Attention Deficit Hyperactivity Disorder Classification Based on Deep Learning

Donglin Wang, Don Hong, Qiang Wu

Abstract—Attention Deficit Hyperactivity Disorder (ADHD) is a type of mental health disorder that can be seen from children to adults and affects patients' normal life. Accurate diagnosis of ADHD as early as possible is very important for the treatment of patients in clinical applications. Some traditional classification methods, although having been shown powerful in many other classification tasks, are not as successful in the application of ADHD classification. In this paper, we propose two novel deep learning approaches for ADHD classification based on functional magnetic resonance imaging. The first method incorporates independent component analysis with convolutional neural network. It first extracts independent components from each subject. The independent components are then fed into a convolutional neural network as input features to classify the ADHD patient from typical controls. The second method, called the correlation autoencoder method, uses correlations between regions of interest of the brain as the input of an autoencoder to learn latent features, which are then used in the classification task by a new neural network. These two methods use different ways to extract the inter-voxel information from fMRI, but both use convolutional neural networks to further extract predictive features for the classification task. Empirical experiments show that both methods are able to outperform the classical methods such as logistic regression, support vector machines, and other methods used in previous studies.

Index Terms—Attention Deficit Hyperactivity Disorder, Functional Magnetic Resonance Imaging, Convolutional Neural Network, Independent Component Analysis, Correlation-Autoencoder

1 Introduction

Attention Deficit and Hyperactivity Disorder (ADHD) is a mental disorder that commonly occurs in childhood. Based on American Psychiatric Association, the symptoms are characterized mainly in concentration with obvious difficulty, short duration of attention, and hasty acting compared with children of the same age. ADHD occurs not only in children but also in adults. An estimated 8.4% of children and 2.5% of adults have ADHD [1]. The study [2] mentioned that the estimated annual national extra cost for ADHD ranges from $143 billion to $266 billion. There are three main types of ADHD: inattentive type, hyperactive/impulsive type and combined type. ADHD negatively affects patients both mentally and physically. It has become one of the major problems in public health system. It is important to diagnose ADHD as early and as accurately. The common diagnosis of ADHD is often based on the Diagnostic and Statistical Manual of Mental Disorders (DSM), published by the American Psychiatric Association, and the symptoms are consistently exhibited for at least six months. Such diagnosis of ADHD needs the clinician to make the decision based on the interview to the patient and is somehow subjective [3]. The functional magnetic resonance imaging (fMRI) technique offers a novel way to assess the patients with ADHD status. It has been used to study ADHD in different ways recently and obtained exciting results. There is no doubt that classification of ADHD is playing an essential role in the diagnosis applications.

Deep learning, as a branch of machine learning, has been used in many fields and has made many excited achievements. In recent years deep learning was also used in classification of mental disorder based on fMRI data. Two major deep learning models, fully connected neural network and convolutional neural network, were often used in previous studies. In the article [4] the authors used an autoencoder model to help examining the relations among the regions of brain to classify Mild Cognitive Impairment (MCI) patient. The authors of [5] used two encoders to classify autism spectrum disorder (ASD) patients from the database known as Autism Brain Imaging Data Exchange (ABIDE). In [6] the authors proposed to use CNN model and the architecture LeNet-5 to classify Alzheimer's disease (AD) patients from normal controls. The article [7] detected significant differences in cerebellum, motor cortex and temporal lobe between ADHD and normal humans between 73 children with ADHD and 76 normal children, and it used linear discriminant analysis classifier to distinguish health controls from ADHD children with an accuracy rate of 80.08% through 50 times of 2-fold validation. In the paper [8] the authors explored specific regions of interest (ROIs) to distinguish ADHD children from control children and reached a classification performance with a sensitivity rate of 90%. The paper [9] used both structural and functional magnetic resonance imaging data to predict diagnostic status of individuals with ADHD and got an accurate rate of 55%. The article [10] used a machine learning approach called Gaussian process classifiers to distinguish thirty MDD from thirty normal controls, and got an accuracy rate of 77%.

The ADHD-200 dataset (http://fcon_1000.projects.nitrc.org/indi/adhd200/) launched in 2011 for global competition, is a popular ADHD dataset for classification. Except for the classical methods, such as logistic regression model and

• *The authors are with the Department of Mathematical Sciences, Middle Tennessee State University, 1301 E Main Street, Murfreesboro, TN 37132. E-mails: {dwang, dhong, qwu}@mtsu.edu*

Manuscript received XXXX; revised XXXX.

