Deep Forest in ADHD Data Classification

LIZHEN SHAO[ID], DONGHUI ZHANG, HAIPENG DU, AND DONGMEI FU
Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

Corresponding author: Lizhen Shao (lshao@ustb.edu.cn)

ABSTRACT Attention deficit hyperactivity disorder (ADHD) is a kind of mental disease which often appears among young children. Various machine learning techniques including deep neural networks have been used to classify ADHD. As an alternative of deep neural networks, the deep forest or gcForest recently proposed by Zhou and Feng has demonstrated excellent performance on many imaging tasks. Therefore, in this paper, we are going to investigate using fMRI data and gcForest to discriminate ADHD subjects against normal controls. Two types of features are extracted from the fMRI data, they are 1-D functional connectivity (FC) feature and 3-D amplitude of low frequency fluctuations (ALFF) feature. We propose a revised gcForest method which uses a combined multi-grained scanning structure to fuse the two features together, thus a new concatenated feature vector can be formed for each sample. Moreover, considering the imbalanced property of ADHD data, we utilize synthetic minority over-sampling technique combined with edited-nearest neighbor to form synthetic minority concatenated feature vector samples for data balancing. Finally cascade forest is used to take the concatenated feature vector samples as input for classification. We test our method on the ADHD-200 public data sets and evaluate its performance on the hold-out testing data. We compare our method with several methods in the literature. The experiment illustrates that our method performs better than the reported methods.

INDEX TERMS Attention deficit hyperactivity disorder, classification, deep forest, fMRI.

I. INTRODUCTION

Attention deficit hyperactivity disorder (ADHD) is a kind of mental disease which often appears among young children. ADHD is characterized by poor concentration, over activity or lack of self-control. It is reported that millions of people have been affected and about half of the grown patients are still influenced by the disease diagnosed in their childhood. However, the etiology is still unknown in most cases, without a clear diagnostic criteria, many children can not receive timely and proper treatment in the early stage of ADHD. Effective methods are urgently needed to assist the diagnosis of ADHD.

As neuroimaging technology, functional magnetic resonance imaging (fMRI) has been widely used to examine ADHD. FMRI measures brain activity by detecting changes associated with blood flow [1]. This technique relies on the fact that cerebral blood flow and neuronal activation are coupled. When an area of the brain is in use, blood flow to that region also increases. By detecting specific

encephalic regions, like dorsal anterior cingulate cortex (dACC), the ventrolateral prefrontal cortex (VLPFC) and the putamen, abnormal brain activations can be found. Cao et al. (2006) [2] found boys with ADHD has changed regional homogeneity (ReHo) in the frontal-striatal-cerebellar circuits region and occipital cortex. Castellanos (2008) [3] found ADHD-related diseases by examining functional connectivity (FC) between anterior cingulate and precuneus/posterior cingulate cortex regions. Zang (2007) [4] proved that children with ADHD have changed the amplitude of low frequency fluctuations (ALFF) in the right inferior frontal cortex, left sensorimotor cortex, and bilateral cerebellum. Yang et al. (2011) [5] investigated the amplitude of low frequency fluctuations (ALFF) of fMRI and demonstrated abnormal frontal activity in ADHD patient brain area. Tang et al. (2017) [6] used fractional amplitude of low frequency fluctuation (fALFF) to find the change of bilateral superior frontal cortex, anterior cingulate cortex (ACC), and several other brain areas in children with ADHD.

By utilizing fMRI data, various machine learning techniques have been used to diagnose ADHD. Riaz et al. (2017) [7] proposed a machine learning framework based on

---

The associate editor coordinating the review of this manuscript and approving it for publication was Omid Kavehei.

