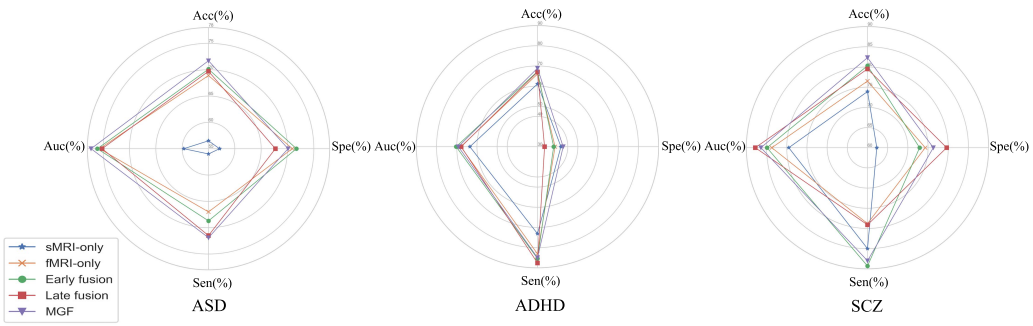
where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative values, respectively. The receiver operating characteristic curve is created via plotting all possible pairs of TP rate and FP rate at various threshold settings. It is good to use AUC value to show the relationship between sensitivity and specificity for each possible cut-off.

C. Performance Comparison

In Table II, we summarize the results of state-of-the-art methods reported in the recent surveys [16], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48] for ASD, ADHD, and SCZ classification on ABIDE I, ADHD-200, and COBRE, respectively.

In addition to 10-fold CV, the existing methods also applied leave-one-out cross-validation (LOOCV) and independent sets of training/testing (IS) for performance evaluation. As ADHD-200 is officially divided into training and validation sets for the ADHD-200 Global Competition, IS is commonly used on ADHD-200 for validation. However, the performance of IS is sensitive to the splitting of the training set and validation set. By contrast, 10-fold CV focuses more on the capability of generalization and reliability. Moreover, previous works tend to adopt a small set of data due to rigid sampling strategies (e.g., IQ-matched participants and similar MRI acquisition protocols). Therefore, their inflated classification accuracy could decline significantly as sample sizes increase. To address this issue, the simulation experiments are conducted with 10-fold CV on the whole datasets of ABIDE, ADHD-200, and COBRE.

As shown in Table II, the proposed model obtains the reliable accuracies of 75.1% (SPE: 72.3%, SEN: 77.7%), 72.9% (SPE: 60.2%, SEN: 81.7%), and 87.2% (SPE: 85.2%, SEN: 88.9%) on ABIDE, ADHD-200, and COBRE, respectively. Considering the natural data split of ADHD-200 for the global competition, the proposed model is also evaluated via IS validation, where the validation set accounts for 20% of the training set. As a result, AMFDA achieves a reliable accuracy of 71.9% (SPE: 50.0%, SEN: 90.3%) using the IS of ADHD-200. Although traditional machine learning methods are dominated in most previous works for mental disorders diagnosis (e.g., SVM in [49], random forest in [45]), deep learning-based methods outperform all compared conventional machine learning-based methods. It could be attributed to the effectiveness of deep learning to learn discriminative nonlinear feature representation. The proposed model obtains supreme



Method	ASD				ADHD				SCZ			
	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC
sMRI-only(0/12)	56.5	57.1	56.0	59.7	61.1	41.6	73.1	63.5	73.9	62.3	85.0	79.5
fMRI-only(0/12)	68.9	70.9	67.0	75.7	66.0	38.3	83.1	68.0	76.5	74.2	78.8	83.7
Early fusion(3/12)	70.1	**71.7**	68.7	76.1	67.3	38.0	85.4	**70.3**	80.3	72.9	**89.3**	84.9
Late fusion (3/12)	69.7	67.7	71.5	75.2	67.0	33.5	**87.6**	67.8	79.5	**79.6**	79.1	**87.8**
MGF (6/12)	**71.1**	70.1	**71.9**	**77.3**	**68.9**	**42.7**	85.0	69.2	**82.3**	76.3	88.0	86.4

Note: no feature selection methods are employed in compared methods.

Fig. 5. Performance comparison between MGF and other multimodal fusion methods.

