A Multimodal Approach for Identifying Autism Spectrum Disorders in Children

Junxia Han, Guoqian Jiang, *Member, IEEE*, Gaoxiang Ouyang, *Member, IEEE*, and Xiaoli Li

***Abstract*— Identification of autism spectrum disorder (ASD) in children is challenging due to the complexity and heterogeneity of ASD. Currently, most existing methods mainly rely on a single modality with limited information and often cannot achieve satisfactory performance. To address this issue, this paper investigates from internal neurophysiological and external behavior perspectives simultaneously and proposes a new multimodal diagnosis framework for identifying ASD in children with fusion of electroencephalogram (EEG) and eye-tracking (ET) data. Specifically, we designed a two-step multimodal feature learning and fusion model based on a typical deep learning algorithm, stacked denoising autoencoder (SDAE). In the first step, two SDAE models are designed for feature learning for EEG and ET modality, respectively. Then, a third SDAE model in the second step is designed to perform multimodal fusion with learned EEG and ET features in a concatenated way. Our designed multimodal identification model can automatically capture correlations and complementarity from behavior modality and neurophysiological modality in a latent feature space, and generate informative feature representations with better discriminability and generalization for enhanced identification performance. We collected a multimodal dataset containing 40 ASD children and 50 typically developing (TD) children to evaluate our proposed method. Experimental results showed that our proposed method achieved superior performance compared with two unimodal methods and a simple feature-level fusion method, which has promising potential to provide an objective and accurate diagnosis to assist clinicians.**

***Index Terms*— Autism spectrum disorders (ASD), multimodal fusion, electroencephalogram (EEG), eye-tracking (ET), stacked denoising autoencoders, classification.**

Manuscript received 16 September 2021; revised 5 May 2022 and 6 June 2022; accepted 15 July 2022. Date of publication 19 July 2022; date of current version 22 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62003228 and Grant 61761166003, in part by the Science and Technology Development Project of Beijing Municipal Education Commission of China under Grant KM202010028019, and in part by the National Key Research and Development Program of China under Grant 2017YFC0820205. *(Junxia Han and Guoqian Jiang contributed equally to this work.) (Corresponding author: Xiaoli Li.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the School of Psychology Research Ethics Committee of Beijing Normal University.

Junxia Han is with the Beijing Key Laboratory of Learning and Cognition, School of Psychology, Capital Normal University, Beijing 100048, China (e-mail: hanjunxia@cnu.edu.cn).

Guoqian Jiang is with the School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China (e-mail: jiangguoqian@ysu.edu.cn).

Gaoxiang Ouyang and Xiaoli Li are with the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China (e-mail: xiaoli@bnu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3192431

I. Introduction

AUTISM spectrum disorder (ASD) is a complex neurodevelopmental disorder characterized by social and communication impairments and restricted and stereotyped behaviors [1]. ASD typically appears during early childhood and affects the child's cognitive ability, social emotion, sensory and motor functioning, and social interaction. It is becoming more widespread. According to the estimates from the Centers for Disease Control and Prevention (CDC) in United States, about 1 in 54 children has been identified with ASD [2]. Consequently, the diagnosis and treatment of ASD have become a worldwide public health concern and attracted considerable attention. However, the cause of ASD still remains unclear. Currently, clinical diagnosis of ASD mainly relies on behavior diagnosis and scales assessmen [3]. However, there is limited understanding of the neural patterns behind ASD and the severity of the disease. Also, there is a lack of experienced experts for ASD diagnosis [4]. Therefore, there is an increasing need to develop objective and effective tools to identify and diagnose ASD in children and assist clinicians to make accurate diagnosis result.

In recent years, different neuroimaging techniques, including functional magnetic resonance imaging (fMRI) [5], [6], magnetoencephalography (MEG) [7], and electroencephalogram (EEG) [8], have been used to explore the characteristics of brain structure and function associated with ASD. Among these neuroimaging techniques, EEG is a relatively easy-to-use, low-cost brain measurement tool that has been widely used for monitoring atypical brain development. Previous studies have shown that patients with ASD have abnormalities in neural oscillations and brain functional connections at different developmental stages [8]–[10], and accordingly, various EEG-based features or indicators were extracted from different aspects such as neural oscillation rhythm, functional connectivity, and nonlinear information dynamics to quantitatively describe the differences between ASD children and TD children. To further facilitate an automated diagnosis, machine learning techniques have been used to develop diagnosis models with extracted EEG features. For example, Wadhera *et al.* developed a support vector machine (SVM) classification model with a combination of two features, average weighted degree and mutual information, and obtained a detection accuracy of 92.34% [11]. Mehmet Baygin *et al.* proposed a new hybrid deep lightweight feature extractor to extract deep features using a combination of pre-trained models and achieved 96.44% accuracy with an SVM classifier [12].

Moreover, there are not only brain abnormalities in individuals with ASD, but also atypical eye gaze patterns such as eye contact avoidance and altered joint attention in social activities. Eye-tracking (ET) technology can provide a direct measure of gaze allocation and goal-directed looking behaviors and has been primarily used to study the attention allocation of ASD population [13], [14]. Nakano and Hosozawa *et al.* demonstrated that children with ASD spent less time looking at faces and social interactions than TD children [15]. Liu *et al.* presented a machine learning framework to deal with an ET dataset in a face recognition task to classify ASD children and TD children, and the maximum classification accuracy reached 88.51% [16]. Wan *et al.* employed a linear discriminant analysis using the fixation time of children watching a 10-second video of a female speaking and the classification accuracy can be achieved 85.1% [17].

