IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 31, 2017, accepted September 30, 2017, date of publication October 13, 2017,
date of current version November 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2762703

3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI

LIANG ZOU[1,2], JIANNAN ZHENG[1], CHUNYAN MIAO[3], MARTIN J. MCKEOWN[4], AND Z. JANE WANG[1], (Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[2]School of Computing, Queen's University, Kingston, ON K7L 2N8, Canada
[3]School of Computer Engineering, Nanyang Technological University, Singapore 639798
[4]Pacific Parkinsons Research Centre, Department of Medicine (Neurology), The University of British Columbia, Vancouver, BC V6T 2B5, Canada

Corresponding author: Jiannan Zheng (jiannanz@ece.ubc.ca)

This work was supported by NPRP through the Qatar National Research Fund (a member of Qatar Foundation) under Grant 7-684-1-127. The statements made herein are solely the responsibility of the authors.

ABSTRACT Attention deficit hyperactivity disorder (ADHD) is one of the most common mental-health disorders. As a neurodevelopment disorder, neuroimaging technologies, such as magnetic resonance imaging (MRI), coupled with machine learning algorithms, are being increasingly explored as biomarkers in ADHD. Among various machine learning methods, deep learning has demonstrated excellent performance on many imaging tasks. With the availability of publically-available, large neuroimaging data sets for training purposes, deep learning-based automatic diagnosis of psychiatric disorders can become feasible. In this paper, we develop a deep learning-based ADHD classification method via 3-D convolutional neural networks (CNNs) applied to MRI scans. Since deep neural networks may utilize millions of parameters, even the large number of MRI samples in pooled data sets is still relatively limited if one is to learn discriminative features from the raw data. Instead, here we propose to first extract meaningful 3-D low-level features from functional MRI (fMRI) and structural MRI (sMRI) data. Furthermore, inspired by radiologists' typical approach for examining brain images, we design a 3-D CNN model to investigate the local spatial patterns of MRI features. Finally, we discover that brain functional and structural information are complementary, and design a multi-modality CNN architecture to combine fMRI and sMRI features. Evaluations on the hold-out testing data of the ADHD-200 global competition shows that the proposed multi-modality 3-D CNN approach achieves the state-of-the-art accuracy of 69.15% and outperforms reported classifiers in the literature, even with fewer training samples. We suggest that multi-modality classification will be a promising direction to find potential neuroimaging biomarkers of neurodevelopment disorders.

INDEX TERMS Attention deficit hyperactive disorder, 3D CNN, magnetic resonance imaging, multi-modality analysis.

I. INTRODUCTION

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common mental-health disorders, affecting around 5%-10% of school-age children [1]. ADHD can be characterized by excessive impulsive, hyperactive or inattention behaviors. These symptoms begin at an early age and may continue through to adulthood, leading to serious impairments, as well as inducing substantial burdens for families and society. The traditional diagnosis of ADHD mainly depends on clinical ratings of behavioral symptoms, which can be unreliable [2], [3]. For instance, the diagnosis criteria for children is mainly based on report of behavior from parents or teachers. Therefore, an accurate biomarker based on non-invasive imaging would be highly valuable.

In order to automatically diagnose neurological and psychiatric disorders, such as ADHD, a multitude of features extracted from fMRI have been proposed. These features can be categorized into voxel-level features and region-level features. Yang *et al.* investigated the Amplitude of Low Frequency Fluctuations (ALFF) [4] and demonstrated abnormal frontal activity in ADHD [5]. Long *et al.* extracted Regional Homogeneity (ReHo) [6] and ALFF features from fMRI data and employed these features to classify early Parkinson's disease [7]. Although these voxel-level features

2169-3536 © 2017 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

are simple and intuitive to extract, these features usually have very high dimensionality, thus feature selection is generally needed before classification [8]. Alternatively, many researchers employ a hypothesis-driven approach and consider only certain predefined regions and extract region-level features from these regions. For instance, Eloyan *et al.* investigated the functional connectivity (FC) between five regions of the motor cortex and analyzed the connectivity matrix to diagnose ADHD [2]. Unfortunately, compared with voxel-level features, these low dimensional region-level features are generally insensitive to subtle changes involved in neurological disorders. In addition, disease-related changes may occur in part of a region or across multiple regions. Therefore, examining solely simple voxel-level or region-level features may not effectively capture disease-specific pathologies.

Although ADHD is not associated with gross morphological changes in the brain, several studies have shown that subtle anatomical differences associated with ADHD can be found in MR images [9]–[11]. For instance, Kobel *et al.* reported significant changes in volume of the cerebral cortex between children with ADHD and typically developing children (TDC) [10]. The differences between sMRI of ADHD and TDC also suggest that sMRI may be an important classification feature to diagnose ADHD. Compared with fMRI, sMRI is less sensitive to noise and requires simple preprocessing steps. In addition, sMRI can be acquired with better spatial resolution. Thus sMRI and fMRI may provide complementary information about brain changes in pathology. Therefore, to further improve the classification accuracy, we set forth to develop an ADHD classification method using both sMRI and fMRI jointly. The combination of different measures (i.e., fMRI and sMRI) may increase the reliability of classification [12].

Previous studies have explored deep learning methods on ADHD classification. Kuang *et al.* first introduced a Deep Belief Network (DBN) with three hidden layers to discriminate ADHD, utilizing frequency domain features in fMRI [13]. However, the conventional one-dimension neural network (e.g., DBN), which employs a vector as the input, generally neglects the topological information of the input data. Yet radiologists typically navigate through 2D planes for diagnostic purposes, using local 3D patterns of neural images across the brain to formulate a diagnosis [8]. Thus automated approaches which utilize local 3D patterns from the whole brain, rather than from an individual voxel or predefined region, may contribute to the diagnosis of neurological disorders [14], [15]. Inspired by the way that radiologists examine brain images, in this paper we design a 3D convolutional neural network (CNN) model to learn hierarchical spatial patterns to diagnose ADHD from fMRI and sMRI features. We first extracted 6 types of 3D features, including 3 types of functional features and 3 types of morphological features. More specifically, we extract 3 low-level features from fMRI data: ReHo, fractional ALFF (fALFF) [16] and Voxel-Mirrored Homotopic Connectivity (VMHC) [17]; and 3 voxel-based morphometry features: gray matter (GM), white matter(WM) and cerebrospinal fluid (CSF) probabilities of each voxel in Montreal Neurological Institute (MNI) space from sMRI data. Furthermore, we employ the 3D CNN model [18], [19] to learn latent 3D local patterns from individual 3D low-level features or the combination of these features to boost classification performance. Finally, we demonstrate that fMRI and sMRI features are complementary, and design a multi-modality architecture to optimize classification accuracy. The performance on the independent hold-out testing dataset demonstrates that the proposed 3D CNN approach outperforms state-of-the-art methods described in the literature, even with fewer training samples.