TABLE II
PERFORMANCE COMPARISON ON ABIDE I, ADHD-200, AND COBRE

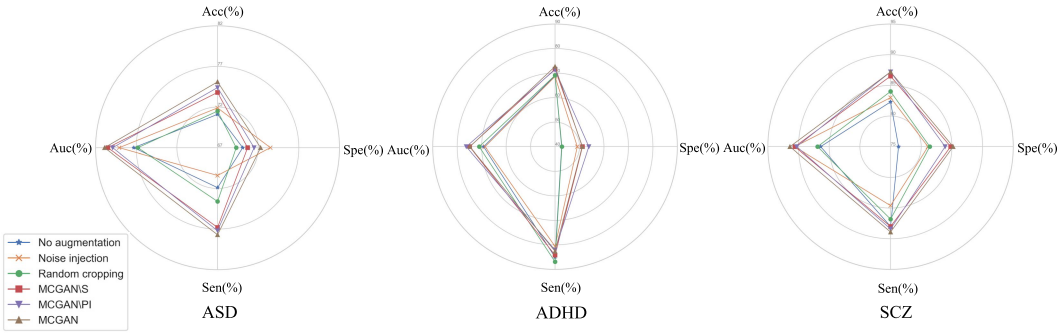
Model	Classifier	Validation	Simple #	Atlas	Accuracy(STD)
For ABIDE I (ASD)					
Ours	MGF	10-fold CV	1035	CC200	**75.1(1.5)**
Huang 2020 [41]	DBN	10-fold CV	1035	CC200	72.5(3.7)
Heinsfeld 2018 [40]	AE+ANN	10-fold CV	1035	CC200	70.0(N.A.)
Dvornek 2017 [39]	LSTM	10-fold CV	1035	CC200	68.5(5.5)
Plitt 2015 [49]	L-SVMs	10-fold CV	178	Others	69.7(2.7)
Aghdam 2018 [22]	DBN	10-fold CV	185	AAL	65.6(N.A.)
For ADHD-200 (ADHD)					
Ours	MGF	10-fold CV	939	CC200	**72.9(1.1)**
Ours	MGF	IS	939	CC200	71.9(N.A.)
Dou 2020 [50]	EM-MI	IS	782	AAL	70.2(N.A.)
Mao 2019 [16]	4D-CNN	IS	788	Others	71.3(N.A.)
Zou 2017 [43]	3D-CNN	IS	730	Others	69.2(N.A.)
Tan 2017 [42]	Linear SVM	10-fold CV	217	CC400	68.6(1.7)
Ghiassian 2016 [42]	RBM-SVM	IS	940	Others	70.0(N.A.)
For COBRE (SCZ)					
Ours	MGF	10-fold CV	146	CC200	87.2(1.5)
Savio 2015 [45]	RF	10-fold CV	146	Others	80.0(2.0)
Zou 2020 [44]	SVM	20-fold CV	130	AAL	78.0(2.3)
Kim 2016 [47]	DNN	5-fold CV	146	AAL	85.8(N.A.)
Latha 2019 [46]	DBN	3-fold CV	146	Others	**90.0(N.A)**
Hsieh 2014 [48]	SVM	LOOCV	141	AAL	71.6(N.A.)

N.A.: means no STD information is available in the paper; AAL: Automated Anatomical Labeling; CC200: Craddock 200; CC400: Craddock 400; Others: means the models adopted the personalized atlas or were not applicable for any atlas.

accuracy among all competitors by using the whole ABIDE and ADHD-200 datasets, respectively. In Huang's work [41], the additional Bayesian optimization-based hyperparameter tuning technique is used to pursue an optimal classification model. For a fair comparison, the best accuracy without using Bayesian optimization in [41] is compared in this article. Generally speaking, the proposed model can achieve competitive results on the whole datasets of ABIDE, ADHD-200, and COBRE.

1) MGF With Other Multimodal Fusion Methods: To evaluate the effectiveness of the attention-like fusion strategy, MGF is compared with four common models, i.e., early fusion, late fusion, sMRI-only, and fMRI-only (as baseline). It needs to note that no data augmentation methods are employed in this section. All compared models are set up with similar ANN-based network architecture as shown in Table I. Early fusion in this work is implemented by concatenating the preprocessed fMRI and sMRI features. Late fusion performs the weighted sum of final decisions from two modalities. The comparison results are shown in Fig. 5. In the single-modality mode for classification, the fMRI-only model outstrips the sMRI-only model in terms of the most evaluation metrics (91.7%). This result is consistent with the finding of recent multimodal fusion studies that fMRI data are prone to contain more informative features than sMRI data for classification. We can also observe that 50% (6/12) of evaluation metrics are dominated by MGF. Particularly, MGF shows supreme classification accuracies in three mental disorders. Compared with the baseline, the early fusion and late fusion models both achieve an improvement in most evaluation metrics. Generally, early fusion performs better than late fusion in this work. In other words, the feature-level fusion strategy is shown to have more advantages over the decision-level fusion strategy. As a variant of the feature-level fusion model, the proposed MGF obtains further promotions in terms of accuracy on ABIDE (1.0%), ADHD-200 (2.3%), and COBRE (2.3%). The results demonstrate that the proposed multiheaded gating fusion network is helpful to learn important mixture patterns by bridging the gap between inter-modality feature representations. Moreover, MGF can manage to strike a balance between specificity and sensitivity and thus achieve a higher AUC value.

2) MCGAN With Other Data Augmentation Methods: Likewise, the proposed MCGAN and the other two extensively used data augmentation methods (i.e., noise injection and random cropping) are compared based on MGF. Considering an SVM model is used to select the generated data with high



