A Dual-Stage Pseudo-Labeling Method for the Diagnosis of Mental Disorder on MRI Scans

1st Yao Hu
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
y.hu@my.cityu.edu.hk

2nd Zhi-An Huang*
Center for Computer Science and Information Technology
City University of Hong Kong
Dongguan Research Institute
Dongguan, China
huang.za@cityu.edu.cn

3rd Rui Liu
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
rliu38-c@my.cityu.edu.hk

4th Xiaoming Xue
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
xming.hsueh@my.cityu.edu.hk

5th Linqi Song*
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
City University of Hong Kong
Shenzhen Research Institute
Shenzhen, China
linqi.song@cityu.edu.hk

6th Kay Chen Tan*
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR
kctan@polyu.edu.hk

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892792

***Abstract*—The high prevalence of mental disorders gradually poses a huge pressure on the public healthcare services. Recently, deep learning-based computer-aided diagnosis has been introduced to relieve the tension in healthcare institutions by automatically detecting abnormal neuroimaging-derived phenotypes in patients. However, the training of deep learning models relies on sufficiently large annotated datasets, which can be costly, time-consuming, and laborious. Semi-supervised learning (SSL) can mitigate this challenge by leveraging both labeled and unlabeled samples. In this work, an effective dual-stage pseudo-labeling based classification framework dubbed DSPL is proposed to diagnose mental disorders on functional magnetic resonance imaging data. A bicriteria-based pseudo-labels selection method is developed to filter out inferior pseudo-labeled samples. Subsequently, we further propose a self-mutual learning enhanced pseudo-labeling generation approach to mitigate the adverse effects bought by the noisy pseudo-labeled samples. On real-world datasets, the proposed method achieves diagnosis accuracies of 68.09%, 67.94%, and 68.13% on ABIDE-I, ABIDE-II, and ADHD-200, respectively. Ablation study suggests that each component in DSPL makes a great contribution to performance improvement.**

***Index Terms*—Semi-supervised learning, pseudo-labeling, computer-aided diagnosis, autism spectrum disorder (ASD), attention deficit/hyperactivity disorder (ADHD), functional magnetic resonance imaging (fMRI).**

I. Introduction

Mental disorders, such as autism spectrum disorder (ASD), attention deficit/hyperactivity disorder (ADHD), have become a worldwide health concern. Their high prevalence gradually poses a huge pressure on the health center services [1], [2]. Given the shortage of psychiatrist, computer-aided diagnosis (CAD) approaches are developed to assist medical decision making by automatically analyzing the high-resolution medical images, e.g., functional magnetic resonance imaging (fMRI) [3]. fMRI can investigate aberrant neurobiological functions in mental disorders by detecting tiny changes in blood flow [2]. With the development of graphics processing units (GPUs), some deep learning-based approaches, e.g., convolutional neural network [4], long short-term memory network (LSTM) [5], and hopfield neural network [6], and et al., achieve decent performance in disorder diagnosis and evaluation. However, the successful training of deep learning models tend to require a large amount of annotated samples.

In real world, it is impractical to harvest sufficiently large amount of labeled fMRI samples. Even with adequate equipment, medical experts struggle to mark annotations to the accumulated raw fMRI data in time. In other words, within an individual institution, there may exist a portion of samples that do not get annotations. Lacking labeled data may weaken the effectiveness of deep learning models. Semi-Supervised Learning (SSL) is a promising solution to solve this problem by leveraging both labeled and unlabeled datasets [7]. With the acquisition of unlabeled data, SSL aims to explore the underlying data distribution rather than the well-learned distribution from annotated data.

Pseudo-labeling and consistency regularization-based approaches are two dominant approaches in SSL [8]. Under the assumption that the decision boundaries should lie in low density regions, consistency regularization-based approaches pursue the invariant outputs under tiny input per-

*Corresponding author.

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

turbations/augmentations [9]. Data augmentation and perturbation play an essential role in the performance of consistency regularization-based methods [10], [11]. However, common image augmentation/perturbation methodologies, such as shearing and concealing, are not feasible for fMRI scans since they cannot yield realistic brain appearance and morphology [12]. Some deep learning models, such as generative adversarial network [13] and variational auto-encoders [14], are designed for fMRI data augmentation, but they still require sufficient labeled training samples. By contrast, pseudo-labeling approach reduces the distribution density around the decision boundary by selecting the unlabeled samples with high confidence [15]. Pseudo-labeling approach is more suitable for processing fMRI samples as it does not require domain-specific data augmentations.

The successful application of pseudo-labeling methods depends on the accurate label generation [16]. Generally, the pseudo-labels are assigned for unlabeled samples in a supervised manner. In the iterative training process, pseudo-labeled samples are combined with labeled data to augment the training set so as to improve the generalization capability of models. Clearly, false-labeled examples could lead to information mismatch by introducing unnecessary noisy data. Therefore, correct label assignment is the key to the success of pseudo-labeling approaches.

In this study, a Dual Stage Pseudo-Labeling (DSPL) semi-supervised classification framework is developed for mental disorder diagnosis using fMRI datasets. In DSPL, a bicriteria-based pseudo-labels selection approach is proposed to improve the quality of generated pseudo-labels by filtering inferior unlabeled samples. Subsequently, we develop a self-mutual learning model to eliminate the influence of noisy labeled samples in the training process. On real-world fMRI datasets, the experimental results indicate that DSPL can achieve promising accuracies of 68.09%, 67.94%, and 68.13% on ABIDE-I, ABIDE-II, and ADHD-200, respectively. The main contributions of our work are summed up into three-folds. First, we propose a novel bicriteria-based pseudo-labeling sample selection method to improve the quality of pseudo-labels by selecting "easy discriminated" samples. Second, a self-mutual learning model is developed to reduce the influence of incorrect pseudo-labels. Third, comprehensive experiments demonstrate the state of the art of DSPL on three real world fMRI datasets.

