Multi-Task Learning for Efficient Diagnosis of ASD and ADHD using Resting-State fMRI Data

1st Zhi-An Huang
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
zahuang2-c@my.cityu.edu.hk

2nd Rui Liu
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
rliu38-c@my.cityu.edu.hk

3rd Kay Chen Tan, *Fellow, IEEE*
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
kaytan@cityu.edu.hk

***Abstract*—Increasing mental disorders have emerged as an urgent public health concern such as autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD). Related mental disorders may share high overlap in clinical symptoms. Therefore, their diagnosis can be challenging to merely rely on the observation of cognitive phenotypes and behavioral manifestations. Unfortunately, there is no additional support of biochemical markers, laboratory tests, or neuroimaging analysis, which can be used as a diagnostic gold standard currently. Over the past decades, resting-state functional magnetic resonance imaging (rs-fMRI) has been considered as one of the most promising modality to capture the intrinsic neural activation patterns between regions in the brain. In this work, we focus on ASD and ADHD due to their high prevalence and relevance with the aim to exploit the multi-task learning (MTL) paradigm for their diagnosis. To the best of our knowledge, this is the first time to make use of the disease-specific heterogeneities for the MTL classification of ASD and ADHD via rs-fMRI signal. We propose a novel graph-based feature selection method to filter out irrelevant functional connectivity features. Then an efficient structure of multi-gate mixture-of-experts (MMoE) is applied to the MTL classification framework. Finally, the experiment results demonstrate that the proposed model can achieve a reliable classification performance in a short term, yielding the mean accuracies of 0.687$\pm$0.005 and 0.650$\pm$0.014 in ASD and ADHD datasets, respectively. The graph-based feature selection method and MMoE model are demonstrated to make great contribution to performance improvement.**

***Index Terms*—Autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), multi-task learning (MTL), functional magnetic resonance imaging (fMRI), functional connectivity (FC)**

I. Introduction

Mental disorders such as Autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) are considered to involve disturbances in the normal-range activities of brain functional regions. Nevertheless, the diagnosis of mental disorders is determined merely by the symptom-based clinical interview. No existing gold standards can be offered for definitive validation so far. With the advance of brain functional neuroimaging techniques, fMRI has rapidly emerged as a promising tool to effectively evaluate the dynamic and robust changes among functionally interconnected regions in brain with high spatial resolution. As one of the fMRI paradigm, resting-state functional magnetic resonance imaging (rs-fMRI) has gained widespread application in neuroscience research by exploring the modular nature of cortical function, based on the spontaneous low frequency fluctuations in blood-oxygen-level dependent (BOLD) signal. Investigating the alteration of functional connectivity (FC) among regions of interest (ROIs) in brain between disorders such as ASD and ADHD provides new insights into their underlying mechanisms [1]. Therefore, the remarkable changes in FC can be identified as a powerful biomarker for classifying individual patients thanks to the big data analytics in neuroimaging and data-driven methods in artificial intelligence.

There has been a fair amount of work using machine learning approaches to distinguish mental patients and typical controls (TCs) based on the FC features [2]–[4]. Particularly, multi-task learning (MTL) as an emerging subfield of machine learning that aims to solve multiple related problems (”tasks”) simultaneously, can lead to substantial improvements in classification performance. Considering the shared knowledge exploited by MTL, it is quite beneficial in situations where the integration of data shared highly consistent feature patterns in conditions of between-subject [5], between-modality [6] and between-site [7]. For examples, Marquand *et al*. developed a MTL method to model the relationships between a group of subjects using Gaussian process priors and then extracted the subject-specific features to generate more accurate models [5]. Hu and Zeng [8] presented a MTL framework to simultaneously capture the site-shared and site-specific features from three data sites for discriminating schizophrenic patients from TCs. However, all these MTL models are learned based on one single disorder's data source and thus fail to take advantage of the shared information among related mental disorders. In other words, sufficient statistical power in deciphering subtle but significant patterns in FC, might be difficult to obtain in multiple homogeneous tasks only focusing on one disorder. On the other hand, learning the pattern of one disorder is at least somewhat noisy and thus bears the risk of overfitting to this disorder. Through averaging the noise patterns, the utilization of disease-specific heterogeneities across multiple related disorders can ideally ignore the data-dependent noise and then learn a more general representation.