support vector machine (SVM) with integration of imaging data and non-imaging data to investigate functional connectivity alterations between ADHD and control subjects. Miao and Zhang (2017) [8] proposed a feature selection algorithm based on relief algorithm and verification accuracy (VA-Relief), which uses the feature subset obtained by pre-processing and feature selection of fractional amplitude of low-frequency fluctuation (fALFF) in resting-state functional magnetic resonance imaging (rs-fMRI). Du et al. (2016) [9] proposed a discriminative subnetwork selection method to mine frequent and discriminative subnetworks from ADHD and control group. The main features extracted from these discriminative subnetworks by using kernel principal component (PCA) were applied to the classification of ADHD. Qureshi et al. (2017) [10] calculated the global connectivity maps from the fMRI images and used the average of the connectivity measure of each atlas-based cortical parcellation as a feature for hierarchical extreme learning machine classifier input. Sen et al. (2018) [11] studied structural MRI image features and fMRI features, utilizing them to train a linear support vector machine (SVM) classifier for ADHD classification. Shao et al. (2018) [12] proposed a bi-objective ADHD classification scheme based on L1-norm support vector machine (SVM) model to achieve ADHD classification.

Recently, deep learning has also been used in ADHD classification. Kuang et al. (2014) [13] created a deep belief network with three layers to discriminate ADHD and normal control group. Zou et al. (2017) [14] built a 3-D convolutional neural networks (CNN) model to investigate the local spatial patterns of MRI features and designed a multi-modality CNN architecture to combine fMRI and sMRI features. Atif et al. (2018) [15] proposed a deep fMRI model which consists of three networks taking fMRI raw time-series signals as input. However, deep neural networks (DNNs) are with too many hyper-parameters and the performance depends heavily on parameter tuning.

As an alternative to deep neural networks, the deep forest or gcForest recently was proposed by Zhou and Feng [16]. It has been shown that the deep forest approach is highly competitive to deep neural networks. The deep forest uses a multiple layer structure where each layer contains many random forests. It is actually an ensemble of decision tree ensembles. In contrast to deep neural networks which require large-scale training data and great effort in hyper-parameter tuning, gcForest is easier to train and it works well even with small-scale training data. These characteristics make gcForest a suitable classifier for ADHD diagnosis.

Therefore, in this paper, we are going to investigate using gcForest to aid the diagnosis of ADHD. Our main contributions are as follows.

1) We propose a revised gcForest method which fuse 1-D FC feature and 3-D ALFF feature with multi-grained scanning to generate a concatenated transformed feature for classification. Compared with only using 1-D

TABLE 1. Description of the data sets of ADHD-200 competition.

Data sets	Peking	KKI	NYU	NI
Training subjects	194	83	216	48
ADHD subjects	78	22	118	25
Control subjects	116	61	98	23
Hold-out testing subjects	51	11	41	25
ADHD subjects	24	3	29	11
Control subjects	27	8	12	14

FC feature or 3-D ALFF feature, using the fused feature can improve the performance of classification.

2) We test our revised gcForest method on the ADHD global competition data sets, experimental results on the public hold-out testing data sets show that our method outperforms the reported methods in the literature.

The rest of the paper is organized as follows. In Section II, the procedure of data preprocessing and methods for computing functional connection and ALFF of fMRI are introduced. In section III, we make a brief description of gcForest, and then we propose our revised gcForest method for ADHD classification with fMRI data. Moreover, for data balancing, synthetic minority over-sampling technique (SMOTE) combined with edited-nearest neighbor (ENN) is also introduced to generate synthetic minority samples. In section IV, we show some experimental results and we compare our method with several reported methods in the literature. In section V, we draw the conclusion.

II. MATERIALS
The ADHD fMRI data we used are from ADHD-200 Global competition (http://fcon_1000.projects.nitrc.org/indi/adhd 200/index.html). We do experiments on four data sets, namely Peking (Peking University), KKI (Kennedy Krieger Institute), NYU (New York University Child Study Center) and NI (Neuro Image Sample). It needs to be noted that these data sets from different centers were collected with different parameter settings. A brief overview of these data sets is shown in Table 1.

We use DPARSF toolbox (http://rfmri.org/DPARSF) to perform data preprocessing. The preprocessing includes removing of the first ten images, slice time correction, motion correction, normalization, band pass filtration and smoothing.

