Discriminating ADHD from healthy controls using a novel feature selection method based on relative importance and ensemble learning

Dongren Yao, Xiaojie Guo, Qihua Zhao, Lu Liu, Qingjiu Cao, Yufeng Wang, Vince D Calhoun, *Fellow IEEE*, Li Sun*, Jing Sui*, *Senior Member IEEE*

***Abstract*— Attention-deficit/hyperactivity disorder (ADHD) is a childhood-onset neurodevelopmental disorder that often persists into adulthood, resulting in adverse effects on work performance and social function. The current diagnosis of ADHD primarily depends on the judgment of clinical symptoms, which highlights the need for objective imaging biomarkers. In this study, we aim to classify ADHD (both children and adults [34/112]) from age-matched healthy controls (HCs [28/77]) with functional connectivity (FCs) pattern derived from resting-state functional magnetic resonance imaging (rs-fMRI) data. However, the neuroimaging classification of brain disorders often meets a situation of high dimensional features were presented with limited sample size. Thus an efficient method that is able to reduce original feature dimension into a much more refined subspace is highly desired. Here we proposed a novel Feature Selection method based on Relative Importance and Ensemble Learning (FS_RIEL). Compared with traditional feature selection methods, FS_RIEL algorithm improved the ADHD classification by about 15% in both child and adult ADHD classification, achieving 80-86% accuracy. Moreover, we found the most frequently selected FCs were mainly involved in frontoparietal network, default network, salience network, basal ganglia network and cerebellum network in both child and adult ADHD cohorts, which indicates that ADHD is characterized by a widely-impaired brain connectivity profile that may serve as potential biomarkers for its early diagnosis.**

I. Introduction

Attention-deficit/hyperactivity disorder (ADHD), the most common childhood-onset neurodevelopmental disorder, defined by a persistent pattern of inattention, hyperactivity, and impulsivity [1]. ADHD affects 5% children and adolescents and 2%-4% adults in the word [2, 3]. Throughout an individual's lifetime, ADHD patients are likely to increase the risk of other psychiatric disorders such as oppositional defiant disorder, conduct disorder or substance misuse which increase mortality. However, as for most mental disorders, the etiological bases and neural substrates of ADHD are far from being fully understood. Furthermore, current diagnosis of ADHD primarily depends on the judgment of clinical symptoms, and the misdiagnosis rate could be high [4]. Therefore, a more accurate discriminative method of ADHD based on objective imaging biomarkers is crucial for its early diagnosis and may facilitate better intervention and more effective treatment.

In literature, functional magnetic resonance imaging (fMRI) techniques have been extensively used in the quantitative analysis of the brain in healthy individuals and patients with psychiatric disorders in an attempt to increase our understanding of human brain functional networks [5, 6]. Following these studies, numbers of classification methods have been proposed for the diagnosis of ADHD with fMRI data [6-9]. Functional neuroimaging studies using resting-state fMRI have implicated alterations in functional connectivities (FCs) between multiple brain regions in ADHD [10-12]. However, FCs with the high dimensional small sample issue that treats original features as input space directly would degrade the sorting performance on distinguishing ADHD from healthy controls (HCs).

To overcome this drawback, in this study, we proposed a new feature selection method based on relative importance, to reduce high dimensional feature space into a much more refined subspace. The relative importance of features is calculated from decision trees. The value reveals the degree that every feature (i.e., node) contributes to the target label [13]. Features at the top of the tree get higher values since they make greater effects on the final prediction. Thus the relative importance of features thus can be used as an estimate of the relationship to the target label after average values on different trees Different ensemble methods calculate not the same relative importance of features. After combing these features, a forward-backward selection algorithm is employed to increase the diversity of new feature space while it still can maintain the low dimensionality of FCs feature space. Therefore, the final refined feature space will be constructed with selected features. A 10 fold cross validation strategy is used to estimate the performance of discriminating ADHD from HCs with the selected FCs resulted from the proposed method.