In summary, there are three main contributions of this paper:

1) We retain MRI spatial information throughout the learning process. Rather than representing low-level features (including ReHo, fALFF and VMHC, as well as the density of GM, WM and CSF in MNI space) as vectors, we keep these low-level features as 3-order tensors (also called 3-dimensional arrays). Inspired by the way that radiologists examine brain images, we design 3D CNN models to learn hierarchical 3D patterns to classify ADHD and show promising results.
2) We investigate and summarize both fMRI and sMRI features' strength in the diagnosis of ADHD. We find that 3D CNN using GM density from sMRI achieves the highest classification accuracy on a test dataset.
3) We find that fMRI and sMRI features are complementary, and design a multi-modality 3D CNN architecture to combine features from both fMRI and sMRI. The proposed multi-modality 3D CNN approach achieves state-of-the-art accuracy of 69.15% on testing data from the ADHD-200 global competition, demonstrating the importance of incorporating both structural and functional images for diagnosis of neurodevelopment disorders.

The rest of this paper is organized as follows. Section II will discuss related works on ADHD diagnosis. Section III will introduce the ADHD-200 dataset. Section IV will discuss the proposed methods. Section V will show the experiment results of the proposed methods and comparison with existing methods. Section VI and Section VII draw the discussion and the conclusion of the paper.

II. RELATED WORK

MRI, including fMRI and sMRI, has been investigated for ADHD diagnosis in many studies [20]–[22]. For instance, Zhu *et al.* trained a classifier based on Fisher-discriminant-analysis (FDA) using fMRI scans from 24 subjects (12 TDC and 12 ADHD) and achieved a leave-one-out cross-validation accuracy of 85%. However, the number of samples utilized in these studies is relatively small, possibly affecting the generalizability of the findings [2]. In order to accelerate the understanding of the neural basis of ADHD and obtain objective diagnosis methods, the ADHD-200 consortium publicly released a large-scale neuroimaging dataset along with associated phenotypic information. They further released a hold-out testing dataset and held the ADHD-200 global competition in 2011 [23]. Twenty-one international teams, from different scientific disciplines, joined the competition and submitted their diagnostic labels. Accuracies derived by internal cross-validation ranged from 55%-78%, however, the accuracies reported on the external hold-out test dataset were substantially lower. Teams were ranked based on the diagnosis accuracy on the hold-out testing dataset, and out of these 21 teams, the best binary classifier based on the neuroimages achieved a diagnostic accuracy of 61.54% [24].

Subsequent to this competition, researchers have continually worked on automatic diagnosis of ADHD based on the ADHD-200 competition dataset. In [25], Dai *et al.* employed different image processing techniques to extract multimodal features, including features from both structural MRI and fMRI. For sMRI, they extracted Cortical Thickness (CT), and Gray Matter Probability (GMP) while for fMRI, ReHo, and FC were extracted. They compared the effects of using different features against each other. In addition, they further integrated multimodal image features using MKL and obtained the diagnosis accuracy of 61.54%, which is comparable to the best result in the competition. Ghiassian *et al.* introduced histogram of oriented gradients (HOG) in visual object recognition to the study of ADHD diagnosis [26]. To avoid overfitting in the training of classifiers, they selected the most relevant 211 features by MRMR (Maximum Relevance Minimum Redundancy) from 116480 possible features. They evaluated several classifiers and found that a Support Vector Machine (SVM) achieved the best classification performance of 62.57%. Dey *et al.* proposed a novel framework for the automatic diagnosis of ADHD based on brain functional connectivity networks. They firstly selected a sequence of highly active voxels and construct the connectivity network between them. They obtained an average accuracy of 62.81% on the hold-out testing dataset when classification was performed on all the subjects. They also concluded that the performance can be improved by incorporating gender information. In [27], Guo *et al.* explored the functional connectivity between voxels and obtained an average accuracy of 63.75% based on a social network analytic method. These results represent the highest diagnostic performance to date on the ADHD-200 hold-out test dataset.

III. EXPERIMENT DATASET

The fMRI data analyzed in this paper is from the ADHD-200 consortium. Initially, they made available a large training dataset consisting of 776 fMRI scans and associated T1-weighted structural scans. Among them, 491 were obtained from typically developing individuals and 285 from patients with ADHD (ages: 7-21 years old). Characteristic information of subjects were also provided, including age, gender, handedness and IQ scores. The data were collected by 8 institutions around the world and were shared anonymously without any protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) guidelines and the 1000 Functional Connectomes Project (FCP) protocols [28]. We refer to this dataset as the "original training dataset" excluding 108 subjects whose fMRI data were regarded as with the 'questionable' quality by the data curators. For the ADHD-200 global competition, the ADHD-200 consortium released a hold-out dataset from 94 TDC and 77 ADHD patients as well as 26 participants without diagnostic information. We refer the subset of this dataset as the "hold-out testing dataset" consisting of 171 subjects for whom diagnostic data were released. Details of scan parameters, diagnostics criteria and other site-specific protocols are available at http://fcon_1000.projects.nitrc.org/indi/adhd200/.

Raw data sharing demands intensive coordinating efforts, huge manpower and large data storing/management facilities. In addition, the preprocessing of medical images frequently requires professional medical knowledge which may be a barrier for other scientific communities (such as machine learning experts) to join in the field of neuroimaging. To address these concerns, Chaogan *et al.* initiated the R-fMRI maps project (http://mrirc.psych.ac.cn/RfMRIMaps) [29] and encouraged scientists to share the preprocessed data through this project. For the ADHD-200 dataset, they preprocessed the entire hold-out testing dataset and a subset of the original training dataset. In this work, we refer this as the "preprocessed training dataset" and train the 3D CNN based on this dataset. For details of these datasets, please see Table 1 below.

All resting-state functional MRI images were preprocessed using Data Processing Assistant for RestingState fMRI (DPARSF) programs [30]. The following steps were performed:

1) Slice timing correction;
2) Head motion correction by realigning for each volume relative to initial one;
3) Regress out the nuisance covariates, such as regressing out head motion effects from the realigned data;

TABLE 1. Some details of the datasets utilized in this paper.

Dataset	ADHD (male)	TDC (male)	In Total (male)
Original training dataset	239 (188)	429 (225)	668 (413)
Hold-out testing dataset	77 (60)	94 (46)	171 (106)
Preprocessed training dataset	197 (158)	362 (190)	559 (348)

4) Spatially coregistered (normalized) to standardized space;
5) Voxel-wise band pass filtering (0.01-0.1Hz, which is regarded as the traditional bandpass frequency range for resting state fMRI);
6) Normalization of anatomic images to MNI template space using unified segmentation of anatomic images;
7) Smoothing with a 4mm Full Width at Half Maximum (FWHM) Gaussian kernel.