The rest of this paper is organized as follows. Section II illustrates the proposed framework in detail. Section III conducts a series of simulation experiments to evaluate the effectiveness of DSPL. Finally, section IV concludes this paper.

II. Methodology

In this section, we introduce the detail implementation of the proposed DSPL framework. First, we present the bicriteria-based pseudo-label selection (BPLS) algorithm, which aims to improve the quality of pseudo-labels by filtering inferior samples. Then, the self-mutual learning enhanced pseudo-label generation (SMLPL) algorithm is developed to ease the negative effects of noisy pseudo-labeled samples. Finally, we specify the workflow of DSPL framework.

A. Problem Formulation

The primary goal of this work is to improve the performance of deep learning models by leveraging both labeled and unlabeled fMRI samples. Let $\mathcal{D}_l = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^{n_l}$ and $\mathcal{D}_u = \left\{\mathbf{x}^{(i)}\right\}_{i=1}^{n_u}$ be the labeled and unlabeled examples with sizes of n_l and n_u, respectively. We define the deep learning model f_θ through two modules: feature extraction module g_θ and classification module h_θ, i.e., $f_\theta(\mathbf{x}) = h_\theta(g_\theta(\mathbf{x}))$. The optimization objective can be reached as

$$\theta^* = \arg\min_\theta \sum_{i=1}^{n} \mathcal{L}\left(f_\theta; \mathcal{D}_l\right) + \mathcal{R}(f_\theta), \tag{1}$$

where $\mathcal{L}$ is the loss function between the predictions and targets. In addition, we utilize $\mathcal{R}$ to represent the additional regularization item related to D_u.

B. Bicriteria-based Pseudo-Label Selection

The proposed BPLS algorithm adopts two different criteria to improve the accuracy of generated pseudo-labels, i.e., threshold-based filtering and clustering-based assignment. The threshold-based filtering is developed to move the decision boundary from high-density regions to low-density regions by filtering the incompetent unlabeled samples. Specifically, the model f_θ trained on $\mathcal{D}_l$ is used to predict all unlabeled samples. Since fMRI samples are time-series data, we adopt the LSTM [17] as f_θ in this study. LSTM leverages the long term information stored in the time series by the recurrent units and achieves promising results in processing fMRI samples [18]. Additionally, a global average pooling (GAP) is added on the features extracted by LSTM layers to avoid overfitting by enforcing the correspondences between features and label categories [19]. Based on [20], the involved phenotypic information including age, sex, handedness, and IQ, is added to concatenate the output of feature extraction module. The inferior samples with predictive scores below the preset threshold are eliminated. The high effectiveness and low complexity make the threshold-based approach [21] suitable for the preliminary screening. Accordingly, threshold-based filtering can be formulated as

$$\tilde{\boldsymbol{y}}_t = \mathbb{1}_{x_i \in \mathcal{D}_u}\left[f_\theta\left(x_i\right) \geq \gamma\right], \tag{2}$$

where $\tilde{y}_t$ denotes the selected pseudo-labels, and $\gamma \in (0, 1)$ represents the threshold. However, there may still exist incorrect pseudo-labels with high confidence [16], leading to noisy training.

To further polish the assigned pseudo labels, inspired by [22], clustering-based assignment is performed for unlabeled samples at the level of feature distribution. The clustering centroids of healthy controls (HC) and patients (PT) are