Then according to the work of Tzourio-Mazoyer et al. [17], for each fMRI data, we divide the cerebella brain image into 90 brain regions. Each region in the cerebra is used to calculate an average time series of all voxels. For any pair of the average time series, we calculate the Pearson correlation coefficient to form a functional connection (FC) matrix [18]. The flowchart of FC matrix acquisition is shown in Figure 1. Since the FC matrix is a symmetric matrix, we use the lower left triangular of the matrix to form a feature vector. Simply by concatenating the first-row vector to the last-row vector of the lower left triangle, the feature vector
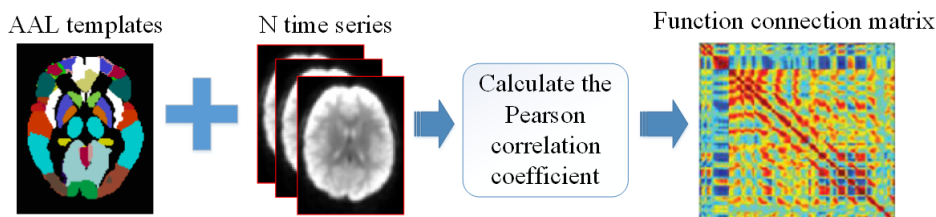
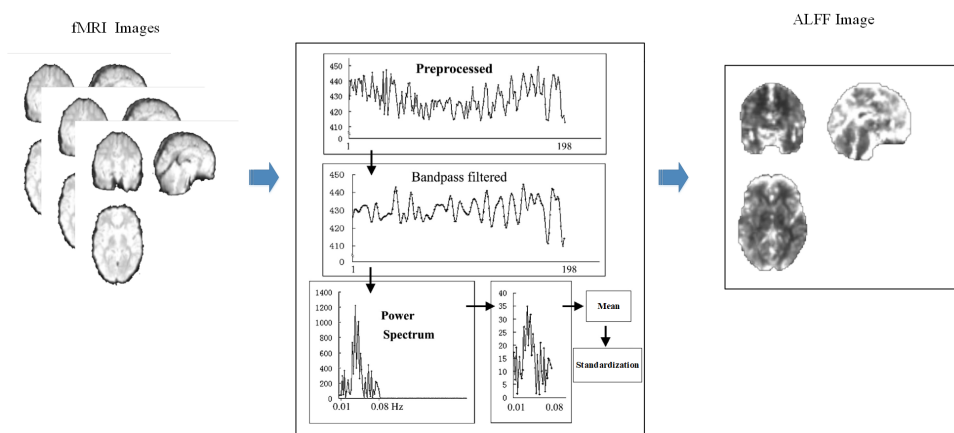**FIGURE 1.** Functional connection matrix acquisition flowchart.

FIGURE 2. ALFF image processing flowchart.

can be obtained. The dimension of the FC feature vector is $((90 \times 90 - 90)/2) = 4005$.

Moreover, we also obtain the ALFF image of an fMRI by REST (http://restfmri.net). The processing procedure for generating ALFF is shown in Figure 2. Firstly, the filtered time series corresponding to the fMRI voxels are transformed into frequency domain signals with fast Fourier transformation in order to obtain the power spectrum. Then the square root of the power spectrum at each frequency is calculated. The averaged square root across $0.01 \sim 0.08$Hz at each voxel is taken as the ALFF value. Finally, for standardization, the ALFF of each voxel is divided by the global mean ALFF value. For each fMRI sample in our experiment, we obtain its three dimensional ALFF image with the size of $61 \times 73 \times 61$.

III. METHODS
In this section, we first review the gcForest method, then we are going to show our revised gcForest method to classify ADHD with fMRI.

A. GCFOREST
Tree-based ensemble machine learning techniques like random forest [19] have advantages in dealing with nonlinear classification problems and overfitting. Recently, a new tree-based ensemble method called gcForest was proposed [16]. GcForest generates a deep forest ensemble and achieves high performance in representation learning and high dimensional data learning problems. GcForest has two major structures, namely multi-grained scanning and cascade forest. Given a set of raw input data, it will be processed by the multi-grained scanning to generate transformed concatenated feature vectors. Then the concatenated feature vectors will be fed into the cascade forest structure to achieve classification task.