Method	ASD				ADHD				SCZ			
	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC
Section A: Performance comparison between MCGAN and other data augmentation methods												
No Augmentation(0/12)	71.1	70.1	71.9	77.3	68.9	42.7	85.0	69.2	82.3	76.3	88.0	86.4
Noise Injection(1/12)	71.9	**73.5**	70.4	79.1	68.6	49.1	80.6	68.6	83.0	81.1	84.6	90.6
Random Cropping(1/12)	71.5	69.3	73.6	76.8	69.1	42.8	**86.9**	71.0	84.0	81.4	86.8	86.9
MCGAN\S (0/12)	73.8	70.7	76.8	80.5	71.6	51.2	84.2	74.9	86.5	84.8	88.0	90.6
MCGAN\PI (1/12)	74.4	71 5	77.2	79.9	71.4	53.8	82.1	76.3	**87.2**	83.9	88.4	90.3
MCGAN (10/12)	**75.1**	72.3	**77.7**	**80.9**	**72.9**†	**60.2**†	81.7†	**76.8**†	**87.2**	**85.2**	**88.9**	**91.4**
Section B: Performance comparison on different fusion methods with (w/) or without (w/o) MCGAN												
Early Fusion	70.1	**71.7**	68.7	76.1	67.3	38.0	**85.4**	**70.3**	80.3	72.9	**89.3**	84.9
Early Fusion w/ MCGAN	**72.5**	69.7	**75.1**	**79.4**	**68.7**	**45.5**	83.0	69.9	**83.7**	**80.5**	86.6	**88.4**
Late Fusion	69.7	67.7	71.5	75.2	67.0	33.5	**87.6**	67.8	79.5	79.6	79.1	87.8
Late Fusion w/ MCGAN	**72.6**	**70.3**	**74.7**	**80.3**	**69.6**	**45.6**	84.3	**72.2**	**83.1**	**83.6**	**82.5**	**90.5**
MGF	71.1	70.1	71.9	75.2	68.9	42.7	**85.0**	69.2	82.3	76.3	88.0	86.4
MGF w/ MCGAN	**75.1**	**72.3**	**77.7**	**80.9**	**72.9**	**60.2**	81.7	**76.8**	**87.2**	**85.2**	**88.9**	**91.4**

†:To address the class imbalance problem on ADHD, MCGAN is used to generate sample data with a reversed ratio (i.e., 6:4).

Fig. 6. Performance comparison between data augmentation methods.

quality as the final step in MCGAN, we attempt to exclude the data selection by SVM in MCGAN (denoted as MCGAN\S) for evaluation. Similarly, phenotypic information is also excluded in MCGAN (denoted as MCGAN\PI) for evaluation. The proposed AMFDA without data augmentation is considered as the baseline. All compared data augmentation methods are set to the best mode with the optimal augmented factor ranging from 2X to 10X. The augmented factor for MCGAN is set to 2 in this work. To address the class-imbalance problem in the ADHD dataset, we simply use MCGAN to generate corresponding positive and negative samples in a ratio of 6:4 as augmented data to maintain the final training dataset class balance. As we can see in Fig. 6 (Section-A), 83.3% (10/12) of the evaluation metrics are dominated by the proposed MCGAN on all three real-world datasets. It is shown that MCGAN without data selection of SVM (i.e., MCGAN\S) loses a night-item lead accordingly. It is a simple yet effective step to help MCGAN better find the optimal decision boundaries by high-quality generated data. As compared with the baseline, the data augmentation methods all have positive effects on classification performance. However, the transformation-based methods (i.e., noise injection and random cropping) struggle to handle the complexity in applying multimodality fusion. Worse, the feature inconsistency between modalities in generated data could be detrimental. As a deep generative modeling method, MCGAN is proficient in exploring the multimodal distribution and coordinating the outputs based on the conditional inputs. The results also show that, compared with MCGAN\PI, the proposed framework with phenotypic information achieves a slight improvement in the classification performance of three disorders. Specifically, the diagnosis accuracies of ASD and ADHD are promoted by 0.7% and 1.5%, respectively. Compared with the baseline, MCGAN achieves significant improvements on the classification for ASD (4.0%), ADHD (4.0%), and SCZ (4.9%) in terms of accuracy as expected.

TABLE III

PERFORMANCE COMPARISON (ACC%) TO EVALUATE THE EFFECTIVENESS OF SVM SELECTION IN MCGAN

Method	ASD	ADHD	SCZ
No augmentation	71.1	68.9	82.3
MCGAN\S	73.8	71.6	86.5
MCGAN w/ $c < 0.55$	71.8	69.7	81.8
MCGAN w/ $c \geq 0.55$	**75.1**	**72.9**	**87.2**

As we mentioned in Section III-B, an SVM classifier was trained on the training dataset to select high-quality synthetic samples with high classification confidence (e.g., confidence score $c \geq 0.55$). To explore how low-quality synthetic data (denoted as "MCGAN w/ $c < 0.55$") affect the diagnosis performance in our framework, a comparison experiment is performed in this section. From Table III, we can observe that MCGAN w/ $c < 0.55$ still outperforms the baseline model (with no data augmentation) except on SCZ dataset. Not surprisingly, the low-quality synthetic data is a detriment to the effectiveness of MCGAN. The result demonstrates that the pretrained SVM is a simple but helpful filter method to get rid of the low-quality synthetic data. Moreover, a supplementary experiment was conducted to evaluate the effectiveness of MCGAN on the other commonly used fusion methods, i.e., early fusion and late fusion. The results are shown in Fig. 6 (Section-B). In addition to the proposed MGF, early fusion

TABLE IV
PERFORMANCE COMPARISON TO EVALUATE THE EFFECTIVENESS OF GFU AND MHGN

Method	ASD		ADHD		SCZ	
	ACC	AUC	ACC	AUC	ACC	AUC
AMFDA w/ addition	73.1	80.2	70.3	72.5	85.0	90.7
AMFDA w/ concatenation	73.4	79.3	71.0	73.4	85.7	91.0
AMFDA w/ GFU	**75.1**	**80.9**	**72.9**	**76.8**	**87.2**	**91.4**
MGF w/ self-attention	68.6	73.3	66.5	68.6	79.6	86.1
MGF w/ Transformer	70.6	76.7	68.5	**70.0**	80.9	85.2
MGF w/ MHGN	**71.1**	**77.3**	**68.9**	69.2	**82.3**	**86.4**