support vector machine [11], [12], deep learning techniques also have been used on this dataset for ADHD classification since the dataset was released. The paper [13] could be the first to use the deep learning technique for the classification of ADHD with fMRI data. The authors proposed to use the deep belief network (DBN) including a stack of restricted Boltzmann machines (RBM) to do the classification task on ADHD-200 dataset, and the accuracy rate improved comapred with the results from the competition. In [14] the authors proposed a deep learning architecture called DeepFMRI, which includes three sequential networks: a feature extractor network, a functional connectivity network, and a classification network. The DeepFMRI model got an accuracy rate of 73.1% for the holdout dataset collected from the New York University imaging site. In [15] the authors proposed a 3D convolutional neural networks including single-modality and multi-modality forms, which employed the fractional ALFF (fALFF) [16] and density of gray matters as the features. The proposed method reached an accuracy rate of 72.82% for the holdout dataset collected from the Kennedy Krieger Institute site. In [17] the authors proposed an algorithm based on convolutional denoising autoencoder (CDAE) and adaptive boosting decision trees (AdaDT) for classification on the ADHD-200 test dataset. The results not only improved classification accuracy but also maintain a certain balance between specificity and sensitivity. In [18] the authors proposed an ensemble hybrid features selection method including a 3D DenseNet and a XGBoost for the neuropsychiatric disorder classification based on the Neuropsychiatric Phenomics (CNP) dataset in the Consortium. The accuracy rate based on binary and multi-class classification can reach 91.22% and 78.62%, respectively. Notice that most of these existing studies applied deep neural networks directly to the fMRI data.

It is widely accepted that ADHD and other mental disorder problems usually are caused by abnormality of functional connectivity abnormality of human brains. It looks more reasonable to seek appropriate measures to characterize functional connectivity from fMRI data first before applying feature extraction and classification tools. In this paper, we propose two deep learning strategies to improve the accuracy rate for classification on ADHD. One is to incorporate independent component analysis with convolutional neural network (ICA-CNN) based on rs-fMRI data and the other is to incorporate correlation between nodes with one-dimension deep convolutional autoencoder model (Correlation-Autoencoder) based on rs-fMRI. Note that both ICA and correlation are able to capture the inter-voxel relationships and characterizes the functional connectivity to some extent. The two proposed strategies are applied to ADHD-200 database, and the classification accuracy is improved compared to the results from competition and other previous literature.

The rest of the paper is arranged as follows. In Section 2 we introduced the two strategies in detail. In Section 3 the ADHD-200 data set collected from Peking University site is analyzed by the two proposed methods. Section 4 conclude the paper with discussions.

2 METHODS

2.1 ICA-CNN model

Independent component analysis (ICA) has been becoming one of the most popular decomposition techniques for fMRI data analysis since it was first used in [19]. It assumes the observed data is a linear combination of sources which are statistically independent [20] and the purpose of ICA is to extract these independent components. Let $\mathbf{X} \in \mathbb{R}^{s\times n}$ be a matrix with dimension $n \times s$, where s is the number of voxels and n presents the number of time points. The goal of ICA is to solve the equation

$$\mathbf{X} = \mathbf{WS} \tag{1}$$

to find $\mathbf{W}, \mathbf{S}$, where $\mathbf{W} \in \mathbb{R}^{s\times m}$ is called the matrix of mixing coefficients and $\mathbf{S} \in \mathbb{R}^{m\times n}$ is called source matrix. Each row of the matrix $\mathbf{S}$ represents one spatially independent component. Several optimization algorithms have been proposed to implement such a decomposition, including the maximum likelihood estimation [21], minimum of the mutual information between sources [22], maximum of the non-gaussianity between sources [23], and the canonical ICA algorithm [24], to name just a few.

Convolutional neural network (CNN), a popular deep learning technique, is used in many fields and has made many excited achievements especially in image recognition since it was proposed in [25]. One of the big advantages of CNN is the shared convolution kernel to extract features faster and effectively.

For a given ADHD data set, suppose there are N subjects and each subject has n volumes with dimension of $l \times w \times h$. The ICA-CNN model integrates the ICA and CNN and is implemented via the following steps.

1) First, m independent components are extracted from each of N subjects using the canonical ICA algorithm [24].
2) After extracting m components from all subjects, a masker based on the m components is made to extract the corresponding signals for each subject. It is a template used to extract signals of the regions of interest of brains for different subjects.
3) After having extracted the signal based on the masker, each subject has a matrix with dimension $n \times m$, where n is the number of volumes. Totally, there are N matrices with dimension $n \times m$.
4) A convolutional neural network (CNN) is built with several convolutional layers, pooling layers, dropout layers, batch-normalization layers, flatten layers and dense layers. Each convolutional layer has activation of rectified linear activation unit (ReLU). The last layer is dense layer with the sigmoid activation function. The N matrices are as input for this CNN model and the output is the medical status of subjects. The value of 0 means normal control and 1 means ADHD status.
5) The model is trained by back-propagation algorithm with mini-batch method. The parameters w and bias b are updated according to the following rules until the convergence,

$$w^l \to w^l - \frac{\lambda}{k}\sum_x \frac{\partial C(w,b)}{\partial w^l} \tag{2}$$