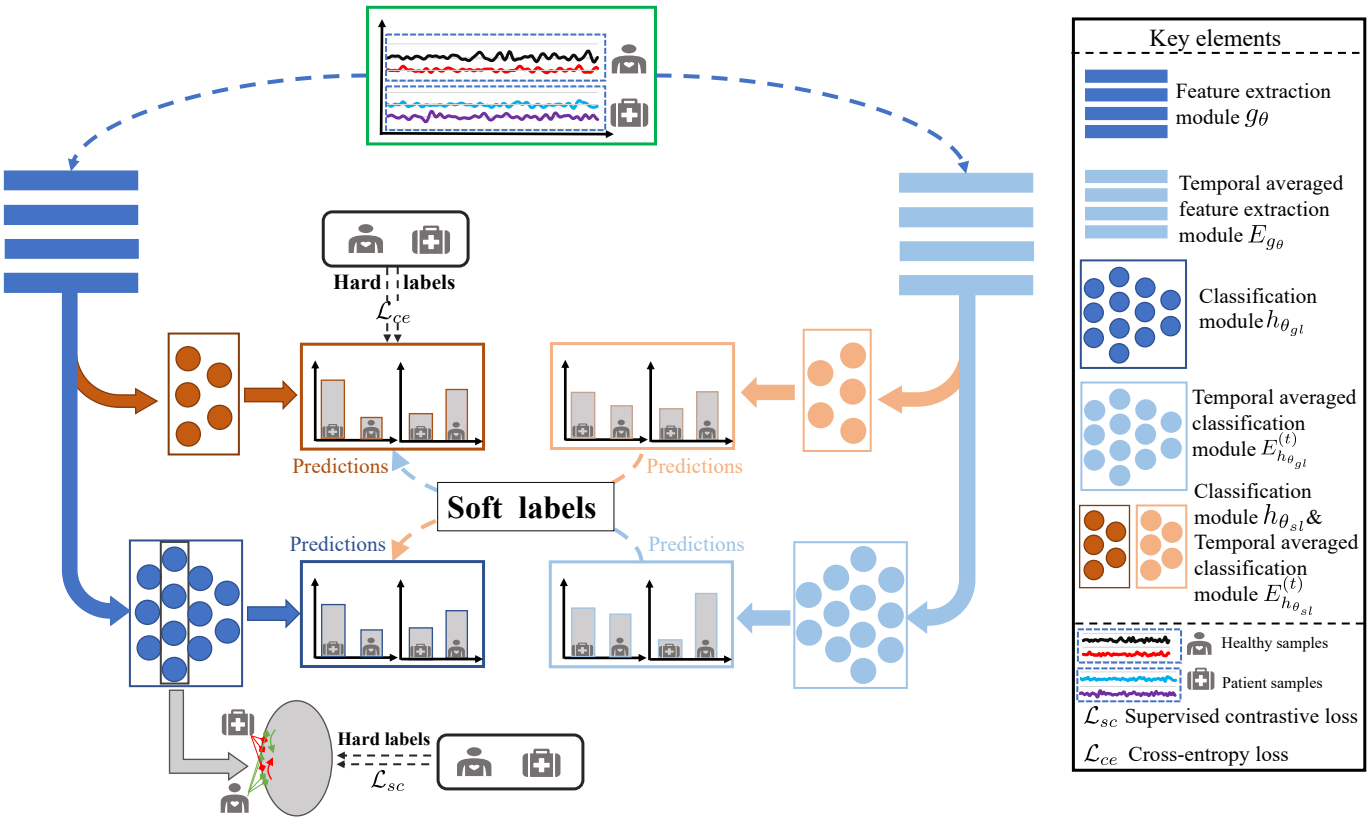


Fig. 1. The proposed SMLPL algorithm. Two classification modules with distinct structures are built upon the shared feature extraction module to support self-mutual learning. Various loss functions are adopted on the training of various classifiers for avoiding collaborative biases.

determined by the extracted feature distribution, which can be obtained through

$$c_m = \frac{\sum_{\mathbf{x}_u \in \mathcal{D}_u} \delta_m \left(f_\theta \left(\mathbf{x}_u \right) \right) g_\theta \left(\mathbf{x}_u \right)}{\sum_{\mathbf{x}_u \in \mathcal{D}_u} \delta_m \left(f_\theta \left(\mathbf{x}_u \right) \right)} \quad m \in \{\mathrm{HC}, \mathrm{PT}\}, \quad (3)$$

where δ_m indicates the score of softmax function on HC or PT. Pseudo-labels are assigned according to the minimum distance between extracted features and each centroid as:

$$\hat{\boldsymbol{y}}_c = \mathbb{1}_{x_i \in \mathcal{D}_u} \left[\arg\min_k D_{cos} \left(g \left(\mathbf{x}_i \right), c_k \right) \right], \quad (4)$$

where $g_\theta \left(\mathbf{x}_i \right)$ represents the extracted features from $\mathbf{x}_i$ and c_m denotes the centroid of m-th category obtained from Eq. (3). D_{cos} measures the cosine distance between $g_\theta \left(\mathbf{x}_i \right)$ and c_m.

Based on the principle that lower entropy values usually correspond to less uncertainty [23] [24], we propose to only assign pseudo-labels for those easily discriminated samples with low entropy values. We first estimate the averaged entropy values of all unlabeled samples via $\zeta = \left[\sum_{\mathbf{x}_u \in \mathcal{D}_u} \boldsymbol{E} \left(\delta \left(f_\theta \left(x_u \right) \right) \right) \right] / n_u$, where $\boldsymbol{E}$ calculates the entropy values. Then, the unlabeled samples can be assigned into either the high-entropy group ($\mathcal{G}_h$) or the low-entropy group ($\mathcal{G}_l$) based on ζ. The center of each group is then obtained by the mean entropy values of the assigned samples. Subsequently, we iteratively update the samples assignments and group centers until they remain unchanged. Only the unlabeled samples belong to $\mathcal{G}_l$ are assigned with pseudo-labels, i.e., $\tilde{\boldsymbol{y}}_c = \hat{\boldsymbol{y}}_c \left[\mathcal{G}_l \right]$. Finally, the intersections of $\tilde{\boldsymbol{y}}_c$ and $\tilde{\boldsymbol{y}}_t$ are taken as the outputs of BPLS algorithm, which are combined with $\mathcal{D}_l$ as an integrated dataset $\tilde{\mathcal{D}}_l$.

C. Self-Mutual Learning Enhanced Pseudo-Label Generation

Although BPLS improves the accuracy of pseudo-labeling, there still exist some incorrect pseudo-labeled samples in $\tilde{\mathcal{D}}_l$, leading to performance degradation. To eliminate the influence of noisy labels, co-teaching strategy [25] is used to build two peer networks with varied structures and utilize their different learning capabilities to filter out noisy samples. The training of individual peer network depends on the qualified samples selected by another network. Based on this idea, we develop a self-mutual learning method to further promote the pseudo-labeling accuracy. Our method iteratively generates pseudo-labels via a mutual learning-like mode, and the pseudo-labeled samples further support the training of networks. Typical mutual learning technique is mostly used to enhance the training of both networks by matching their predictions [26]. In this work, the proposed self-mutual learning method aims to ease the influence of noisy samples by leveraging the different representation capacities of different classification modules.

As illustrated in Fig. 1, a common feature extraction module (g_θ) is used for two different classification modules. Since different learning capabilities are the keys for filtering out the noisy information, two multi-layer perceptrons with different structures serve as classification modules are thus built upon the outputs of g_θ. Two distinct loss functions are adopted to evaluate the learning effects from group-level and subject-level perspectives, respectively.