TABLE V
PERFORMANCE COMPARISON BETWEEN DIFFERENT FEATURE SELECTION ALGORITHMS

Method		ASD		ADHD		SCZ	
		ACC	AUC	ACC	AUC	ACC	AUC
Ours	No Feature Selection	71.2	76.3	68.2	70.6	81.8	84.1
	Genetic Alogrithm	**75.1**	**80.9**	**72.9**	**76.8**	**87.2**	91.4
Machine Learning	Extra-tree Classifier	72.5	77.4	69.7	71.5	82.7	86.2
	SVM-RFE Classifier	72.8	77.3	70.2	72.8	83.5	88.3
Traditional Optimization	Greedy Search	72.4	76.3	69.5	72.4	82.9	86.8
	Tabu Search	72.7	78.1	70.0	72.7	83.2	89.7
Meta-heuristic Search	Simulated Annealing	73.1	78.6	71.7	73.5	84.9	91.4
	Ant Colony Optimization	73.7	79.8	72.1	74.9	85.8	**91.7**

and late fusion methods can also greatly benefit from the improved performance delivered by MCGAN. It is suggested that MCGAN could become a powerful generic data augmentation tool to address the scarcity of multimodal data by maintaining strong robustness and versatility generalization in a wide range of multimodel fusion methods. By comparison, MGF with MCGAN contributes to a larger promotion of classification performance, especially in terms of ACC and AUC. That is to say, the proposed framework can be more beneficial from the combination and complementarity of MCGAN and MGF.

3) Other Ablation Studies: To evaluate the effectiveness of the proposed GFU, we replaced the GFU in AMFDA with simple concatenation (denoted as AMFDA w/ concatenation) and elementwise addition operations (denoted as AMFDA w/ addition). Concatenation operation jointly links multimodal features in the first layer of deep learning model, while elementwise addition operation directly adds up feature vectors together. The corresponding experimental results quantified by ACC and AUC are summarized in Table IV. As expected, both addition and concatenation operations carry a performance penalty on AMFDA, leading to the average ACC decreases of 2.3% and 1.7% on three datasets, respectively. This is probably because they equally treat all multimodal features in model learning rather than flexibly handling different characteristics of each modality. AMFDA w/ GFU shows a superior classification performance, dominating all the metrics in the table. The proposed GFU method can dynamically adjust the contribution of each multimodal feature for modeling the joint feature embedding space. This fusion strategy is demonstrated to enhance the feature representation learning in AMFDA.

To evaluate the multiheaded attention of MHGN, we have implemented the self-attention mechanism and transformer (a deep learning model that also adopts self-attention for tracking sequential data [51]) to separately reconstruct our MGF as a substitute for MHGN. From Table IV, we can observe that the self-attention mechanism adopted in MGF does not achieve the desired effect. It only achieved similar performance compared with the common early fusion and late fusion strategies (shown in Fig. 5). After a thorough investigation, we believe that the self-attention mechanism could be more effective at handling sequential data (e.g., text or video [52]) rather than the extracted FC features employed in this work. Self-attention is a well-known attention mechanism relating to different positions of single sequence for modeling a representation of the entire sequence. FC features are extracted from the ROIs' mean time-series data by Pearson's correlation coefficient. Each FC feature represents the connectivity strength between different ROIs. However, different positions of FC features cannot provide any helpful information toward the task by extracting the interdependencies among FCs. Therefore, we turned to use raw time-series MRI (sequential data) as inputs fed to transformer. To apply Transformer in our framework, we made appropriate modifications to multimodality fusion network based on the encoder of Transformer. As we expected, the transformer-based model outperforms the self-attention-based model in most evaluation metrics. Although the overall performance of transformer-based model is slightly inferior to our MHGN-based model, this experiment result demonstrates that this article has great potential that can be extended to effectively handle time-series fMRI data.

Since we applied GA-based feature selection method in our AMFDA framework, the influence of feature selection methods on the diagnostic performance needs to be investigated. Three types of feature selection algorithms, including machine learning-based feature selection methods (e.g., Extra-tree and SVM-RFE), traditional discrete optimization algorithms (e.g., greedy search and tabu search), and metaheuristic search algorithms (e.g., simulated annealing and ant colony optimization), are considered for comparison on three diagnostic tasks. From Table V, we can observe that our framework with GA-based feature selection achieves the best overall performance dominating almost each evaluation metric on three datasets, except for the highest AUC value on SCZ reached by ant colony optimization. It is interesting to note that metaheuristic search algorithms are also effective in feature selection achieving the second-best results. The main advantage of metaheuristic algorithms is that if used correctly, they can take advantage of feedback from the outcome to direct the search path. The two remaining categories, i.e., machine learning-based algorithms and traditional optimization algorithms, manifest a performance degradation while maintaining competitive performance compared with the state-of-the-art methods shown in Table II. Furthermore, without feature selection, the classification performance of our framework is restricted to the noisy, redundant features and thus shows

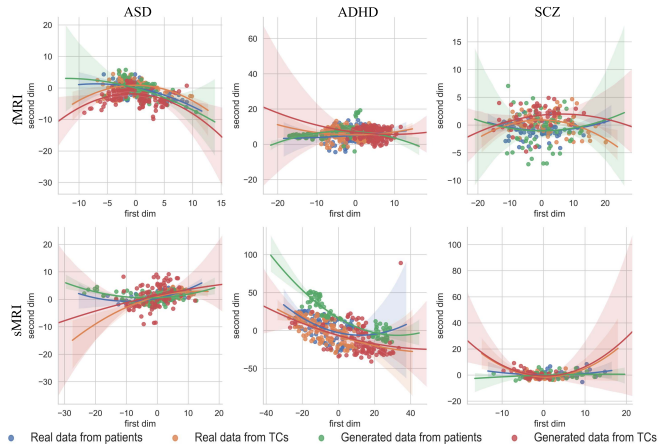


Fig. 7. 2-D PCA visualizations of the real and generated features in ABIDE, ADHD-200, and COBRE dataset.