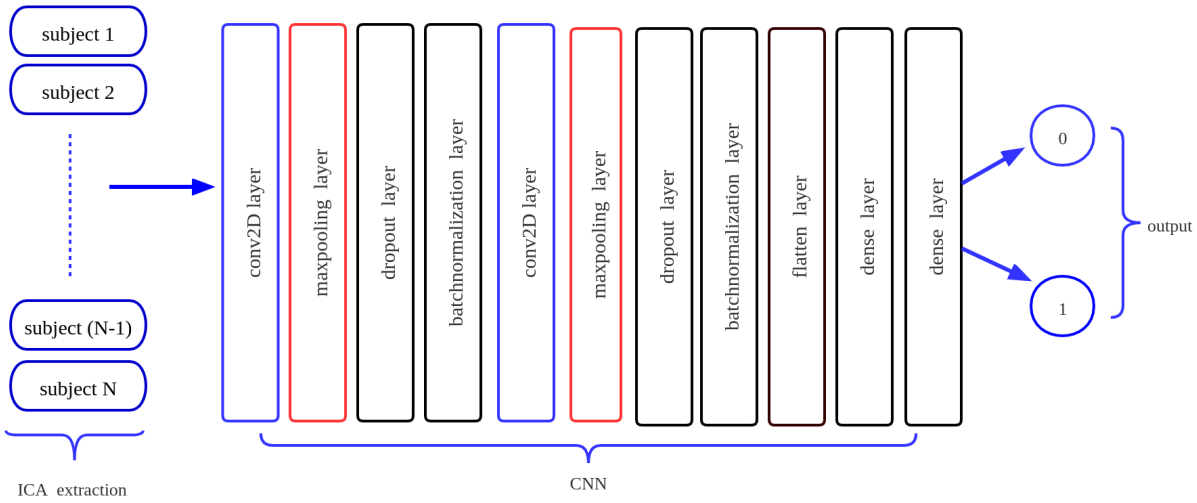


Fig. 1. Framework of ICA-CNN method: *each subject is extracted a matrix with dimension of $n \times m$ as inputs for CNN model, n is the number of volumes and m is the number of components; the last dense layer has two nodes with sigmoid activation function.*

$$b^l \to b^l - \frac{\lambda}{k} \sum_x \frac{\partial C(w,b)}{\partial b^l} \tag{3}$$

where λ is the learning rate, k is the number of batch size, l is the lth layer, x is the data in a batch size, $C(w,b)$ is the binary cost function of the form:

$$C(w,b) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

where $\hat{y}$=$f(w,b,x)$. The architecture of this proposed method is shown in Figure 1.

2.2 Correlation-autoencoder model

Autoencoder [26], [27] is an unsupervised learning method. It consists of two neural networks: an encoder and a decoder. The main purpose of an autoencoder is to learn the latent representations denoted by μ in a lower dimension from the input denoted by x. Generally the encoder network compresses x into μ and the decoder network rebuilds μ into $\hat{x}$ so that $x \approx \hat{x}$. The autoencoder model is often used in dimension reduction [28], [29], image reconstruction [30] and noise reduction [31], [32].

For a given ADHD data set, the data includes N subjects and each has n volumes. A brain atlas is needed for partitioning the brain into the regions of interest (ROIs). In this application the Schaefer atlas [33] is used to partition the brain to extract the signal from the rs-fMRI data, and the algorithm follows the steps below.

1) First the Schaefer atlas [33] is used to partition the brain into r regions of interest (ROIs) and help extract the signal from rs-fMRI for each subject. Each ROI corresponds a time series data with length t, t is the number of time points of each subject.
2) For each subject the correlation is calculated between every two ROIs. This results in a square matrix with dimension $r \times r$ for each subject. Since this matrix is symmetric, only the upper triangular matrix is reserved and flattened into a vector of length $(r^2+r)/2$. As a result the original rs-fMRI data of N subjects with n volumes for each subject is transferred into N observations with each having $(r^2+r)/2$ features.
3) The N observations, with $(r^2+r)/2$ features for each, are the input denoted by x for the encoder model and the input shape is $(r^2+r)/2$ by 1. The encoder model includes convolutional 1D layers, maxpooling 1D layers and batch-normalization layers. The purpose of encoder model is to extract the representation of input in a lower dimension $a \ll (r^2+r)/2$.
4) The extracted representation $a \times 1$ is as input for the decoder model. The decoder model includes convolutional 1D layers, upsampling 1D layers and batch-normalization layers. The purpose of the decoder model is to rebuild the extracted representation a into $\hat{x}$ with dimension $(r^2+r)/2$. The encoder and the decoder are trained together as an autoencoder model and the loss function is mean square error:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x_i})^2 \tag{4}$$

5) After the autoencoder model is trained, the encoder model is used to build the new network for binary classification. This new model first keeps all the layers from encoder model and the weights of the encoder model are fixed. In addition, the flatten layer and dense layers are added in the model. This new model is trained and used to do classification task with binary loss function:

$$C(w,b) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \tag{5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2022.3170527, IEEE/ACM Transactions on Computational Biology and Bioinformatics