To solve the above issues, we attempt to model the relationships between ASD and ADHD using a MTL framework based on the fact that both of them are highly relevant

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

sharing similar patterns in several ways. First, the symptoms of ASD and ADHD overlap and between 30% and 50% of individuals with ASD are observed to manifest ADHD symptoms [9]. Furthermore, both ASD and ADHD are highly heritable. Satterstrom's research [10] supports the idea that they both share a similar burden of variants in high risk genes, contributing to the underlying biological mechanisms being involved. We expect that the MTL model for classifying multi-disease neuroimaging data could improve the diagnosis of related diseases by providing a more complete picture of the their hidden common causes.

In this work, we first propose a MTL classification framework to identify ASD and ADHD subjects from TCs based on the structure of multi-gate mixture-of-experts (MMoE) [11], which has a gating network trained to optimize each expert submodels across all tasks. To select the remarkable FC features distinguishing between ASD and ADHD, a graph-based feature selection (GBFS) method is developed based on both external and internal measures. As MTL is the concept of knowledge transfer implying a sequentially shared representation with different levels of abstraction, the pre-trained model without further adjustment can be used as a promising feature extractor. We demonstrated that the pre-trained model can be used to efficiently learn a new classification task with local training in smaller task-specific tower networks. Compared to training the whole MTL network, such local training can spend only one-eighth of time to achieve convergence, yielding a comparable classification performance. In the era of "big data", the proposed framework is anticipated to provide valuable insights into the auxiliary diagnosis of highly relevant mental disorders using rs-fMRI signal.

The rest of this paper is organized as follows. Section II describes the datasets and their preprocessing pipelines. Section III explains the main idea of the proposed GBFS method and the MTL classification framework based on the structure of MMoE. Section IV specifies the settings of model structure and parameters, and shows the experimental studies for verifying the proposed model. Some discussions are also provided for some extensions of this work. Section V is the conclusion of the paper.

II. Datasets and preprocessing

To reveal the complex brain mechanisms underlying ASD and ADHD, the scientific community dedicated to aggregate and release two large-scale rs-fMRI collections from laboratories around the world.

The Autism Brain Imaging Data Exchange I (ABIDE) database [12] includes 505 ASD samples and 530 TC samples from 17 independent international sites. There is no clear consensus on a best preprocessing pipeline of raw rs-fMRI data. In an effort to open share the preprocessed ABIDE data, the Preprocessed Connectomes Project (http://preprocessedconnectomes-project.org/abide/) announced the public release from five different teams using their preferred tools. We select the data preprocessed through the Configurable Pipeline for the Analysis of Connectomes (C-PAC). The preprocessing with C-PAC makes use of AFNI, FSL, ANTs software libraries and custom python code involving the following procedures: correction of slice time and motion, skull stripping, global mean intensity normalization, regressing out nuisance signal (24 motion parameters and 5 principal components of CompCor [13]) as well as linear and quadratic trends, band-pass filtering (0.01-0.1 Hz), functional image transformation, and spatially smoothing with a 6-mm Gaussian kernel of full width at half maximum (FWHW).

The ADHD-200 Consortium provides 285 ADHD samples and 491 TC samples aggregated from 8 independent imaging sites (http://fcon_1000.projects.nitrc.org/indi/adhd200/). Through efforts of the ADHD-200 consortium, Neuro Bureau and Virginia Tech's ARC, the preprocessing strategy for the Athena was performed based on AFNI and FSL software libraries to achieve a high quality transformation between MNI space and subject space. This pipeline is implemented as following: removal of first 4 echo-planar image (EPI) volumes, slice-timing correction, deoblique dataset, motion correction, masking the dataset to exclude voxels at non-brain regions, averaging the EPI volumes for a mean functional image, co-registration of mean EPI image to the respective anatomic image, spatial transformation of rs-fMRI data and mean image into template space at $4 \times 4 \times 4$ mm^3 resolution, extraction of the time-courses of rs-fMRI from white matter (WM) and cerebrospinal-fluid (CSF), regressing out WM, CSF, head motion and a low order polynomial from EPI data, band-pass filtering (0.009-0.08 Hz), and spatial smoothing using a 6-mm FWHW Gaussian filter.

Since the identification of nodes in brain functional networks is of great significance, appropriate atlases can facilitate the quantification of brain networks by parceling the brain into a certain number of ROIs. In this work, the Craddock 200 (CC200) atlas [14] is used to extract the mean time-series for a set of 200 ROIs by normalized cut spectral clustering. Such that, the four-dimensional raw rs-fMRI data can be downsampled to a two-dimensional feature matrix $\boldsymbol{T}$, where T_{ij} represents the ith ROI's mean time series of jth timestamp.