IV. PROPOSED METHOD

In order to automatically classify ADHD, after data preprocessing steps, our framework starts by extracting low-level 3D fMRI and sMRI features as illustrated in Fig 2. The CNN networks and softmax classifier are then trained to distinguish ADHD cases from TDC cases. In this section, we will present our automatic ADHD classification framework in detail.

A. LOW-LEVEL FEATURE EXTRACTION BASED ON PRIOR KNOWLEDGE

Considering the fact that the number of subject samples is still relatively fewer than the potential millions of parameters in a DNN, we first encode prior knowledge and extract 3 types of popular low-level features from fMRI scans, including ReHo, fALFF and VMHC. We further exclude boundary areas of these three types of features (which are filled with zeros) and extract a cube with the size of 47 × 60 × 46. All voxels within the brain are presented graphically in Fig. 1 by color-coding the region that each voxel belongs to based on the Automated Anatomical Labeling (AAL) atlas [31].

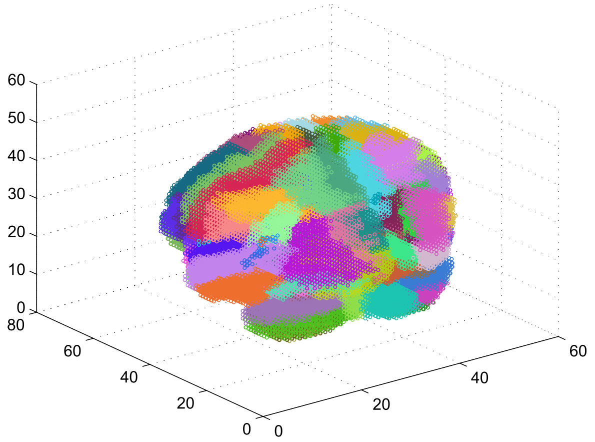

Fig. 1. Illustration of the voxels within the whole brain. Each color represents a specific brain region defined by the Automated Anatomical Labeling (AAL) atlas.

Regional Homogeneity (ReHo) maps local brain activity across the whole brain and has been used to detect abnormal neural activity in children with ADHD [25]. It measures the functional synchronization of a given voxel with its nearest neighbors.

fALFF has been successfully utilized to detect the abnormal spontaneous brain activity of various neuropsychiatric disorders, such as ADHD, Parkinson's disease and schizophrenia. It measures the ratio of power in the low-frequency (0.01Hz-0.1Hz) range to that of the entire detectable frequency range. fALFF is the normalized amplitude of low-frequency fluctuations (ALFF). It provides a more specific measure in detecting spontaneous brain activity [16].

Functional homotopy is a fundamental characteristic of the brain's functional architecture. In this paper, voxel-mirrored homotopic connectivity (VMHC) is evaluated, which quantifies functional homotopy by providing a voxel-wise measure of connectivity between hemispheres. Recently, VMHC was used to analyze the group difference between children with and without ADHD [32].

In addition, 3D low-level morphological features are extracted through voxel-based morphometry (VBM) analysis of high-resolution T1-weighted images. In this paper, we also employ the relative probability densities of GM, WM and CSF in MNI space as inputs to the 3D CNN. These three kinds of morphological features are derived from image segmentation [33]. After segmentation, each voxel contains three measures of the probabilities, according which it belongs to specific segmentation classes, corresponding to GM, WM and CSF respectively. We further exclude boundary areas of these three types of features (which are filled with zeros) and extract a cube with the size of 90 × 117 × 100.

MRI data preprocessing and feature extraction were performed with DPARSF. All the data and features used in this work are publicly available and can be downloaded through the R-fMRI maps project [29]. While other types of features can also be incorporated into the proposed approach, we focussed on the above frequently-used 6 types of low-level features to illustrate the proposed method. Fig. 2 shows the flowchart for ADHD classification based on fMRI and sMRI using 3D CNN.

B. 3D CONVOLUTIONAL NEURAL NETWORKS

Similar to traditional deep learning architectures, CNN models are hierarchical architectures where several convolutional layers are stacked on top of each other. Traditional CNNs have 2D convolutional kernels for applications on 2D images. However, it is challenging to apply 2D CNNs on 3D data because convolutions in a 2D CNN only can capture 2-dimensional spatial information, and neglect the information along the third dimension. To address this concern, Ji *et al.* extended the idea of 2D CNN used for 2D images to a 3D convolution in both space (2D) and time for video classification [19]. This approach can effectively incorporate motion information in video analysis [18], [19]. Similar to video data (x,y,t), the extracted low level features mentioned in Section IV-A have 3 dimensions (x,y,z). Therefore, in this paper, we employ 3D convolutions to learn the 3D local patterns across the whole brain to assist the diagnosis of ADHD.

Compared to fully-connected DBNs, convolutional layers have two main properties: partial connectedness and weight

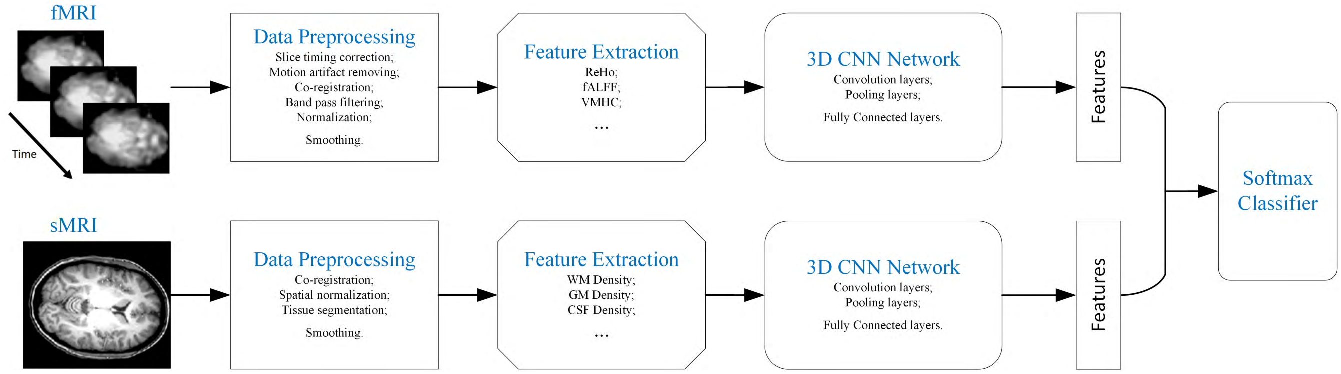

Fig. 2. A flowchart for ADHD classification based on fMRI and sMRI using 3D CNN.