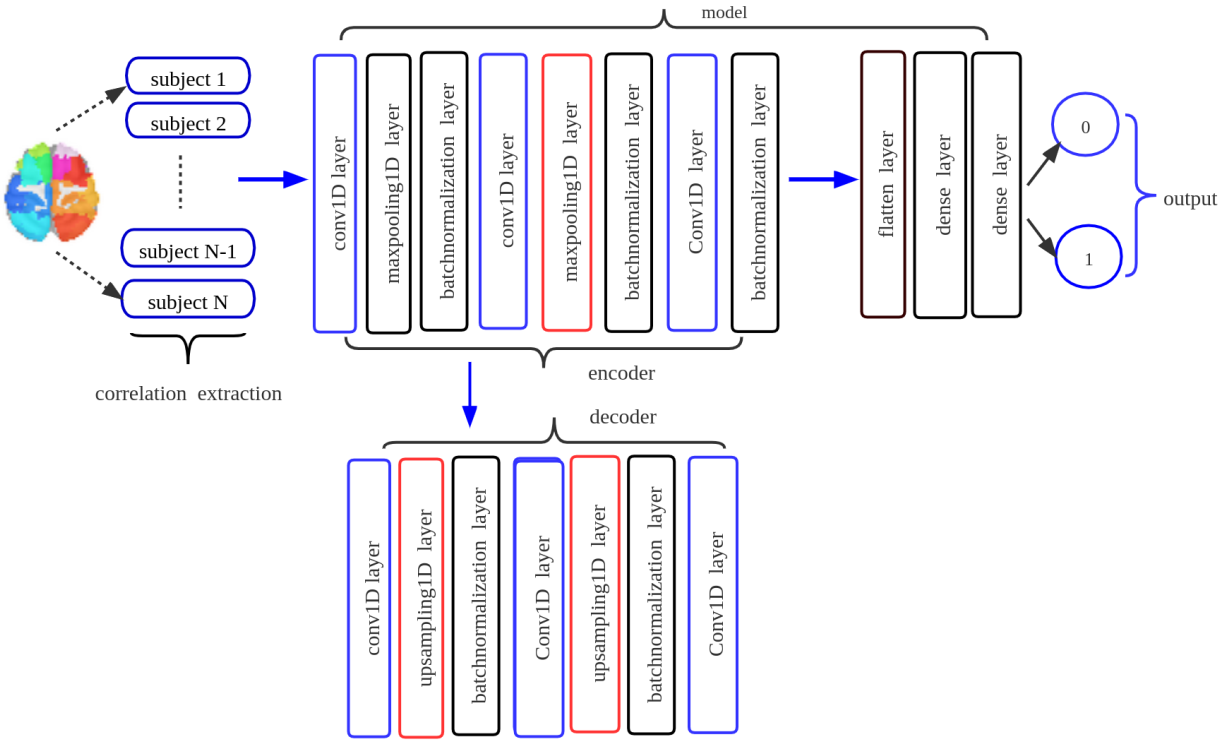


Fig. 2. Framework of correlation autoencoder method: *each subject is partitioned into r ROIs, that is, $n \times r$ matrix for each subject, n is the number of volumes; then the correlation between the r ROIs is calculated, then the upper triangular matrix of $r \times r$ is flattened as input with length $(r^2+r)/2$ for the autoencoder model; the last dense layer in the model has two nodes with sigmoid activation function.*

where $\hat{y}=f(w,b,x)$. The architecture of this proposed method is shown in Figure 2.

3 Case Study

3.1 Data description

There is a publicly available ADHD-200 database, the 1000 Functional Connectomes Project, which is provided by http://fcon_1000.projects.nitrc.org/indi/adhd200/. The dataset has 973 resting state fMRI acquisitions labeled with ADHD or typically developing controls (TDC) from subjects aging from 7 to 21. The dataset is collected from eight different international sites, which are NeuroImage group (NeuroImage), New York University Child Study Center (NYU), Peking University (Peking), Brown University (BU), Kennedy Krieger Institute (KKI), University of Pittsburgh (Pitts), Oregon Health and Science University (OHSU) and Washington University in St. Louis (WU). The ADHD-200 dataset was used in a competition to classify ADHD in 2011 and then the Preprocessed Connectomes Project (PCP) made the competition accessible to a broader range by preprocessing the data and sharing the results to public. Three different pipelines were used to preprocess the fMRI data. The preprocess data is used in the subsequent analysis based on Athena pipeline [34] which was performed by Cameron Craddock using AFNI and FSL running on the Athena computer cluster at Virginia Tech's ARC. There are more details on preprocessing for both structural data and functional data through the website of https://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline. The functional image is using a 6mm FWHM Gaussian filter and band pass filter is between 0.009 and 0.08 Hz.

The data was used for global competition of classification task in 2011, and each data was also provided a holdout test data. In the following analysis, the data from the site of Peking University is used. There are totally 198 subjects in the training data, including 123 typical controls and 75 ADHD (ADHD-Hyperactive/Impulsive, ADHD-inattentive and ADHD-combined) subjects, and the test data has 51 subjects including 24 ADHD subjects and 27 typical controls. In practice the training data only has 168 subjects due to preprocessing and the test data is still the same as the original test data. The final data to be used is summarized in Table 1. The functional brain of one random chosen subject with dimension $49 \times 58 \times 47 \times 232$ is shown in Figure 3.