The remaining of this paper is organized as follows: section II mainly presents adults ADHD dataset and our proposed feature selection algorithm. In section III, both adults ADHD dataset and children ADHD dataset with age-matched HCs are used to test the proposed feature selection algorithm while compared with Lasso and ElasticNet methods. In section IV, we present our conclusion and discuss possible future research directions.

II. Materials and Methods

A. Participants

The first dataset includes 112 ADHD patients and 77 age-matched HCs who were recruited from clinics of Peking University Sixth Hospital (PKU6) or Beijing Normal

*Research supported by the Chinese National Natural Science Foundation and Chinese Academy of Sciences (CAS)

Dongren Yao and Jing Sui are with the Brainnetome center & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and University of Chinese Academy of Sciences. (Corresponding author: kittysj@gmail.com).

Vince D Calhoun is with the Mind Research Network, and Dept. of ECE, University of New Mexico, Albuquerque, NM, USA.

Xiaojie Guo, Qihua Zhao, Lu Liu, Qingjiu Cao, Yufeng Wang and Li Sun are with the National Clinical Research Center for Mental Disorders & Key Laboratory of Mental Health, Ministry of Health, Peking University (corresponding author: sunlioh@bjmu.edu.cn)

University (BNU). In the second dataset, 34 drug-naïve (stimulants and other psychotropic drugs) right-handed boys with ADHD and full-scale IQ score>80 were recruited from the child and adolescent psychiatric clinics of Peking University Sixth Hospital. 28 age- and sex-matched HCs were recruited from local primary schools. Demographic data were provided in Table 1.

TABLE I. Demographic Characteristics

	First Dataset (Adults)		Second Dataset (Children)	
	HC	**ADHD**	**HC**	**ADHD**
Site PKU6	43	73	28	34
Site BNU	34	39	NA	
Demographic				
Number	77	112	28	34
Gender (F/M)	34F/43M	37F/75M	All Boys	
Age (Mean±sd)	26.04±3.94	25.93±4.86	10.29±1.67	9.79±1.86

B. MRI Data Acquisition

Data were acquired on a Siemens Trio 3 T scanner (Siemens, Erlangen, Germany) at BNU while a GE Signa 3T Horizon HDx system (General Electric, Milwaukee, WI) at PKU6. Hospital. Functional (rs-fMRI) images were acquired using an echo-planar imaging sequence with the following parameters on the Siemens scanner: repetition time (TR) = 2,000 ms, echo time (TE) = 30 ms, flip angle= 90°, thickness/skip =3.5/0.7mm, matrix = 64 × 64, field of view (FOV) = 200 mm × 200 mm, 33 axial slices, and 240 volumes. And the parameters on the GE scanner were: TR = 2,000 ms, TE = 30 ms, flip angle = 90°, matrix = 64 × 64, FOV = 200 mm × 200 mm, 43 axial slices, slice thickness = 3.2 mm, slice gap = 0 mm in rs-fMRI.

C. Resting-State Functional Connectivity Analysis

Two datasets were both preprocessed using the Data Processing Assistant for Resting-State fMRI (DPARSFA, http://rfmri.org/DPARSF [14]). The first ten volumes were discarded to allow for magnetization equilibrium. Subsequent data preprocessing included slice timing correction, head motion correction, spatial normalization to the MNI template, resampling to 3 × 3 × 3 mm^3 , smoothing using a 4 mm Gaussian kernel, temporal band-pass filtering (0.01 Hz to 0.1 Hz), and regressing out nuisance signals of head motion parameters, white matter, CSF, and global signals. The registered functional MRI volumes with the MNI template were divided into 246 regions according to the Brainnetome Atlas [15] incorporating 210 cortical, 36 subcortical and 27 cerebellar regions.

