Deep Spatio-Temporal Representation and Ensemble Classification for Attention Deficit/Hyperactivity Disorder

Shuaiqi Liu, Ling Zhao, Xu Wang, Qi Xin, Jie Zhao, David S. Guttery, and Yu-Dong Zhang, *Senior Member, IEEE*

***Abstract*— Attention deficit/Hyperactivity disorder (ADHD) is a complex, universal and heterogeneous neurodevelopmental disease. The traditional diagnosis of ADHD relies on the long-term analysis of complex information such as clinical data (electroencephalogram, etc.), patients' behavior and psychological tests by professional doctors. In recent years, functional magnetic resonance imaging (fMRI) has been developing rapidly and is widely employed in the study of brain cognition due to its non-invasive and non-radiation characteristics. We propose an algorithm based on convolutional denoising autoencoder (CDAE) and adaptive boosting decision trees (AdaDT) to improve the results of ADHD classification. Firstly, combining the advantages of convolutional neural networks (CNNs) and the denoising autoencoder (DAE), we developed a convolutional denoising autoencoder to extract the spatial features of fMRI data and obtain spatial features sorted by time. Then, AdaDT was exploited to classify the features extracted by CDAE. Finally, we validate the algorithm on the ADHD-200 test dataset. The experimental results show that our method offers improved classification compared with state-of-the-art methods in terms of the average accuracy of each individual site and all sites, meanwhile, our algorithm can maintain a certain balance between specificity and sensitivity.**

***Index Terms*— Adaptive boosting decision tree, ADHD, convolutional denoising autoencoder, fMRI classification.**

Manuscript received July 13, 2020; revised August 9, 2020 and August 14, 2020; accepted August 19, 2020. Date of publication August 24, 2020; date of current version February 25, 2021. This work was supported in part by the Natural Science Foundation of China under Grant 61401308 and Grant 61572063; in part by the Natural Science Foundation of Hebei Province under Grant F2020201025, Grant F2016201187, and Grant F2018210148; in part by the Science Research Project of Hebei Province under Grant QN2020030 and Grant QN2016085; and in part by the Natural Science Foundation of Hebei University under Grant 2014-303. *(Corresponding authors: Ling Zhao; Yu-Dong Zhang.)*

Shuaiqi Liu, Ling Zhao, Xu Wang, and Jie Zhao are with the College of Electronic and Information Engineering, Hebei University, Baoding 071002, China, also with the Key Laboratory of Digital Medical Engineering of Hebei Province, Hebei University, Baoding 071002, China, and also with the Machine Vision Engineering Research Center of Hebei Province, Hebei University, Baoding 071002, China (e-mail: shdkj-1918@163.com; lingzhao_hbu@163.com; xuwang_hbu@163.com; jzhao_hbu@126.com).

Qi Xin is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: qxin@bjtu.edu.cn).

David S. Guttery is with the Leicester Cancer Research Institute, Leicester LE2 7LX, U.K. (e-mail: dsg6@le.ac.uk).

Yu-Dong Zhang is with the Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: yudongzhang@ieee.org).

Digital Object Identifier 10.1109/TNSRE.2020.3019063

I. Introduction

ATTENTION deficit/Hyperactivity disorder (ADHD) is a neurodevelopmental condition characterized by core symptoms such as inattention, hyperactivity, and impulsivity [1]. ADHD is one of the most debilitating childhood illnesses. Approximately 65% of cases will last to adulthood [2] and seriously affect the study and work of patients, causing a heavy burden to families and society. Mental health experts often use the *Diagnostic and Statistical Manual of mental disorders (DSM)* developed by the *American Psychiatric Association* to help diagnose ADHD [3] in clinical practice. At present, ADHD is only diagnosed after clinical review by an experienced child psychiatrist, in addition to discussions with the child's parents and teachers. However, diagnoses are often inconsistent since the diagnostic process is greatly affected by subjective assessment. Therefore, it is essential to find a consensus method to diagnose ADHD according to the existing medical means [4].

Functional magnetic resonance imaging (fMRI) is a widely used noninvasive tool to measure brain activity and highlight the slow fluctuation of blood oxygen level dependence (BOLD) between brain regions during task states or resting states [5]. With the development of machine learning, scholars have paid more attention to the prediction of neurodevelopmental diseases with fMRI data like Alzheimer's disease [6] (AD), Autism spectrum disorders [7] (ASD), ADHD [8], etc.

To promote research in disease imaging of ADHD, the ADHD-200 consortium held the ADHD-200 global competition in 2011 supported by the *International Neuroimaging Data-sharing Initiative* (INDI). The competition aimed to develop imaging classification methods of patients with ADHD. The ADHD-200 dataset consists of rs-fMRI and structural magnetic resonance imaging (sMRI) images of approximately 800 subjects, which are collectively provided by eight scientific research institutions, such as *Kennedy Krieger Institute* (KKI), *New York University Medical Center* (NYU), *Oregon Health and Science University* (OHSU), *Neuroimage*

Sample (NeuroImage) and *Peking University* (Peking), etc. The competition aimed to determine the prediction accuracy of each team for typically-developing (TD) and ADHD patients (including the prediction accuracy of ADHD subcategories), and J-statistics (including sensitivity and specificity). ADHD-200 also trained an image-based classifier to distinguish three types: mixed type (ADHD-I), inattentive type (ADHD-II) and TD [9]. The highest accuracy achieved using the imaging data was 60.51% in 2011.

Many researchers have exploited the data from this competition to carry out various studies on ADHD. For instance, Dai *et al.* used the cortical thickness (CT), gray matter probability (GMP) extracted from sMRI and ReHo, and functional connectivity (FC) extracted from fMRI as features to improve the classification accuracy of ADHD [10]. The authors not only compared the impact of each feature on classification but also fused the features through multi-kernel learning, with the classification accuracy reaching 61.5%. The same year, Sidhu *et al.* used the fast Fourier transform and kernel principal component based on phenotypic and imaging which yielded accuracies of 76.0% on two class diagnosis [11]. In addition, Zou *et al.* [12], proposed a 3D-convolutional neural network (CNN) deep learning classification method based on fMRI and sMRI. Firstly, ReHo, fractional amplitude of low-frequency oscillation (fALFF), and voxel mirrored homotropy connectivity (VMHC) were extracted manually from fMRI. Then, gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) were extracted from sMRI. Finally, a 3D-CNN classifier was employed to evaluate the performance of each feature and the classification performance of a multi-feature combination given. The study showed that the combination of fALFF and GM yields the best result and the accuracy of ADHD classification is 69.2%. Complementing this, Riaz *et al.* [13] created an end-to-end network for ADHD classification, which consists of a feature extraction layer, similar network, and classification network. The network first extracted 90 features from 90 brain regions of fMRI after preprocessing and the similarity between features was calculated. The classification accuracy of this algorithm in the ADHD dataset of Peking, NeuroImage, and NYU reached 62.7%, 67.9%, and 73.1%, respectively. In addition, Kuang *et al.* [14] proposed an ADHD classification algorithm based on fast Fourier transform and deep belief network. The classification accuracy of this algorithm in the ADHD dataset of NYU, Peking, and KKI reached 37.41%, 54.00%,and 71.82%, respectively., Mao *et al.* [15] obtained good results using spatial information of each frame from fMRI images extracted by 3DCNN and the temporal information of fMRI time-series images extracted by feature pooling and long short-term memory (LSTM) models. Finally, the proposed 4D-CNN extracting the spatial and time information of fMRI at the same time achieved the highest accuracy of 71.3% in the application to ADHD classification.