3.2 Data analysis

The data is analysed by the traditional logistic regression model and support vector machine (SVM) model as well as the two proposed methods. The extracted features are used in all the four methods. The two proposed methods are also

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2022.3170527, IEEE/ACM Transactions on Computational Biology and Bioinformatics

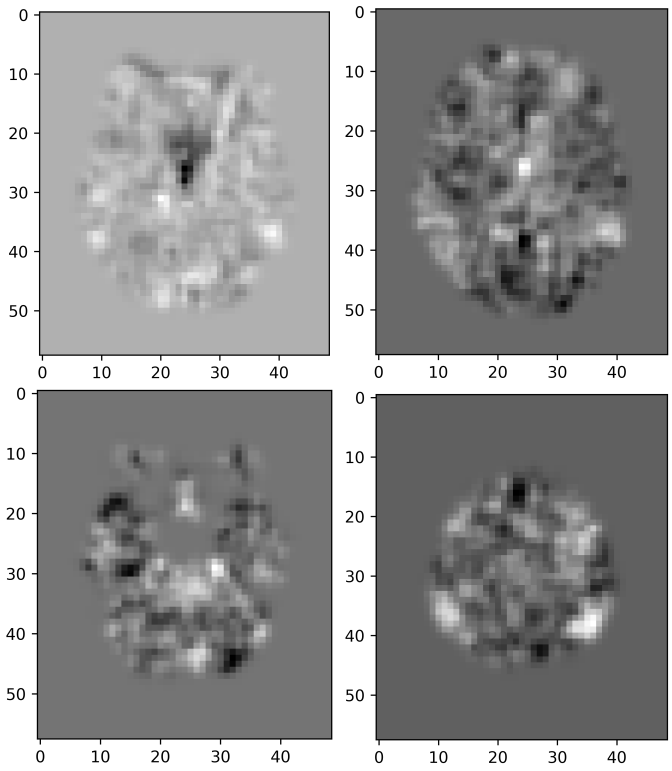


Fig. 3. The functional brain of a random chosen subject

TABLE 1
Summary of original data and used data

	Training data			Test data		
	ADHD	control	Total	ADHD	control	Total
Original	75	123	198	24	27	51
Used	69	99	168	24	27	51

compared with the traditional methods and other methods in the previous literature.

3.3 Analysis I

In this dataset some subjects have 231 volumes while some others have 232 volumes. In order to extract the same number of independent components all subjects are set to have 231 volumes by slicing. A number of 100 independent components are extracted from each volume for each subject, so each subject has 100 times series of length 231, represented by a 231×100 matrix. So there are 168 matrices of dimension 231×100 for the training data and 51 matrices for the testing data.

When logistic regression and support vector machine are used to classify the data, the average of 231 volumes, which produces an vector of 100 feature values corresponding to the 100 independent component for each subject, will be the input features. The logistic model gives an accuracy rate of 61%, sensitivity 42%, specificity 74%, precision 63%. The area under receiver operating characteristics curve (AUC) is 0.54. The support vector machine with Gaussian kernel gives an accuracy rate of 57%, sensitivity 8%, specificity 100%, precision 78%, and AUC 0.52.

For ICA-CNN model, a convolutional neural network is used to classify the ADHD patients from health controls. Its input is the matrices of dimension 231×100, each of which represents the 100 independent components extracted for each subject by ICA. The two convolutional layers use 32 and 64 filters, respectively, with kernel size 3×3 and activation function ReLU. The maxplooing size is 2×2, the stride step is 1, and the dropout rate is 0.2. The last dense layer has two nodes with sigmoid function. The optimization uses Adam algorithm [35] with a learning rate 0.001. The number of epochs are set to 300, early stopping is triggered when there is no improvement in the loss for ten consecutive epochs. The fine-tuned model gives an accuracy rate of 67%, sensitivity 42%, specificity rate 89%, precision 77%, and AUC 0.65.

For comparison purpose, the ROC curves for logistic regression, support vector machine and ICA-CNN are shown in Figure 4. For the accuracy rate, sensitivity (which is also known as recall), specificity, precision, and AUC, the results are also compared with other studies based on machine learning and deep learning algorithms in the literature. See Table 2. We see ICA-CNN outperforms the results in [14], [15], [36] while logistic regression and SVM underperform. This means after ICA has extract sufficient information from fMRI data, CNN are able to extract predictive features for ADHD classification while averaging the volumes suffers from information loss.

ICA summarizes the fMRI data of $49 \times 58 \times 47$ voxels by 100 independent components. This simultaneous denoising and feature extraction process considers individual differences and captures the relationships between voxels of each subject to some extent. The resulted matrix is time related in one dimension and spatially related in the other dimension. Predictive information extraction from both type of data is thought the strength of convolution filters. Therefore, ICA-CNN model is able to combine the benefits of both ICA and CNN, which plausibly explains its superior performance.

TABLE 2
Classification performance of different models (Accuracy, Specificity, Sensitivity, and Precision are abbreviated as Acc., Spec., Sens., and Prec., respectively, to condense space. The values from [14], [15], [36] are rounded to align with other numbers.)