Regional mean time series were obtained for each by averaging the functional MRI time series over all voxels in each of the 273 regions. Pearson correlation coefficients between pairs of node time courses were calculated and normalized to *z* score using Fisher transformation, resulting in a 273 × 273 symmetric connectivity matrix for each subject. Removing 273 diagonal elements, we extracted the upper triangle elements of the functional connectivity matrix as prediction features, i.e., the feature space for prediction was spanned by the (273 × 272)/2 = 37128 dimensional feature vectors.

D. Ensemble Feature Selection Algorithm

The success of machine learning algorithms in many areas, such as computer vision, speech recognition and so on, brings more probabilities and development space in computer-aided diagnosis system. With these algorithms, some patterns hidden in different subjects who suffer from the same disease can be found out easier. FC matrixes are of huge dimensionality, and the direct use of these features for classification often leads to low performance due to the "curse of dimensionality" [16]. To address this critical issue, ensemble learning methods are employed to establish a better-refined feature space. A benefit of using such methods like random forest or gradient boosting is that, after constructed those decision trees, it is relatively straight-forward to retrieve importance scores for each feature. In general, importance with a score [13] is calculated using formula (1) that indicates how informative or valuable each feature was averaged across all of the decision trees in the constructions of the classifier.

$$\pounds_l^2 = \frac{1}{M}\sum_{m=1}^{M}\sum_{t_m=1}^{J-1}\hat{l}_t^2 I(v(t_m)=l) \qquad (1)$$

t_m represents all J-1 internal nodes of m tree in all M trees. The function I returns one if condition true otherwise returns zero.

The high interpretability and generalization of feature space lead to build more refined subspace. Based on them, we proposed a novel feature selection algorithm named FS_RIEL. The flowchart of FS_RIEL algorithm is shown in Figure 1. As we mentioned in introduction, different ensemble methods do not generate the same relative importance of features. Hence, an ensemble learning thought was employed to maintain the low dimensionality of FCs feature space while it still via increasing the diversity of new feature space to ensure a better generalization.

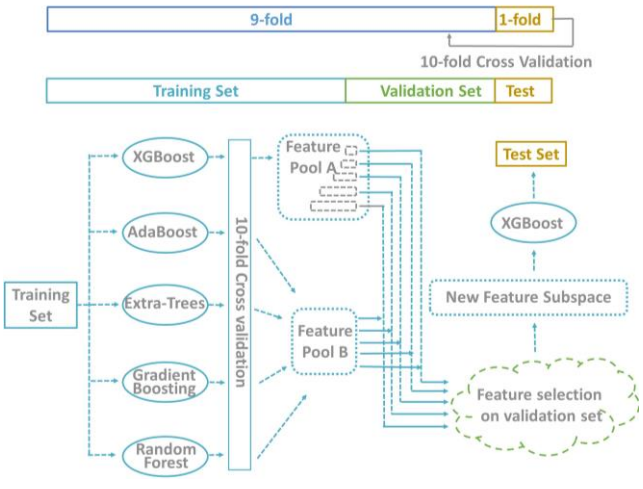

Figure 1 Flowchart of our proposed FS_RIEL algorithm

Extreme gradient boosting (XGBoost) [17] generates the features pool A with sorted features importance based on training data. To be clarified, Features pool A was composed of five different feature space that represents different numbers of features we employed (from 0.05% to 0.2% with the size of original features). In order to generate the features pool B, Four different algorithm including AdaBoost [18],

randomized decision trees (a.k.a. Extra-Trees) [19], Gradient Boosting [20] and Random Forest [21] were employed. Top 10 percent from different models' importance features were assembled without repetition. After combing these features, a forward-backward selection algorithm shown in Figure 2 will be used.