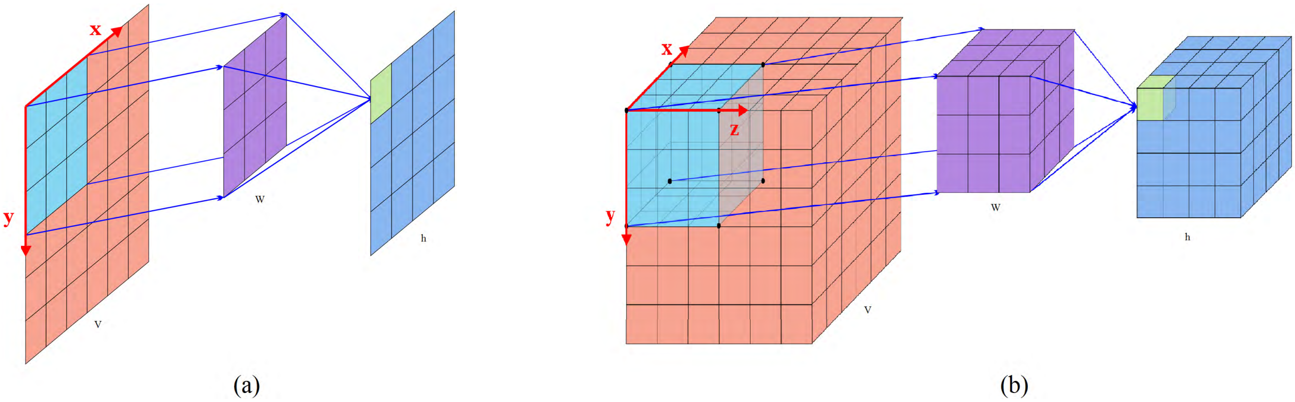

Fig. 3. Differences between the 2D convolution and the 3D convolution. (a) 2D convolution: $h_{1,1} = \sum_{x=1}^{3}\sum_{y=1}^{3} W_{x,y}V_{x,y} + b$; (b) 3D convolution: $h_{1,1,1} = \sum_{x=1}^{3}\sum_{y=1}^{3}\sum_{z=1}^{3} W_{x,y,z}V_{x,y,z} + b$, where W is the weight of the kernel, V is the feature map in the previous layer and b is the bias term.

sharing [15]. In the convolutional layer, unlike in the DBN, an output neuron is connected only to a local region of the input feature maps. This property reduces the number of parameters, thus making the CNN less prone to overfitting. Another benefit of this approach is that the convolutional layer can retain local spatial patterns which may be appropriate for image related tasks. The weights sharing property means weights in convolutional kernels are shared across the whole spatial region of the feature maps which further reduces the number of parameters and increases the generalization capability of the network. It is common to periodically insert a pooling layer between successive convolutional layers in a CNN. The pooling operation reduces the spatial size of the feature maps and the number of parameters. As shown in Fig.3, the 2D CNN are applied on 2D features maps to extract the spatial features, whereas, to detect the 3D local patterns in our case, 3D kernels are convolved over 3D feature cubes. More specifically, for the case of 3D CNN, the value at position (x, y, z) on the j^{th} feature map in the i^{th} layer is obtained as follows,

$$h^{i,j}_{x,y,z} = f((W_{i,j} * V^{i-1})_{x,y,z} + b_{i,j}), \tag{1}$$

where $W_{i,j}$ and $b_{i,j}$ are the weights and the bias for j^{th} feature map respectively, V^{i-1} denotes the sets of input feature maps from the $(i-1)^{th}$ layer connected to the current layer, f is the non-linear function and $*$ is the convolution operation. In the training process, all the weights of these convolutional kernels in the CNN, **W**, together with the bias **b** are optimized with respect to a given loss function.

A 3D convolutional layer is effective in learning local patterns and exploring spatial information across 3D input images [14]. The complexity of learned local patterns is closely related to the numbers of 3D convolutional kernels in the network. With more kernels, the network can learn deeper and more powerful features, but on the other hand will be more susceptible to overfitting. A general principle is that a network should have sufficient convolutional layers to learn deeper features, and fewer numbers of feature maps in each layer to limit overall complexity [34], [35].

C. SINGLE MODALITY 3D CNN ARCHITECTURE

For the three fMRI features and three sMRI features mentioned in Section IV-A, we design a universal single modality 3D CNN architecture to classify ADHD. The architecture takes either fMRI features (including ReHo, VMHC and fALFF) with the dimension of $47 \times 60 \times 46$, or sMRI features (including GM, WM and CSF) with the dimension of $90 \times 117 \times 100$ as input. In this paper, we first reduce the feature

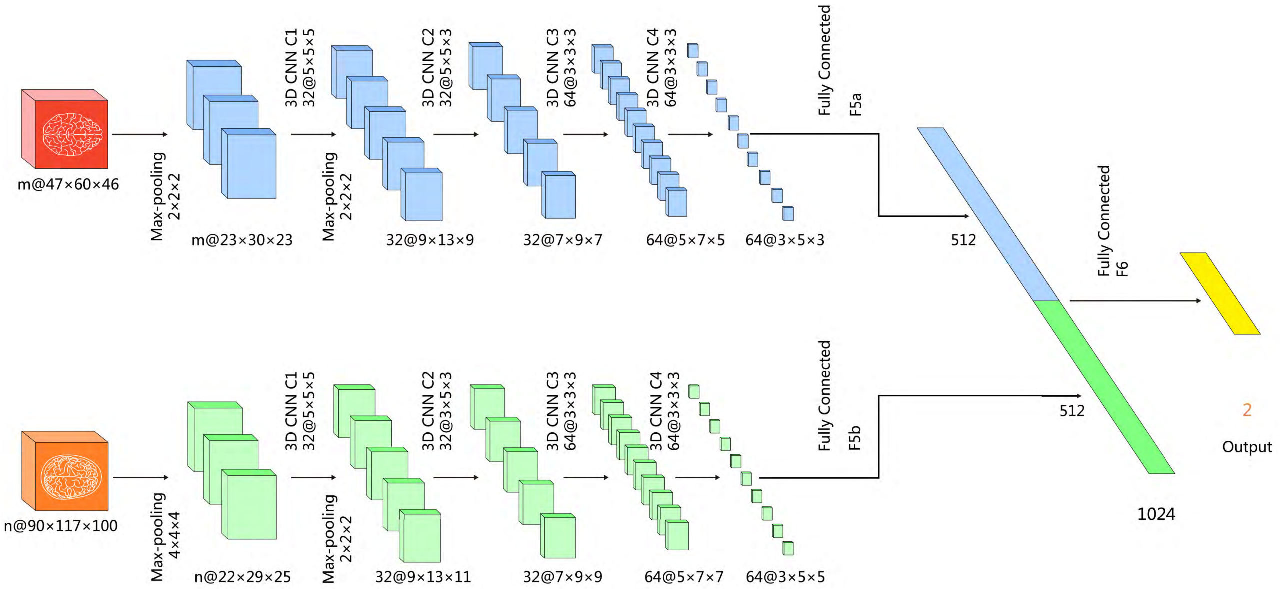

Fig. 4. Architecture of the proposed 3D CNN for diagnosing ADHD. We utilizes six types of 3D features across the whole brain as the inputs, including ReHo, fALFF and VMHC from fMRI as well as the density of GM, WM and CSF in MNI space from sMRI. This architecture contains 6 layers, including four convolutional layers and two fully connected layers.