Model	Acc.	Spec.	Sens.	Prec.	AUC
Competition [37]	51%	-	-	-	-
Logistic	61%	78%	42%	63%	0.54
SVM-G	57%	100%	8%	78%	0.52
DeepFMRI [14]	63%	79%	48%	-	-
3D CNN [15]	63%	-	-	-	-
Deep Forest [36]	65%	-	-	-	-
ICA-CNN	**67%**	**89%**	**42%**	**77%**	**0.65**

3.4 Analysis II

In the second analysis each subject is partitioned into 300 regions of interest (ROIs) based on the Schaefer atlas [33]. Since each subject is set to 231 volumes, there is a time series of length 231 for each ROI. The correlation between each pair of ROIs is calculated. For each subject this gives us a square matrix of correlation with dimension 300×300. Since

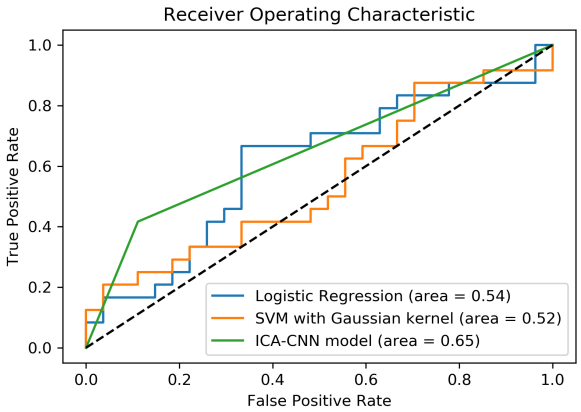


Fig. 4. ROC curve for logistic regression, SVM and ICA-CNN models

this matrix is symmetric, only the upper triangular part is used and is flattened into a vector with the length of 45150. Totally there are 168 subjects with 45150 features for training data and 51 subjects with 45150 features for testing data.

For logistic regression model, a variable selection algorithm is used to choose the most important features from the 45150 features. The Recursive Feature Elimination algorithm (RFE) [38] is used to select the most important features. The basic idea of RFE is to build a full model at the beginning and drop the least important feature, then a new model is built again with the rest features, and the process is repeated until the necessary number of features are reached. The RFE algorithm is a backward feature selection method. Two logistic regression models were built. One used the most important 100 features and the other used the most important 200 features. The model with the most important 100 features gives an accurate rate of 61%, specificity rate 85%, sensitivity rate 33%, precision 67%, and AUC 0.59. The model with the most important 200 features gives an accurate rate of 63%, specificity 85%, sensitivity 38%, precision 69%, and AUC 0.63.

For SVM with Gaussian kernel, the RFE algorithm again is used to choose the most important features from the 45150 features. With the most important 100 features, SVM has the same performance as the logistic model. With the most important 200 features, SVM gives an accurate rate of 65%, specificity 89%, sensitivity 38%, precision 75%, and AUC 0.63. It is slightly better than the logistic model.

The correlation-autoencoder (CorrAE) model includes three parts: encoder, decoder and prediction. The encoder and the decoder make an autoencoder model which takes all the 45,150 features as input. The purpose of this autoencoder model is to extract the latent features which can represent the input as much as possible. The encoder includes convolutional layers with 1024, 512 and 128 filters, respectively. The kernel size is 3, maxpooling size is 2, and all the activation functions are ReLU. The purpose of the decoder is to rebuilt the input features as much as possible based on the latent features. The decoder model includes convolutional layers with 256, 512 and 256 filters, respectively. The kernel size is 3, upsampling size is 2, and all the activation functions are ReLU. The learning rate is 0.001. After the autoencoder is trained and the encoder is to hold fixed, that is, all the weights from the encoder part are fixed and used for the prediction part. The prediction model includes the encoder part and an extra part with a flatten layer and two dense layers. For this prediction model the input is the same as autoencoder model and contains all the 45150 features. The response is ADHD patients with values 1 and typical controls with values 0. In this prediction model only the last two dense layers are trained, first dense layer has 128 nodes and the last dense layer has two nodes with the activation function of sigmoid. The number of epochs are set to 400, early stopping is triggered when there is no improvement in the loss for ten consecutive epochs. The fine-tuned model gives an accuracy rate of 69%, specificity 89%, sensitivity 46%, precision 79%, and AUC 0.67. The results are also compared with studies in the literature. Details are shown in Table 3. The ROC curves for logistic regression, support vector machine, and correlation-encoder models are shown in Figure 5.

Mental disorder problems including ADHD are usually thought to be results of brain functional connectivity abnormality. When inter-voxel correlations are used as measures of the brain functional connectivity, we see that even the classical logistic regression and SVM are able to perform comparably with the most important 200 features. Unlike the studies in the literature such as [14], [15], [17] the neural networks were applied directly to the fMRI data, in the correlation autoencoder model the inter-voxel correlations were first created and the autoencoder then reduces noises and extract predictive features. The superior classification performance indicates this process is effective. It is worth remarking that further improvement may be possible if better functional connectivity measures become available.

TABLE 3
Classification performance of different models based on correlation (Accuracy, Specificity, Sensitivity, and Precision are abbreviated as Acc., Spec., Sens., and Prec., respectively, to condense space.)