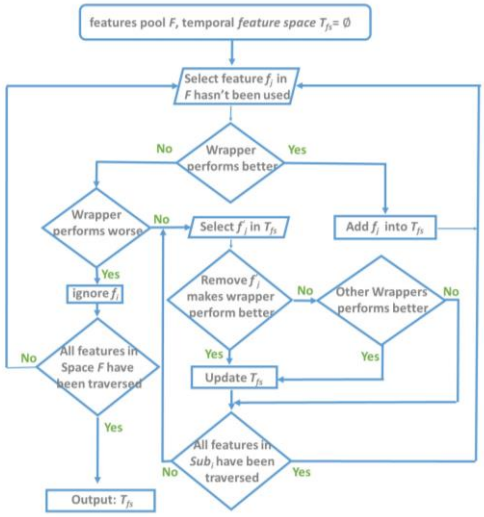

Figure 2 Details on forward-backward feature selection algorithm

XGBoost is used to measure the performance of the related feature subspace with validation set as the wrapper in our proposed forward-backward feature selection method. Other wrappers are employed to guarantee more informative features would take into consideration. The final results come from a voting strategy that each wrapper keeps the same weights. Then the final refined feature subspace will be constructed with selected features.

III. Results And Analysis

In this study, a ten-fold cross-validation strategy was used to estimate the performance of discriminating both adults and children ADHD from age-matched HCs with the selected FCs selected from the proposed method. As shown in Figure 1, for each loop, 10% subjects were left untouched for testing, the remaining data were further split into a training set (2/3, 60%) and a validation set (1/3, 30%). Each loop generates a new robust feature subspace for classification to predict labels of the test data. Then the loops go on ten times until all subjects were tested. FS_RIEL algorithm is also compared with traditional Lasso [22] and ElasticNet [23] methods by using metrics of *accuracy* (*ACC*), *sensitivity* (*SEN*) and *specificity* (*SPE*). Denote TP, TN, FP and FN as true positive, true negative, false positive and false negative, respectively. These evaluation metrics are defined as: *ACC* = (TP+TN)/ (TP+TN+FP+FN), *SEN*= TP/ (TP+FN), *SPE*=TN/ (TN+FP)

Lasso and ElasticNet are tuned their parameter with grid search method. Lasso uses parameter alpha to control sparsity, and the values of alpha are chosen from {0.005, 0.006, … , 0.1}. ElasticNet has parameters l1_ratio and alpha to control penalty terms. The values of alpha are varied from 0.1 to 2 with step 0.1while l1_ratio changing from 0.02 to 0.7 also with step 0.01.

A. adults ADHD vs. age-matched HCs

After optimizing the parameters of Lasso and ElasticNet model, the parameter alpha in Lasso is 0.03. The l1_ratio and alpha in ElasticNet equal 0.1, 0.35 respectively. Lasso and ElasticNet Method hold stable performance while our proposed algorithm cannot guarantee this. Figure 3 with FS_RIEL algorithm illustrates the average of accuracy, sensitivity, and specificity in 10 times for each fold while others present the best results. The results indicate mean accuracy is **80.0%** with our algorithm better than Lasso (67.80%) and ElasticNet (66.67%). The mean sensitivity is **90.83%** vs. 76.74%, and 77.58% while mean specificity is **64.89%** vs. 54.82% and 50.71%. Meanwhile, less half FCs (43.02 vs. 104.9/120.5) were employed in FS_RIEL algorithm.

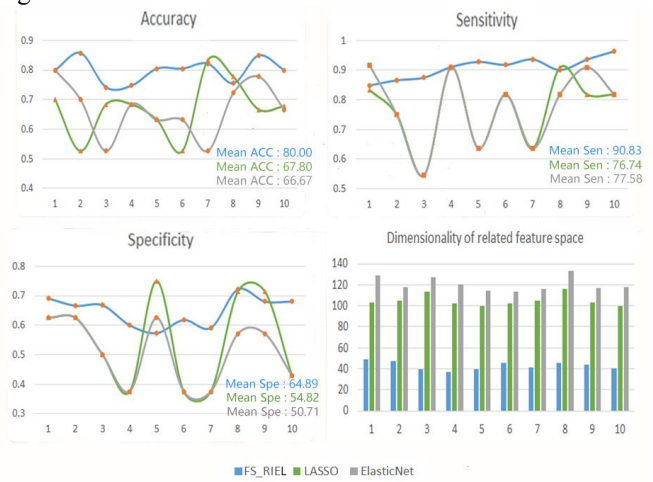

Figure 3 Four metrics for performance evaluation on different feature selection methods with adults ADHD.