III. Model design

As shown in Fig. 1, the proposed model can be divided into two steps: 1) We conduct the GBFS to select the remarkable connections of ASD and ADHD distinguishing from each other as well as TC group, using both external and internal measures; 2) By filtering out the irrelevant features, the input data is fed to the MTL classification framework based on MMoE structure.

A. Graph-Based Feature Selection

To quantitatively characterize the correlations between time courses of functionally linked ROIs in the brain, it is widely to adopt the Pearson correlation coefficient (PCC) thanks to its efficiency. In this work, 190 ROIs in CC200 atlas are selected in order to harmonize the features setting. The possible FC features/connections between theses 190 ROIs add

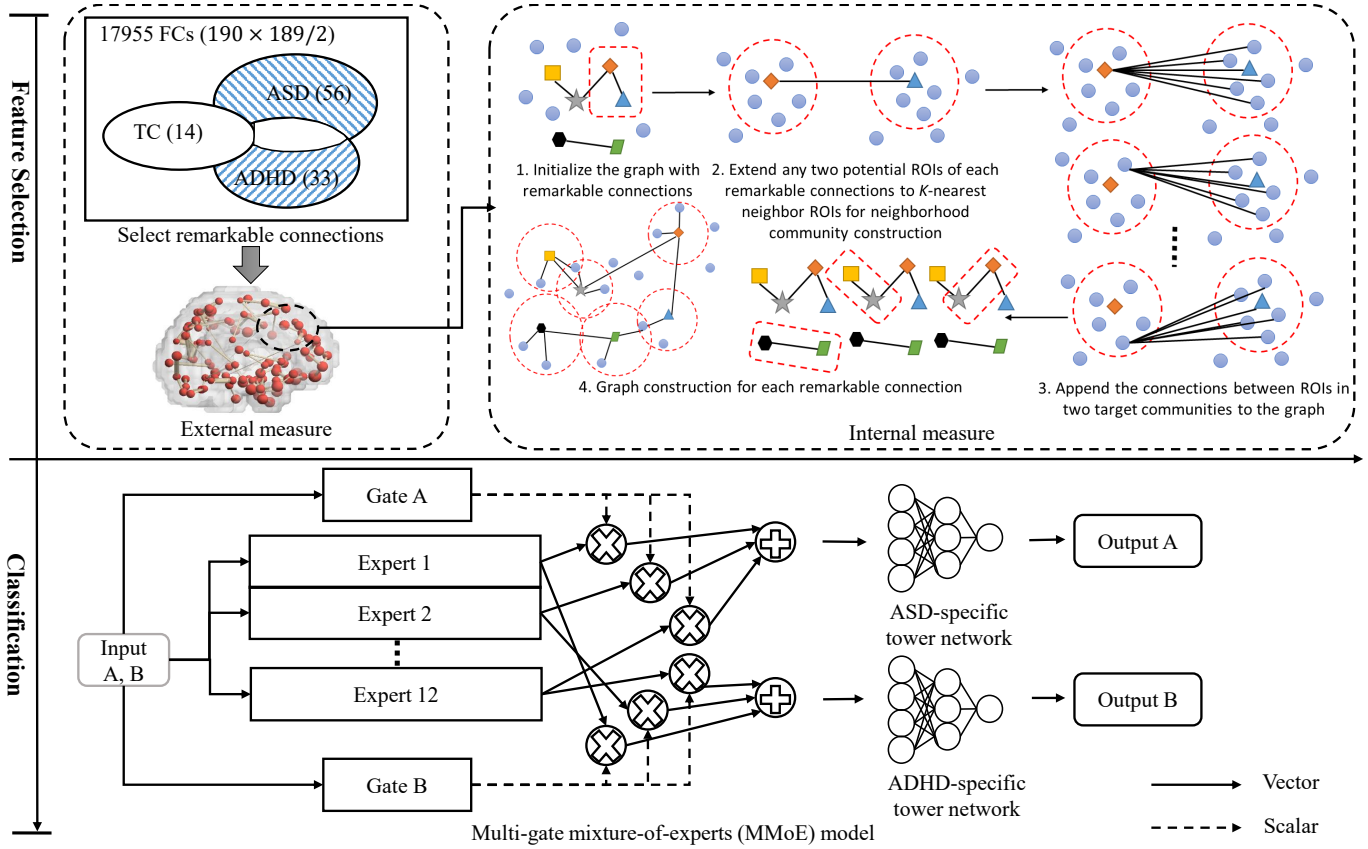

Fig. 1. The flowchart of the proposed model.

up to 17,955 ($= \frac{190\times189}{2}$), retrieving from the upper triangle values of a correlation matrix. We develop a novel graph-based feature selection (GBFS) method to incorporate remarkable FC features in terms of both external and internal measures.