Recently, the popularity of deep learning methods has resulted in their extensive application to various phenomena including as image denoising [16], image fusion [17], image recognition [18] and image classification [19]. As one of the most commonly used deep learning methods, CNNs can obtain the features of the input data through automatic learning, especially for high-dimensional data. However, as a supervised learning method, CNN needs a lot of labeled data in the training stage, which is not only time-consuming and labor-intensive but also prone to over-fitting. Therefore, an unsupervised deep learning method is selected to perform the process of extracting features.

The autoencoder is a practical unsupervised learning model in deep learning and consists of an encoder and decoder. The former is employed to encode the original representation into the hidden layer representation while the latter is used to decode the hidden layer representation into the original representation. The training target minimizes the reconstruction error function via backpropagation. Generally speaking, the dimension of the hidden layer is lower than the original feature [20].

Since the autoencoder is just a concept, the encoder and decoder can be composed of a variety of deep learning models, such as a fully connected layer, convolution layer, and LSTM. CNN has advantages in image processing due to the ability to extract the spatial information hidden in the image. It is instinctively assumed that CNNs can work better than other autoencoders when constructing an encoder and decoder network, hence why the convolutional autoencoder (CAE) is generated [21].

To solve the problem of ADHD classification based on fMRI images, the convolution denoising autoencoder is proposed as the feature extractor in the feature extraction stage. CAE has the structure of CNN and autoencoder as well as the corresponding advantages. As a simple and efficient neural network, CAE can effectively extract useful feature information from the data for classification without massive labels [22], [23].

We adopt the convolutional denoising autoencoder (CDAE) for mining spatial features to fully extract 3D spatial information of fMRI data. The 3D convolutional denoising autoencoder was applied to train each frame of fMRI image in the feature extraction stage, after that the pre-trained encoder was used to extract the spatial features of fMRI. Considering the small amount of fMRI image data in the ADHD-200 dataset, we utilized the fMRI spatial features extracted in time order to perform dimension reduction processing again based on principal component analysis (PCA) to avoid over-fitting caused by "small sample and high-dimension". The data after dimension reduction was processed as the features of ADHD classification. We employed AdaDT as a classifier and the experimental results show that this algorithm can effectively classify the ADHD in the test set. The overall flow of the proposed method is shown in Fig. 1.

The main contributions of this article are as follows:

(1) In this article, CDAE was employed to automatically extract the features of fMRI data, which can fully extract the 3D spatial information of fMRI data and avoid the unreliability and instability brought by hand-crafted features.

(2) The spatial features of fMRI extracted in time were reduced by PCA, which effectively avoids the over-fitting phenomenon caused by small samples of high-dimensional data.

(3) The adaptive boosting decision tree (AdaDT) can turn the weak classifier set of the trained decision tree into a strong

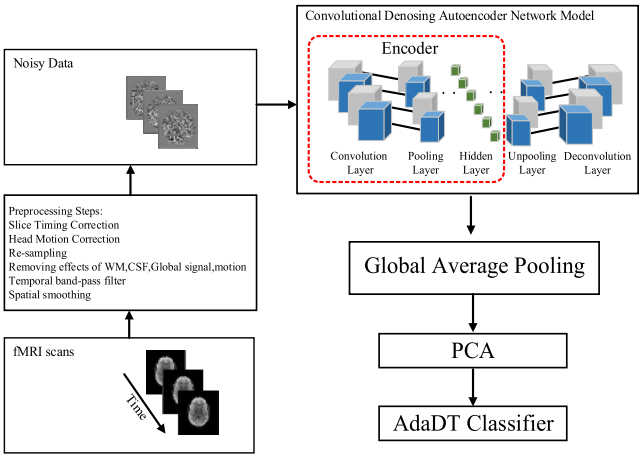

Fig. 1. The flow diagram of this article.

classifier and effectively avoid the under-learning phenomenon caused by insufficient learning data to classify ADHD.

The remainder of this article is arranged as follows: the second section introduces the theoretical background of the proposed CDAE-AdaDT algorithm in detail; the third section is the experimental setup, including data processing and training details of the CDAE-AdaDT algorithm model; the fourth section describes and discusses the experimental results; the last section summarizes the algorithm and experimental results of this article.

II. Methods

In recent years, extracting features of unlabeled samples through autoencoder has achieved encouraging results with the rapid development of unsupervised learning [24], [25]. Therefore, 3D convolutional denoising autoencoder was used to extract the features of fMRI in this article. The following describes the feature extraction algorithm used in our method.

A. Feature Extraction

As a kind of artificial neural network, deep neural networks (DNN) have attracted attention due to its improved performance. As a special structure of DNN, CNN [26], [27] has the advantages of local connectivity and parameter sharing. It can extract spatial information from the original data without other complex preprocessing. The CNN structure used in this article includes three basic layers: convolution layer, pooling layer and global average pooling layer. The common structure of CNN is shown in Fig. 2.

Traditional CNN employs a 2D convolution kernel in a 2D image. While fMRI data is a three-dimensional structure in space, a 3D convolution kernel is used in this article to make better use of the spatial structure characteristics of the fMRI image. In the convolution layer, a series of 3D convolution kernels are convoluted with the receiving domain of the input image or the feature map of the previous layer in the sliding window to learn the features of the data [28]. Let the output of neurons v_{ij}^{xyz} in (x, y, z) of the j-th feature map of the i-th

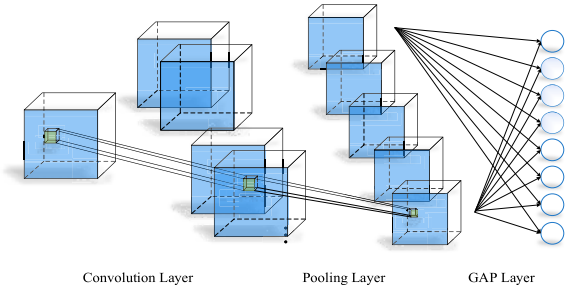

Fig. 2. Schematic diagram of CNN.

layer be defined as

$$v_{ij}^{xyz} = f\left(b_{ij} + \sum_{n}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijn}^{pqr} v_{(i-1)n}^{(x+p)(y+q)(z+r)}\right) \tag{1}$$

where n is the index of the i-1 feature map, b_{ij} denotes the bias, and P_i, Q_i and R_i are the length, width, and height of the convolution kernel respectively, w_{ijn}^{pqr} is the value of the convolution kernel connected to the n-th feature map, and f is the nonlinear activation function.