In summary, EEG and ET have been independently applied in ASD studies to identify effective biomarkers and then to design diagnosis models with advanced machine learning algorithms. Notably, these existing studies mainly focus on single modality data analysis. However, ASD is a complex and heterogeneous disease with abnormal manifestations from the cellular level to the behavioral level, and therefore, it is difficult to accurately and effectively identify ASD solely relying on unimodal data, such as EEG or ET. EEG and ET data are completely different modalities and can be viewed from internal neurophysiological and external behavioral perspectives, respectively. These two modalities contain rich and complementary information associated with ASD [18], [19]. Due to the data heterogeneity of neurophysiological and behavioral modalities, it is still challenging to explore hidden correlations and complementarity directly from the original data. To address this challenge, multimodal fusion is a great option. In recent years, multimodal fusion has attracted considerable attention, especially in the medical context, which has been applied in the diagnosis of ASD [20], [21] as well as other diseases, such as Parkinson [22], Alzheimer [23] and Depression [24]. In a recent study, Cociu *et al.* integrated three different neuroimaging techniques, EEG, fMRI, and diffusion tensor imaging (DTI) to characterize an autistic brain and provided a better understanding of the neurobiological basis of ASD [20]. Mash *et al.* concentrated on multimodal analysis to explore the relation in ASD between fMRI and EEG measures of spontaneous brain activity [21]. In Vasquez-Correa *et al.* [22], a deep learning-based multimodal diagnosis model was proposed to classify patients with Parkinson's disease in different stages of the disease by integrating different information from speech, handwriting, and gait signals. In Shi *et al.* [23], multimodal neuroimaging data, magnetic resonance imaging (MRI) and positron emission tomography (PET), were fused to perform the diagnosis of Alzheimer's disease and achieved superior performance in binary and multi-class classification tasks. These studies have proven that multimodal information fusion can take full advantage of the strengths of individual modality data and overcome their respective weakness, yielding an enhanced performance.

Motivated by previous studies, this study aims to develop a reliable and accurate diagnosis model with fusion of EEG and ET data and proposes a new multimodal diagnostic framework to identify ASD in children. Specifically, we attempt to design a deep learning-based fusion model to capture correlations and complementarity from EEG and ET data. To evaluate the performance of our proposed framework, we collect a multimodal dataset containing the resting-state EEG data and task-state ET data from 40 children with ASD and 50 TD children. Experimental results demonstrate the superior performance of our proposed method.

The rest of this paper is organized as follows. Section II details our proposed multimodal framework for identification of children with ASD. Section III presents detailed experimental and performance evaluation results. Lastly, Section IV concludes this paper.

II. Methodology

The pipeline of the proposed multimodal identification framework for children with ASD is shown in Fig. 1, where EEG and ET data are used as two individual input modalities. It aims to integrate the complementary information in both modalities to enhance identification performance. It mainly consists of three sequential steps: data acquisition, feature extraction, and multimodal identification. Firstly, resting-state EEG data and ET data were acquired and preprocessed to remove unrelated noise signals, respectively. Then, in the feature extraction stage, we extracted typical features from each modality respectively as the initial EEG feature set and ET feature set, which contains rich but redundant diagnosis information related to ASD. In the multimodal identification stage, a two-step multimodal fusion network based on stacked denoising autoencoders (SDAE) is designed to learn useful EEG and ET representation from two initial feature sets and further fuse learned multimodal information for final classification between ASD and TD. Our multimodal fusion network can capture the complementary characteristics of neurophysiological (EEG) and behavioral (ET) modalities and enhance identification performance, which is evaluated through a comparative study with unimodal data in Section III. The details of each part will be described in the following subsections.

A. Data Acquisition and Preprocessing

1) Subjects: In our study, a total of 90 subjects, including 40 ASD children and 50 typically developing (TD) children aged 3-6 years, were enrolled. The detailed demographics of all subjects are listed in Table I. All ASD children were recruited and received diagnostic confirmation based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) [1]. However, we had limited access to reliable information in the school site. Thus, the children were assessed by using the parent report on the Autism Behavior Checklist (ABC), Social Communication Questionnaire (SCQ), Social Responsiveness Scale (SRS), and Clancy Behavior Scale (CABS). Details of the sample demographics and behavior scores are shown in Table I. All TD children

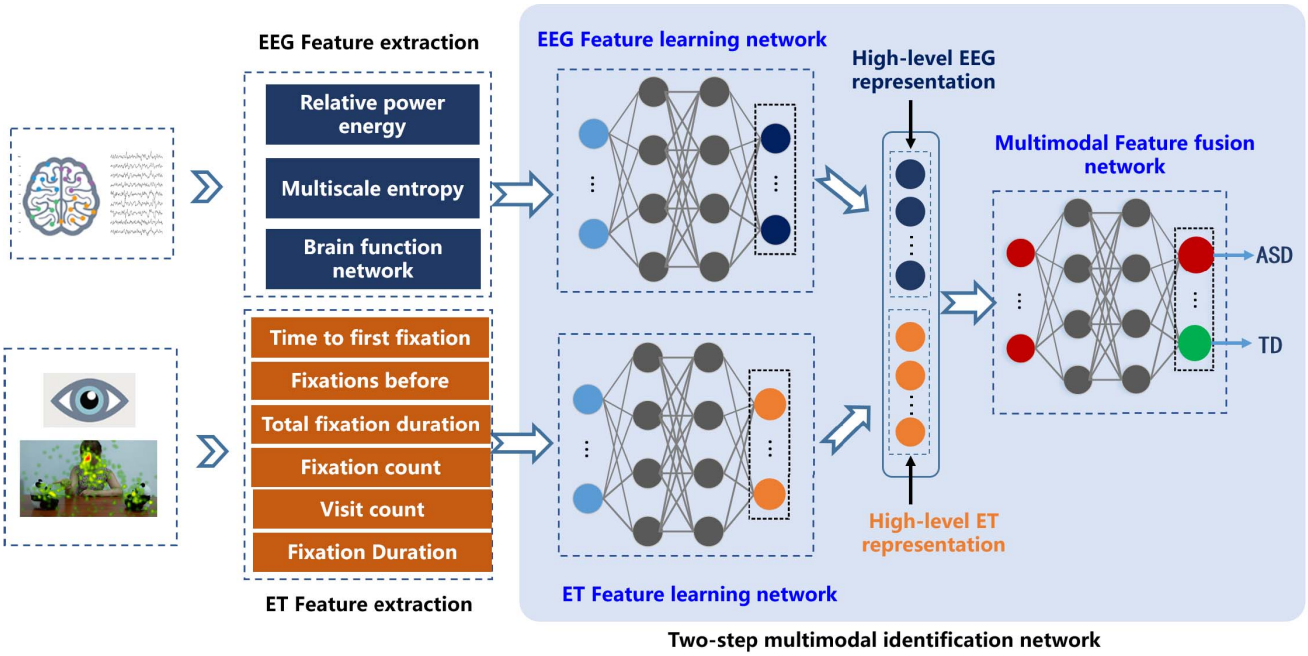

Fig. 1. Proposed multimodal identification framework of children with ASD by fusion of multimodal EEG and ET data, which are from neurophysiological and behavioral view, respectively.