First, as a data-drive approach, the external measure is proposed to globally identify the unique neural patterns associated with ASD and ADHD respectively. We divide the datasets into three groups, i.e., ASD group, ADHD group and TC group. Their means and standard deviations (STDs) are then calculated. The global measure selects the remarkable connections based on two following criteria: 1) The mean values of remarkable connections (denoted as $mean(FC)$) should be significantly distinctive from the average levels in ASD group or ADHD group. 2) The selected remarkable connections should be simultaneously distinguishable from ASD group, ADHD group and TC group. This means that they should be only remarkable in ASD/ADHD group while general in TC group and ADHD/ASD group. Accordingly, the remarkable connections for ASD group can be mathematically subject to

$$\begin{cases} \|mean(FC_{ASD}(i,j)) - mean_{ASD}\| > \alpha * STD_{ASD} \\ \|mean(FC_{TC}(i,j)) - mean_{TC}\| \leq \alpha * STD_{TC} \\ \|mean(FC_{ADHD}(i,j)) - mean_{ADHD}\| \leq \alpha * STD_{ADHD} \end{cases} \tag{1}$$

where the filter factor α determines how the selected connections should be highly remarkable, and the $mean(FC_{ASD}(i,j))$ represents the global mean of a given connection between the *i*th ROI and *j*th ROI in ASD group. Likewise, the remarkable connections for ADHD group can be mathematically selected as follows.

$$\begin{cases} \|mean(FC_{ADHD}(i,j)) - mean_{ADHD}\| > \alpha * STD_{ADHD} \\ \|mean(FC_{TC}(i,j)) - mean_{TC}\| \leq \alpha * STD_{TC} \\ \|mean(FC_{ASD}) - mean_{ASD}\| \leq \alpha * STD_{ASD} \end{cases} \tag{2}$$

In this way, 56 ASD remarkable connections and 33 ADHD remarkable connections are finally chosen for the following internal measure.

We assume that mental disorders could dysfunction with the collaborative activation patterns in potential associated areas, instead of the single ROI-to-ROI connectivity interaction. To make use of the spatial distribution information, we design a graph normalization of *K*-nearest neighbors as internal measure to detect relevances between two ROIs' subgraphs of each remarkable connections derived from the external measure. The subgraph normalization is performed by incorporating those *K*-nearest neighbor ROIs according to Euclidean distance. This can be considered as the "receptive field" to append their connections between ROIs to the whole graph. As a consequence, we only consider the edges in graph as the final remarkable connections. Their FC features calculated by PCC are used as an input for the classification framework of MMoE model.

B. MMoE Model

We present a MTL classification framework to identify ASD and ADHD subjects from TCs using MMoE structure, which is inspired by the recent advance in [11], [15], [16]. As we can see in Fig. 1, MMoE structure consists of a group of expert networks, as well as the gating networks and individual tower networks for each task. The main idea of MMoE is to model the task relationships in a sophisticated way by assembling the expert networks with different weights, allowing to learn different mixture patterns for different tasks. In this work, the expert networks have the same structure as a mixture-of-expert (MoE) layer [15] with a certain number of neuron units, which is stacked as a basic block of artificial neural networks (ANNs) and trained in an end-to-end way. The gating network g^k for each task k performs a linear transformation of the input x with the *softmax* function as follows.