The convolution layer is connected with the pooling layer. Generally speaking, there are two kinds of pooling: max pooling and average pooling. The pooling operation down-samples the feature map to reduce the network parameters which can lessen the amount of computation while the characteristic of space invariance [29] can preserve the spatial relationships. In this article, we choose the max pooling operation and the last network we use is the global average pooling (GAP) [30]. Unlike the traditional fully connected layer, GAP is used to combine the feature map in a non-linear way, which can not only reduce the number of network parameters and improve the training speed but also effectively prevent the occurrence of over-fitting.

Whereas the great success CNNs have achieved in various fields, especially in image classification [31], [32], it cannot be ignored that the classification algorithm based on CNN needs a lot of manually marking data since it is a type of supervised learning [33], [34]. Nevertheless, manually marking workload is time-consuming in ADHD classification, which brings great difficulty to the application of CNNs. Consequently, the classification algorithm based on unsupervised learning has attracted attention in recent years in view of the advantages of requiring no labels.

Autoencoder (AE) is an unsupervised algorithm that can learn from data automatically. The purpose of the AE is to select encoder and decoder functions so that the image can be encoded with the least information and be reconstructed on the other side [35]. As an unsupervised learning method, AE can reconstruct the output data into the input data without labels while preserving the dimensions of the original data [36]. Fig. 3 is a schematic diagram of the autoencoder.

It can be seen from Fig. 3 that there are two parts in the autoencoder: encoder and decoder. In the structure of the autoencoder, each layer is fully connected with the next layer

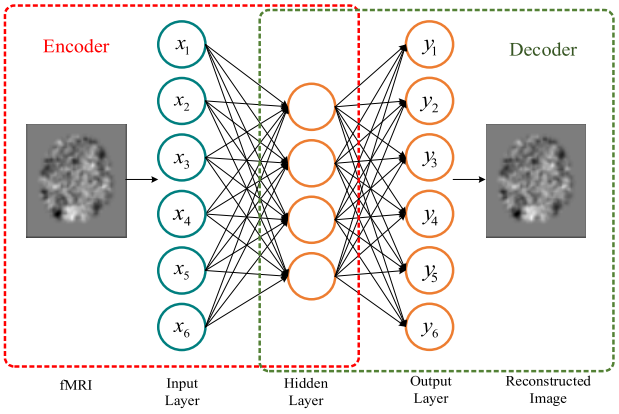

Fig. 3. Schematic diagram of the autoencoder.

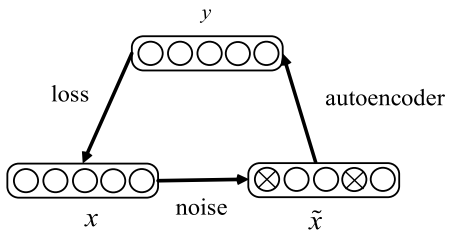

Fig. 4. Schematic diagram of denoising autoencoder.

with an activation function. However, each training of the autoencoder is a comparison of the original data itself, which will increase the similarity between input and output, but not sensitive to other images of the same kind. This is especially obvious in fMRI images that have nuances between different frames.

Vincent *et al.* [22], [23] proposed a denoising autoencoder aiming to solve the aforementioned problems. Random noise is added to the images to make each input vary slightly before applying them to the network. Finally, the output of the autoencoder is compared with the clean image before adding noise to optimize the network. In this way, the network will have better generalization ability when processing data, with little difference between different frames [37]. Fig. 4 is a schematic diagram of a denoising autoencoder.

The denoising autoencoder can not only extract the low-dimensional representative features from the original data but also recover the clean image from the noisy data, which can effectively prevent the over-fitting problem while retaining robustness [38]. The excellent performance of CNN directly promotes the generation of CDAE. Strictly speaking, CDAE is a special case of traditional denoising autoencoder, which uses the convolution layer and pooling layer instead of a fully connected layer. CDAE combines the merits of CNN and denoising autoencoder and can not only acquire the robust spatial characteristics of input data through learning but also effectively prevent overfitting.

Firstly, data with random noise is used as input into the neural network. The reconstructed data should be as close as possible to the original data instead of the noisy data, that is, the clean input is recovered from the corrupted data. Let $x = \left[x_1, \ldots x_p\right]$ represent the raw data, where x_i denotes the voxel of the fMRI and $p \in [0, 60 \times 72 \times 60]$. Let the data with random noise be $\tilde{x} = \left[\tilde{x}_1, \ldots \tilde{x}_p\right]$, where x_i is the voxel added random noise and $p \in [0, 60 \times 72 \times 60]$. The noisy data is sent to the encoder network of CDAE to obtain the hidden layer data, and the hidden layer data can be obtained as

$$h = g\left(\tilde{x}\right) = g\left(W * \tilde{x} + b\right) \tag{2}$$

where W is the weight matrix, b denotes the bias vector, $*$ represents the convolution operation and g represents the nonlinear activation function. The decoder can be regarded as the "mirror" of the encoder to some extent. The decoder recovers the same amount as the original data by using the max unpooling layer which adopts nearest-neighbor interpolation after each deconvolution layer. Accordingly, the decoder restores y from the hidden layer h, that is

$$y_i = g\left(h\right) = g\left(W^T * h + b^T\right) \tag{3}$$

where W^T and b^T are the transposition of W and b respectively. Thus, there is a certain relationship between the weights of the autoencoder, which will reduce the parameters by half and effectively decrease the complexity of the network [39]. The denoising autoencoder optimizes the network by minimizing the reconstruction error of y and x. Compared with the autoencoder, the denoising autoencoder can not only reliably capture the main change factors from the noise dataset without assuming linearity but is also robust and can effectively prevent over-fitting [40].

In this article, CDAE is employed for feature extraction and only the encoder part of the trained CDAE model is adopted for feature extraction of fMRI sequences. To improve the performance of extracted feature classification, we add a global average pooling layer after the encoder of the CDAE to convert the obtained fMRI features into one-dimension feature vectors:

$$O_c = \frac{1}{K} \sum_{x,y,z} h_{x,y,z} \tag{4}$$

where K is the number of activation values and $c \in \{1, 2, \ldots, n\}$ denotes the frames of the fMRI. We then form the initial feature vectors by connecting the data of one-dimension feature vectors end-to-end according to the time dimension.

$$O = \left(O_1, O_2, \ldots, O_n\right) \tag{5}$$

The classifier is prone to over-fitting due to the small amount of fMRI data and the correlation between the feature vectors extracted by CDAE. Therefore, PCA is used to decorrelate the initial feature vectors of fMRI to solve the problem.