TABLE I
DEMOGRAPHIC INFORMATION OF ALL SUBJECTS, WHERE *p* VALUES WERE OBTAINED BY A TWO-INDEPENDENT SAMPLE *t* TEST. (STD: STANDARD DEVIATION; SCQ: SOCIAL COMMUNICATION QUESTIONNAIRE; SRS: SOCIAL RESPONSIVENESS SCALE; ABC: AUTISM BEHAVIOR CHECKLIST; CABS: CLANCY BEHAVIOR SCALE)

Characteristic	ASD (n=40)	TD (n=50)	p value
Gender (Male/Female)	33/7	32/18	NA
Age (mean±std)	4.13±1.03	3.96±0.90	$p = 0.735$
ABC score	46.23±18.12	6.36±3.86	$p < 0.001$
CABS score	13.57±15.07	NA	NA
SRS score	90.57±26.77	40.22±12.32	$p < 0.001$
SCQ score	19.93±4.76	4.64±2.91	$p < 0.001$

were recruited from a local kindergarten. We also employed these ABC, SRS, and SCQ reported by their teachers to examine if there were any autistic symptoms in the TD group. No TD children reached the cut-off score of ABC, SRS and SCQ. The present study has been approved by the School of Psychology Research Ethics Committee of Beijing Normal University and informed consent was obtained from all children with the permission of their parents before subject enrollment.

2) EEG Data: In our study, continuous open-eye resting-state EEG signals were recorded with a high-density array of 128 Ag/AgCl passive electrodes (Electrical Geodesics Inc., EGI) with a sampling rate of 1000 Hz for at least five minutes, as shown in Fig. 2 (a). All participated children were instructed to be seated comfortably on an armchair usually accompanied by their caregivers in a quiet room. Before the EEG recording, scalp impedance was checked online by employing Net Station (EGI, Inc.) and was reduced below 50 kilo-ohm. The EEG data were referenced online to Cz.

EEG data preprocessing was done using EEGLAB [25] and MATLAB. The EEG signal was first downsampled to 250 Hz. A notch filter centered at 50 Hz was employed to remove the line noise, and the data were then band-pass filtered (0.5-45 Hz). EEG Data were divided into 4 s segments with no overlap. An artifact detection algorithm proposed in [26] was utilized to select the segments without artifact involvement, including eye movements, eye-blinks, power supply, breathing, muscle movements, abrupt slopes, and outlier values. After that, a visual inspection was performed to reject those segments containing noise. During individual recording segments or throughout the entire recording, sensors were marked as bad channels by using a 200 μV threshold, which were interpolated from neighboring channels as described in our previous study [8]. Finally, 4.5 ± 2.2 (mean ± std) bad channels were identified and processed, leaving 25.11 ± 5.95 segments for further analysis (ASD: 20.4 ± 4.9 versus TD: 27.2±5.3). In this study, we selected 62 electrodes of interest from the 128-channel GSN to ensure maximal spatial coverage of the frontal, central, temporal, and occipital regions. The whole brain was divided into ten regions as shown in Fig. 2 (b).

3) ET Data: ET experiments were carried out after that each subject finished EEG data collection and had a break. Fig. 3 shows the ET data acquisition experimental setup. The Tobii TX300 eye tracker was used to record the gaze behavior of each subject with a sampling frequency of 300 Hz. The screen resolution was set to 1024 pixels × 768 pixels. Before the formal experiment, a five-point calibration program was performed and the experiment proceeded after all 5 points were captured with small error vectors. All subjects were

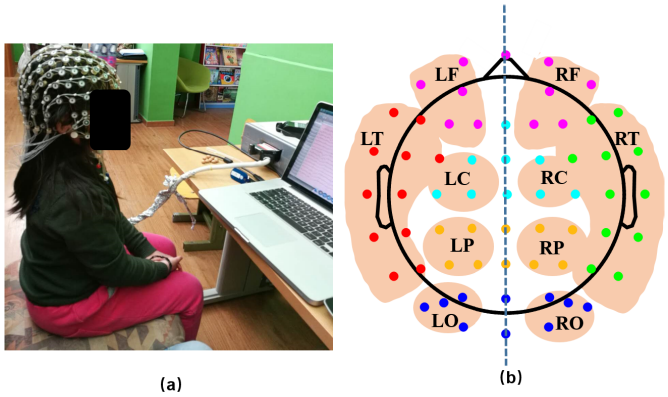

Fig. 2. EEG experimental setup. (a) Resting-state EEG data acquisition; (b) Electrodes of interest with different colors representing different brain regions, where LF: left frontal, RF: right frontal, LC: left central, RC: right central, LT: left temporal, RT: right temporal, LO: left occipital, RO: right occipital, LP: left parietal, and RP: right parietal.

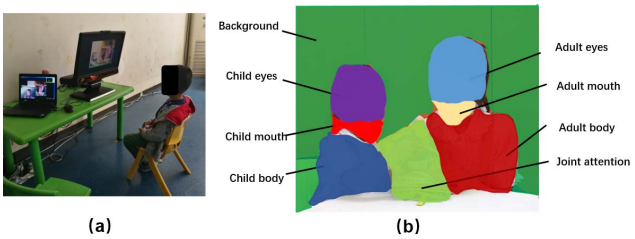

Fig. 3. Eye-tracking experiment setup: (a) eye-tracking data acquisition and (b) a visual stimuli video clip and its ROIs.

required to be seated in front of a monitor with an eye-to-monitor distance of about 60 cm, and they were instructed to watch the dynamic visual stimuli. The video stimuli material was selected from the *Tiger Qiao Hu*[1] containing social interaction between a child and an adult as shown in Fig. 3 (b). In each experiment, the video stimuli clips were presented and each clip lasted about 30 seconds. During the interval between trials, a dynamic kitten with sound was presented in the middle of the screen to attract the attention of children. A total of 2 trials were displayed in a random order for each child. During the whole experiment, no response was required from the children. To explore the child's engagement with each AOI, we processed the eye-tracking data according to the Tobii fixation filter. Linear interpolation with a maximum gap length of 100 ms was used to fill the missing gaze data. Eye-tracking sample data were calculated with averaged gaze positions of the left and right eyes. All selected subjects had more than 70% screen-looking time captured by the eye-tracking equipment.