Here we take a binary classification problem as an example to show how gcForest works. Suppose we have a data set with 100 dimensional features, i.e. the size of the data is 100×1. For each training sample, it will go through the multi-grained scanning structure (see Figure 3) to form a new concatenated feature vector. As it can be seen in multi-grained scanning, sliding window technology is used to construct new instances from the original data. Suppose that the sliding window size is 10×1 and the step size is 1, then for each sample, 91 of 10-dimensional instances would be generated. Next each instance is fed into two different forests respectively to output a 2-dimensional vector as class distribution vector. Furthermore we concatenate all the output class vectors as one. Therefore a new $91 \times 2 \times 2 = 364$-dimension transformed feature vector is obtained as the output of multi-grained scanning.

In cascade forest structure (see Figure 3), gcForest employs multilevel ensembles of decision tree forests. Each cascade level contains several forests, each forest will output a class distribution vector. Then we concatenate all the class vectors generated by the forests in the same level with the output of multi-grained scanning as the input vector to the next level.
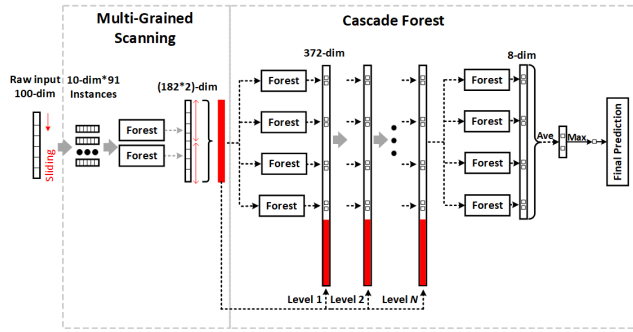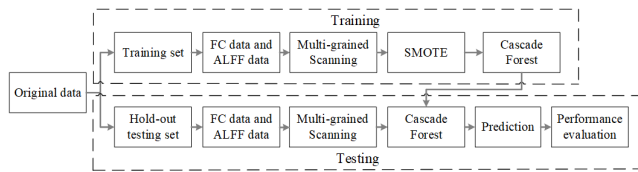
FIGURE 3. Structure of gcForest.

FIGURE 4. Flowchart for model training and testing.

For the binary classification example we considered, suppose there are four random forests in each level, we can see that each forest outputs a 2-dimensional vector and the output of each cascade level is a $4 \times 2 = 8$-dimensional vector. Concatenating the 8-dimensional vector with the the original 364-dimension feature, the input of the following cascade level is a 372-dimensional vector. Cascade levels are increased gradually until the convergence of validation performance. The final prediction would be the the max value in the averaged class vector obtained from the last cascade level. It should be noticed that forests in gcForest are not limited to normal random forests, which can be replaced by completely- random forests or other classifiers that can output class distribution vectors.

B. REVISED GCFOREST FOR FMRI CLASSIFICATION
The framework of the revised gcForest for fMRI data classification can be seen in Figure 4. For all the fMRI samples, we extract their FC and ALFF features. Then in the training process, the training data will go through multi-grained scanning, SMOTE with ENN for data balancing and cascade forest to train the model parameters. After the training, the testing data will go through the multi-grained scanning and the trained cascade forest to evaluate the performance of the classifier.

1) FEATURE-FUSED MULTI-GRAINED SCANNING STRUCTURE
In order to fuse the 1-D FC feature with the 3-D ALFF feature, in this section, we are going to propose a feature fusion structure, which consists of two multi-grained scanning, see Figure 5.

For each sample x_i, its FC feature is an 1-D array. A 1-D sliding window with fixed step size will be used for scanning the FC data into multiple instances. All these instances will be fed into a fixed number of random forests to output
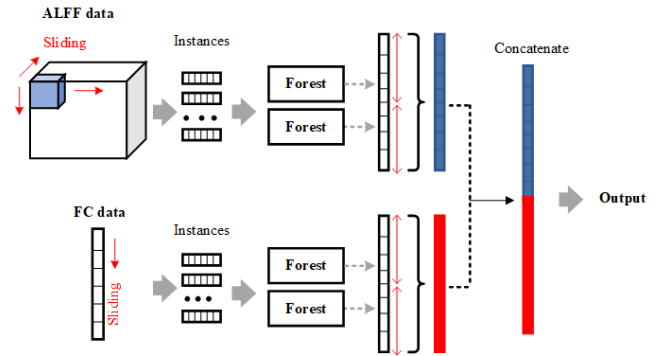

FIGURE 5. Feature-fusion multi-grained scanning.

class vectors. The concatenation of these class vectors is the final output of 1-D scanning.