$$g^k(x) = \textit{softmax}(W_k x) \tag{3}$$

$$\text{s.t.} \sum_{i=1}^{m} g_i^k(x) = 1 \tag{4}$$

where $W_k \in \mathbb{R}^{m \times n}$ is a trainable matrix to be learned, m and n represent the numbers of expert networks and feature dimension of x, respectively. It is easy to see that g_i^k indicates the confidence (probability) for the *i*th expert network. Then g^k can be integrated with the MoE model for the output of task k as:

$$\mathcal{H}^k(x) = \sum_{i=1}^{m} g_i^k(x) h_i(x) \tag{5}$$

where h_i is the output of the *i*th expert network. Such that, $\mathcal{H}^k$ can be fed to the corresponding task-specific tower network $\mathcal{F}$. The task-specific tower networks are multi-layer ANNs with a dropout regularization. Consequently, the final MMoE model can be formulated for a given task *k* as,

$$y^k = \mathcal{F}^k(\mathcal{H}^k(x)). \tag{6}$$

IV. Experiments & Results

A. Model Structure and Parameters

In this work, as the filter factor α and the number of nearest neighbor ROIs *K* are empirically set to 4 and 6 respectively, the input dimension for the MMoE model is $(33+56)\times(6+1)^2 = 4361$. The MoE model have 12 expert networks where each network is implemented as a single MoE layer with size = 400. On the top of the bottom MoE model, the task-specific tower networks are multi-layer ANN models with the configuration of 400-64-10-1, where a dropout factor of 0.5 is applied to the first layer to regularize the network. The linear activation function is used for the final outputs. We apply the rectified linear unit (ReLU) activation function to hidden layers of both the MoE model and tower networks. The whole MMoE model is trained using Adam optimizer with a learning rate of 0.001.

TABLE I
Performance comparison on ABIDE I.

Model	Classifier	Validation	Sample #	Acc (STD)
Heinsfeld *et al.* 2018	DNN	10-fold CV	1035	0.700 (N.A.)
Dvornek *et al.* 2017	LSTM	10-fold CV	1035	0.685 (0.055)
Plitt *et al.* 2015	L-SVMs	10-fold CV	178	0.697 (0.027)
Chen *et al.* 2015	RFE-SVM	IS	252	0.660 (N.A.)
Abraham *et al.*2017	ℓ_2-SVC	10-fold CV	871	0.669 (0.027)
Nielsen *et al.* 2013	LOO linear	LOOCV	964	0.600 (N.A.)
Ghiassian *et al.* 2016	RBF-SVM	IS	1035	0.592 (N.A.)
Our model (Mean)	MMoE	10-fold CV	1035	0.687 (0.005)
Our model (Best)	MMoE	10-fold CV	1035	0.694 (N.A.)
Our model (Worst)	MMoE	10-fold CV	1035	0.675 (N.A.)

TABLE II
Performance comparison on ADHD-200.

Model	Classifier	Validation	Sample #	Acc (STD)
Tan *et al.* 2017	Linear SVM	10-fold CV	217	0.687 (N.A.)
Chang *et al.* 2012	Linear SVM	10-fold CV	436	0.700 (N.A.)
Du *et al.* 2016	SVM	10-fold CV	216	0.950 (N.A.)
Ghiassian *et al.* 2013	RBF-SVM	IS	1069	0.630 (N.A.)
Daietal *et al.*2012	RBF-SVM	IS	776	0.590 (N.A.)
Eloyanetal *et al.* 2012	Various	IS	776	0.610 (N.A.)
Ghiassian *et al.* 2016	RBF-SVM	IS	776	0.700 (N.A.)
Dey *et al.* 2012	PCA-LDA	IS	734	0.700 (N.A.)
Fair *et al.* 2013	SVM	IS	668	0.710 (N.A.)
Colby *et al.* 2012	RBF-SVM	IS	776	0.550 (N.A.)
Siqueira *et al.* 2014	Linear SVM	LOOCV	609	0.730 (N.A.)
Our model (Mean)	MMoE	10-fold CV	776	0.650 (0.014)
Our model (Best)	MMoE	10-fold CV	776	0.674 (N.A.)
Our model (Worst)	MMoE	10-fold CV	776	0.638 (N.A.)

B. Results

To evaluate the classification performance for ABIDE I and ADHD-200 databases, 10-fold cross validation (CV) and independent sets of training/validation (IS) are widely used for the validation.

10-fold CV: First the original datasets are randomly divided into 10 equal subsets. We keep 1 subset as validation set and train the model using all the remaining 9 subsets. This process is repeated 10 rounds until each subset is used as validation set in turns. The results from all the 10 rounds are then averaged to estimate the model's accuracy.