B. Feature Extraction

1) EEG Features: In this study, multi-domain features from different analytic perspectives are extracted to highlight the characteristics of EEG signals associated with ASD.

1) *Relative power energy features:* Previous studies showed atypical activity in multiple EEG oscillatory measures in ASD [27], [28]. Resting-state EEG studies of ASD showed reduced alpha power in individuals with ASD and increased power in low-frequency bands (delta band and theta band) [29]. Therefore, power spectral density (PSD) analysis was used to calculate spectral features. The spectral power was computed by employing a Hanning window on each 4-second segment using the fast Fourier transform (FFT). Relative power energy is defined as the ratio of the power within each frequency band to the total power over the whole power spectrum. For ten brain regions, we first calculated the relative power energy of each channel, and then the mean relative power energy was calculated in five frequency bands: delta (1-4Hz), theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz), and gamma (30-45Hz), thus yielding $5 \times 10 = 50$ spectral features in total.
2) *Multiscale entropy features*: Many recent reports have described exploration of abnormal brain signal complexity in ASD with multiscale entropy (MSE) [10], [30]. MSE is a method to describe the complexity of signals on multiple time scales by introducing a multiscale coarse-grain process. For a given time series $X = [x_1, x_2, x_3, \cdots, x_N]$, where N is the length of the time series, it was first coarse-grained using the scale factor s, with non-overlapping windows as follows:

$$y_j^{(s)} = \frac{1}{s} \sum_{i=(j-1)s+1}^{js} x_i, \quad 1 \le j \le \frac{N}{s} \tag{1}$$

Then, the sample entropy for each coarse-grained time series was calculated to describe the complexity of the original time series at different time scales s. In this study, the scale factor s was set to 20 suggested by previous studies [10]. With the MSE method, we calculated the averaged signal complexity on four scale ranges (scales 1-5, scales 6-10, scales 11-15, and scales 16-20) for ten brain regions shown in Fig. 2 (b). Finally, a total of $10 \times 4 = 40$ complexity features were obtained.
3) *Brain network features*: Previous studies have shown that the brain network in children with ASD was disrupted [8], [31]. In this study, a complex network with 62 nodes was constructed based on graph theory analysis to explore the brain network features of ASD children and TD children. Functional connectivity quantifies the relationship of EEG oscillatory activities between two nodes. Specifically, we calculated phase lag index (PLI) between two nodes as the edge weights of the connectivity matrix of the brain network. To describe the differences in the brain function network of ASD and TD children, we computed the following seven network metrics [8] of the whole brain, including global efficiency, clustering coefficient, path length, normalized clustering coefficient, normalized path length, small-worldness, and transitivity. Note that these features are global not local. In summary, a total of $7 \times 5 = 35$ network features over five frequency bands were extracted.

[1]https://kids.qiaohu.com/

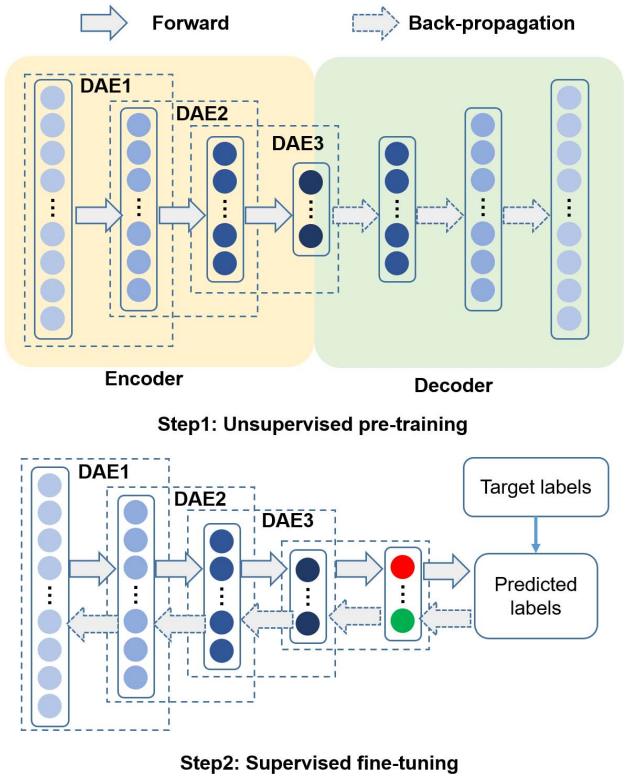

Fig. 4. Illustration of an SDAE model with three DAEs, which consists of two steps: unsupervised pre-training and supervised fine-tuning.

2) ET Features: For ET data analysis, we consider eight areas of interest (AOIs) shown in Fig. 3 (b): (1) background, (2) adult body, (3) child body, (4) adult eyes, (5) child eyes, (6) adult mouth, (7) child mouth, and (8) joint attention. For each AOI, we extracted six statistical indicators, including time to first fixation, fixations before, total fixation duration, fixation count, fixation duration, and visit count. We analyzed two different dynamic video clips in the experiment. To sum up, we finally obtained a total of $6 \times 8 \times 2 = 96$ ET features for each subject.

After performing the above initial feature extraction for EEG and ET data, respectively, a multimodal dataset containing 125 EEG features and 96 ET features is generated, which will be used for multimodal feature learning and fusion for ASD identification. Considering that all EEG and ET features have different scales which will impact the subsequent model training performance, each feature is then linearly normalized in the range of [0 1] using Eq. (2).

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

where x and x_{norm} are the original feature value and the normalized value, respectively and x_{min} and x_{max} are the minimum and maximum value of the original features.

It should be noted that feature normalization is first performed on the training set and accordingly, and the testing set is then rescaled according to the maximum and the minimum value of the training set, thus ensuring that both data sets are in the similar range.

C. Multimodal Identification Model

EEG and ET data are collected from neurophysiological and behavioral perspectives, respectively, and they belong to two different modalities and contain rich and complementary information associated with ASD. Due to the data heterogeneity of two modalities, it is still difficult to explore hidden correlations and complementarity directly from the original data. To effectively fuse both modalities for improved performance, we designed a two-step multimodal feature learning and fusion model, multimodal stacked denoising autoencoder (MMSDAE), as shown in Fig. 1. It consists of two core modules: *unimodal feature learning module* and *multimodal feature fusion module*. The first one is used to learn the high-level and compact features in a latent space through multiple nonlinear transformations with a designed SDAE multilayer network structure from high-dimensional and correlated EEG features and ET features with information redundancy. The second one aims to fuse complementary information between the high-level EEG-based and ET-based representations learned from the previous stage.