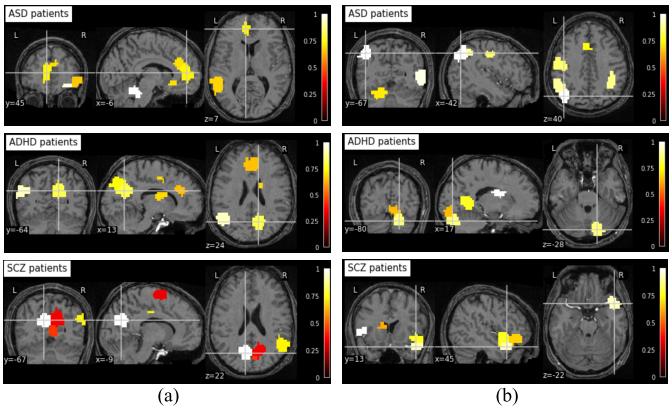


Fig. 8. Data-driven analysis of the top-10 remarkable ROIs for ASD group, ADHD group, and SCZ group. (a) Remarkable ROIs of patients in fMRI. (b) Remarkable ROIs of patients in sMRI.

a significant decrease. In comparison, our framework with GA-based feature selection can also significantly reduce the number of trainable parameters, elapsed time, and model size by 67.1%, 63.3%, and 67.1%, respectively (averaged across the three datasets). In a nutshell, the proposed GA-based feature selection method is targeted to provide possible solutions resistance to local optimums, and we demonstrated its effectiveness and robustness in our framework.

D. Visualization of the Generated Data

To further analyze the quality of generated data from MCGAN, we plot the distributions of the real and generated features by principal component analysis (PCA) in Fig. 7. For better visualization, PCA is commonly used to convert features from high-dimensional space into a 2-D plane. Polynomial regression is also used to fit sample points corresponding to the given category for auxiliary analysis. The boundaries of real data and generated data from both patient and TC groups are approximated by polynomial functions with curves in different colors. It is shown that the real data from patients and TCs are fairly close to each other. It is challenging to discriminate between the patient group and TC group from real data. According to the curves plotted by polynomial functions, the boundaries of both groups in the real data manifold are not distinct. To address this issue, the generated data by MCGAN supplement the training data manifold from two perspectives, i.e., vraisemblance and diversity. It is observed that the polynomial curves of the real data and the generated data have similar trends and the same convexity/concavity property. Besides, the sample distributions of the real data and the generated data are generally mixed and indistinguishable. These observations suggest that the generated data show similar distributions as the real data do, whatever the real distribution is sparse or dense. Furthermore, the polynomial curves of the generated data have a large confidence interval (the shadow region of polynomial curves in Fig. 7), covering the confidence interval of the real data in most cases. This implies that the generated data are helpful to enlarge the margin around the decision boundary in the hyperplane for classifiers. In other words, the generated data can contain not only realistic data information but also carry diverse information in the problem space. Therefore, the data augmentation of MCGAN can lead to better performance and robustness of the classifier.

E. Model Interpretation

Based on the guided back-propagation method [53], the remarkable neural patterns for different mental disorders can be identified for further interpretation. First, the gradient-based score is calculated to indicate the importance of each FC feature input based on the final probability output of testing sets. Then we calculate the weighted sum of each ROI by accumulating the scores from their associated FC features. Following the rule of 10-fold CV, the final scores of ROIs are averaged and normalized. The top-10 remarkable ROIs for ASD, ADHD, and SCZ are shown in Fig. 8. It is worth noting that 20%, 30%, and 30% of top-10 ROIs are presented in cortical areas. This observation is consistent with the theoretical evidence in previous studies [54] that the common disruption pattern across mental disorders in cortical areas could result in gray matter reduction, vulnerable to broad-spectrum psychopathology.

Moreover, we visualized top-10 FC features/connections with the highest scores between regions in Fig 9. Some of these results have been confirmed by the existing literature. For example, the volumes of the left superior temporal gyrus (4) are smaller in autism subjects as compared with healthy controls [55]. Both boys and girls with ADHD have smaller lobules of the cerebellar vermis (152) than healthy children [56]. By comparing variances between SCZ groups and controls, a significant difference in the amplitude of low-frequency fluctuations is observed in the left postcentral gyrus (8) [57]. We hope the identified dysfunction in these ROIs could be leveraged to decipher an intermediate transdiagnostic phenotype, advancing understanding and therapeutics for those mental disorders.

F. Limitations and Future Work

While our proposed framework performs well in mental disorders diagnosis, several limitations should be carefully addressed to improve its performance and practical values.