B. Classifier

There are many kinds of classifiers, such as linear discriminant, naive Bayesian classification, k-nearest neighbor, support vector machine, random forest, decision tree, etc. [41], [42], and [43]. In this article, we adopted AdaDT for ADHD classification. The decision tree is a tree structure, in which each internal node represents a judgment on each attribute while each branch represents an output of judgment, and

finally each leaf node represents a classification result [44]. It is a very common and supervised learning classification method. Supervised learning means that the classification results are known and a decision tree is obtained by learning these samples. In this way, the decision tree can classify the new data correctly.

The classifier used in this article is the classification and regression trees (CART) algorithm. CART is a binary tree, where the data is cut into two parts each time by using a binary segmentation method and sent into the left and right subtree [45] respectively. Each non-leaf node has two children, so there are more leaf nodes in CART than non-leaf nodes. In CART classification, the Gini index, namely Gini impurity, is used to select the optimal data segmentation feature, which is similar to the meaning of information entropy. Each iteration in CART will reduce the Gini impurity. The smaller the Gini impurity is, the higher the purity is, and the better the classification is. The definition of Gini impurity (G) is shown as

$$G(S) = 1 - \sum_{i=1}^{k} (p_i)^2 \tag{6}$$

where S represents all samples, p_i represents the probability of the i-th category, and k represents the total number of categories.

The decision tree is powerful but unstable. The decision tree will change greatly when the training data varies [46]. Compared with the single decision tree algorithm, the integrated tree algorithm has a higher prediction ability and can overcome the problem that is difficult for a single decision tree. The integration algorithm trains multiple learners to solve the same problem and the commonly used combination methods are bagging and boosting [47]. Boosting is used in this article. Adaptive boosting (AdaBoost) is one of the most popular reinforcement algorithms as a supervised learning method. It combines weak classifiers with certain rules to build a strong classifier [48], [49]. AdaBoost determines the weight of each sample according to the classification in each training and the accuracy obtained in the last overall classification, and then the data with new weight is transferred to the next classifier for training. Finally, the classifier obtained in each training is fused and the classifier obtained by fusion is the final decision classifier to achieve the target classification. Compared with other machine learning algorithms, the AdaDT classifier will not reduce the generalization ability of the classifier with the increasing number of iterations and can avoid over-fitting at the same time, which makes AdaDT more suitable for medical images with fewer samples. Specifically, the implementation steps of AdaDT are shown in algorithm 1.

Algorithm 1 AdaDT Algorithm

Input: $\left(O^1_{PCA}, y_1\right), \left(O^2_{PCA}, y_2\right) \ldots, \left(O^N_{PCA}, y_N\right)$, where $O^i_{PCA} \in O_{PCA} O_{PCA}$ is the training dataset, $y_i \in \{0, 1\}$ is the label.

Step:

1. Initial the weight distribution of training data $W_1(i) = \frac{1}{N}$ where $i = 1, 2, \ldots, N$, denotes the total number of the samples, $t = 1, \ldots, T$ is the number of iterations.
2. Training weak classifier $h_t = \xi(W_t)$ based on sample distribution W_t cyclically.
3. Calculating the weak classifier corresponding to the j-th feature, the error ε_i is calculated as

$$\varepsilon_i = \frac{1}{N}\left[\sum_{i=1}^{n} W_t(i)\, I\left(h_i(x_i) \neq y_i\right)\right]$$

where $I\left(h_i(x_i) \neq y_i\right)$ represents the indicating function which is

$$I\left(h_i(x_i) \neq y_i\right) = \begin{cases} 0, & h_i(x_i) = y_i \\ 1, & h_i(x_i) \neq y_i \end{cases}$$

4. Update and adjust the sample distribution

$$W_{t+1}(i) = \frac{W_t(i)\, e^{-\varepsilon_t y_i h_t(x)}}{Z_t}$$

where Z_t is the normalization factor.

5. Repeat Step 2-4 until $T \geq t$.

Output:Final classification results $H(x) = \text{sign}(\sum_{t=1}^{T} \varepsilon_t h_t(x))$.

III. Experimental Setup

A. Data and Preprocessing

The data we used is from the ADHD-200 public dataset. The dataset consists of eight international imaging sites, including 973 individuals' rs-fMRI, sMRI and basic phenotypic information (age, gender, dominant hand and intelligence quotient (IQ)), which contains 362 children and adolescents diagnosed as ADHD, 585 TD and 26 unknown individuals [9]. We only used five sites including Peking, KKI, NeuroImage, NYU and OHSU. The other three sites were not used in this experiment because *Brown University* (Brown) lacks the diagnostic information of each subject, *the University of Pittsburgh* (Pittsburgh) and *Washington University* (WashU) only have TD subjects in the training set and lack ADHD subjects. In conclusion, we decided to exclude these three sites and only use the data of the remaining five sites for testing since the classification is related to the proportion of data.

In this article, the Data Processing Assistant for Resting State fMRI (DPARSF) toolbox in [50], [51] was used to process the raw fMRI data, with the processing flow as follows: (1) To achieve data balance, the first four-time points of training data and the first three-time points of test data were removed to eliminate the influence of instability; (2) Slice-timing correction; (3) Head correction; (4) Normalized into the Montreal Neurological Institute (MNI) space, resampled to 3-mm isotropic voxels; (5) Band-pass filtered; (6) Linear detrended, remove the nuisance covariates including WM, CSF, global signal and six head motion parameters; (7) Smooth using a Gaussian filter with Full Width Half Height (FWHM = 4 mm). Before the experiment, samples without