1) SDAE: SDAE is a typical deep neural network architecture, which consists of multiple denoising autoencoders (DAEs) in a stacked way [32]. Fig. 4 illustrates a typical SDAE model with three DAEs. Specifically, a DAE consists of an encoder network and a decoder network. Note that DAE has the same number of neurons in the input layer and the output layer and reproduces its inputs at its output layer. In other words, it attempts to reconstruct itself only with input data while without extra label information. DAE aims to recover a data sample x from its corrupted version $\tilde{x}$ with a typical zero masking strategy [32]. In doing so, it can prevent the autoencoder from simply learning the identity mapping and help obtain robust representations from noisy data.

For the encoder network of DAE, it aims to transform the original corrupted version $\tilde{x}$ by a nonlinear mapping function f into a hidden representation h as (3)

$$h = f(\mathbf{W}_1\tilde{x} + b) \tag{3}$$

where $\mathbf{W}_1$ is the weight matrix and b is the bias vector. In this study, we use the sigmoid function $f(x) = 1/(1 + \exp(-x))$ [32] for the nonlinear mapping purpose. The learned latent representation h can be viewed as a compression of input data with some loss when the number of hidden units is less than the number of input units. It can capture the main variations in the high-dimensional input data and eliminate those less important information through dimension reduction, as demonstrated in Section III-B.

A decoder network then maps the hidden representation h back to a reconstruction output $\hat{x}$ as

$$\hat{x} = g(\mathbf{W}_2 h + c) \tag{4}$$

where $\mathbf{W}_2$ is the weight matrix, c is the bias vector, and g is the activation function. Likely, sigmoid function is chosen here. The training process of DAE is to find optimal parameters $\theta = \{\mathbf{W}_1, \mathbf{W}_2, b, c\}$ by minimizing the mean square error between the original input and the reconstructed output, which

is performed in an unsupervised manner only with input data while without any label information.

As shown in Fig. 4, the training process of an SDAE model consists of two steps: an unsupervised pre-training step and a supervised fine-tuning step. Given a training set data, the learning of SDAE is started by a greedy layer-wise pre-training procedure which learns a stack of DAEs one by one in an unsupervised learning manner. The key concept in greedy layer-wise learning is to train one layer every time. In this way, the network parameters are initialized, reducing the problem of local minima. In the fine-tuning phase, all the learned hidden layers from several DAEs are stacked to form a deep network, the decoder networks of each DAE are removed, and a softmax layer is added in the top of the encoder network, as shown in Fig. 4 (b). Then, the whole network parameters can be jointly optimized and fine-tuned using the back-propagation (BP) algorithm with the label information in a supervised manner. Thus, the learned representations from the unsupervised learning step can be improved with better intra-class compactness and inter-class discriminability.

Recent studies have shown that SDAE has powerful non-linear feature learning ability and can capture more hidden information and high-level features with compactness. It has been widely used in many challenging tasks, such as diagnosis of Alzheimer's disease [33], [34], classification of attention deficit/hyperactivity disorder (ADHD) [35], and machinery fault diagnosis [36], and gained successful achievements. Motivated by its excellent property in feature presentation learning, in this study, SDAE is used for EEG and ET feature learning and fusion.

2) Unimodal Feature Learning Module: To reduce the high-dimensionality and redundancy of the extracted initial EEG and ET features, we designed an EEG-SDAE and an ET-SDAE to learn the high-level EEG and ET feature representations, respectively. For two input modalities, $\mathbf{X}_{EEG}$ and $\mathbf{X}_{ET}$, two feature learning models $\mathbf{f}_{EEG}$ and $\mathbf{f}_{ET}$, are trained through an unsupervised pre-training followed by a supervised fine-tuning shown in Fig. 4, respectively. Once two models are trained, the top layer (i.e. classification layer) obtained in the supervised step can be removed and the output of the last second layer can be treated as the learned high-level representation, which can be denoted as follows:

$$\mathbf{H}_{EEG} = \mathbf{f}_{EEG}(\mathbf{X}_{EEG}) \tag{5}$$

$$\mathbf{H}_{ET} = \mathbf{f}_{ET}(\mathbf{X}_{ET}) \tag{6}$$

3) Multimodal Feature Fusion Module: Considering that two different modalities provide different and complementary discriminability for TD and ASD, we designed another Fusion-SDAE to fuse the high-level representations $\mathbf{H}_{EEG}$ and $\mathbf{H}_{ET}$, which are concatenated to form a multimodal feature vector $\mathbf{H}_{Fusion}$ as follows:

$$\mathbf{H}_{Fusion} = [\mathbf{H}_{EEG}; \mathbf{H}_{ET}] \tag{7}$$

Using $\mathbf{H}_{Fusion}$ as the input, the Fusion-SDAE model is trained to learn unified representations hidden in two different modalities.

Once the proposed framework is trained, feature extraction, unimodal feature learning via EEG-SDAE and ET-SDAE and multimodal feature fusion via Fusion-SDAE are performed, and finally give the identification label (ASD or TD).

III. Experiments and Resuls

To demonstrate the effectiveness and superiority of our proposed method, we compare it with two unimodal-based methods, EEG-SDAE and ET-SDAE. For EEG-SDAE, unimodal EEG data is used for training an SDAE-based identification model. For ET-SDAE, an SDAE-based model is trained with unimodal ET data. Also, a simple feature-level fusion method named CONCAT-SDAE is compared, where EEG data and ET data are simply concatenated to train an SDAE model. All compared methods are evaluated using the same dataset.

A. Evaluation Metric

To evaluate the performance of our proposed method, we used three common classification metrics [23], including accuracy, sensitivity, and specificity, which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. In our study, we define ASD children as the positive class and TD children as the negative class. For a comprehensive evaluation, the Receiver Operating Characteristic (ROC) curve and the resulting Area Under Curve (AUC) [22] are also used.