Similarly, for the ALFF feature of sample x_i, since it is a 3-D array, a 3-D sliding window with fixed step size is used to generate multiple instances. All these instances will be input into a fixed number of random forests to output class vectors. Therefore, the final output of 3-D scanning for each sample is the concatenation of these class vectors.

Finally the output of the 1-D and 3-D multi-grained scanning are concatenated together to form a new vector as the transformed fused feature of FC and ALFF.

2) DATA BALANCING
The ADHD data sets we utilized in our experiments are imbalanced data sets, i.e., the positive and negative samples are not balanced. Since standard learning algorithm may generate suboptimal classifiers [20], it is necessary to deal with the imbalanced problem.

SMOTE algorithm is a kind of random over sampling algorithm which generates new synthetic samples by analyzing neighbors of minority samples. Suppose $S_A \in S$, where S is the set of all samples and S_A is the set of minority samples. SMOTE algorithm works as follows. For each sample $x_i \in S_A$, its k nearest in S_A is calculated. Then one of the k nearest sample y_i is randomly chosen, and the new synthetic minority sample can be calculated as:

$$x_s = x_i + (y_i - x_i) * r \tag{1}$$

where r is a random number between [0, 1] and x_s is the new synthetic sample. However, SMOTE algorithm generates new samples by using the original minority samples without considering its neighboring samples. It may generate minority samples that lie among majority samples. This may increase the overlapping between different classes, thus lead to poor classification results.

In order to improve the performance, in our study, SMOTE with ENN is adopted to balance the training sets. ENN is used to remove new synthetic samples that differ from two of its three nearest neighbors [21]. Figure 6 clearly shows the difference between SMOTE algorithm and SMOTE with ENN algorithm. Since the size of the original fMRI data
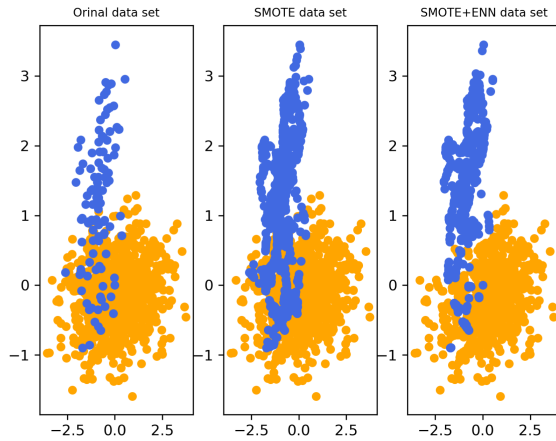
FIGURE 6. Comparison of SMOTE and SMOTE with ENN.

is too big, to save computational time, in our experiment, SMOTE with ENN is used after the multi-grained scanning to directly produce new synthetic transformed concatenated fused features for minority samples.

3) CASCADE FOREST

In the experiment, we used two random forests and two completely-random forests at each level of the cascade forest. During the training process, the transformed feature vector will be fed into the cascade forest to train the model parameters. During the testing process, the transformed feature vectors will be fed into the trained cascade forest model to output the prediction results.

IV. EXPERIMENT AND DISCUSSION

We have downloaded the original gcForest algorithm from the website http://lamda.nju.edu.cn/MainPage.ashx, and revised it for ADHD classification with fMRI data. We conduct the experiments on Peking, KKI, NYU, NI data sets. Different experiments as follows have been carried out to show the performance of the revised gcForest method.

A. RESULTS OF THE REVISED GCFOREST METHOD

For each data set, we use the training data set to train the model and the hold-out testing set to test the model. We have tried different parameter settings, and finally we select the one which gives the best performance.