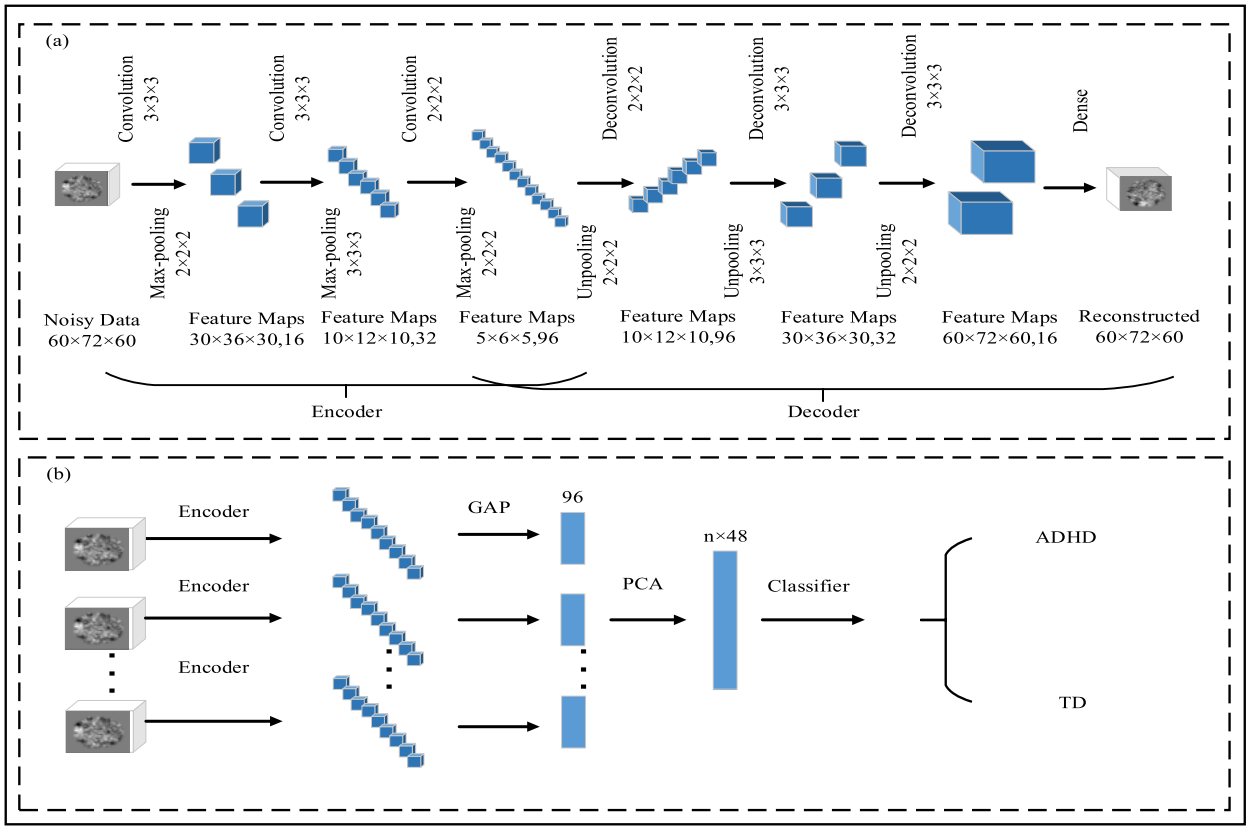

Fig. 5. The model of CDAE-AdaDT.

TABLE I
SUMMARY OF THE TRAINING AND TEST DATA-SETS FROM FIVE SITES WHICH ARE USED IN OUR WORK

	Training Set			Test Set		
	ADHD	TD	Total	ADHD	TD	Total
NYU	105	90	195	29	12	41
Peking	61	108	169	24	27	51
KKI	20	53	73	3	8	11
NeuroImage	12	16	28	5	14	19
OHSU	24	33	57	6	28	34
Total	222	300	522	67	88	156

Peking: Peking University; KKI: Kennedy Krieger Institute; NYU: New York University Child Study Center; OHSU: Oregon Health and Science University. ADHD is the number of ADHD and TD is the number of TD.

the corresponding rs-fMRI data were checked and deleted. To prevent the noise in the scanning process from interfering with the fMRI data, the data of the subjects whose head movement is more than 3mm or rotation is more than 3 degrees were removed after preprocessing. At the same time, the data of the subjects with artifacts and poor registration effects were removed through visual inspection. Finally, the data composition used in this article is shown in Table I. The number of fMRI frames participating in CDAE model training was 93650.

B. Model Training

The deep learning model is implemented by the keras framework with tensorflow as the back-end. The optimizer adopts the Adam optimizer with a learning rate of 0.0001 and a batchsize of 50. CDAE consists of two parts: convolution encoder and deconvolution decoder. The encoder part is employed to extract the feature map of the frame of the fMRI while the decoder is used to reconstruct the image from the feature map. Fig. 5 shows the CDAE-AdaDT model proposed in this article.

Fig. 5 (a) shows the training process of the CDAE model with 15 layers. The first layer and the last layer are input and output layer respectively. The second layer to the seventh layer belongs to the encoder and the eighth layer to the fourteenth layer belongs to the decoder. Each layer is connected to the next layer by linear multiplication and activation function. The network is optimized by minimizing the loss function. Fig. 5 (b) shows the ADHD classification flow chart based on the CDAE-AdaDT model.

The detailed training steps of CDAE-AdaDT model are as follows:

First, every single fMRI image with random noise was used as an input with the size of $60 \times 72 \times 60$. The encoder consists of three layers of the convolution layer and each convolution layer is connected with a max-pooling layer. The kernels in the convolution layers are $3 \times 3 \times 3$, $2 \times 2 \times 2$ and $3 \times 3 \times 3$ respectively while the window size of max-pooling layer are $2 \times 2 \times 2$, $3 \times 3 \times 3$ and $2 \times 2 \times 2$ respectively. The max-pooling layer can reduce the size of the feature map and the parameters of the network. The kernel size and window size of the pooling layer in the decoder are symmetric to the encoder. At the end of the encoder network, a dense layer was added to realize the nonlinear combination of features and increase the representativeness of the extracted features. The model learned

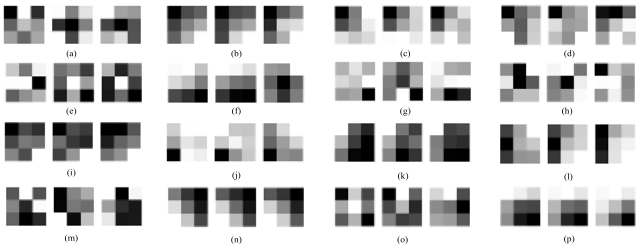

Fig. 6. Weights in the first convolution layer.

the abstract features of the image during training. As shown in Fig. 5 (a), the output of the previous layer was the input of the next layer. The activation function is the corrected linear units (ReLU), that is

$$f(z) = \max(0, z) \tag{7}$$

where z is the input of the next layer.

In this article, the adaptive moment estimation (Adam) algorithm was adopted to optimize the network by minimizing the error between the clean image and the reconstructed image, that is

$$loss = (y_i - x_i)^2 \Big/ x_i^2 \tag{8}$$

where x_i denotes the original clean image and y_i is the reconstructed image.

Finally, the pre-trained encoder was used as the initial feature extractor of fMRI data. It is worth noting that the scrambled fMRI frame data is used in the training of CDAE, while the fMRI data is passed through the encoder in chronological order to obtain the time characteristics in the feature extraction stage. The output of the encoder was connected to a global average pooling layer which transforms the features of the extracted fMRI sequences into one-dimension vectors. The global average pooling layer has fewer parameters compared with the traditional fully connected layer. Because of the characteristics of "small sample with high-dimension" of fMRI data, PCA was adopted to reduce the dimensionality of the extracted time series to reducing the occurrence of the overfitting problem. PCA is a kind of data dimensionality reduction method widely used in data analysis. The initial feature vector of fMRI after PCA is $n \times p_c$, where n is the time points of each fMRI and p_c is the number of features left after dimension reduction. In this article, the selected value of p_c is 48. And n varies according to the subjects of different sites.