Due to the limited samples, we perform a 10-fold subject-independent cross-validation to evaluate the performance of the proposed model, where each subject is regarded as a sample. For TD and ASD subjects, we perform a stratified sampling strategy to ensure that each fold contains TD and ASD samples with a similar ratio. All samples are divided into 10 folds, nine folds of which are trained as the training set and the remaining one fold for testing. Thus, each fold is tested once and the evaluation process is repeated 10 times. Finally, the average testing performance of ten folds is reported.

B. Parameter Setup

In our proposed framework, there are three SDAE models for unimodal feature learning and multimodal feature fusion of EEG and ET, respectively. For each SDAE, we designed a 4-layer network structure, consisting of one input layer, two hidden layers, and one output layer. Specifically, three network structures are set as follows: 125-64-50-2 for EEG feature learning, 96-64-50-2 for ET feature learning, and 100-50-20-2 for multimodal feature fusion. For simplicity, other parameters for model training are set as the same values for three SDAEs. The noise level of each DAE is set to 0.1. The learning rate for pretraining and fine-tuning are set as 0.1 and 0.2, respectively. The iterations for pretraining and fine-tuning are 100 and 200, respectively. During the training phase, we minimize the

TABLE II
PERFORMANCE (%) OF DIFFERENT METHODS

Methods	Accuracy	Sensitivity	Specificity
EEG-SDAE	81.11±7.50	82.50±16.87	80.00±16.33
ET-SDAE	86.67±12.61	77.50±18.45	94.00±9.66
CONCAT-SDAE	93.33±10.73	92.50±12.08	94.00±13.50
MMSDAE	**95.56±5.74**	**92.50±12.08**	**98.00±6.32**

cost function by using the stochastic gradient descent (SGD) optimization algorithm. All experiments are implemented with a deep learning toolbox developed with MATLAB, which can be available online .[2]

C. Overall Performance Comparison

Table II gives the comparative results with different models in terms of accuracy, sensitivity, and specificity, where the average and standard deviation of three metrics over ten folds are reported for each method. It is obvious that our proposed MMSDAE model achieved the best performance in terms of 95.56% accuracy, 92.5% sensitivity, and 98% specificity, which significantly outperformed two unimodal methods (EEG-SDAE and ET-SDAE). Also, our MMSDAE presents the best stability and robustness with the smallest standard deviation among all compared methods. By comparing two multimodal fusion methods, CONCAT-SDAE performs worse than our MMSDAE. This result demonstrates that the complex relations between EEG and ET modalities are difficult to be captured. Different from a simple feature-level fusion, our MMSDAE can learn the shared representations between two modalities at a higher level via a multilayer network architecture. In more detail, compared with two unimodal methods, we find that ET features obtained higher average classification accuracy (86.67%) than EEG features (81.11%), and this means that ET modality has better discriminative ability between ASD and TD children.

Fig. 5 shows the ROC curves of different models with their corresponding AUC values. It can be clearly found that our MMSDAE obtains the best performance with an AUC value of 0.984, significantly higher than those of the other three methods. Also, the true positive rate of our MMSDAE increases at the beginning of the ROC curve, which means its higher identification rate with a lower misdiagnosis rate for ASD subjects. This will provide accurate and reliable diagnosis results in clinical applications. Notably, the performance of ET exceeds the result obtained with EEG, indicating that ET features have the advantage of classifying ASD. In more detail, it can be observed that two multimodal fusion methods (CONCAT-SDAE and MMSDAE) significantly outperform two unimodal methods (EEG-SDAE and ET-SDAE). This can be explained that multimodal fusion can combine the complementary information in each modality and effectively enhance the performance. These results demonstrate the effectiveness of multimodal information fusion combining EEG and ET data for ASD identification and diagnosis.

It should be noted that from model structure complexity, our proposed MMSDAE needs to train three SDAE models,

[2]https://github.com/rasmusbergpalm/DeepLearnToolbox

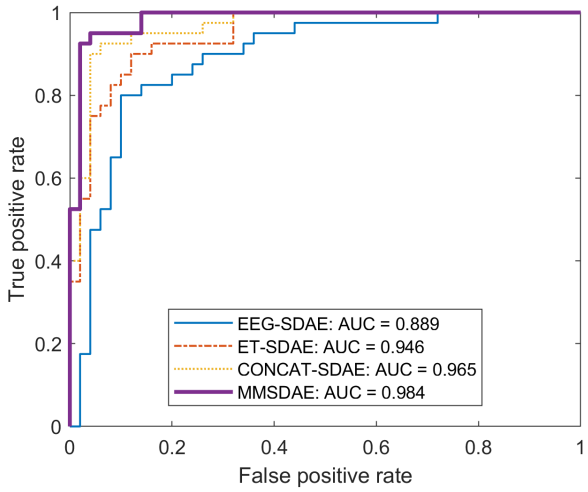

Fig. 5. ROC curves of different models with the corresponding AUC values. The horizontal axis of the ROC curve is the false positive rate (FPR), also called the misdiagnosis rate, which is defined as the ratio between the number of falsely identified ASD children (FP) and the total number of all true TD children (TN + FP). The vertical axis is the true positive rate (TPR), also called the diagnosis rate, which is defined as the ratio between the number of accurately identified ASD children (TP) and the total number of all true ASD children (TP + FN).

while CONCAT-SDAE model only requires one SDAE to be trained. Therefore, our proposed MMSDAE model has much more computational cost, especially during the model training phase. In practice, we should make a trade-off between the computation cost and the identification performance.

D. Investigation of Complementary Characteristics of EEG and ET Data

To further investigate the complementary characteristics of EEG and ET data, we calculate the confusion matrix of four different models, which reveals the classification ability of each modality. Fig. 6 shows the confusion matrix results. Comparing Fig. 6 (a) and (b), we can conclude that EEG and ET data have important complementary characteristics. Specifically, EEG-SDAE obtained a higher classification accuracy (82.5%) than ET-SDAE (77.7%) for ASD subjects, whereas for TD subjects, ET significantly outperforms EEG (94.0% versus 80%). This proves that EEG and ET data contain complementary information and have different classification abilities for ASD and TD children. Fig. 6 (c) and (d) show that two multimodal models achieved significant performance improvements compared to two unimodal models (EEG-SDAE and ET-SDAE). Multimodal fusion methods achieved the enhanced performance of identifying ASD children with an accuracy of 92.5% (10% improvement), where only three ASD children are not accurately identified and wrongly classified as TD children. Also, our proposed MMSDAE model improved the classification accuracy from 94% to 98% with 4% performance improvement for identifying TD children, which means the lowest misdiagnosis rate (only 1 child is wrongly classified as ASD). As shown in Table II, in terms of overall accuracy, MMSDAE gained an improvement of 14.45% and 8.89% against EEG-SDAE and ET-SDAE, respectively. These results reveal the contribution of each modality for ASD identification and diagnosis and why the fusion of both modalities can