IS: There are plenty of ways to split the original datasets into independent training and validation sets. The common way for ABIDE database is to randomly assign 80% of the data for training and leave the remaining 20% of the data as a validation set. As ADHD-200 database is originally released for the ADHD-200 Global Competition, the datasets have been officially divided into training and validation sets.

Here we choose 10-fold CV to assess the proposed model. To reduce the bias caused by random sampling, ten times of 10-fold CV are conducted. Our model is compared with previous studies (including seven ABIDE-based models and eleven ADHD-200-based models), whose results are derived from the recent reviews [17]–[19]. Because of the sophisticated

TABLE III
COMPARISON WITH HEINSFELD'S WORK.

Model	Heinsfeld *et al*. 2018	Our model
Acc	0.700 (SEN 74.0%, SPE 63.0%)	0.687 (SEN 68.9%, SPE 68.6%)
Time	32h52m36s	7m12s
CPU	2 Intel Cores Xeon E5-2620@2GHz	1 Intel Core i7-8700K@3.7GHz
GPU	1 NVIDIA Tesla K40	1 NVIDIA RTX 2080 Ti
RAM	48 GB	32GB

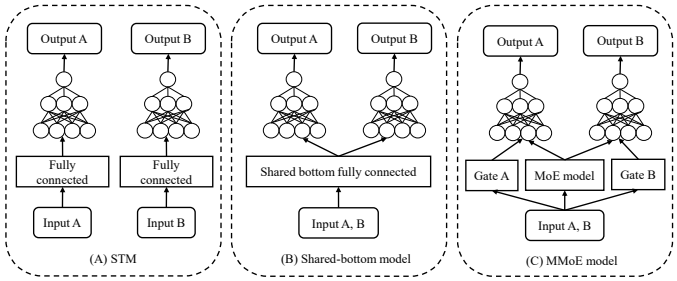

Fig. 2. The architecture of STM, shared-bottom model and MMoE model.

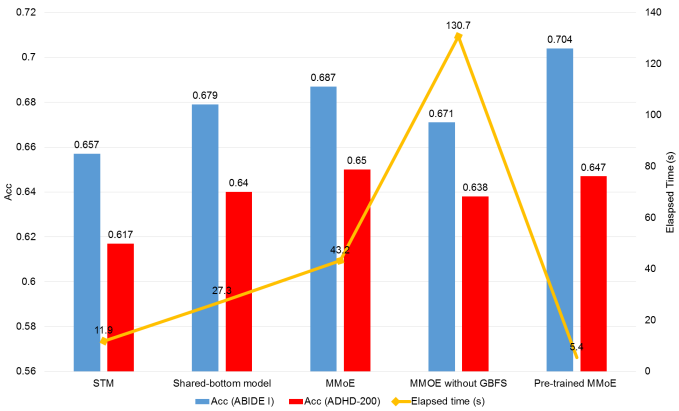

Fig. 3. The performance comparison to evaluate GBFS and MMoE in terms of accuracy and elapsed time.

sampling strategies such as IQ-matched participants and similar MRI acquisition protocols, a part of previous works are restricted to use relatively small datasets. Arbabshirani *et al*. [20] demonstrated that the reliable, robust classification accuracies they achieved degrade significantly as sample data increases. The performance comparison on ABIDE I and ADHD-200 are shown in Table. I and Table. II, respectively. As we can see that, the results of most previous models are given from the best case based on one-time evaluation. However, the effect of random sampling should be considered. For a fair comparison, our model's performance is analyzed in three scenarios, i.e., the mean-case, best-case and worst-case scenarios. Besides 10-fold CV and IS, some researchers utilized leave-one-out cross validation (LOOCV) for validation.