The most frequently selected FCs via feature selection method with relative importance show in figure 4(a). Lines with more width denotes more frequency were used in new space, and two pink lines are the top two FCs. Figure 4(c) demonstrates the new feature subspace with FCs in 4(a) reduced related dimensions into two by T-SNE[24].

The most frequently selected FCs were mainly involved in frontoparietal network, default network, salience network, basal ganglia network and cerebellum network, consistent with previous findings that large-scale brain networks were impaired in ADHD [25].

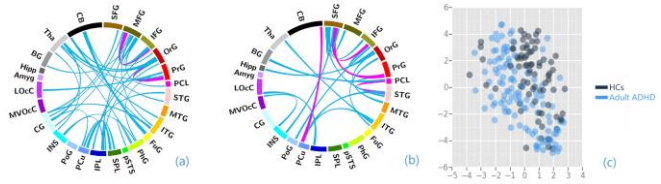

Figure 4 Related FCs in adults, children ADHD and 2D visualization after t-sne algorithm on adult ADHD.

B. Children ADHD vs. age-matched HCs

The same methods are also used in children ADHD dataset. We still had optimized related feature selection models. The parameter alpha in Lasso is 0.0018. The l1_ratio and alpha in ElasticNet is equal to 0.12, and 1.6, respectively. 5-fold cross-validation is used to validate methods' performance as the children dataset is not big enough. The final results are

showed in Figure 5. The most frequently selected FCs via FS_RIEL are showed in figure 4(b).

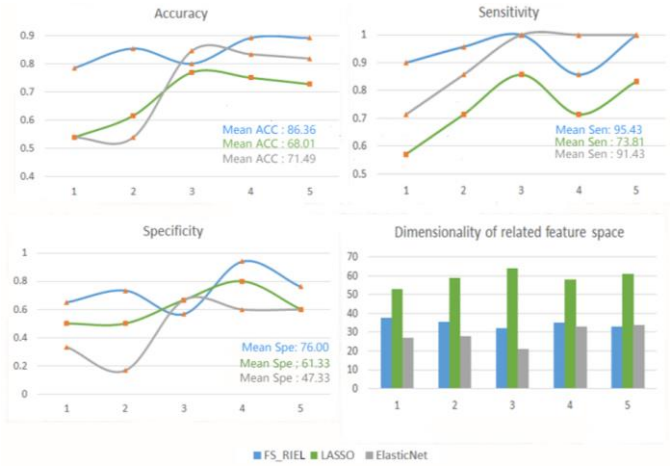

Figure 5 Four metrics for performance evaluation on different feature selection methods with children ADHD.

C. *Future directions*

Overall, our studies offer a new feature selection strategy to tackle high dimensional with small samples on neuroimage classification issues. Because FCs features would not the only way describing fMRI data, other heterogeneous fMRI features like ReHo, fALFF should be considered into our proposed method. Furthermore, structure MRI and Diffusion Tensor Imaging (DTI) features depicted from different views when compares with fMRI. Fusion multi-modality image features might improve performance on classification. Therefore, we plan to pursue these possibilities in our future work.

IV. CONCLUSION

In summary，we proposed a novel feature selection method named FS_RIEL that is able to reduce original feature space into a much more refined subspace when facing high dimensional features with limited sample size. Compared with traditional feature selection methods, FS_RIEL performed much higher accuracy on ADHD classification issue. To the best of our knowledge, this is the first attempt using feature selection method on both child, and adult ADHD fMRI datasets achieved high accuracy. Furthermore, the most frequently selected FCs consistent with previous findings that large-scale brain networks were impaired in ADHD.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation No. 61773380, 81471367, 81771479, 81471382, and 81641163, the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB02060005), “100 Talents Plan” of Chinese Academy of Sciences, and NIH grants P20GM103472, R01EB006841 and R01EB005846.