The data of different sites after dimension reduction are sent to the classifier for training. The scanning parameters are different among sites and testing at all sites may aggravate the impact of data heterogeneity, hence we chose to train classifiers and test them on individual sites. It is very important to choose two parameters when training AdaDT classifier: the number of weak classifiers and the number of nodes in each tree. In this article, the optimal parameters were selected for each site classifier to achieve the best classification results through a large number of experiments.

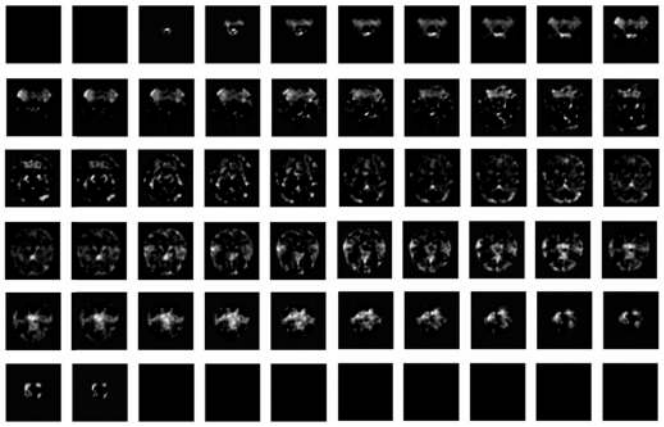

Fig. 7. Feature map of the first layer convolution layer in the first filter.

IV. Results and Discussion

A. Visualization

The traditional fully connected network is a "black box" algorithm. To better understand the features learned by CDAE, the weight of the convolution layer and the feature map are visualized. Weight plays an important role in the neural networks, and the Xavier is used to initialize the weight, that is, the weight is initialized to a uniform distribution, keeping the variance of information flowing in the neural network unchanged. Fig. 6 shows the visualization of 16 weights of the first convolution layer. Fig. 6 shows that the weights of the first convolution layer changed differently compared with the weights of the initial state, which had a uniform distribution. Different weights of convolution kernels mean that different convolution kernels can extract features from different angles, so they can effectively learn and process the fMRI image.

Fig. 7 shows the output of the first convolution layer on the first filter.

Fig. 8 shows the difference of the feature map of the first filter in the third convolution layer of ADHD and TD randomly selected. As can be seen from Fig. 7 and Fig. 8, the features learned by the model became more abstract with the increase of the convolution layer.

B. Comparison of Different Parameter Values

In order to choose the best p_c value and classifier, we compare the combination of different p_c values and classifiers. We adopt the grid search to select the optimal parameters of the classifiers. The comparison of classifiers includes linear support vector machine (L-SVM), radial basis function kernel support vector machine (RBF-SVM) and random forest (RF). The PCA values are 12, 24, 48 and 96. The accuracy, sensitivity and specificity are used as evaluation indices. The accuracy represents the ability of the model to distinguish ADHD and TD correctly. It is given by the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Sensitivity represents the ability to distinguish ADHD correctly, which is estimated by the following formula:

$$Sensibility = \frac{TP}{TP + FN} \tag{10}$$

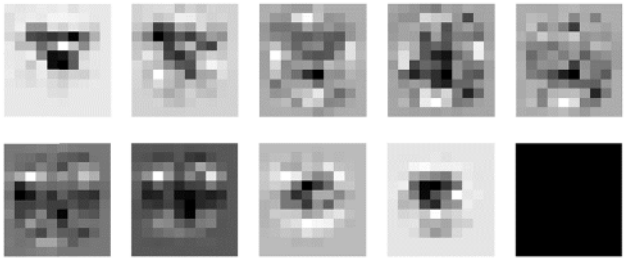

Fig. 8. Difference feature map of the first filter in the third convolution layer of ADHD and TD.

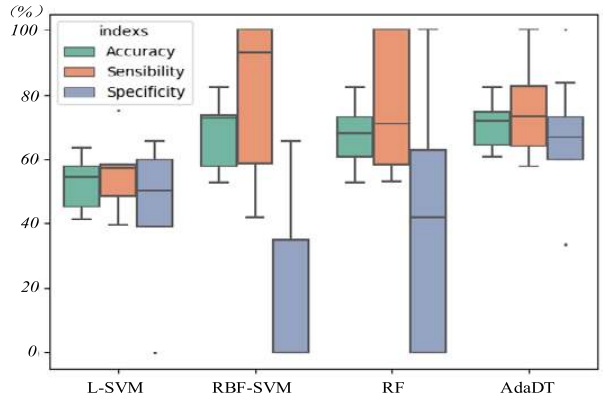

Fig. 9. The comparison of classifiers.

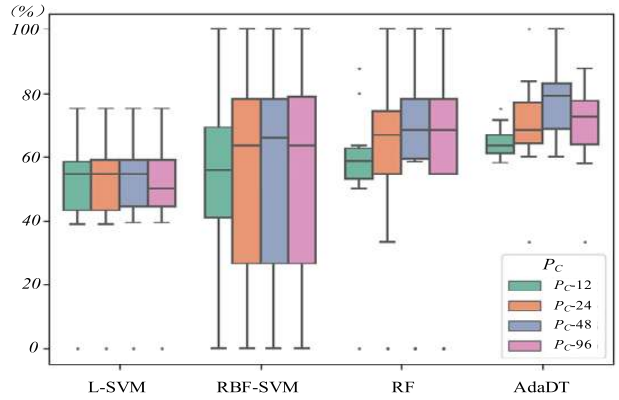

Fig. 10. Comparison of combinations of different classifiers and p_c.

Specificity describes the ability of the model to distinguish TD correctly, which can be obtained from the following formula.

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

where TP is the true positive rate (the number of correctly classified as ADHD), FP is the false positive rate (the number of correctly classified as ADHD), TN is the true negative rate (the number of correctly classified as TD), TP is the false negative rate (the number of wrongly classified as TD in ADHD patients).

Fig.9 shows the boxplot of three evaluation indices obtained by different classifiers based on all test sets. It indicates that the ensemble classifiers (RF and AdaDT) yield better results than the SVM classifiers. What's more, the AdaDT can keep the balance among all the three indices.

TABLE II
COMPARISON OF DIFFERENT ALGORITHMS IN DIFFERENT SITES

Methods	NYU	Peking	KKI	NeuroImage	OHSU
ADHD-2017	35.19%	51.05%	61.90%	56.95%	65.37%
DBN	37.41%	54.00%	71.82%	-	-
3D-CNN	70.50%	62.95%	72.82%	-	-
R-RELIEF	70.73%	68.63%	**81.82%**	76%	-
Our method	**73.17%**	**70.59%**	**81.82%**	**78.95%**	**82.35%**

Since the Accuracy, Sensitivity and Specificity are of equal importance, we adopt the three indices as the data of the boxplot on the all test sets to select the best value of p_c. Fig.10 shows the results with different combination of classifiers and p_c. Upon inspecting Fig. 10, we see that the indices increase the value of p_c when it is less than 48 and achieve the best performance when p_c is 48. However, the performance decreases when p_c is greater than 48. And the AdaDT yields the best results when p_c is 48. On the basis of the above research results, we adopt AdaDT as the classifier and 48 as the value of p_c to perform classification.