In regards to the comparison on ABIDE I, our model obtains a reliable mean accuracy of 0.687±0.005 (sensitivity 68.9%, specificity 68.6%) comparable to Heinsfeld's model [19], which achieves the best accuracy of 0.700 (sensitivity 74.0%, specificity 63.0%) reported to data using the whole ABIDE I datasets. Thanks to feature selection and simplified model design, it allows to increase efficiency by greatly shortening the training time. As we can see in Table. III, based on the similar runtime environment, the average elapsed time for training a 10-fold CV in Heinsfeld's work is 32 hours 52 minutes 36 seconds for ABIDE I classification, while the proposed model requires 7 minutes 12 seconds to complete a 10-fold CV for MTL classification of ABIDE I and ADHD-200, showing a 270X speed up. On this point, there is great potential to further develop our model in the future. We can take advantage of speed to allow sufficient iterations for exploring the optimal hyperparameter settings of model structure and training algorithm via automatic hyperparameter tuning techniques (e.g. Bayesian optimization [21] and genetic programming [22]). As expected, deep learning approaches outperforms the traditional machine learning approaches. It would benefit from their hierarchical structure with different levels of complexity as well as non-linear transformations. As for the comparison on ADHD-200, previous works tend to employ support vector machine (SVM) to achieve high accuracies. The proposed model also achieve a high mean accuracy of 0.650±0.014 (sensitivity 76.7%, specificity 58.1%) based on 10-fold CV using the whole ADHD-200 datasets. The unbalanced sensitivity and specificity could be attributed to the inherent data quality of ADHD-200. This similar phenomenon can also be observed in several previous works. Tan's model [23] achieves mean accuracy of 0.69 with 78% sensitivity and 57% specificity in 10-fold CV and Dey's model [24] achieves the best accuracy of 70% with 87% sensitivity and 49% specificity in IS.

C. Effects of Features Selection and MMoE Model

The main contribution of this work is to propose a novel GBFS method and apply MMoE model to the classification framework. In this section, we aim to evaluate their effects based on 10-fold CV. To better demonstrate the effectiveness of MMoE model, the similar architecture of single task model (STM) and the most commonly used shared-bottom MTL ANN model are utilized for the comparison as baseline (see Fig. 2). Similarly, they are configured with a single fully connected layer with size of 400. Since using MTL profits from a regularization effect, i.e., making the learned representations general across tasks, we also try to fix the pre-trained MoE model as a feature extractor to further fine-tune the individual towers respectively. The comparison result is shown in Fig. 3. We can observe that the shared-bottom model is more accurate using approximately double elapsed time compared to STM. It reveals the fact that the MTL paradigm jointly solves multiple tasks to achieve performance improvement by sharing inductive bias between them. By replacing a fully connected layer in shared bottom with a double-gate MoE layer, the model achieves higher accuracies of 0.687 in ABIDE and 0.650 in

ADHD-200, which demonstrates the success of MMoE. The more sophisticated structure of MMoE contains larger amount of trainable parameters within, resulting in a longer training time (43.2s). We also use the whole FC features to train the MMoE model without GBFS. As expected, the higher input dimension (17,955) leads to more training time (130.7) and poorer performance. Finally, the pre-trained MMoE model shows a slight improvement in accuracy spending only one-eighth of time in local fine-tuning in tower networks.

This simulation result demonstrates great application potential in common situations, where we do not have the computational resources to train the model on large datasets. In a practical way, off-line learning can be performed to pre-train the MTL bottom network on large datasets in advance. Such that, we can make a classification for a small subset in real time by on-line learning. Namely, we use the compressed format of the well-trained bottom network like hdf5 (instead of loading the entire one in memory) to fine-tune the tower networks on small training samples, which should be task specific. It is anticipated that, the performance of MMoE could be further improved as we incorporate more datasets of other related mental disorders, allowing to capture subtle but valuable patterns in feature space. Furthermore, this idea can also offers a feasible solution to address the cold-start issue for a new task having small sample data.

V. Conclusions

In this paper, we proposed a MTL classification framework for auxiliary diagnosis of ASD and ADHD based on the rs-fMRI data of ABIDE I and ADHD-200, which are two world-wide multi-site functional and structural brain imaging data aggregations. First, we presented a graph-based feature selection method to find a subset of remarkable FC features that are discriminative across ASD and ADHD using both external and internal measures. Then an efficient MTL structure of multi-gate mixture-of-experts was applied to simultaneously classify ASD and ADHD subjects from TCs. Compared with the state-of-the-art methods, the proposed model was demonstrate to achieve a reliable classification performance in terms of 10-fold CV. Based on the similar runtime environment, the training speed of our model is about 270 times faster than the competitor's. We conducted a series of experiments to estimate the effects of using graph-based feature selection method and MMoE model. Our experiment results demonstrated their effectiveness as well as great potential in practical use, typical for the scenario when we need to estimate a small subset in real time with small training samples. The current work is expected to provide insights into building a MTL-based intelligent auxiliary diagnosis system for large-scale classification of relevant mental disorders using rs-fMRI data.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC) under grant No. 61876162, by the Shenzhen Scientific Research and Development Funding Program under grant JCYJ20180307123637294, and by the Research Grants Council of the Hong Kong SAR under grant No. CityU11202418 and CityU11209219.