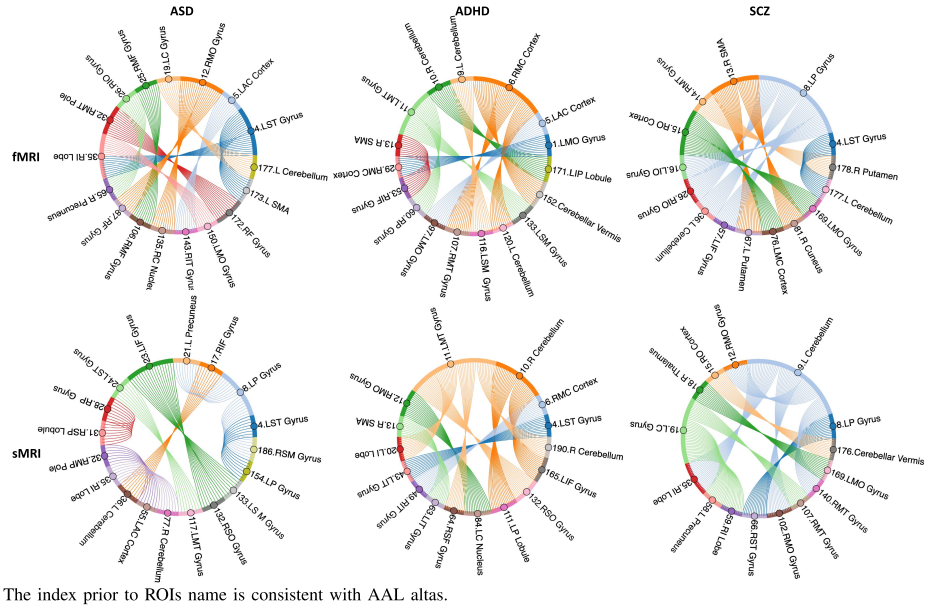


Fig. 9. Data-driven analysis of the top-10 FC connections between ROIs for ASD group, ADHD group, and SCZ group.

First, as mentioned in Section IV-A, only the CC200 atlas is empirically adopted to partition the whole brain into 200 ROIs for extracting the mean time-series data. However, such ROI-based connectivity analysis can be largely sensitive to the empirical selection and prior knowledge, which poses a challenge to decide a reasonable set of brain regions/voxels. Since different ROI atlases can provide complementary information at different resolutions for detecting brain abnormalities, multiatlas combination strategy could be adopted to address this issue and further improve the generalizability of our model [58]. Besides, not every subject can offer multimodal data due to various practical and technical reasons. Disposing of missing modalities is less tractable in this work. To this end, how to synthesize missing modalities with feature consistency constraints or utilize semisupervised learning on scarce samples deserves more research effort to be studied in depth [11], [59].

V. Conclusion

In this work, an effective AMFDA framework has been presented for the identification of mental disorders from multimodal MRI data. MCGAN has been proposed to simultaneously synthesize high-quality multimodal data. An attention-like MGF model has also been proposed to extract complementary information from multimodal data for classification via the learned flexible feature fusion patterns. The proposed model effectively leverages the attention mechanism to extract different levels of intra- and inter-modality interactions in multimodality fusion. Through a series of simulation experiments in three mental disorders, the effectiveness and reliability of AMFDA have been demonstrated. This article is expected to provide insights into developing effective CAD methods based on multimodal fusion techniques.

References

[1] C. Kieling et al., "Child and adolescent mental health worldwide: Evidence for action," *Lancet*, vol. 378, no. 9801, pp. 1515–1525, 2011.

[2] T. Chung, J. Cornelius, D. Clark, and C. Martin, "Greater prevalence of proposed ICD-11 alcohol and cannabis dependence compared to ICD-10, DSM-IV, and DSM-5 in treated adolescents," *Alcoholism, Clin. Experim. Res.*, vol. 41, no. 9, pp. 1584–1592, Sep. 2017.

[3] J. Ji, A. Zou, J. Liu, C. Yang, X. Zhang, and Y. Song, "A survey on brain effective connectivity network learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 1, 2021, doi: 10.1109/TNNLS.2021.3106299.

[4] Z.-A. Huang et al., "Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans," *IEEE Trans. Biomed. Eng.*, early access, Sep. 30, 2022, doi: 10.1109/TBME.2022.3210940.

[5] L. Velayudhan et al., "Entorhinal cortex thickness predicts cognitive decline in Alzheimer's disease," *J. Alzheimer's Disease*, vol. 33, no. 3, pp. 755–766, 2013.

[6] J. H. Callicott et al., "Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 24, pp. 8627–8632, 2005.

[7] C. Lian, M. Liu, L. Wang, and D. Shen, "Multi-task weakly-supervised attention network for dementia status estimation with structural MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 4056–4068, Aug. 2022, doi: 10.1109/TNNLS.2021.3055772.

[8] D. Wang, T. Zhao, W. Yu, N. V. Chawla, and M. Jiang, "Deep multimodal complementarity learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 18, 2022, doi: 10.1109/TNNLS.2022.3165180.

[9] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, "MATNet: Motion-attentive transition network for zero-shot video object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 8326–8338, 2020.

[10] L. Jin, Z. Li, and J. Tang, "Deep semantic multimodal hashing network for scalable image-text and video-text retrievals," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 5, 2020, doi: 10.1109/TNNLS.2020.2997020.

[11] Y. Pan, M. Liu, Y. Xia, and D. Shen, "Disease-image-specific learning for diagnosis-oriented NeuroImage synthesis with incomplete multi-modality data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6839–6853, Oct. 2022, doi: 10.1109/TPAMI.2021.3091214.