map size with max-pooling (2 × 2 × 2 for fMRI features and 4 × 4 × 4 for sMRI features), which reduces the three-dimension spatial resolution of the input to 23 × 30 × 23 for fMRI features and 22×29×25 for sMRI features respectively. In our preliminary experiments, we observed that this setting can boost performance, improve network generalization by greatly reducing the number of parameters, and dramatically reduce computational load. We then train 32 different 3D kernels with size of 5×5×5 on all three channels as at the first convolutional layer $C1$. We further down-sample the feature map size with max-pooling. The output feature maps after these layers are made up of 32 feature maps of size 9 × 13 × 9 and 9 × 13 × 11 for fMRI and sMRI features respectively. Additional 3 convolutional layers, $C2$, $C3$ and $C4$, are further employed to learn deeper feature with 64 output maps. After these four convolutional layers, the output feature maps are fully-connected to 512 neurons in $F5$. $F6$ is the last layer and topped with a softmax activation function to output the probabilities of two classes, i.e., ADHD and TDC.

Similar to most deep learning problems, the choice of the specific network architecture here is problem-dependent. In our preliminary work, we have tested a variety of 3D architectures with different number of convolutional layers and kernel sizes. The 3D architecture described above yields the best performance on the ADHD-200 Dataset.

D. MULTI-MODALITY 3D CNN ARCHITECTURE

Although single modality 3D CNN on fMRI or sMRI improves the classification performance over the existing methods in our experiments, one should note that fMRI and sMRI carries significantly different information. Considering the complicated pathologic process of ADHD, it is reasonable to assume morphometric and functional changes simultaneously in the brain of ADHD children. Therefore, these two types of features can be complementary and combining them could boost enhance the ability to classify ADHD. In this section we present a multi-modality 3D CNN architecture which incorporates both fMRI and sMRI features as input to a 3D CNN training framework.

Fig. 4 demonstrates the proposed multi-modality 3D CNN architecture. The architecture contains two separate branches for fMRI features (top) and sMRI features (bottom). Each branch takes MRI features as inputs and learns a 512-dimension feature vector through back-propagation. The branches have the same CNN structure as in the previous single modality CNN architecture: 4 convolutional layers, 2 max-pooling layers and 1 fully-connected layer. The output 2 512-dimension feature vectors are then concatenated and fed to a fully-connected layer with output size 2 (2 classes – ADHD and TDC). This multi-modality architecture has three advantages: 1) fMRI and sMRI features have different size of feature maps, thus a single modality architecture is not able to combine these two modalities; 2) since fMRI and sMRI carries different information, the two branches are able to learn hierarchical CNN features for fMRI and sMRI separately without interfering with one another; 3) this architecture also enables joint training of feature extractor and classifier, which has proven to be more effective than training them separately. As a result, the proposed multi-modality 3D CNN architecture yields better performance.

E. TRAINING OF THE 3D CNN ARCHITECTURE

The training of the above architectures is carried out by optimizing a loss function via updating the network's parameters $\{\mathbf{W}, \mathbf{b}\}$. In this paper, we select the cross-entropy as the loss function, which is defined as follows,

$$L(\mathbf{W}, \mathbf{b}) = -\frac{1}{N}(\sum_{n=1}^{N} y_n \ln H_{\mathbf{W},\mathbf{b}}(x_n) + (1 - y_n) \ln(1 - H_{\mathbf{W},\mathbf{b}}(x_n))), \quad (2)$$

where N is the number of samples, x_n and y_n are the input and corresponding label of the n^{th} sample, $H_{\mathbf{W},\mathbf{b}}(\cdot)$ is the function learned by the network and $H_{\mathbf{W},\mathbf{b}}(x_n)$ represents the output of the neural network given the input x_n. The weights of the 3D convolutional networks are randomly initialized based on the Xavier initialization strategy [36]. The 3D CNN architecture then is trained via stochastic gradient descent (SGD) with mini-batches of 20 training samples. The weights $\mathbf{W}$ are updated for every mini-batch as:

$$\begin{aligned} \nabla_{\mathbf{W}_t} &= \langle \nabla_{\mathbf{W}_t} L(\mathbf{W}_t) \rangle_{mini-bacth} \\ \mathbf{v}_{t+1} &= \gamma \mathbf{v}_t - \alpha \nabla_{W_t} \\ \mathbf{W}_{t+1} &= \mathbf{W}_t + \mathbf{v}_{t+1} \end{aligned} \quad (3)$$

where $\mathbf{v}_t$ is the current velocity vector, α is the learning rate and γ is the momentum. Momentum accumulates the velocity vector in directions of persistent reduction in the objective across iterations and accelerates the training process.

The large number of parameters existing in our network made it susceptible to overfitting. Besides taking advantage of the intrinsic features of 3D CNN architecture, such as partial connectivity, weights sharing and pooling, we additionally adopt several methods to further avoid overfitting. We use the dropout technique [37] with a probability of 0.5 in the fully connected layers. During drop-out, the inputs of layers $F5$ and $F6$ are randomly set to 0 with a probability 0.5. This drop-out procedure is a variant of data augmentation and has been proved to be an effective way to reduce overfitting in deep neural networks [37]. Batch normalization (BN), which is a regularization technique, can guarantee faster and better convergence of network training. We added BN layers after every convolutional layer and fully-connected layer in our architectures.