The supervised contrastive learning loss ($\mathcal{L}_{sc}$) [27] is adopted to train the group level-based classifier ($h_{\theta_{gl}}$). In the embedding space, $\mathcal{L}_{sc}$ aims to gather samples from the same category into same group while amplifying the distances between samples from different categories. A data augmentation method using the Gaussian noise is first conducted on each sample of $\tilde{\mathcal{D}}_l$, resulting in $2|\tilde{\mathcal{D}}_l|$ pairs [27]. The intermediate dataset is named $\mathcal{D}'_l$. The distances between the augmented samples from the same category are iteratively minimized by leveraging the label information. At t-th iteration, the $\mathcal{L}_i^{sc}$ of the i-th augmented sample within mini-batch $\boldsymbol{B}$ ($\boldsymbol{B} \subset \mathcal{D}'_l$) is formulated as follows:

$$\begin{aligned}&\mathcal{L}_{sc}^{i,t}\left(g_\theta^t, h_{\theta_{gl}}^t; \boldsymbol{B}\right)\\&= \frac{-1}{|P(i)|}\sum_{j=1}^{|\boldsymbol{B}|} 1_{i\neq j}\cdot 1_{\tilde{y}_i=\tilde{y}_j}\cdot \log\frac{\exp(z_i\cdot z_j/\tau)}{\sum_{k=1}^{2n} 1_{i\neq k}\cdot\exp(z_i\cdot z_k/\tau)},\end{aligned} \tag{5}$$

where z_i denotes the output which is normalized into a unit sphere, and $P(i)$ represents the set of indices of the i-th augmented samples with same label in $\boldsymbol{B}$. The scalar temperature parameter $\tau \in \mathbb{R}^+$ is set to 0.1 following the setting of [27]. The overall objective is to iteratively minimize $\mathcal{L}_{sc}^t = \sum_{i=1}^{|\boldsymbol{B}|}\mathcal{L}_{sc}^{i,t}\left(g_\theta^t, h_{\theta_{gl}}^t; \boldsymbol{B}\right)$. After training, $h_{\theta_{gl}}$ can build decision boundaries based on the clustering group features of HC and TP samples. Another cross-entropy loss ($\mathcal{L}_{ce}$) is adopted to train the subject-level classification module ($h_{\theta_{sl}}$). $\mathcal{L}_{ce}$ measures the divergence between predicted probability of individual subject and its actual label. Thus, $h_{\theta_{sl}}$ forms decision boundaries based on the final outputs probabilities of each subject. Overall, $h_{\theta_{gl}}$ and $h_{\theta_{sl}}$ build decision boundaries from two complementary views. Their different learning capacities can be considered as supervisions for further mutual learning.

The soft labels generated by classification modules, i.e., the output probabilities, serve as supervisions for mutual training. Inspired by the temporally averaged parameter method in [28], [29], we calculate a weighted average of the parameters between previous iterations and the current iteration to improve the robustness of generated soft labels. We utilize $E_{g_\theta}^{(t)}$, $E_{h_{\theta_{gl}}}^{(t)}$, and $E_{h_{\theta_{sl}}}^{(t)}$ to denote the temporally averaged parameters of g_θ, $h_{\theta_{gl}}$, and $h_{\theta_{sl}}$ at t-th iteration, respectively. Given the calculation of $E_{g_\theta}^{(t)}$, it is obtained via $E_{g_\theta}^{(t)} = \alpha E_{g_\theta}^{(t-1)} + (1-\alpha)g_\theta$, where $E_{g_\theta}^{(t-1)}$ denotes temporally averaged parameters in previous $(t-1)$ iterations and α is an ensembling momentum parameter. Specially, $E_{g_\theta}^0 = g_\theta$. The calculation of $E_{h_{\theta_{sl}}}^{(t)}$ also follows the similar procedure. The soft labels yielded from temporally averaged model are fed to current training classifiers for supervised training. The mutual training loss at t-th iteration with mini-batch $\boldsymbol{B}$ can be written as:

$$\begin{aligned}&\mathcal{L}_{mt}^t\left(g_\theta^t, p^t; E_{g_\theta}^{(t)}, E_q^{(t)}, \boldsymbol{B}\right)\\&= -\mathbb{E}_{x\in\boldsymbol{B}}\sum_{i=1}^{|\boldsymbol{B}|} E_q^{(t)}\left(E_{g_\theta}^{(t)}(x_i)\right)\cdot\log p\left(g_\theta^t(x_i)\right),\end{aligned} \tag{6}$$

where p and q denote the indexes of classification modules, i.e., $p, q \in \left\{h_{\theta_{gl}}, h_{\theta_{sl}}\right\}$ and $p \neq q$. When $p = \theta_{gl}$ and θ_{sl}, the mutual loss function is abbreviated to $\mathcal{L}_{mt}^{t,gl}$ and $\mathcal{L}_{mt}^{t,sl}$, respectively.

At t-th iteration, the overall loss for self-mutual learning $\mathcal{L}_{sml}^t$ can be obtained by simultaneously optimizing the classification modules and the shared feature extraction module, i.e., the whole SMLPL network f_{sml}, as follows:

$$\begin{aligned}\mathcal{L}_{sml}^t(f_{sml}; \mathcal{D}'_l) = \varphi_1\mathcal{L}_{sc}^t + (1-\varphi_1)\mathcal{L}_{mt}^{t,gl}\\ + \varphi_2\mathcal{L}_{ce}^t + (1-\varphi_2)\mathcal{L}_{mt}^{t,sl},\end{aligned} \tag{7}$$

where φ_1 and φ_2 are regulation parameters of $h_{\theta_{gl}}$ and $h_{\theta_{sl}}$, respectively. The averaged predictions of $h_{\theta_{gl}}$ and $h_{\theta_{sl}}$ are taken as the ultimate outputs of SMLPL algorithm.