References

[1] Y. Du, Z. Fu, and V. D. Calhoun, "Classification and prediction of brain disorders using functional connectivity: promising but challenging," *Frontiers in neuroscience*, vol. 12, 2018.

[2] V. D. Calhoun and N. de Lacy, "Ten key observations on the analysis of resting-state functional mr imaging data using independent component analysis," *Neuroimaging Clinics*, vol. 27, no. 4, pp. 561–579, 2017.

[3] S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, M. W. Woolrich *et al.*, "Functional connectomics from resting-state fmri," *Trends in cognitive sciences*, vol. 17, no. 12, pp. 666–682, 2013.

[4] A. El Gazzar, L. Cerliani, G. van Wingen, and R. M. Thomas, "Simple 1-d convolutional networks for resting-state fmri based classification in autism," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.

[5] A. F. Marquand, M. Brammer, S. C. Williams, and O. M. Doyle, "Bayesian multi-task learning for decoding multi-subject neuroimaging data," *NeuroImage*, vol. 92, pp. 298–311, 2014.

[6] L. Xiao, J. M. Stephen, T. W. Wilson, V. D. Calhoun, and Y.-P. Wang, "A manifold regularized multi-task learning model for iq prediction from two fmri paradigms," *IEEE Transactions on Biomedical Engineering*, 2019.

[7] Q. Ma, T. Zhang, M. V. Zanetti, H. Shen, T. D. Satterthwaite, D. H. Wolf, R. E. Gur, Y. Fan, D. Hu, G. F. Busatto *et al.*, "Classification of multi-site mr images in the presence of heterogeneity using multi-task learning," *NeuroImage: Clinical*, vol. 19, pp. 476–486, 2018.

[8] D. Hu and L.-L. Zeng, "Multi-task learning of structural mri for multi-site classification," in *Pattern Analysis of the Human Connectome*. Springer, 2019, pp. 205–226.

[9] N. O. Davis and S. H. Kollins, "Treatment for co-occurring attention deficit/hyperactivity disorder and autism spectrum disorder," *Neurotherapeutics*, vol. 9, no. 3, pp. 518–530, 2012.

[10] F. K. Satterstrom, R. K. Walters, T. Singh, E. M. Wigdor, F. Lescai, D. Demontis, J. A. Kosmicki, J. Grove, C. Stevens, J. Bybjerg-Grauholm *et al.*, "Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants," *Nature neuroscience*, pp. 1–5, 2019.

[11] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1930–1939.

[12] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto *et al.*, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, no. 6, p. 659, 2014.

[13] Y. Behzadi, K. Restom, J. Liau, and T. T. Liu, "A component based noise correction method (compcor) for bold and perfusion based fmri," *Neuroimage*, vol. 37, no. 1, pp. 90–101, 2007.

[14] R. C. Craddock, G. A. James, P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg, "A whole brain fmri atlas generated via spatially constrained spectral clustering," *Human brain mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[15] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," *arXiv preprint arXiv:1312.4314*, 2013.

[16] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[17] A. A. Pulini, W. T. Kerr, S. K. Loo, and A. Lenartowicz, "Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 4, no. 2, pp. 108–120, 2019.

[18] N. C. Dvornek, P. Ventola, K. A. Pelphrey, and J. S. Duncan, "Identifying autism from resting-state fmri using long short-term memory networks," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 362–370.

[19] 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.

[20] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: promises and pitfalls," *Neuroimage*, vol. 145, pp. 137–165, 2017.

[21] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, and R. Adams, "Scalable bayesian optimization using deep neural networks," in *International conference on machine learning*, 2015, pp. 2171–2180.

[22] M. Suganuma, S. Shirakawa, and T. Nagao, "A genetic programming approach to designing convolutional neural network architectures," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2017, pp. 497–504.

[23] L. Tan, X. Guo, S. Ren, J. N. Epstein, and L. J. Lu, "A computational model for the automatic diagnosis of attention deficit hyperactivity disorder based on functional brain volume," *Frontiers in computational neuroscience*, vol. 11, p. 75, 2017.

[24] S. Dey, A. R. Rao, and M. Shah, "Exploiting the brain's network structure in identifying adhd subjects," *Frontiers in systems neuroscience*, vol. 6, p. 75, 2012.