The fused multi-grained scanning and data balancing with SMOTE with ENN are used. The sliding windows and step sizes used in the multi-grained scanning for different data sets are listed in Table 2. For all the data sets, one random forest which consists of 50 trees is used in the multi-grained structure, whereas two random forests and two completely-random forests, each contains 101 trees, are used in the cascade forest structure. It needs to be noted that different parameter settings are selected for different data sets, this may be due to that these data sets are collected from different centers.

TABLE 2. Sliding windows and step sizes of FC and ALFF features for the four data sets.

	Peking	KKI	NYU	NI
FC sliding window	100×1	200×1	200×1	200×1
FC step size	5	1	5	5
ALFF sliding window	$5 \times 5 \times 5$	$5 \times 5 \times 5$	$3 \times 3 \times 3$	$9 \times 9 \times 9$
ALFF step size	3	3	3	3

TABLE 3. The results of the revised gcForest method for different data sets.

Data sets	Peking	KKI	NYU	NI
ave	64.87%	82.73%	73.17%	72.00%
best	78.05%	90.91%	78.05%	80.00%

TABLE 4. Comparison results with/without data balancing.

Performance	Data balancing	Peking	KKI	NYU	NI
ACC	Yes	64.87%	82.73%	73.17%	72.00%
	No	63.92%	72.73%	72.93%	71.60%
SEN	Yes	64.72%	77.67%	83.00%	96.07%
	No	65.49%	0%	78.87%	93.57%
SPE	Yes	66.69%	87.23%	54.33%	67.21%
	No	63.31%	72.73%	54.91%	67.39%
G-means	Yes	65.52%	81.37%	67.07%	80.26%
	No	64.37%	0%	65.72%	79.19%

For each data set, the experiment was repeated ten times, the average and best accuracy value for each data set obtained by the revised gcForest method are shown in Table 3.

B. IMPACT OF DATA BALANCING

To investigate the impact of data balancing, experiments with/without SMOTE with ENN after the fused multi-grained scanning have been carried out. For each data set, the experiments with/without data balancing are repeated ten times. We calculate the average accuracy (ACC), sensitivity (SEN), specificity (SPE), and g-means values on the hold-out testing data sets, and list them in Table 4.

From Table 4, we can see that for KKI data set, the sensitivity and g-means values without data balancing are zeros. This may be due to that there are 22 ADHD subjects and 61 control subjects in the KKI training data set, which leads to the trained classifier overfitting. Thus the classifier predicts all samples as positive samples. Therefore, comparing the two results, it is obvious that data balancing can improve the performance of classification and thus effectively avoid overfitting.

C. COMPARISON RESULTS WITH FUSED FEATURES AND ONLY FC/ALFF FEATURES

We also compare the results of using fused features with the results of using only FC or ALFF feature. For each data set, we carry out the experiment ten times using only FC, only ALFF and the fused feature, respectively. The average accuracy and the best accuracy in the ten experiments are shown in Table 5.

TABLE 5. Average and best accuracy with different features for the four data sets.

Data sets	FC		ALFF		FC and ALFF	
	average	best	average	best	average	best
Peking	62.94%	68.63%	63.73%	70.59%	64.87%	78.05%
KKI	74.55%	81.82%	77.27%	90.91%	82.73%	90.91%
NYU	53.66%	60.98%	67.81%	73.17%	73.17%	78.05%
NI	58.00%	72.00%	68.00%	80.00%	72.00%	80.00%

TABLE 6. Average accuracy values for FC feature vectors with different orderings.

FC feature vector / Data sets	Peking	KKI	NYU	NI
original	62.94%	72.73%	53.66%	58.00%
ordered	62.75%	72.73%	54.39%	58.80%
R1	62.75%	72.73%	53.90%	58.40%
R2	62.09%	70.91%	53.17%	57.60%
R3	62.09%	71.82%	53.17%	58.00%
R4	62.75%	72.73%	53.90%	57.60%
R5	62.75%	72.73%	53.17%	58.40%

Comparing the three experimental results in Table 5, we can see that using ALFF feature obtains higher average (best) accuracy value than using FC feature, and using the fused feature obtains the highest average (best) accuracy value for all the data sets. The results also indicate that the training model can learn more useful information and be trained to be a better classifier using the fused feature compared with using only FC or ALFF features. It also needs to be noted that there are differences between the average accuracy values and the best accuracy values. This is due to that in each experiment multi-grained scanning will generate different feature vectors, therefore different training models and different accuracy values will be obtained.