REFERENCES

[1] Biederman J, Faraone SV. Attention-deficit hyperactivity disorder. [J]. CNS Drugs, 2005, 20(2):107-123.

[2] Kooij S J, Bejerot S, Blackwell A, et al. European consensus statement on diagnosis and treatment of adult ADHD: The European Network Adult ADHD[J]. BMC Psychiatry, 10, 1(2010-09-03), 2010, 10(1):67.

[3] Kessler R C, Adler L, Barkley R, et al. The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication [J]. American Journal of Psychiatry, 2006, 163(4):716.

[4] Willcutt E G. The Prevalence of DSM-IV Attention Deficit/ Hyperactivity Disorder: A Meta-Analytic Review [J]. Neurotherape -utics the Journal of the American Society for Experimental Neurothera -peutics, 2012, 9(3):490-499.

[5] Biswal BB, Mennes M, Zuo XN, et al. Toward discovery science of human brain function. [J]. Neuroscience Research, 2010, 71(10): e30-e31.

[6] Sato J R, Hoexter M Q, Fujita A, et al. Evaluation of Pattern Recognition and Feature Extraction Methods in ADHD Prediction[J]. Frontiers in Systems Neuroscience, 2012, 6(4):68.

[7] Craddock R C, James G A, Iii P E H, et al. A whole brain fMRI atlas generated via spatially constrained spectral clustering [J]. Human Brain Mapping, 2012, 33(8):1914.

[8] Brown M R, Sidhu G S, Greiner R, et al. ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements [J]. Frontiers in Systems Neuroscience, 2012, 6:69.

[9] Du J, Wang L, Jie B, et al. Network-based classification of ADHD patients using discriminative subnetwork selection and graph kernel PCA[J]. Computerized Medical Imaging & Graphics the Official Journal of the Computerized Medical Imaging Society, 2016, 52:82-88.

[10] Castellanos F X, Kelly C, Milham M P. The restless brain: attention-deficit hyperactivity disorder, resting-state functional connectivity, and intrasubject variability [J]. Canadian Journal of Psychiatry Revue Canadienne De Psychiatrie, 2009, 54(10):665-72.

[11] Damien A. Fair, Jonathan Posner, Bonnie J. Nagel, et al. Atypical Default Network Connectivity in Youth with ADHD [J]. Biol Psychiatry. 2010, 68(12):1084-1091.

[12] Bush G. Cingulate, Frontal and Parietal Cortical Dysfunction in Attention-Deficit/Hyperactivity Disorder [J]. Biological Psychiatry, 2011, 69(12):1160-7.

[13] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees [M]. CRC press, 1984. 6, pp. 2619-31

[14] Chao-Gan Y, Yu-Feng Z. DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI [J]. Frontiers in Systems Neuroscience, 2010, 4(13):13.

[15] Fan, Lingzhong, et al. "The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture:" Cerebral Cortex 26.8 (2016):3508.

[16] Duda R O, Hart P E, Stork D G. Pattern Classification (2nd Edition) [J]. En Broeck the Statistical Mechanics of Learning Rsity, 2000.

[17] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-94.

[18] Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer & System Sciences, 2010; 55: 119-39.

[19] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning, 2006; 63: 3-42.

[20] J. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, the Annals of Statistics, Vol. 29, No. 5, 2001.

[21] Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1):5-32.

[22] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.3(2011):273-282.

[23] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005): 301-320.

[24] Laurens V D M. Accelerating t-SNE using tree-based algorithms [J]. Journal of Machine Learning Research, 2014, 15(1):3221-3245.

[25] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005): 301-320.