[12] P. Zhuang, A. G. Schwing, and O. Koyejo, "FMRI data augmentation via synthesis," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1783–1787.

[13] S. Liu, H. Zheng, Y. Feng, and W. Li, "Prostate cancer diagnosis using deep learning with 3D multiparametric MRI," *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 1013428.

[14] R. Hao, K. Namdar, L. Liu, M. A. Haider, and F. Khalvati, "A comprehensive study of data augmentation strategies for prostate cancer detection in diffusion-weighted MRI using convolutional neural networks," *J. Digit. Imag.*, vol. 34, no. 4, pp. 862–876, Aug. 2021.

[15] N. C. Dvornek, P. Ventola, and J. S. Duncan, "Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 725–728.

[16] Z. Mao et al., "Spatio–temporal deep learning method for ADHD fMRI classification," *Inf. Sci.*, vol. 499, pp. 1–11, Oct. 2019.

[17] K. P. Nguyen, C. C. Fatt, A. Treacher, C. Mellema, M. H. Trivedi, and A. Montillo, "Anatomically-informed data augmentation for functional MRI with applications to deep learning," *Proc. SPIE*, vol. 11313, Mar. 2020, Art. no. 113130T.

[18] Q. Yao and H. Lu, "Brain functional connectivity augmentation method for mental disease classification with generative adversarial network," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*. Cham, Switzerland: Springer, 2019, pp. 444–455.

[19] L. Bi, J. Kim, A. Kumar, D. Feng, and M. Fulham, "Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs)," in *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*. Cham, Switzerland: Springer, 2017, pp. 43–51.

[20] Y. Wang et al., "3D auto-context-based locality adaptive multi-modality GANs for PET synthesis," *IEEE Trans. Med. Imag.*, vol. 38, no. 6, pp. 1328–1339, Jun. 2019.

[21] S. Guo et al., "Combining multi-modality data for searching biomarkers in schizophrenia," *PLoS ONE*, vol. 13, no. 2, 2018, Art. no. e0191202.

[22] M. A. Aghdam, A. Sharifi, and M. M. Pedram, "Combination of rs-fMRI and sMRI data to discriminate autism spectrum disorders in young children using deep belief network," *J. Digit. Imag.*, vol. 31, no. 6, pp. 895–903, 2018.

[23] O. B. Ahmed, J. Benois-Pineau, M. Allard, C. B. Amar, and G. Catheline, "Classification of Alzheimer's disease subjects from MRI using hippocampal visual features," *Multimedia Tools Appl.*, vol. 74, no. 4, pp. 1249–1266, Feb. 2015.

[24] Y. Shi, H.-I. Suk, Y. Gao, S.-W. Lee, and D. Shen, "Leveraging coupled interaction for multimodal Alzheimer's disease diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 186–200, Jan. 2020.

[25] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *NeuroImage*, vol. 59, pp. 895–907, Jan. 2012.

[26] S. Abdolmaleki and M. S. Abadeh, "Brain MR image classification for ADHD diagnosis using deep neural networks," in *Proc. Int. Conf. Mach. Vis. Image Process. (MVIP)*, Feb. 2020, pp. 1–5.

[27] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[28] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[29] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 299–307.

[30] X. Zhao et al., "Investigating the correspondence of clinical diagnostic grouping with underlying neurobiological and phenotypic clusters using unsupervised machine learning," *Frontiers Appl. Math. Statist.*, vol. 4, p. 25, Jan. 2018.

[31] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, Jun. 2013.

[32] F. Sadjadi, "Comparison of fitness scaling functions in genetic algorithms with applications to optical processing," *Proc. SPIE*, vol. 5557, pp. 356–364, Oct. 2004.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 214–223.

[35] M. C. Mukkamala and M. Hein, "Variants of RMSProp and Adagrad with logarithmic regret bounds," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2545–2553.

[36] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1930–1939.

[37] S. Brieber et al., "Structural brain abnormalities in adolescents with autism spectrum disorder and patients with attention deficit/hyperactivity disorder," *J. Child Psychol. Psychiatry*, vol. 48, no. 12, pp. 1251–1258, 2007.

[38] R. E. Gur et al., "Clinical subtypes of schizophrenia: Differences in brain and CSF volume," *Amer. J. Psychiatry*, vol. 151, no. 3, pp. 343–350, 1994.

[39] N. C. Dvornek, P. Ventola, K. A. Pelphrey, and J. S. Duncan, "Identifying autism from resting-state fMRI using long short-term memory networks," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 362–370.

[40] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage, Clin.*, vol. 17, pp. 16–23, Jan. 2018.

[41] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, "Identifying autism spectrum disorder from resting-state fMRI using deep belief network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2847–2861, Jul. 2021.

[42] A. A. Pulini, W. T. Kerr, S. K. Loo, and A. Lenartowicz, "Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis," *Biol. Psychiatry, Cognit. Neurosci. Neuroimag.*, vol. 4, no. 2, pp. 108–120, Feb. 2019.

[43] L. Zou, J. Zheng, C. Miao, M. J. Mckeown, and Z. J. Wang, "3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI," *IEEE Access*, vol. 5, pp. 23626–23636, 2017.

[44] H. Zou and J. Yang, "Multiple functional connectivity networks fusion for schizophrenia diagnosis," *Med. Biol. Eng. Comput.*, vol. 58, no. 8, pp. 1779–1790, Aug. 2020.