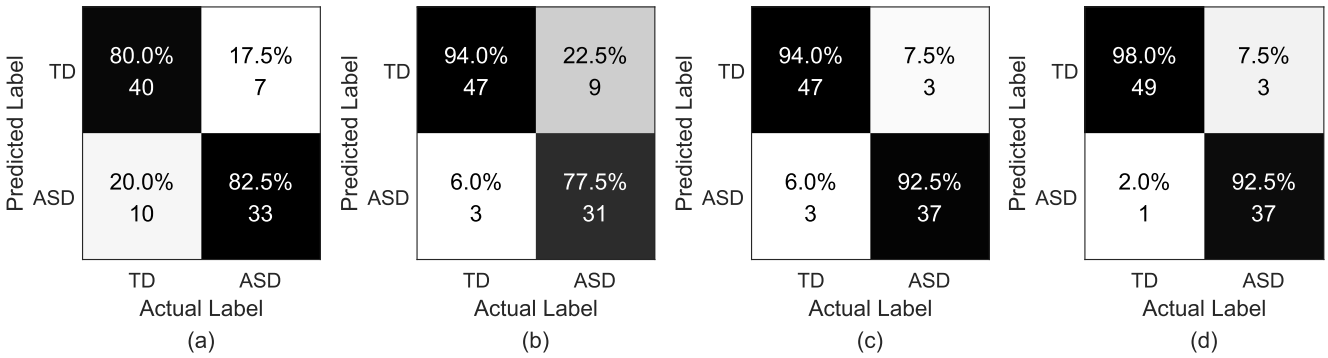

Fig. 6. Confusion matrices of different methods: (a) EEG-SDAE; (b) ET-SDAE; (c) CONCAT-SDAE and (d) MMSDAE. Each row and each column represent the predicted labels and the true labels, respectively.

enhance the identification performance. Multimodal fusion methods integrate the advantages of EEG for classifying ASD children and the advantages of ET for classifying TD children and take full use of the complementary information between EEG and ET to enhance the classification accuracies for each modality.

E. Feature Visualization

To explicitly show the better feature learning and fusion ability of our proposed method, we adopt the t-SNE technique [37] to do a 2-D feature visualization. Specifically, we reduce unimodal original EEG and ET features, unimodal learned EEG and ET features via SDAE, concatenated multimodal features, and fused multimodal features via MMSDAE into 2-D maps for visualization. The scatterplot of different features is shown in Fig.7. It can be seen that original features (EEG, ET, and concatenated) are randomly distributed in the 2-D mapping with a larger overlap between ASD and TD, which indicates the difficulty in identifying ASD when directly using original features. Also, we can observe that the learned EEG features are not separated well between TD and ASD. A possible reason is that our proposed SDAE model performs dimension reduction on raw EEG features, which will lose some useful information and therefore result in poor performance. From Fig. 7, the learned ET features via SDAE and fused multimodal features via MMSDAE exhibit better intra-class cluster performance and inter-class discriminability. Our proposed MMSDAE presents the best performance, which supports the highest overall classification accuracy listed in Table II.

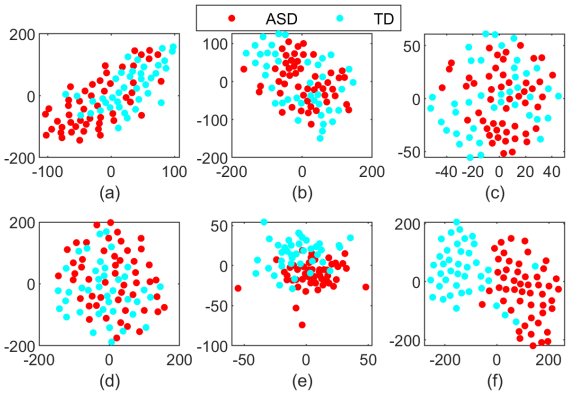

Fig. 7. Feature visualization: (a) original EEG features; (b) original ET features; (c) concatenated EEG and ET features; (d) learned EEG features via EEG-SDAE; (e) learned ET features via ET-SDAE and (f) fused multimodal features via MMSDAE.

IV. Conclusion and Future Works

In this paper, we proposed a new multimodal feature learning and fusion framework for identifying ASD in children. Its core idea is a two-step multimodal learning model, where at the first step the high-level EEG and ET feature representations are learned via an EEG-SDAE and an ET-SDAE respectively from initial high-dimensional features with information redundancy, and then the learned EEG and ET feature representations are further fused for final classification via a Fusion-SDAE at the second stage. Our proposed model realizes the joint modeling and analysis of EEG and ET data and can learn complementary information between two different modalities and enhance identification performance. Experimental results have demonstrated that our proposed method achieved better identification performance with an overall accuracy of 95.56% than unimodal methods and a simple feature-level fusion method. It should be noted that our proposed framework is data-driven and can automatically learn and fuse useful information from neurophysiological and behavioral modalities to identify ASD without the need of much more diagnostic expert experience. It provides a new tool for an easier and more objective diagnosis of ASD in children, which can assist clinicians to make a precise diagnosis decision and improve diagnosis efficiency, suggesting its great potential in clinical applications.

It is worth noting that our developed method has the following limitations. On the one hand, to deploy our model in practice, multimodal data from EEG and ET modalities must be available simultaneously. On the other hand, our model is a two-step approach containing three SDAE models, which requires much more computational costs than only one SDAE model needs to be trained. To address these limitations, in our future work, we will investigate more efficient models by introducing advanced neural network algorithms, such as

convolution neural networks (CNN) and attention networks, to fuse multimodal data, even especially in the absence of one modality during the model training. In addition, we will attempt to explore more effective features associated with ASD using advanced signal processing methods.

REFERENCES

[1] C. Sarmiento and C. Lau, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Hoboken, NJ, USA: Wiley, 2020, pp. 125–129.