V. EXPERIMENTS AND RESULTS

A. EXPERIMENT SETUP

To evaluate the proposed single and multi-modality 3D CNN architectures, we used the ADHD-200 dataset where 559 subjects are used for training the 3D CNNs. We then tested the performance of trained models on the hold-out testing dataset with 171 subjects. Percentage prediction accuracy of the two-class diagnosis (TDC vs. ADHD) was used for evaluation and comparison of the proposed methods and previously-reported methods. In the experiment, we report 6 single feature approaches for each of the fMRI features (ReHo, fALFF and VMHC) and sMRI features (GM, WM and CSF); 2 combined approaches with 3 fMRI features and 3 sMRI features separately (fMRI-all and sMRI-all); and 2 multi-modality approaches with both fMRI features and sMRI features (All and fALFF+GM). We split the training dataset into 4 folds for cross validation. We set the learning rate to be 0.0001, and decayed the learning rate after every 20 epochs of training with a factor of 0.5. We set the batch size to be 20 and trained each approach for 100 epochs to ensure training was converged at the end. With the random effects of the dropout technique as well as the random initialization of the network parameters, we repeated the experiments for 50 times and report the average accuracy.

B. COMPARISON OF SINGLE AND MULTI-MODALITY ARCHITECTURES

Fig. 5 shows the statistical results of 3D CNN approaches. First, we compared the results of single modality approaches. The single modality 3D CNN architecture with fALFF achieved a mean accuracy of 66.04% while the model with GM achieved 65.86%. Surprisingly, combining all 3 fMRI features or all 3 sMRI features did not yield better performance than fALFF and GM. One possible explanation of this phenomenon is that ReHo, VMHC, WM and CSF may not contain additional complementary information for fALFF and GM, and combining different fMRI features or sMRI features within a single modality 3D CNN architecture will not benefit ADHD classification.

We then evaluated the multi-modality approaches. Two approaches were evaluated in this test: 1) all 3 fMRI features and 3 sMRI features were used as input; 2) since fALFF and GM achieved superior performance, we combined only these two features. In fact, fALFF+GM improved over fALFF and GM by a large margin, and achieved state-of-the-art performance on the ADHD-200 dataset with an average accuracy of 69.15%, with the best accuracy of 71.49%. The reduced variance value also indicates that the training of the multi-modality 3D CNN is more stable than single modality CNNs. As stated in Section III E, the multi-modality 3D CNN architecture is able to learn fMRI and sMRI convolutional features separately though two separate CNN branches, and combine the learned high-level feature to boost classification accuracy. The results demonstrate the superior performance of the proposed multi-modality 3D CNN architecture.

C. COMPARISON WITH EXISTING METHODS

Table 2 shows the results of related works and the proposed approaches when they are evaluated on the ADHD-200 hold-out testing dataset. In [25], Dai *et al.* employed features from structural MRI and fMRI, including ReHo, FC, GM and Cortical Thickness (CT). They integrated multimodal image features using MKL and obtained a diagnostic accuracy of 61.54%. Ghiassian *et al.* adopted a histogram of oriented gradients (HOG) and a feature selection process and then evaluated several classifiers. They found that the best performance, 62.57%, was achieved with a Support Vector Machine (SVM) [26]. Dey *et al.* first selected a sequence of highly active voxels and construct the connectivity network between them. They obtained an average accuracy of 62.81% on the hold-out testing dataset [38]. In [27], Guo *et al.* explored the

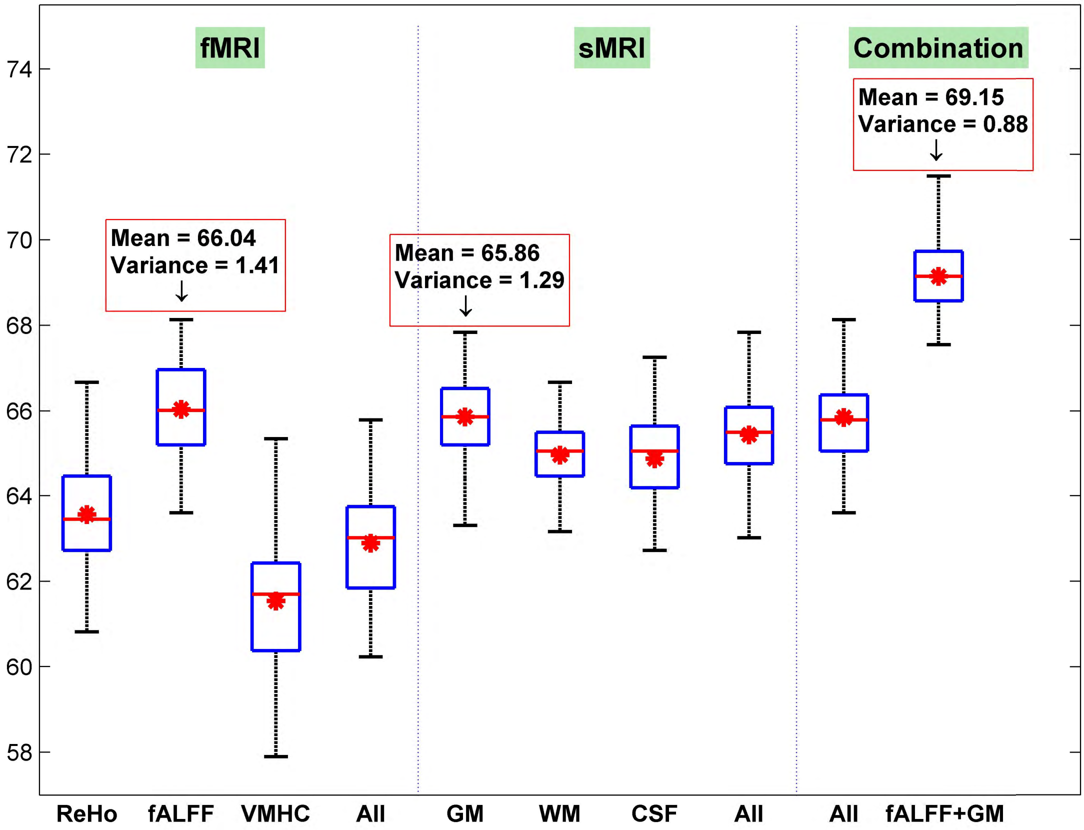

Fig. 5. Statistical results of 3D CNN approaches corresponding to different features over 50 individual runs. First, we evaluate single modality approaches where fMRI features and sMRI features are utilized individually. We further test the performance of multi-modality approaches where fMRI and sMRI features are combined via the proposed multi-modality 3D CNN architecture. The red asterisks and lines represent the average and median values respectively. The edges of the box are the lower and upper quartiles.

TABLE 2. Diagnosis performance comparisons between the proposed method and state-of-the-art methods based on the ADHD-200 dataset.