C. Comparison of Classification Results Among Different Sites

In this article, the test dataset of ADHD-200 is used to evaluate the performance of the model. There are two different ways to compare the results in the literature using ADHD-200 data: classification comparison among different sites and classification comparison in comprehensive sites. The vast majority of literature only choose one method to explain the effectiveness of the experiment, meanwhile the evaluation indicators vary according to the comparison methods. In order to make the proposed method more persuasive, we employ each of the two methods to explain the experimental results.

1) Comparison of Classification Results Among Different Sites: The following ADHD classification algorithms were selected for comparison: (1) the 2017 ADHD-200 global competition champion algorithm (ADHD-2017) provided in [9]; (2) the deep belief network based ADHD classification algorithm (DBN) proposed in [14]; (3) the 3D-CNN based ADHD classification algorithm (3D-CNN) proposed in [12] (4) an R-RELIEF based ADHD classification algorithm (R-RELIEF) proposed in [52]. Table II shows the results of the proposed method and comparison algorithms on the test set, where "-" represents that the corresponding site had not been adopted, so the corresponding experimental results are none.

Table II shows that the accuracy of our method is the highest in different sites which is 70.59%-83.33%. Compared with ADHD-2017, the accuracy of different sites increased by 16.98-40.42%; the accuracy of DBN in NYU is only 37.41% while the accuracy in NYU is increased by 37.98% in our algorithm. And in KKI, the accuracy of our method is the same as R-RELIEF while the OHSU is added in our method and the rest of the sites are better than R-RELIEF.

In order to comprehensively demonstrate the effect of the model, the ROC curve and (Area Under the Curve) AUC are employed to evaluate the model. The ROC curve takes the false positive rate (i.e. specificity) as the abscissa and the true

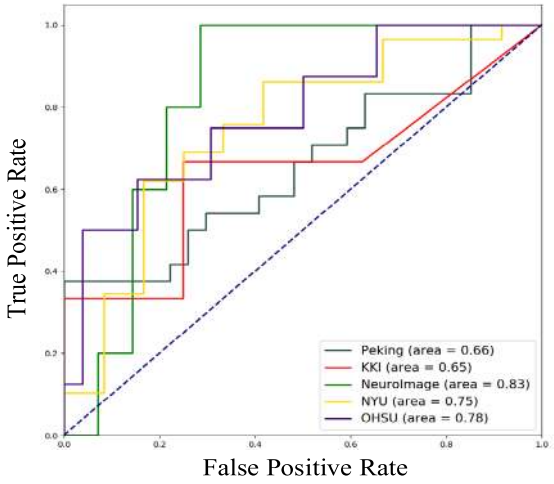

Fig. 11. ROC comparison of different sites.

TABLE III
COMPARISON OF DIFFERENT CLASSIFICATION METHODS

Methods	Accuracy	Sensitivity	Specificity
MKL	61.54%	41.33%	77.66%
MDA-SVM	62.81%	27.27%	83.12%
3D-CNN	69.15%	-	-
4-D CNN	71.30%	73.2%	69.7%
Our method	**75.64%**	**76.922%**	**73.08%**

positive rate (i.e. sensitivity) as the ordinate, which can reflect the trend of sensitivity (FPR) and specificity (TPR) of the model when selecting different thresholds. Compared with P-R curves (accuracy and recall rate), ROC curve has a huge advantage that when the distribution of positive and negative samples changes, its shape can basically remain unchanged, while the shape of P-R curve generally changes dramatically. This evaluation method can reduce the interference caused by different test sets and more objectively measure the performance of the model itself. AUC is the area under ROC curve. The larger the AUC value is, the better the model classification effect is. Fig. 11 shows the ROC curves and AUC values of the algorithm at different sites.

2) Comparison of Classification Results in Comprehensive Sites: In order to make the experiment more intact, we also attempt to test our model in all site datasets. The comparison methods are as follows: (1) The classification algorithm (denoted as MKL) by using multi-kernel learning fusion multimodal MRI features is proposed in [10]; (2) The classification algorithm (denoted as MDS-SVM) by using support vector machine after multi-dimensional scaling of functional connection network is proposed in [53]; (3) The algorithm of ADHD classification (denoted as 3D-CNN) based on 3D-CNN proposed in [12]; (4) The algorithm (denoted as 4D-CNN) based on 4D-CNN proposed in [15]. Table III shows that the proposed method yields the best results compared with others.

V. CONCLUSION

In this article, a new ADHD classification method based on fMRI is proposed, which can directly extract features from fMRI images to classify ADHD and TD. The experimental results at different sites show that the proposed method is superior to the existing methods in accuracy and can maintain a certain balance between specificity and sensitivity. Visualizing the feature maps of the middle layers shows that CDAE can effectively extract local information from spatial dimensions, which is helpful for classification. Although the pretraining of CDAE will increase the computational complexity of training and storage, it can effectively improve the performance in the classification of ADHD. In future work, we will focus on how to eliminate the impact of data heterogeneity on classification results as much as possible. Given the lack of utilization of the fMRI data as a time series (thereby implicitly ignoring time as an independent dimension), we will try to explore a better model and method to extract time dimension features.