Algorithm 1: The pseudo-code of DSPL

Data: Labeled dataset $\mathcal{D}_l$, unlabeled dataset $\mathcal{D}_u$, test dataset $\mathcal{D}_{te}$, phenotypic traits of labeled, unlabeled, and test samples $\mathbf{X}_l^p$, $\mathbf{X}_u^p$ and $\mathbf{X}_{te}^p$. Batch-size B, and pseudo-labeling round E.
Result: Predicted probabilities of test set Y_{pred}^{te}.

```
Initialize f_θ;
Train network f_θ using D_l;
while e ≤ E do
    if e = 1 then
        X'_u^{tr}, X'_u^{p_tr}, Y'_u^{tr} ← BPLS.pred (f_θ, D_u, X_u^p);
    else
        X'_u^{tr}, X'_u^{p_tr}, Y'_u^{tr} ← BPLS.pred(f_sml, D_u, X_u^p);
    D̃_l ← [X'_u^{tr}, X'_u^{p_tr}, Y'_u^{tr}] ∪ [D_l, X_l^p];
    Initialize f_sml;
    while t ≤ |D̃_l|//B, each mini-batch B ⊂ D̃_l do
        f_sml ← SMLPL.fit (f_sml, B) via Eq. (7);
Y_pred^te ← f_sml.pred(D_te, X_te^p).
```

D. Workflow of DSPL

The workflow of DSPL is composed by BPLS and SMLPL modules, and the specific procedure is described in Algorithm 1. First, a LSTM network f_θ within BPLS is trained on $\mathcal{D}_l$. Then an iterative optimization process is conducted to refine the pseudo-labels and reduce the adverse effect of noisy pseudo-labels. Based on the well-trained model, the assigned pseudo-labeled samples are combined with $\mathcal{D}_l$ as the $\tilde{\mathcal{D}}_l$. We conduct SMLPL with $\tilde{\mathcal{D}}_l$ to iteratively update f_{sml}. After E pseudo-labeling rounds, the final f_{sml} is utilized to make predictions for $\mathcal{D}_{te}$.

III. Experiments and Results

A. Configuring Simulation Settings

In this work, three publicly released resting-state fMRI aggregation datasets, namely the Autism Brain Imaging Data Exchange I[1] (ABIDE-I for ASD), Autism Brain Imaging Data Exchange II[2] (ABIDE-II for ASD) and the ADHD-200 Competition[3] (ADHD-200 for ADHD) are adopted to evaluate the effectiveness of the proposed method. The ABIDE-I dataset gathers 1035 valid samples (aged 7 to 64) from 17 international brain imaging sites involving 530 individuals with ASD and 530 health controls (HCs). ABIDE-II dataset contributes 1113 valid samples aged 5 to 64 from 521 individuals with ASD and 592 HCs. The Configurable Pipeline for the Analysis of Connectomes (CPAC)[4] [30] and the pipeline of Data Processing Assistant for rs-fMRI [31] are adopted to preprocess ABIDE-I and ABIDE-II, respectively. ADHD-200 dataset was originally collected from 7 sites for the identification of ADHD in a global competition. ADHD-200 dataset archives 939 valid samples aged 7 to 21 from 358 ADHD patients and 581 HCs. The ADHD-200 samples are processed by the pipeline of Athena[5]. In this work, the Craddock 200 (CC200) atlas is used to extract the mean time-series for a set of 190 ROIs. Since there is no consensus on the optimal sampling protocol for fMRI data acquisition, the time dimensions of data samples collected by various institutions are different. To solve this issue, a random cropping is performed on the time-series of each subject to keep the time course dimensions identical [20]. Specifically, we fix the sequence length $T = 90$ and randomly crop 10 sequences for each sample. In this way, each sampled sequence is cropped into a space-time matrix with the fixed size of 90×190. The cropped samples with identical shape can be taken as inputs of DSPL. Specific parameter settings of DSPL are tabulated in Table I.

Five-fold cross validation (CV) is adopted to evaluate the effectiveness of proposed framework. In each round of CV, 20% of the training data are set as labeled samples, while the rest of training data are taken as unlabeled samples. The classification performance is validated through four evaluation metrics, i.e., accuracy (ACC), specificity (SPE), sensitivity (SEN), and the area under the receiver operating characteristic curve (AUC).

B. Results

The uncertainty-aware pseudo-label selection (UPS) [8] is used for comparison, representing the state-of-the-art performance in SSL. UPS algorithm leverages uncertainty to negate the effects of incorrect calibration for improving pseudo labeling accuracy via the confidence thresholds. Pseudo-labeling round T is an important hyperparameter according to the results shown in the paper [8] (where T is set to 10). Here, the achieved results of DSPL and UPS algorithms within 5 and 10 rounds are presented in Table II. "Oracle" denotes results of the model trained in supervised learning (i.e., all unlabeled samples are assigned with correct labels), representing the upper-bound performance of DSPL in theory.

TABLE I
Parameter Settings of Each Component within DSPL Framwork

For BPLS			
Components of feature extraction module		LSTM & GAP & FC	
Configuration of classification module		10-2	
LSTM hidden size	32	GAP size	90
FC configuration	5	Activation function	Softmax
Batch_first	True	Training epochs	30
Batch size	64	Learning rate	0.001
Optimizer	Adam	Threshold value γ	0.65
For SMLPL			
Configuration of $h_{\theta_g l}$		10-20-10-2	
Configuration of $h_{\theta_s l}$	10-2	Training epoch	30
Activation function	Softmax	Batch size	128
Learning rate	0.001	Optimizer	Adam
α	0.9	φ_1 & φ_2	0.3 & 0.5

Note: FC denotes fully connected layer

As we can see within 10 rounds, DSPL achieves the promising accuracies of 68.09% (SPE: 69.28%, SEN: 67.50%, AUC: 69.55), 67.94% (SPE: 69.68%, SEN: 66.79%, AUC: 69.33%), and 68.13% (SPE: 75.17%, SEN: 63.18, AUC: 70.47) on ABIDE-I, ABIDE-II, and ADHD-200, respectively. Only 1.34% and 2.38% average gaps compared with the Oracle results in terms of ACC and AUC, respectively. The performance of UPS algorithm is inferior to that of of our DSPL framework. This is mainly because the UPS algorithm fails to ease the negative effects of the incorrect pseudo-labeled samples, causing the error accumulation during training. By contrast, the SMLPL algorithm can effectively mitigate the influence of incorrect pseudo-labeled samples. It is shown that UPS cannot manage a trade-off between SPE and SEN in ADHD-200 dataset because of the issues of skewed data distribution (358 vs 581). The experiment results suggest that DSPL can alleviate this issue to some extent. As expected, pseudo-labeling round T has a positive effect on the performance of both UPS and DSPL. We further conduct a parameter analysis of T in next paragraph.