[45] A. Savio and M. Graña, "Local activity features for computer aided diagnosis of schizophrenia on resting-state fMRI," *Neurocomputing*, vol. 164, pp. 154–161, Sep. 2015.

[46] M. Latha and G. Kavitha, "Detection of schizophrenia in brain MR images based on segmented ventricle region and deep belief networks," *Neural Comput. Appl.*, vol. 31, no. 9, pp. 5195–5206, Sep. 2019.

[47] J. Kim, V. D. Calhoun, E. Shim, and J.-H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia," *NeuroImage*, vol. 124, pp. 127–146, Jan. 2016.

[48] T.-H. Hsieh, M.-J. Sun, and S.-F. Liang, "Diagnosis of schizophrenia patients based on brain network complexity analysis of resting-state fMRI," in *Proc. 15th Int. Conf. Biomed. Eng.* Cham, Switzerland: Springer, 2014, pp. 203–206.

[49] M. Plitt, K. A. Barnes, and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *NeuroImage, Clin.*, vol. 7, pp. 359–366, Jan. 2015.

[50] C. Dou, S. Zhang, H. Wang, L. Sun, Y. Huang, and W. Yue, "ADHD fMRI short-time analysis method for edge computing based on multi-instance learning," *J. Syst. Archit.*, vol. 111, Dec. 2020, Art. no. 101834.

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

[52] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 18, 2022, doi: 10.1109/TPAMI.2022.3152247.

[53] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[54] L. M. McTeague, J. Huemer, D. M. Carreon, Y. Jiang, S. B. Eickhoff, and A. Etkin, "Identification of common neural circuit disruptions in cognitive control across psychiatric disorders," *Amer. J. Psychiatry*, vol. 174, no. 7, pp. 676–685, Jul. 2017.

[55] G. D. Waiter, J. H. G. Williams, A. D. Murray, A. Gilchrist, D. I. Perrett, and A. Whiten, "Structural white matter deficits in high-functioning individuals with autistic spectrum disorder: A voxel-based investigation," *NeuroImage*, vol. 24, no. 2, pp. 455–461, Jan. 2005.

[56] C. M. Anderson, A. Polcari, S. B. Lowen, P. F. Renshaw, and M. H. Teicher, "Effects of methylphenidate on functional magnetic resonance relaxometry of the cerebellar vermis in boys with ADHD," *Amer. J. Psychiatry*, vol. 159, no. 8, pp. 1322–1328, Aug. 2002.

[57] S. Lui et al., "Resting-state brain function in schizophrenia and psychotic bipolar probands and their first-degree relatives," *Psychol. Med.*, vol. 45, no. 1, p. 97, 2015.
[58] P. Dong, Y. Guo, Y. Gao, P. Liang, Y. Shi, and G. Wu, "Multi-atlas segmentation of anatomical brain structures using hierarchical hypergraph learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 3061–3072, Aug. 2020.
[59] Y. Hu, Z.-A. Huang, R. Liu, X. Xue, L. Song, and K. C. Tan, "A dual-stage pseudo-labeling method for the diagnosis of mental disorder on MRI scans," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2022, pp. 1–8.

Rui Liu received the B.S. degree in intelligence science and technology from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong.

His current research interests include machine learning, multitask/modality learning, and medical images diagnosis and applied deep learning.

Zhi-An Huang received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2021.

He is currently a Research Fellow with the City University of Hong Kong Dongguan Research Institute (CityU DGRI), Hong Kong. His research interests include artificial intelligence, machine learning, bioinformatics, and medical imaging analysis.

Yao Hu received the B.S. degree in mining engineering and the M.Sc. degree in control science and engineering from the China University of Mining and Technology, Xuzhou, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong.

His current research interests include machine learning, transfer learning, and federated learning.

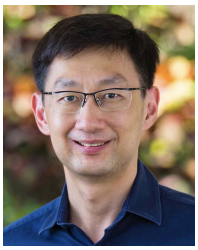
Zexuan Zhu (Senior Member, IEEE) received the B.S. degree in computer science and technology from Fudan University, Shanghai, China, in 2003, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2008.

He is currently a Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include computational intelligence, machine learning, and bioinformatics.

Dr. Zhu is an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE. He is also the Chair of the IEEE CIS Emergent Technologies Task Force on Memetic Computing.

Ka-Chun Wong received the B.Eng. degree in computer engineering and the M.Phil. degree from the Chinese University of Hong Kong, Hong Kong, in 2008 and 2010, respectively, and the Ph.D. degree from the Department of Computer Science, University of Toronto, Toronto, ON, Canada, in 2015.

He was an Associate Professor with the City University of Hong Kong, Hong Kong. His current research interests include bioinformatics, computational biology, evolutionary computation, data mining, machine learning, and interdisciplinary research.



Kay Chen Tan (Fellow, IEEE) received the B.Eng. (Hons.) and the Ph.D. degrees from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively.

He is currently a Chair Professor with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong.

Dr. Tan was the Editor-in-Chief of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. He currently serves on the Editorial Board member of 10+ journals. He is currently the Vice-President (Publications) of the IEEE Computational Intelligence Society, an IEEE Distinguished Lecturer Program (DLP) Speaker, an Honorary Professor with the University of Nottingham, Nottingham, U.K., and the Chief Co-Editor of Springer Book Series on *Machine Learning: Foundations, Methodologies, and Applications*.