[2] G. Xu, L. Strathearn, B. Liu, and W. Bao, "Prevalence of autism spectrum disorder among US children and adolescents, 2014–2016," *Jama*, vol. 319, no. 1, pp. 81–82, 2018.

[3] 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.

[4] R. P. Goin-Kochel, V. H. Mackintosh, and B. J. Myers, "How many doctors does it take to make an autism spectrum diagnosis?" *Autism*, vol. 10, no. 5, pp. 439–451, Sep. 2006.

[5] H. C. Hazlett *et al.*, "Early brain development in infants at high risk for autism spectrum disorder," *Nature*, vol. 542, no. 7641, pp. 348–351, 2017.

[6] H. Zhang, R. Li, X. Wen, Q. Li, and X. Wu, "Altered time-frequency feature in default mode network of autism based on improved Hilbert–Huang transform," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 485–492, Feb. 2020.

[7] M. Kikuchi, Y. Yoshimura, K. Mutou, and Y. Minabe, "Magnetoencephalography in the study of children with autism spectrum disorder," *Psychiatry Clin. Neurosci.*, vol. 70, no. 2, pp. 74–88, Feb. 2016.

[8] J. Han *et al.*, "Development of brain network in children with autism from early childhood to late childhood," *Neuroscience*, vol. 367, pp. 134–146, Dec. 2017.

[9] T.-M. Heunis, C. Aldrich, and P. J. De Vries, "Recent advances in resting-state electroencephalography biomarkers for autism spectrum disorder—A review of methodological and clinical challenges," *Pediatric Neurol.*, vol. 61, pp. 28–37, Aug. 2016.

[10] T. Takahashi *et al.*, "Enhanced brain signal variability in children with autism spectrum disorder during early childhood," *Human Brain Mapping*, vol. 37, no. 3, pp. 1038–1050, Mar. 2016.

[11] T. Wadhera and D. Kakkar, "Social cognition and functional brain network in autism spectrum disorder: Insights from EEG graph-theoretic measures," *Biomed. Signal Process. Control*, vol. 67, May 2021, Art. no. 102556.

[12] M. Baygin *et al.*, "Automated ASD detection using hybrid deep lightweight features extracted from EEG signals," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104548.

[13] G. Tan, K. Xu, J. Liu, and H. Liu, "A trend on autism spectrum disorder research: Eye tracking-EEG correlative analytics," *IEEE Trans. Cognit. Develop. Syst.*, early access, Aug. 5, 2021, doi: 10.1109/TCDS.2021.3102646.

[14] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada, and R. Mitkov, "Detecting high-functioning autism in adults using eye tracking and machine learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 6, pp. 1254–1261, Jun. 2020.

[15] T. Nakano *et al.*, "Atypical gaze patterns in children and adults with autism spectrum disorders dissociated from developmental changes in gaze behaviour," *Proc. Roy. Soc. B, Biol. Sci.*, vol. 277, no. 1696, pp. 2935–2943, Oct. 2010.

[16] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Res.*, vol. 9, pp. 888–898, Aug. 2016.

[17] G. Wan *et al.*, "Applying eye tracking to identify autism spectrum disorder in children," *J. Autism Develop. Disorders*, vol. 49, pp. 209–215, Jan. 2019.

[18] J. Kang, X. Han, J. Song, Z. Niu, and X. Li, "The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103722.

[19] S. Zhang, D. Chen, Y. Tang, and L. Zhang, "Children ASD evaluation through joint analysis of EEG and eye-tracking recordings with graph convolution network," *Frontiers Human Neurosci.*, vol. 15, May 2021, Art. no. 651349.

[20] B. A. Cociu *et al.*, "Multimodal functional and structural brain connectivity analysis in autism: A preliminary integrated approach with EEG, fMRI, and DTI," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 2, pp. 213–226, Jun. 2018.

[21] L. E. Mash *et al.*, "Atypical relationships between spontaneous EEG and fMRI activity in autism," *Brain Connectivity*, vol. 10, no. 1, pp. 18–28, Feb. 2020.

[22] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019.

[23] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 173–183, Jan. 2018.

[24] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 525–536, Mar. 2018.

[25] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[26] P. J. Durka, H. Klekowicz, K. J. Blinowska, W. Szelenberger, and S. Niemcewicz, "A simple system for detection of EEG artifacts in polysomnographic recordings," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 4, pp. 526–528, Apr. 2003.

[27] S. Matlis, K. Boric, C. J. Chu, and M. A. Kramer, "Robust disruptions in electroencephalogram cortical oscillations and large-scale functional networks in autism," *BMC Neurol.*, vol. 15, no. 1, p. 97, Dec. 2015.

[28] A. R. Levin, K. J. Varcin, H. M. O'Leary, H. Tager-Flusberg, and C. A. Nelson, "EEG power at 3 months in infants at high familial risk for autism," *J. Neurodevelop. Disorders*, vol. 9, no. 1, p. 34, Dec. 2017.

[29] J. Wang, J. Barstein, L. E. Ethridge, M. W. Mosconi, Y. Takarae, and J. A. Sweeney, "Resting state EEG abnormalities in autism spectrum disorders," *J. Neurodevelopmental Disorders*, vol. 5, no. 1, p. 24, Dec. 2013.

[30] A. Catarino, O. Churches, S. Baron-Cohen, A. Andrade, and H. Ring, "Atypical EEG complexity in autism spectrum conditions: A multiscale entropy analysis," *Clin. Neurophysiol.*, vol. 122, no. 12, pp. 2375–2383, Dec. 2011.

[31] K. Zeng *et al.*, "Disrupted brain network in children with autism spectrum disorder," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, Dec. 2017.

[32] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

[33] S. Liu *et al.*, "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132–1140, Apr. 2015.

[34] R. Ferri *et al.*, "Stacked autoencoders as new models for an accurate Alzheimer's disease classification support using resting-state EEG and MRI measurements," *Clin. Neurophysiol.*, vol. 132, no. 1, pp. 232–245, Jan. 2021.

[35] S. Liu *et al.*, "Deep spatio–temporal representation and ensemble classification for attention deficit/hyperactivity disorder," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1–10, 2021.

[36] G. Jiang, H. He, P. Xie, and Y. Tang, "Stacked multilevel-denoising autoencoders: A new representation learning approach for wind turbine gearbox fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 9, pp. 2391–2402, Sep. 2017.

[37] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, Nov. 2008.