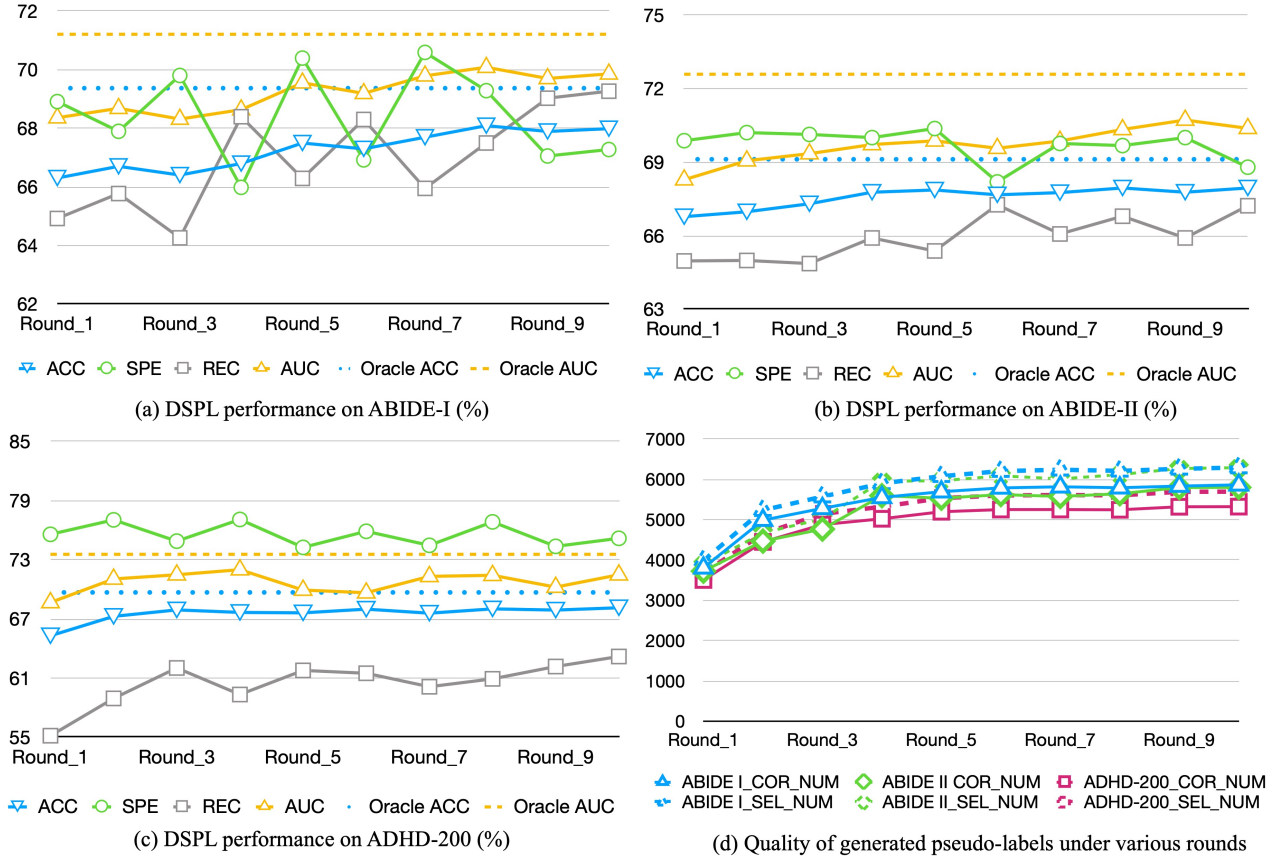
The results of DSPL by changing T are shown in Fig. 2 (a)-(c). We can observe a fluctuation of SPE and SEN in three datasets. By contrast, ACC and AUC remain relatively stable and rise slightly as increasing the number of T. DSPL is able to automatically determine the number of "easily discriminated" samples for pseudo-labels assignments. The quality of generated pseudo-labels, including the number of selected (SEL_NUM) and correct (COR_NUM) pseudo-labels, for three datasets are shown in Fig. 2 (d). Generally, DSPL can yield highly accurate pseudo-labeled samples on the selected samples. At the end of the first round, DSPL selects partitions of samples and assigns pseudo-labels with high accuracy (3804 samples with 95.47% accuracy on ABIDE-I, 3886 samples with 95.7% accuracy on ABIDE-II, 3674 samples with 95.35% accuracy on ADHD-200). As the number of rounds increases,

[1]http://fcon_1000.projects.nitrc.org/indi/abide
[2]http://fcon_1000.projects.nitrc.org/indi/abide/abide_II
[3]http://fcon_1000.projects.nitrc.org/indi/adhd200/
[4]https://fcp-indi.github.io/
[5]https://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline

TABLE II
PERFORMANCE COMPARISON ON ABIDE-I, ABIDE-II AND ADHD-200 (%).