D. IMPACT OF FC FEATURE ORDERINGS
The multi-grained scanning uses a sliding window along the input data. For the vectorized FC data, if the order of two brain regions are swapped, the multi-grained scanning will generate different features. Therefore, to investigate the impact of FC feature orderings to classification results, we test seven different FC feature vectors, namely

- the FC feature vector formed by concatenating the first-row vector to the last-row vector of the FC lower left triangle, labelled original
- the FC feature vector with elements being ordered from big to small, labelled ordered
- five randomly ordered feature vectors obtained by permuting the order of the original vector, labelled from R1 to R5.

In the experiment, instead of fused features, only FC features are used as the input of gcForst. The same hyper parameters as the ones shown in Table 2 are used. For each FC feature vector, the experiment was repeated ten times, the average accuracy value for each data set is shown in the Table 6.

From Table 6, we can see that in general the average accuracy values for the same data set are similar. For PU data set, the average accuracy value with the original FC vector

TABLE 7. Average accuracy on the hold-out testing data sets for different methods.

Data sets	Peking	KKI	NYU	NI
ADHD-200 Competition	51.05%	61.90%	35.19%	56.95%
Riaz et al [7]	64.70%	81.80%	60.90%	44.00%
3D-CNN [14]	62.95%	72.82%	70.50%	-
The proposed method	64.87%	82.73%	73.17%	72.00%

is higher than the ones with the ordered and R1 to R5. For KKI, the highest average accuracy value was achieved on the original, the ordered, R1, R4 and R5. For NYU and NI, the ordered feature vector has the highest accuracy value. Therefore, FC feature orderings do not affect the classification results much.

E. COMPARISON OF DIFFERENT METHODS
Finally, we compare our revised gcForest method that used fused feature with the method of ADHD-200 competition, and the methods of [7] and [14]. We list the average accuracy on the hold-out testing data sets in Table 7. From the results we can see that for all the data sets we tested our proposed method obtains the highest average accuracy value among all the methods. Especially with NI and NYU data sets, our method has obvious advantage.

V. CONCLUSION
In this paper, we have proposed a revised gcForest method to identify ADHD and control subjects. In order to combine FC and ALFF features, we have proposed a combined multi-grained structure so as to fuse ALFF and FC feature. Moreover, in order to handle the data imbalanced property, we used the SMOTE with ENN to generate minority samples. Experimental results on the KKI, Peking, NYU and NI data sets showed that our method did achieve superior performance than the reported methods in the literature on the hold-out testing data sets. Our method can also be applied to other disease diagnosis with fMRI data, such as Alzheimer's disease and autism etc. In this work, we used the average time series for each brain region to calculate FC. In the future, functional PCA type of approaches for fMRI data processing can be considered.

REFERENCES
[1] B. B. Biswal, J. Van Kylen, and J. S. Hyde, "Simultaneous assessment of flow and BOLD signals in resting-state functional connectivity maps," *NMR Biomed.*, vol. 10, nos. 4–5, pp. 165–170, 1997.

[2] Q. Cao, Y. Zang, L. Sun, M. Sui, X. Long, Q. Zou, and Y. Wang, "Abnormal neural activity in children with attention deficit hyperactivity disorder: A resting-state functional magnetic resonance imaging study," *NeuroReport*, vol. 17, no. 10, pp. 1033–1036, 2006.

[3] F. X. Castellanos, D. S. Margulies, C. Kelly, L. Q. Uddin, M. Ghaffari, A. Kirsch, D. Shaw, Z. Shehzad, A. Di Martino, B. Biswal, E. J. S. Sonuga-Barke, J. Rotrosen, L. A. Adler, and M. P. Milham, "Cingulate-precuneus interactions: A new locus of dysfunction in adult attention-deficit/hyperactivity disorder," *Biol. Psychiatry*, vol. 63, no. 3, pp. 332–337, Feb. 2008.