REFERENCES

[1] J. Wiklund, C. Lomberg, L. Alkærsig, and D. Miller, "When ADHD helps and harms in entrepreneurship: An epidemiological approach," *Acad. Manage. Proc.*, vol. 2019, no. 1, p. 17481, 2019.
[2] J. C. Agnew-Blais, G. V. Polanczyk, A. Danese, J. Wertz, T. E. Moffitt, and L. Arseneault, "Young adult mental health and functional outcomes among individuals with remitted, persistent and late-onset ADHD," *Brit. J. Psychiatry*, vol. 213, no. 3, pp. 526–534, Jun. 2018.
[3] American Psychiatric Association A, "Diagnostic and statistical manual of mental disorders," in *Encyclopedia of the Neurological Ences*, vol. 25, no. 2. 1994, pp. 4–8.
[4] Z. Hawi, H. Yates, L. Kent, M. Gill, and M. Bellgrove, "A case-control genome wide association study of childhood attention deficit hyperactivity disorder (ADHD)," *Eur. Neuropsychopharmacol.*, vol. 29, p. 956, Mar. 2019.
[5] K. Rubia *et al.*, "Functional connectivity changes associated with fMRI neurofeedback of right inferior frontal cortex in adolescents with ADHD," *NeuroImage*, vol. 188, pp. 43–58, Mar. 2019.
[6] S. H. Hojjati, A. Ebrahimzadeh, A. Khazaee, and A. Babajani-Feremi, "Predicting conversion from MCI to AD by integrating rs-fMRI and structural MRI," *Comput. Biol. Med.*, vol. 102, pp. 30–39, Nov. 2018.
[7] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data," *Frontiers Neuroinform.*, vol. 13, p. 70, Nov. 2019.
[8] C.-Z. Zhu *et al.*, "Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder," *NeuroImage*, vol. 40, no. 1, pp. 110–120, Mar. 2008.
[9] *The ADHD-200 Global Competition*. Accessed: Oct. 1, 2017. [Online]. Available: http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html
[10] D. Dai, J. Wang, J. Hua, and H. He, "Classification of ADHD children through multimodal magnetic resonance imaging," *Frontiers Syst. Neurosci.*, vol. 6, p. 63, Sep. 2012.
[11] G. S. Sidhu, N. Asgarian, R. Greiner, and M. R. G. Brown, "Kernel principal component analysis for dimensionality reduction in fMRI-based diagnosis of ADHD," *Frontiers Syst. Neurosci.*, vol. 6, p. 74, Nov. 2012.
[12] 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.
[13] A. Riaz *et al.*, "Deep fMRI: AN end-to-end deep network for classification of fMRI data," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Washington, DC, USA, Apr. 2018, pp. 1419–1422.
[14] D. Kuang, X. Guo, X. An, Y. Zhao, and L. He, "Discrimination of ADHD based on fMRI data with deep belief network," in *Proc. Int. Conf. Intell. Comput. (ICIC)*, Cham, Switzerland, 2014, pp. 225–232.
[15] Z. Mao *et al.*, "Spatio-temporal deep learning method for ADHD fMRI classification," *Inf. Sci.*, vol. 499, pp. 1–11, Oct. 2019.
[16] S. Liu, T. Liu, L. Gao, H. Li, Q. Hu, J. Zhao, and C. Wang, "Convolutional neural network and guided filtering for SAR image denoising," *Remote Sens.*, vol. 11, no. 6, pp. 702–720, Mar. 2019.
[17] S. Liu, J. Wang, Y. Lu, S. Hu, X. Ma, and Y. Wu, "Multi-focus image fusion based on residual network in non-subsampled shearlet domain," *IEEE Access*, vol. 7, pp. 152043–152063, 2019.

[18] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 8928–8939.

[19] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," *Med. Image Anal.*, vol. 54, pp. 10–19, May 2019.

[20] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Bellevue, WA, USA, 2012, pp. 37–50.

[21] H. Li, L. Meng, J. Zhang, Y. Tan, Y. Ren, and H. Zhang, "Multiple description coding based on convolutional auto-encoder," *IEEE Access*, vol. 7, pp. 26013–26021, 2019.

[22] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[23] 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.

[24] Y. Feng, L. Zhang, and J. Mo, "Deep manifold preserving autoencoder for classifying breast cancer histopathological images," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 1, pp. 91–101, Jan. 2020.

[25] B. Hou, J. Yang, P. Wang, and R. Yan, "LSTM-based auto-encoder model for ECG arrhythmias classification," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1232–1240, Apr. 2020.

[26] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[27] J. Calvo-Zaragoza and A.-J. Gallego, "A selectional auto-encoder approach for document image binarization," *Pattern Recognit.*, vol. 86, pp. 37–47, Feb. 2019.

[28] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[29] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[32] X. Ouyang *et al.*, "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2595–2605, Aug. 2020.

[33] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, Jan. 2020.

[34] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4604–4616, Jul. 2020.

[35] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[36] S. M. Mousavi, W. Zhu, W. Ellsworth, and G. Beroza, "Unsupervised clustering of seismic signals using deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1693–1697, Nov. 2019.

[37] P. Liang, W. Shi, and X. Zhang, "Remote sensing image classification based on stacked denoising autoencoder," *Remote Sens.*, vol. 10, no. 2, p. 16, Dec. 2017.

[38] W. Jifara, F. Jiang, S. Rho, M. Cheng, and S. Liu, "Medical image denoising using convolutional neural network: A residual learning approach," *J. Supercomput.*, vol. 75, no. 2, pp. 704–718, Feb. 2019.

[39] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, Apr. 2016.

[40] H.-T. Chiang, Y.-Y. Hsieh, S.-W. Fu, K.-H. Hung, Y. Tsao, and S.-Y. Chien, "Noise reduction in ECG signals using fully convolutional denoising autoencoders," *IEEE Access*, vol. 7, pp. 60806–60813, 2019.

[41] U. Shruthi, V. Nagaveni, and B. K. Raghavendra, "A review on machine learning classification techniques for plant disease detection," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Coimbatore, India, Mar. 2019, pp. 281–284.

[42] Y. Chen *et al.*, "KNN-BLOCK DBSCAN: Fast clustering for large-scale data," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Dec. 18, 2019, doi: 10.1109/TSMC.2019.2956527.

[43] Y. Chen *et al.*, "Fast density peak clustering for large scale data based on kNN," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104824.

[44] H. Rao *et al.*, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Appl. Soft Comput.*, vol. 74, pp. 634–642, Jan. 2019.

[45] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, and L. Halawi, "Predictive models using decision trees," in *Applying Predictive Analytics*. Cham, Switzerland: Springer, 2019, pp. 123–144.

[46] P. Pandey and R. Prabhakar, "An analysis of machine learning techniques (J48 & AdaBoost)-for classification," in *Proc. 1st India Int. Conf. Inf. Process. (IICIP)*, Delhi, India, 2016, pp. 1–6.

[47] Y. Singhal, A. Jain, S. Batra, Y. Varshney, and M. Rathi, "Review of bagging and boosting classification performance on unbalanced binary classification," in *Proc. IEEE 8th Int. Advance Comput. Conf. (IACC)*, Greater Noida, India, Dec. 2018, pp. 338–343.

[48] M. Chakraborty, S. K. Biswas, and B. Purkayastha, "A novel ensembling method to boost performance of neural networks," *J. Exp. Theor. Artif. Intell.*, vol. 32, no. 1, pp. 17–29, Jan. 2020.

[49] Z. Yu *et al.*, "Semi-supervised ensemble clustering based on selected constraint projection," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2394–2407, Dec. 2018.

[50] C. Yan and Y. Zang, "DPARSF: A MATLAB toolbox for 'pipeline' data analysis resting-state fMRI," *Frontiers Syst. Neurosci.*, vol. 4, p. 13, May 2010.

[51] C. G. Yan, X. D. Wang, X. N. Zuo, and Y. F. Zang, "DPABI: Data processing & analysis for (resting-state) brain imaging," *Neuroinformatics*, vol. 14, no. 3, pp. 339–351, Apr. 2016.

[52] B. Miao, L. L. Zhang, J. L. Guan, Q. F. Meng, and Y. L. Zhang, "Classification of ADHD individuals and neurotypicals using reliable RELIEF: A resting-state study," *IEEE Access*, vol. 7, pp. 62163–62171, 2019.

[53] S. Dey, A. R. Rao, and M. Shah, "Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects," *Frontiers Neural Circuits*, vol. 8, p. 64, Jun. 2014.