Model	Acc.	Spec.	Sens.	Prec.	AUC
Competition [37]	51%	-	-	-	-
Logistic-100	61%	85%	33%	67%	0.59
SVM-G-100	61%	85%	33%	67%	0.59
Logistic-200	63%	85%	38%	69%	0.63
SVM-G-200	65%	89%	38%	75%	0.63
DeepFMRI [14]	63%	79%	48%	-	-
3D CNN [15]	63%	-	-	-	-
Deep Forest [36]	65%	-	-	-	-
CorrAE	**69%**	**89%**	**46%**	**79%**	**0.67**

4 Discussion and Conclusion

We proposed two new methods, the ICA-CNN and the correlation autoencoder, to classify the ADHD patients from typical controls. Both methods try to first summarize the inter-voxel information from the fMRI data and then apply the convolutional neural networks to extract predictive features. Both methods outperform the existing models. Based on 100 independent components extracted from each subject, the ICA-CNN approach gives an accuracy rate of 67% and AUC 0.65 on the ADHD-200 holdout test data from

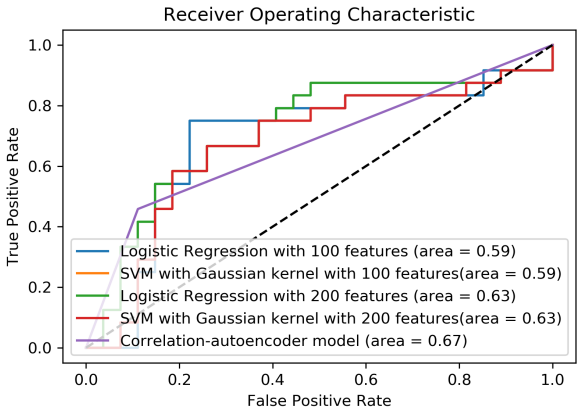


Fig. 5. ROC curve for logistic regression, SVM and Correlation-autoencoder models

the Peking University site. Based on 300 regions of interest extracted from each subject the correlation-autoencoder approach gives an accuracy rate of 69% and AUC 0.67, slightly better than ICA-CNN.

The two proposed methods give the accuracy rate near 70%. To a certain extent, they provide some guidance for clinical applications. Though, many open problems remain. Due to non-linear learning strategies with multiple layers, deep learning methods often give much better performance in many applications. However, they usually need a big sample data to train thousands of parameters to learn the latent features. In our analyses, the training sample contains only 168 observations which are not enough to have the parameters trained very well. It is reasonable to believe that the performance based on the two proposed methods will get better if the training data is large and the computer resources are allowed.

Although both methods proved to be effective and show higher accuracy and AUC scores, there is obvious imbalance between the sensitivity and specificity. Such imbalance not only appeared in our study, but also in previous studies such as [14]. This is probably caused by the imbalance between number of ADHD patients and the number of typical controls in the training set. Note that the CDAE-AdaDT method used in method [17] is able to maintain balance between sensitivity and specificity. Pairwise learning approaches [39] are usually effective to deal with imbalance data classification. They may shed light on future efforts to improve our approaches.

A common criticism on CNN and other deep learning models is their interpretability since the trained models are regarded as black boxes. There are recent research efforts to unpack the black boxes and improve the model interpretability [40], [41]. It will be an interesting research direction to apply these existing techniques to the deep learning models for fMRI data to discover the predictive features that contribute to the classification of ADHD patients from health people and improve our understanding of the mechanism of ADHD.

Acknowledgments

The authors gratefully thank the editor and two anonymous referees for their constructive comments and suggestions, which helped improve the quality of the paper. The work by Qiang Wu is partially supported by Simons Foundation (#712916) and NSF (DMS-2110826).

References

[1] M. L. Danielson, R. H. Bitsko, R. M. Ghandour, J. R. Holbrook, M. D. Kogan, and S. J. Blumberg, "Prevalence of parent-reported adhd diagnosis and associated treatment among us children and adolescents, 2016," *Journal of Clinical Child & Adolescent Psychology*, vol. 47, no. 2, pp. 199–212, 2018.

[2] J. A. Doshi, P. Hodgkins, J. Kahle, V. Sikirica, M. J. Cangelosi, J. Setyawan, M. H. Erder, and P. J. Neumann, "Economic impact of childhood and adult attention-deficit/hyperactivity disorder in the united states," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 51, no. 10, pp. 990–1002, 2012.

[3] E. Parens and J. Johnston, "Facts, values, and attention-deficit hyperactivity disorder (adhd): an update on the controversies," *Child and adolescent psychiatry and mental health*, vol. 3, no. 1, p. 1, 2009.

[4] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, "State-space model with deep learning for functional dynamics estimation in resting-state fmri," *NeuroImage*, vol. 129, pp. 292–307, 2016.

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

[6] S. Sarraf and G. Tofighi, "Deep learning-based pipeline to recognize alzheimer's disease using fmri data," in *2016 Future Technologies Conference (FTC)*. IEEE, 2016, pp. 816–820.

[7] S.-F. Liang, T.-H. Hsieh, P.-T. Chen, M.-L. Wu, C.-C. Kung, C.-Y. Lin, and F.-Z. Shaw, "Differentiation between resting-state fmri data from adhd and normal subjects: based on functional connectivity and machine learning," in *2012 International conference on Fuzzy Theory and Its Applications (iFUZZY2012)*. IEEE, 2012, pp. 294–298.

[8] Y. Monden, I. Dan, M. Nagashima, H. Dan, M. Uga, T. Ikeda, D. Tsuzuki, Y. Kyutoku, Y. Gunji, D. Hirano *et al.*, "Individual classification of adhd children by right prefrontal hemodynamic responses during a go/no-go task as assessed by fnirs," *NeuroImage: Clinical*, vol. 9, pp. 1–12, 2015.