Method	Features	Classifier	Accuracy
[25]	ReHo, FC, GM and cortical thickness	Multi-kernel learning	61.54%
[26]	histogram of oriented gradient	Support vector machine	62.57%
[38]	functional connectivity networks	Support vector machine	62.81%
[27]	FC, assortative mixing and synchronization	Support vector machine	63.75%
The proposed method	fALFF	Single-modality 3D CNN	66.04%
	GM density	Single-modality 3D CNN	65.86%
	fALFF, GM density	Multi-modality 3D CNN	**69.15%**

functional connectivity between voxels and obtained an average accuracy of 63.75% based on a social network method. These results represent the highest diagnostic performance on the ADHD-200 hold-out test dataset. As a comparison, our proposed single modality architecture which only takes fALFF or GM yields better performance than the existing methods on the hold-out testing dataset. When combining fALFF and GM into a multi-modality 3D CNN, the proposed

TABLE 3. The diagnosis performance on the hold-out testing data from different sites.

Site (number of subjects)	ADHD-200 Competition [24]	DBN [13]	The proposed method
PekingU (51)	51.05%	54.00%	**62.95%**
KKI (11)	61.90%	71.82%	**72.82%**
NYU (41)	35.19%	37.41%	**70.50%**

method achieves a classification accuracy of 69.15% which is a significant improvement over existing methods.

As with [13], we also tested the proposed method on the individual subset of ADHD-200 hold-out testing set from PekingU (Peking University), KKI (Kennedy Krieger Institute) and NYU (New York University Child Study Center) over 50 individual runs. As shown in Table 3, the average accuracy of the proposed method is superior to the best result of the ADHD-200 competition in [24] and that of the DBN method [13] especially on PekingU and NYU. We also included the number of subjects in each subset. The notable differences between performances on different subsets also suggests a substantial amount of heterogeneity within the entire dataset. In summary, our proposed 3D CNN based architecture achieves state-of-the-art classification performance in ADHD classification.

VI. DISCUSSION

The classification performance of techniques applied to the ADHD-200 competition dataset seem inferior to the seemingly impressive results of published studies [20]–[22] utilizing much smaller data sets. However, considering the heterogeneity of the clinical manifestation of ADHD, it is always hard to generalize the findings of studies utilizing a small number of samples [2]. The ADHD-200 dataset is probably much more difficult to classify because of its heterogeneity and its relatively large sample size [23]. Taking into account phenotypic information and scanner information may improve diagnostic accuracy in the future. In general, development of automatic ADHD classification tools from MRI scans is challenging work and there is still a long way to go to apply these tools in formulating an ADHD diagnosis.

FMRI and sMRI data are typically analyzed separately and the joint information is not fully explored. To the best of our knowledge, our work is the first study to examine a 3D CNN model on the diagnosis of ADHD utilizing both fMRI and sMRI. Considering the complicated pathologic process of ADHD, functional and anatomical changes may happen simultaneously, and it may prove difficult to make a diagnosis based on a signal modality. Our results suggest that a multiple modalities classification approach will be a promising direction for finding neuroimaging biomarkers of ADHD.

Performance of any classification network is dependent upon the selected features. Several fMRI studies suggest that ADHD is associated with brain sub-network dysfunction [39], [40]. Other types of fMRI features and prior knowledge (e.g., gender information) could be incorporated into the proposed approach to further improve ADHD classification performance. However, a larger number of features requires more training samples and may result in overfitting, especially when the number of training samples is limited.

To avoid overfitting during 3D CNN training, several methods were considered in this paper: 1) a 3D CNN architecture was adopted, taking advantage of its intrinsic features, such as partial connectivity, weights sharing and pooling architectures; 2) we carefully designed the number of layers and feature maps to avoid overfitting while retaining sufficient capacity for the network to solve the complex ADHD classification problem; 3) we performed data augmentation via a dropout technique at the fully connected layers that contain most of the weights in the network. As a result, the 3D CNN models were well trained and yielded state-of-the-art classification accuracy.

VII. CONCLUSION

With the availability of the large scale ADHD-200 dataset and the successes of deep learning in many recognition problems, we were motivated to develop an automatic classification algorithm based on deep learning to classify ADHD vs. TDC using MRI scans. Inspired by the way that radiologists examine 3D brain images, we propose an automatic and effective 3D CNN architecture for ADHD classification which exploits the complementary information gleaned from both fMRI and sMRI. The proposed 3D CNN method is fundamentally different from previous attempts to classify the ADHD using MRI scans. Specifically, we first encoded prior knowledge on six types of 3D low-level features previously used to diagnose ADHD, including ReHo, fALFF and VMHC as well as GM, WM and CSF probability densities in MNI space. Then a 3D CNN based strategy was used to extract the high-level features from each modality. Unlike previous methods that mostly considered low-level features as a vector and hence neglects potential 3D local patterns, we kept these low-level features in 3rd-order tensors and trained the 3D CNN based on them. We further combined the fMRI and sMRI features with a multi-modality 3D CNN architecture which yielded the state-of-the-art performance. Experimental results on the hold-out ADHD-200 testing dataset shows that the proposed 3D CNN is superior in performance to previous approaches, even with a fewer training samples.

REFERENCES

[1] G. V. Polanczyk, E. G. Willcutt, G. A. Salum, C. Kieling, and L. A. Rohde, "ADHD prevalence estimates across three decades: An updated systematic review and meta-regression analysis," *Int. J. Epidemiol.*, vol. 43, no. 2, pp. 434–442, 2014.

[2] A. Eloyan *et al.*, "Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging," *Frontiers Syst. Neurosci.*, vol. 6, p. 61, 2012.

[3] X. Peng, P. Lin, T. Zhang, and J. Wang, "Extreme learning machine-based classification of ADHD using brain structural MRI data," *PLoS ONE*, vol. 8, no. 11, p. e79476, 2013.

[4] Y. Zang *et al.*, "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 *et al.*, "Abnormal spontaneous brain activity in medication-naive ADHD children: A resting state fMRI study," *Neurosci. Lett.*, vol. 502, no. 2, pp. 89–93, 2011.

[6] Y. Zang, T. Jiang, Y. Lu, Y. He, and L. Tian, "Regional homogeneity approach to fMRI data analysis," *Neuroimage*, vol. 22, no. 1, pp. 394–400, 2004.

[7] D. Long *et al.*, "Automatic classification of early Parkinson's disease with multi-modal mr imaging," *PLoS ONE*, vol. 7, no. 11, p. e47714, 2012.

[8] H.-I. Suk, S.-W. Lee, D. Shen, and The Alzheimer's Disease Neuroimaging Initiative, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014.

[9] C.-W. Chang, C.-C. Ho, and J.-H. Chen, "ADHD classification by a texture analysis of anatomical brain MRI data," *Frontiers Syst. Neurosci.*, vol. 6, p. 66, Sep. 2012.

[10] M. Kobel *et al.*, "Structural and functional imaging approaches in attention deficit/hyperactivity disorder: Does the temporal lobe play a key role?" *Psychiatry Res., Neuroimag.*, vol. 183, no. 3, pp. 230–236, 2010.