Frameworks	ABIDE-I (ASD)				ABIDE-II (ASD)				ADHD-200 (ADHD)			
	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC
UPS†	66.60	63.82	**74.55**	67.26	66.35	**71.74**	62.18	67.73	67.14	**79.13**	55.41	68.89
UPS‡	67.39	**72.27**	64.41	68.56	67.11	67.44	**66.85**	68.75	67.83	76.16	59.87	69.42
DSPL†	67.49	70.39	66.29	68.55	67.86	70.37	65.37	68.87	67.91	74.88	62.02	69.14
DSPL‡	**68.09**	69.28	67.50	**69.55**	**67.94**	69.68	66.79	**69.33**	**68.13**	75.17	**63.18**	**70.47**
Oracle	69.37	68.86	72.63	72.21	69.12	69.71	68.01	71.59	69.68	76.91	63.57	72.55

Note: † and ‡ denote the achieved best results within 5 and 10 rounds.



(a) DSPL performance on ABIDE-I (%)

(b) DSPL performance on ABIDE-II (%)

(c) DSPL performance on ADHD-200 (%)

(d) Quality of generated pseudo-labels under various rounds

Fig. 2. Performance evaluation of DSPL on increasing pseudo-labeling rounds in three datasets

the number of selected pseudo-labels gradually rises, and the proportion of correct predictions presents a slight decrease. Despite the pseudo-labeling accuracy decreases, the ACC and AUC still show a slight upward trend in all datasets, illustrating that SMLPL can be a credible solution to mitigate the influences of noisy labels.

C. Ablation Studies

The BPLS and SMLPL algorithms are two important modules in our framework, we conduct an ablation study to evaluate their contributions to the performance improvement. The experimental results are shown in Table III.

First, the proposed DSPL framework without BPLS is evaluated. We can observe that there are 1.29%, 1.59%, and 1.70% of accuracy degradation on the ABIDE-I, ABIDE-II, and ADHD-200 datasets, respectively. This is because the removal of BPLS leads to increasing noisy samples in the training process. All assigned pseudo labels are fed to the model without screening, which inevitably causes the systematic bias. Additionally, we also evaluate the performance of DSPL by excluding SMLPL. We note that the declines of ACC and AUC reach at a ratio of 1.99% and 2.17% on average in three datasets. The absence of SMLPL makes DSPL disable the capability of eliminating the adverse effects of incorrect pseudo labels. The incorrect pseudo labels participate in the training process and then aggregate their mismatched feature representation iteratively. In summary, BPLS and SMLPL are demonstrated to make great contributions to the performance improvement of DSPL.

TABLE III
ABLATION STUDIES ON THE PROPOSED DSPL FRAMEWORK (%).

Methods	Metrics	ABIDE-I	ABIDE-II	ADHD-200
w/o BPLS	ACC	66.70(↓1.29)	66.36(↓1.59)	66.42(↓1.70)
	SPE	68.92(↓1.65)	68.04(↓0.76)	71.23(↓3.94)
	REC	66.38(↓2.88)	64.76(↓2.46)	62.30(↓0.88)
	AUC	68.71(↓1.14)	66.18(↓3.20)	67.72(↓2.75)
w/o SMLPL	ACC	65.88(↓2.11)	66.29(↓1.66)	65.93(↓2.20)
	SPE	65.56(↓1.71)	67.27(↓1.52)	71.79(↓3.38)
	REC	66.68(↓2.58)	63.38(↓3.84)	60.73(↓2.45)
	AUC	67.70(↓2.15)	67.85(↓1.53)	67.63(↓2.84)
Full	ACC	67.98	67.94	68.13
	SPE	67.27	68.79	75.17
	REC	69.26	67.22	63.18
	AUC	69.86	69.38	70.47

Notes: w/o denotes without;
↓ / ↑ represents the worse/better difference compared to the full DSPL.

D. Effects of Labeled Samples

DSPL begins with the initial model training, the performance of initial model f_θ could play an important role in the downstream processing. Given the fact that f_θ is trained on a portion of labeled dataset $\mathcal{D}_l$, the portion size of $\mathcal{D}_l$ could also influence the performance of DSPL. Therefore, in this subsection, we conduct a simulation experiment to evaluate the effects on the portion size of $\mathcal{D}_l$ which are used to train f_θ. Fig 3 presents the results of DSPL using different portion sizes of $\mathcal{D}_l$ (i.e., 10%, 20%, and 30%) in three datasets. With 10% labeled dataset, DSPL can generally maintain decent experiment results by achieving ACCs of 67.1%, 67.5%, and 66.8% in ABIDE-I, ABIDE-II, and ADHD-200, respectively. It needs to note that this result is still better than that of UPS with 20% portion size of $\mathcal{D}_l$ according to Table II. With increasing labeled samples, the performance of DSPL presents a general upward trend as expected. Specifically, DSPL with 20% portion sizes of $\mathcal{D}_l$ achieves more remarkable improvement than that with 30% portion sizes of $\mathcal{D}_l$. Therefore, DSPL with 20% portion sizes of $\mathcal{D}_l$ can achieve a better trade-off between effectiveness and required quantity of labeled samples. Overall, DSPL with relatively small portion sizes of $\mathcal{D}_l$ can maintain robustness and reliability for prediction.