[9] J. B. Colby, J. D. Rudie, J. A. Brown, P. K. Douglas, M. S. Cohen, and Z. Shehzad, "Insights into multimodal imaging classification of adhd," *Frontiers in systems neuroscience*, vol. 6, p. 59, 2012.

[10] H. Hart, K. Chantiluke, A. I. Cubillo, A. B. Smith, A. Simmons, M. J. Brammer, A. F. Marquand, and K. Rubia, "Pattern classification of response inhibition in adhd: toward the development of neurobiological markers for adhd," *Human Brain Mapping*, vol. 35, no. 7, pp. 3083–3094, 2014.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[13] D. Kuang and L. He, "Classification on adhd with deep learning," in *2014 International Conference on Cloud Computing and Big Data*. IEEE, 2014, pp. 27–32.

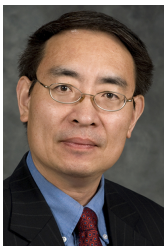
[14] A. Riaz, M. Asad, E. Alonso, and G. Slabaugh, "Deepfmri: End-to-end deep learning for functional connectivity and classification of adhd using fmri," *Journal of Neuroscience Methods*, vol. 335, p. 108506, 2020.

[15] 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. 23 626–23 636, 2017.

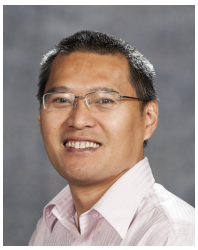
[16] Q.-H. Zou, C.-Z. Zhu, Y. Yang, X.-N. Zuo, X.-Y. Long, Q.-J. Cao, Y.-F. Wang, and Y.-F. Zang, "An improved approach to detection of amplitude of low-frequency fluctuation (alff) for resting-state fmri: fractional alff," *Journal of neuroscience methods*, vol. 172, no. 1, pp. 137–141, 2008.

[17] S. Liu, L. Zhao, X. Wang, Q. Xin, J. Zhao, D. S. Guttery, and Y.-D. Zhang, "Deep spatio-temporal representation and ensemble classification for attention deficit/hyperactivity disorder," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1–10, 2021.
[18] L. Liu, S. Tang, F. Wu, Y.-P. Wang, and J. Wang, "An ensemble hybrid feature selection method for neuropsychiatric disorder classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.
[19] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fmri data by blind separation into independent spatial components," *Human brain mapping*, vol. 6, no. 3, pp. 160–188, 1998.
[20] G. D. Brown, S. Yamada, and T. J. Sejnowski, "Independent component analysis at the neural cocktail party," *Trends in neurosciences*, vol. 24, no. 1, pp. 54–63, 2001.
[21] J. V. Stone, *Independent component analysis: a tutorial introduction*. MIT press, 2004.
[22] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
[23] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
[24] G. Varoquaux, S. Sadaghiani, J. B. Poline, and B. Thirion, "Canica: Model-based extraction of reproducible group-level ica patterns from fmri time series," *arXiv preprint arXiv:0911.4650*, 2009.
[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[26] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, pp. 84–96, 2014.
[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[28] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 490–497.
[29] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.
[30] C. C. Tan and C. Eswaran, "Reconstruction and recognition of face and digit images using autoencoders," *Neural Computing and Applications*, vol. 19, no. 7, pp. 1069–1079, 2010.
[31] C. Xiu and X. Su, "Composite convolutional neural network for noise deduction," *IEEE Access*, vol. 7, pp. 117 814–117 828, 2019.
[32] L. Yasenko, Y. Klyatchenko, and O. Tarasenko-Klyatchenko, "Image noise reduction by denoising autoencoder," in *2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. IEEE, 2020, pp. 351–355.
[33] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.
[34] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock, "The neuro bureau adhd-200 preprocessed repository," *Neuroimage*, vol. 144, pp. 275–286, 2017.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[36] L. Shao, D. Zhang, H. Du, and D. Fu, "Deep forest in adhd data classification," *IEEE Access*, vol. 7, pp. 137 913–137 919, 2019.
[37] A. sample, http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html.
[38] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.
[39] S. Liu and Q. Wu, "Pairwise learning for imbalanced data classification."
[40] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
[41] B. Wang, R. Ma, J. Kuang, and Y. Zhang, "How decisions are made in brains: Unpack "black box" of cnn with ms. pac-man video game," *IEEE Access*, vol. 8, pp. 142 446–142 458, 2020.

Donglin Wang received the PhD degree in computational science from Middle Tennessee State University. His current research interests include the functional magnetic resonance image analysis, pattern recognition, data mining, and machine learning.



Don Hong received his Ph.D. in Mathematics from Texas A&M University. He is a professor of Mathematical Sciences and the Computational Science Ph.D. Program. He is also the director of the Actuarial Science Program at Middle Tennessee State University. His current research interests include computational mathematics and statistics, predictive analytics, and medical image data analysis.



Qiang Wu received the PhD degree from City University of Hong Kong and postdoctoral training at Duke University. He is a professor in the Department of Mathematical Sciences, the director of Data Science Master and Graduate Certificate programs, and a faculty member of the Computational and Data Science PhD program at Middle Tennessee State University. His research interests include statistical learning theory, machine learning, data mining, and their applications.