[4] Y.-F. Zang, H. Yong, Z. Chao-Zhe, C. Qing-Jiu, S. Man-Qiu, L. Meng, T. Li-Xia, J. Tian-Zi, and W. Yu-Feng, "Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI," *Brain Develop.*, vol. 29, no. 2, pp. 83–91, 2007.

[5] H. Yang, Q. Z. Wu, L. T. Guo, Q.-Q. Li, X.-Y. Long, X.-Q. Huang, R. C. K. Chan, and Q.-Y. Gong, "Abnormal spontaneous brain activity in medication-naïve ADHD children: A resting state fMRI study," *Neurosci. Lett.*, vol. 502, no. 2, pp. 89–93, 2011.

[6] C. Tang, Y. Wei, J. Zhao, X. Zhang, and J. Nie, "Abnormal brain activity in ADHD: A study of resting-state fMRI," in *Proc. Int. Conf. Brain Informat.* Cham, Switzerland: Springer, 2017, pp. 181–189.

[7] A. Riaz, M. Asad, E. Alonso, and G. Slabaugh, "Fusion of fMRI and non-imaging data for ADHD classification," *Computerized Med. Imag. Graph.*, vol. 65, pp. 115–128, Aug. 2018.

[8] B. Miao and Y. Zhang, "A feature selection method for classification of ADHD," in *Proc. 4th Int. Conf. Inf., Cybern. Comput. Social Syst. (ICCSS)*, Jul. 2017, pp. 21–25.

[9] J. Du, L. Wang, B. Jie, and D. Zhang, "Network-based classification of ADHD patients using discriminative subnetwork selection and graph kernel PCA," *Comput. Med. Imag. Graph.*, vol. 52, pp. 82–88, Sep. 2016.

[10] M. N. I. Qureshi, H. J. Jo, and B. Lee, "ADHD subgroup discrimination with global connectivity features using hierarchical extreme learning machine: Resting-state fMRI study," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 529–532.

[11] B. Sen, N. C. Borle, R. Greiner, and M. R. G. Brown, "A general prediction model for the detection of ADHD and Autism using structural and functional MRI," *PLoS One*, vol. 13, no. 4, 2018, Art. no. e0194856.

[12] L. Shao, Y. Xu, and D. Fu, "Classification of ADHD with bi-objective optimization," *J. Biomed. Informat.*, vol. 84, pp. 164–170, Aug. 2018.

[13] 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.* Cham, Switzerland: Springer, 2014, pp. 225–232.

[14] 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.

[15] A. Riaz, M. Asad, S. M. M. R. Al Arif, E. Alonso, D. Dima, P. Corr, and G. Slabaugh, "Deep fMRI: An end-to-end deep network for classification of fMRI data," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1419–1422.

[16] Z. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3553–3559.

[17] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[18] K. J. Friston, "Functional and effective connectivity: A review," *Brain Connectivity*, vol. 1, no. 1, pp. 13–36, 2011.

[19] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[20] G. E. Batista, R. C. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.

[21] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

LIZHEN SHAO received the M.Sc. degree in automation from the University of Electronic Science and Technology, China, in 2002, and the Ph.D. degree from the Department of Engineering Science, The University of Auckland, New Zealand, in 2008. She is currently an Associate Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. Her current research interests include optimization methods and application, pattern recognition, and machine learning.

DONGHUI ZHANG is currently pursuing the M.Sc. degree with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. His current research interests include big data analysis, machine learning, and medical imaging.

HAIPENG DU is currently pursuing the M.Sc. degree with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. His current research interests include multi-objective optimization, machine learning, and medical imaging.

DONGMEI FU received the M.Sc. degree from the Northwestern Polytechnical University, in 1984, and the Ph.D. degree from the University of Science and Technology Beijing, China, in 2006, where she is currently a Professor and a Doctoral Supervisor with the School of Automation and Electrical Engineering. Her research interests include automation control theory, image processing, and data mining.

• • •