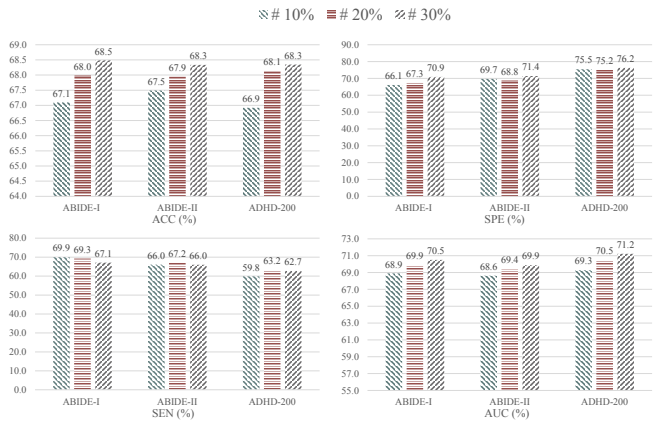


Fig. 3. Effectiveness of DSPL using different portion sizes of $\mathcal{D}_l$.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a dual stage pseudo-labeling framework DSPL to leverage both labeled and unannotated samples to improve diagnosis performance on mental disorders. First, we presented a bicriteria-based pseudo-labeling selection approach to improve the pseudo-labeling accuracy by automatically filtering the inferior pseudo-labels. Then a self-mutual learning network enhanced pseudo-labeling method was developed to mitigate the influence of incorrect pseudo-labeled samples during the training process. Compared with a state-of-the-art method UPS, the effectiveness of DSPL is demonstrated by 5-fold CV. The ablation studies demonstrate the major contributions of BPLS and SMLPL to performance improvement. A simulation experiment also suggests that DSPL can maintain robustness and reliability by using relatively low portion sizes of labeled samples. In the future, we intend to expand the application of DSPL on other deep learning-based CAD tasks, such as tumor segmentation. Furthermore, the DSPL is expected to work in synergy with federated learning under the circumstance that there is only a portion of labeled samples in each client.

V. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61876162, the Hong Kong RGC grant ECS 21212419, the Technological Breakthrough Project of Science, Technology and Innovation Commission of Shenzhen Municipality under Grants JSGG20201102162000001, InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies, the Hong Kong UGC Special Virtual Teaching and Learning (VTL) Grant 6430300, and the Tencent AI Lab Rhino-Bird Gift Fund.

REFERENCES

[1] M. L. Wainberg, P. Scorza, J. M. Shultz, L. Helpman, J. J. Mootz, K. A. Johnson, Y. Neria, J.-M. E. Bradford, M. A. Oquendo, and M. R. Arbuckle, "Challenges and opportunities in global mental health: a research-to-practice perspective," *Current psychiatry reports*, vol. 19, no. 5, p. 28, 2017.

[2] Z.-A. Huang, R. Liu, and K. C. Tan, "Multi-task learning for efficient diagnosis of asd and adhd using resting-state fMRI data," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[3] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, "Identifying autism spectrum disorder from resting-state fMRI using deep belief network," *IEEE Transactions on Neural Networks and Learning Systems*, 2020, early access, July. 21, 2020, doi: 10.1109/TNNLS.2020.3007943.

[4] D. Wen, Z. Wei, Y. Zhou, G. Li, X. Zhang, and W. Han, "Deep learning methods to process fMRI data and their application in the diagnosis of cognitive impairment: a brief overview and our opinion," *Frontiers in neuroinformatics*, vol. 12, p. 23, 2018.

[5] Y. Hu, X. Sun, X. Nie, Y. Li, and L. Liu, "An enhanced lstm for trend following of time series," *IEEE Access*, vol. 7, pp. 34 020–34 030, 2019.

[6] K. C. Tan, H. Tang, and S. S. Ge, "On parameter settings of hopfield networks applied to traveling salesman problems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 5, pp. 994–1002, 2005.

[7] L. Wang, R. Chan, and T. Zeng, "Probabilistic semi-supervised learning via sparse graph structure learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 853–867, 2020.

[8] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=-ODN6SbiUU

[9] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[11] A. Abuduweili, X. Li, H. Shi, C.-Z. Xu, and D. Dou, "Adaptive consistency regularization for semi-supervised transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6923–6932.

[12] K. P. Nguyen, C. C. Fatt, A. Treacher, C. Mellema, M. H. Trivedi, and A. Montillo, "Anatomically informed data augmentation for functional mri with applications to deep learning," in *Medical Imaging 2020: Image Processing*, vol. 11313, 2020, p. 113130T.

[13] P. Zhuang, A. G. Schwing, and O. Koyejo, "FMRI data augmentation via synthesis," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1783–1787.

[14] B. Tajini, H. Richard, and B. Thirion, "Functional magnetic resonance imaging data augmentation through conditional ica," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 491–500.

[15] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[17] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[18] R. Liu, Z.-A. Huang, M. Jiang, and K. C. Tan, "Multi-LSTM networks for accurate classification of attention deficit hyperactivity disorder from resting-state fMRI data," in *2020 2nd International Conference on Industrial Artificial Intelligence (IAI)*. IEEE, 2020, pp. 1–6.

[19] M. Bengs, N. Gessert, and A. Schlaefer, "4D spatio-temporal deep learning with 4d fmri data for autism spectrum disorder classification," *arXiv preprint arXiv:2004.10165*, 2020.

[20] N. C. Dvornek, P. Ventola, and J. S. Duncan, "Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks," in *IEEE 15th International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 725–728.

[21] J. H. Ang, K. C. Tan, and A. Mamun, "An evolutionary memetic algorithm for rule extraction," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1302–1315, 2010.

[22] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6028–6039.

[23] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.

[24] Z. Li, X. Wang, H. Yang, D. Hu, N. M. Robertson, D. A. Clifton, and C. Meinel, "Not all knowledge is created equal," *arXiv preprint arXiv:2106.01489*, 2021.

[25] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *arXiv preprint arXiv:1804.06872*, 2018.

[26] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[27] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[28] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," *arXiv preprint arXiv:2001.01526*, 2020.

[29] Y. Chen, X. Sun, and Y. Jin, "Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 4229–4238, 2019.

[30] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham *et al.*, "The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives," *Frontiers in Neuroinformatics*, vol. 7, 2013.

[31] C. Yan and Y. Zang, "DPARSF: A MATLAB toolbox for" pipeline" data analysis of resting-state fMRI," *Frontiers in Systems Neuroscience*, vol. 4, p. 13, 2010.