[11] M.-G. Qiu, Z. Ye, Q.-Y. Li, G.-J. Liu, B. Xie, and J. Wang, "Changes of brain structure and function in ADHD children," *Brain Topogr.*, vol. 24, nos. 3–4, pp. 243–252, 2011.

[12] M. Angriman, A. Beggiato, and S. Cortese, "Anatomical and functional brain imaging in childhood ADHD: Update 2013," *Current Develop. Disorders Rep.*, vol. 1, no. 1, pp. 29–40, 2014.

[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.*, 2014, pp. 225–232.

[14] J. Kleesiek *et al.*, "Deep MRI brain extraction: A 3D convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, Apr. 2016.

[15] A. Payan and G. Montana. (Feb. 2015). "Predicting Alzheimer's disease: A neuroimaging study with 3D convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1502.02506

[16] Q.-H. Zou *et al.*, "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF," *J. Neurosci. Methods*, vol. 172, no. 1, pp. 137–141, 2008.

[17] X. Zuo *et al.*, "Growing together and growing apart: Regional and sex differences in the lifespan developmental trajectories of functional homotopy," *J. Neurosci.*, vol. 30, no. 45, pp. 15034–15043, 2010.

[18] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 1–7.

[19] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[20] 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, 2008.

[21] B. A. Johnston, B. Mwangi, K. Matthews, D. Coghill, K. Konrad, and J. D. Steele, "Brainstem abnormalities in attention deficit hyperactivity disorder support high accuracy individual diagnostic classification," *Hum. Brain Mapping*, vol. 35, no. 10, pp. 5179–5189, 2014.

[22] C. Z. Zhu *et al.*, "Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2005, pp. 468–475.

[23] M. R. Brown *et al.*, "Adhd-200 global competition: Diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements," *Frontiers Syst. Neurosci.*, vol. 6, p. 69, Sep. 2012.

[24] *The ADHD-200 Global Competition*. Accessed: Oct. 1, 2017. [Online]. Available: http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html

[25] 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.

[26] S. Ghiassian, R. Greiner, P. Jin, and M. Brown, "Learning to classify psychiatric disorders based on fMR images: Autism vs healthy and ADHD vs healthy," in *Proc. 3rd NIPS Workshop Mach. Learn. Interpretation NeuroImag.*, 2013, pp. 1–7.

[27] X. Guo, X. An, D. Kuang, Y. Zhao, and L. He, "ADHD-200 classification based on social network method," in *Proc. Int. Conf. Intell. Comput.*, 2014, pp. 233–240.

[28] Centers for Disease Control and Prevention, "HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services," *MMWR, Morbidity Mortality Weekly Rep.*, vol. 52, no. 1, pp. 1–17, 2003.

[29] *The R-fMRI Maps Project*. Accessed: Oct. 1, 2017. [Online]. Available: http://mrirc.psych.ac.cn/RfMRIMaps

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

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

[32] K. Somandepalli *et al.*, "Short-term test–retest reliability of resting state fMRI metrics in children with and without attention-deficit/hyperactivity disorder," *Develop. Cognit. Neurosci.*, vol. 15, pp. 83–93, Oct. 2015.

[33] E. D. Gennatas *et al.*, "Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to young adulthood," *J. Neurosci.*, vol. 37, no. 20, pp. 5065–5073, 2017.

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

[35] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. (Feb. 2016). "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size." [Online]. Available: https://arxiv.org/abs/1602.07360

[36] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] 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. 94, Jun. 2014.

[39] P. Lin *et al.*, "Global and local brain network reorganization in attention-deficit/hyperactivity disorder," *Brain Imag. Behav.*, vol. 8, no. 4, pp. 558–569, 2014.

[40] M. Greicius, "Resting-state functional connectivity in neuropsychiatric disorders," *Current Opinion Neurol.*, vol. 21, no. 4, pp. 424–430, 2008.

LIANG ZOU received the B.Sc. degree in microelectronics from Anhui University, China, in 2010, the M.Sc. degree in biomedical engineering from the University of Science and Technology of China, in 2013, and the Ph.D. degree in electrical and computer engineering from The University of British Columbia, Canada, in 2017. He is currently a Post-Doctoral Research Fellow with Queen's University, Canada. His current research interest includes statistical signal processing, machine learning, and medical imaging.

JIANNAN ZHENG received the B.Sc. degree in electrical engineering from Xi'an Jiaotong University, China, in 2010, and the M.Sc. degree in mechanical engineering from Concordia University, Canada, in 2012. He is currently pursuing the Ph.D. degree with The University of British Columbia, Canada. His current research interest includes computer vision, deep learning, and medical imaging.

CHUNYAN MIAO is currently a Full Professor with the School of Computer Engineering, Nanyang Technological University. He was involved in new disruptive artificial intelligence (AI) approaches and theories that synergize human intelligence, artificial intelligence and behavior data analytics (AI powered by humans). She has successfully led over ten large national research projects in AI, smart health, personal big data analytics, and aging in place technologies, and has received over $50 Million in research funding. Results from his basic AI research have been used in important application areas, such as technologies for graceful aging, healthy living, and innovative learning. Specific high-impact real world examples include Parkinson's disease predictive analytics, stroke detection, and rehabilitation. She has published over 200 journal and conference papers in top-tier venues. She is an Editor/Associate Editor of the IEEE ACCESS, the IEEE JOURNAL OF INTERNET OF THINGS, and the *Web Intelligence Journal*.

MARTIN J. MCKEOWN is currently the Pacific Parkinson's Research Institute and The University of British Columbia Chair with Parkinson's research, the Director of the Pacific Parkinson's Research Centre, a Full Professor with the Department of Medicine, and an Adjunct Professor with the Department of Electrical and Computer Engineering, The University of British Columbia, Canada. He completed his engineering physics, medicine, and neurology training at McMaster University, the University of Toronto, and the University of Western Ontario, respectively. He has authored over 100 peer-reviewed papers and book chapters. He is a member of the Systems and Clinical Neurosciences–A Canadian Institutes of Health Research scientific peer review committee and a member of the Scientific Advisory Board of the Parkinson's Society of Canada.

Z. JANE WANG (F'17) received the B.S. degree from Tsinghua University, China, in 1996, and the M.S. and Ph.D. degrees from the University of Connecticut in 2000 and 2002, respectively, all in electrical engineering. She was a Research Associate with the Electrical and Computer Engineering Department, University of Maryland, College Park. Since 2004, she has been with the Department of Electrical and Computer Engineering, The University of British Columbia, Canada, where she is currently a Full Professor. Her research interests include statistical signal processing theory and applications with a current focus on brain data analytics. She has been an Associate Editor of the *IEEE Signal Processing Magazine*, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS.