CONDITIONAL DOMAIN ADVERSARIAL TRANSFER FOR ROBUST CROSS-SITE ADHD CLASSIFICATION USING FUNCTIONAL MRI

Ya-Lin Huang, Wan-Ting Hsieh, Hao-Chun Yang, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

ABSTRACT

There is a growing number of large scale cross-site database collection of resting-state functional magnetic resonance imaging (rs-fMRI) for studying neurobehavioral diseases, such as ADHD. Although a large amount of data benefits machine learning-based classification methods, the idiosyncratic variability of each site can deteriorate cross-site generalization ability. This challenge creates a bottleneck in requiring a large number of labeled samples of each site. Hence in this research, we utilize an approach of conditional adversarial domain adaptation network (CDAN) to learn a discriminative fMRI representation that is site-invariant for unsupervised transfer of ADHD classification. We evaluate our framework on a multi-site ADHD dataset and achieve improvement in transferring between sites. Further visualization reveals that there indeed exists a substantial site discrepancy and statistically analysis indicates that male's rs-fMRI could be more vulnerable toward site-specific effects.

Index Terms— ADHD, fMRI, adversarial domain adaptation, multi-site transfer

1. INTRODUCTION

Attention deficit hyperactivity disorder(ADHD) is one of the most common neurobehavioral disorders among children that persists into adulthood resulting in lifelong impaired conditions. Although early diagnosis of ADHD is crucial for effective intervention, there still exits major challenges [1]. The elusive nature of the disease, i.e., ADHD children tend to have comorbid conditions including anxiety, poor concentration, and learning problem [2], creates differential subtypes within ADHD (ADHD combined, hyperactive/impulsive, and inattentive). Further, the diagnosis of an ADHD patient relies on a time-consuming combination of comprehensive medical history examinations, behavior observation reports, and a variety of tests [3] [4]. Recently, many research works have turned to the use of Functional Magnetic Resonance Imaging (fMRI) as a noninvasive method in measuring neural activity [5], and this particular image-based biomarker has opened a new venue in automatic screening of ADHD.

Many studies have applied machine learning techniques on fMRI data to automatically differentiate subjects of ADHD versus control. For example, ReHo features extracted from resting state fMRI were introduced for differentiating ADHD from healthy control [6]; Kuang et al. [7] developed a deep neural network based method in predicting ADHD subtypes. While many of these data-driven methods gradually demonstrate satisfying classification results, retrieving an adequate amount of labeled fMRI data at each clinical site remains to be a challenging prerequisite in real-world practice.

Researchers from across the globe have pooled together fMRI data of ADHD from different sites [8] to provide more data samples to improve the algorithms. However, in order to further extend the use of automated screening of ADHD using fMRI, a key challenge remains in handling the site-specific heterogeneity to enable learning from existing patient's fMRI samples to be used directly in a new clinical site where the collection of labeled data has not been previously occurred. Among each site, there tends to an uncontrolled heterogeneity that emerges and creates unwanted variability deteriorating the classification robustness. These variations are caused by a range of issues, e.g., MRI acquisition protocols (e.g., scanner type, flipping angle, see Friedman et al. [9]), inconsistent instruction to the participants (e.g., eyes closed versus eyes open), recruitment criteria (e.g., age-group, treatment history.). Very limited, if any, researchers have studied these problems except for one of the works done by Heinsfeld et al. that utilized DNN in multi-site ABIDE data for autism spectrum disorder identification [10].

Hence, we argue that to enhance the usability of an fMRI-based ADHD pre-screening model, learning a *site-invariant* fMRI brain image representation is critical to ensure the algorithmic generalization. Specifically, we propose to utilize conditional domain adversarial adaptation (CDAN) to mitigate issues of fMRI data heterogeneity to perform binary classification of ADHD across sites. We evaluate our framework on a public multi-site data source, ADHD-200. Each data site is regarded as a single domain, and our CDAN is trained on one source domain then tested on another unlabeled target site. Our experiments show that the method of CDAN achieves an improved classification result across almost all of the cross-site transfer. Further statistical analysis demonstrates that gender may be one of the key factors that create site-specific variability in fMRI.

Table 1. Demographics and clinical information of dataset.

	P	K	NI	NY	O	Pitt	W
n	245	94	67	256	112	95	74
Age (meanstd)	11.7 1.95	10.2 2.49	17.6 3.03	11.5 2.91	9.1 1.25	15.1 2.77	11.5 3.85
Female	71	38	30	92	52	45	28
Male	174	56	43	171	61	53	33
ADHD	42%	27%	49%	58%	38%	4%	0%
TD	58%	73%	51%	42%	62%	96%	100%

2. RESEARCH METHODOLOGY

2.1. Functional MRI(fMRI) Datasets

We use fMRI data of ADHD-200 Preprocessed dataset [11], comprising 973 individuals, in this work. Each of the participants was scanned from 8 different sites[1]. Along with the scanned results, the research institutes also collect personal attribute data, including age, gender, handedness, ADHD scale, ADHD diagnosis(typically development(TD), ADHD-combined, ADHD-inattentive, ADHD hyperactive/impulsive) [12], and the measurements of intelligence. Since the ADHD diagnosis results from BHBU are pending, and the number of people who are diagnosed as TD and the number of people with ADHD is uneven in Pitt and WUSTL (see Table 1), we exclude those data of BHBU, Pitt, and WUSTL in our experiment. The diagnosis labels are categorized into two classes: TD and ADHD for binary classification.

2.2. Computational Framework

2.2.1. *Data Preprocessing and Feature Extraction*

In this study, we incorporate the pre-processed data from [11], in which the rs-fMRI data has gone through both artifact rejection and spatial/temporal calibration. An AAL90 [13] mask was applied which resulted in 90 regions of interests (ROIs). Several feature descriptors (see Table2.) were extracted.

2.2.2. *Adversarial Domain Adaptation (DANN)*

In this research, each data collection site is regarded as a unique domain $D = (x, y)$ with rs-fMRI features x and ADHD diagnosis label y. Our objective is to learn a model which is trained on source domain $D_s = (x_i^s, y_i^s)_i^{n_s}$ while having the transferability toward other unlabeled target domain $D_t = (x_i^t)_i^{n_t}$. Here we adopt the idea of adversarial training for its effectiveness on different transfer learning applications [14]. Given a model with feature extractor $\mathcal{F}$, source classifier $\mathcal{G}$ and domain discriminator $\mathcal{D}$, the update criteria would be:

$$\min_{\mathcal{G},\mathcal{F}} E_{(x_i^s,y_i^s)\sim\mathcal{D}_s} L(\mathcal{G}(f_i^s, y_i^s))$$
$$+\lambda(E_{(x_i^s)\sim\mathcal{D}_s} log[\mathcal{D}(f_i^s)] + E_{(x_j^t)\sim\mathcal{D}_t} log[1-\mathcal{D}(f_j^t)])$$
$$\max_{\mathcal{D}} E_{(x_i^s)\sim\mathcal{D}_s} log[\mathcal{D}(f_i^s)] + E_{(x_j^t)\sim\mathcal{D}_t} log[1-\mathcal{D}(f_j^t)])$$

note that L is the cross entropy loss which gives supervision of discriminability of the ADHD label, while λ is a hyper-parameter weighting source classifier and domain discriminator. The domain invariance features $f = \mathcal{F}(x)$ are learned through th minimax optimization procedure of the given criteria.

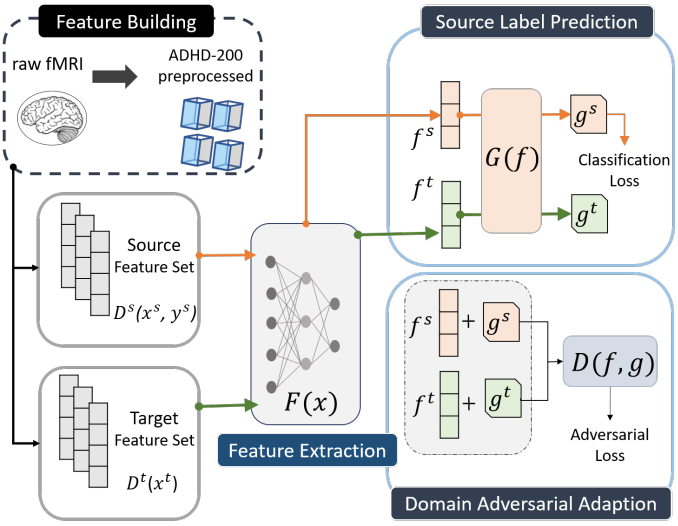


Fig. 1. A schematic of our proposed DNN-based adversarial learning which could extract class conditioned and site-invariant features.

Table 2. Regional and functional connectome based features for fMRI data in ADHD classification

Feature	Description
R90	The temporal mean pooling of each region of interest(ROI).
Func-R	Five functional statistics (maximum, mean, standard deviation, skewness, and kurtosis) of R90.
PCA-R	Principal component analysis with 10 component on R90.
ICA-R	Independent component analysis with 10 component on R90.
FTFC-R	The upper triangular of functional connectivity matrix using Pearson correlation among ROIs.

[1]Peking University(P), Bradley Hospital/Brown University(BHBU), Kenedy Krieger Institue(K), NeuroIMAGE sample(NI), New York University Child Study Center(NY), Oregon Health Sciences University(O), Peking University(P), University of Pittsburgh(Pitt) and Washington University at Saint Louis(W)

Table 3. A summary of ADHD classification results on 5 comparison features and model comparison of prediction results. The UAR results are in percentage. (Models: A:SVM, B:DNN, C:DNN-Z, D:DANN, E:CDAN+E; Features: 1:R90, 2:func-R, 3:PCA-R, 4:ICA-R, 5:FTFC-R)

	K-K	K-P	K-NI	K-NY	K-O	NI-NI	NI-P	NI-K	NI-NY	NI-O	NY-NY	NY-P	NY-K	NY-NI	NY-O	O-O	O-P	O-K	O-NI	O-NY	P-P	P-K	P-NI	P-NY	P-O
A1	48.4	49.7	51.4	50.0	49.6	**62.7**	54.0	46.5	50.0	52.0	**59.5**	50.0	50.0	50.0	50.0	**66.9**	47.8	49.3	39.6	50.0	**63.2**	48.4	52.0	50.0	51.7
A2	**57.1**	50.6	59.8	49.6	50.0	61.0	49.4	52.0	47.8	49.2	50.6	50.0	50.0	50.0	50.0	59.1	51.2	50.0	48.8	54.0	55.2	49.0	61.0	51.2	44.9
A3	40.5	49.4	51.3	50.0	50.4	52.4	50.5	51.6	50.0	48.4	58.5	50.4	39.0	68.4	47.5	62.6	42.7	52.4	41.9	50.0	56.2	50.2	61.9	50.0	42.6
A4	55.3	50.0	50.0	50.0	50.0	44.3	50.0	50.0	50.0	50.0	48.2	50.0	50.0	50.0	50.0	52.9	50.0	50.0	50.0	50.0	44.3	50.0	50.0	50.0	50.0
A5	48.4	49.4	53.1	49.9	52.3	56.7	48.8	52.1	50.6	51.0	57.0	56.0	55.6	45.5	45.6	57.9	51.7	46.2	56.8	50.8	56.4	50.8	57.4	56.2	56.3
B1	64.4	56.2	67.6	52.4	64.1	71.3	60.8	59.3	57.5	61.0	64.0	61.4	57.7	52.4	62.3	66.7	56.6	61.7	56.9	50.8	64.1	55.4	70.9	57.2	56.4
C1	67.2	55.9	68.7	52.9	60.5	65.4	55.3	59.4	54.3	61.0	62.3	58.5	57.2	67.9	51.4	66.9	52.3	60.0	66.7	52.8	65.0	59.9	73.3	60.5	60.2
D1		60.5	**73.7***	**60.8***	67.0		61.0	**63.7***	61.3	62.1		61.2	**64.1***	67.1	58.3		59.1	**67.9***	71.2	**58.7***		**65.9***	71.8	**61.7***	**66.0***
E1		**61.4***	71.8	54.8	**69.8***		**64.0***	63.3	**63.2***	**63.8***		**63.1***	61.0	**76.2***	**63.9***		**61.7***	67.3	**72.5***	53.9		64.1	**75.0***	60.5	65.3

2.2.3. Conditional DANN wiht Entropy (CDAN-E)

In DANN we only focus on the feature invariance between sites but neglect the potential label discrepancy between domains. Hence in [15], an improved conditional adversarial embedding was proposed for joint constraint on both feature spaces and label spaces. We replace f with $h = (f, g)$ which is the outer product of domain-specific feature representation f and the classifier predictions g. This gives us a tight binding between the representational spaces and label information. Besides, the traditional DANN imposes equal importance of different samples, while ignoring that those less informative (highly ambiguous and noisy) samples could harm the model. Thus, we incorporate an entropy criteria $w(H(g)) = 1 + e^{-H(g)}$ for uncertainty regularization where $H(g) = -\sum_{c=1}^{C} g_c \, log \, g_c$ is the prediction entropy acting as confidence level. The final objective would be formulated as:

$$\min_{\mathcal{G},\mathcal{F}} E_{(x_i^s, y_i^s)\sim\mathcal{D}_s} L(\mathcal{G}(f_i^s, y_i^s))$$
$$+\lambda(E_{(x_i^s)\sim\mathcal{D}_s} w_i^s log[\mathcal{D}(h_i^s)] + E_{(x_j^t)\sim\mathcal{D}_t} w_j^t log[1 - \mathcal{D}(h_j^t)])$$
$$\max_{\mathcal{D}} E_{(x_i^s)\sim\mathcal{D}_s} w_i^s log[\mathcal{D}(h_i^s)] + E_{(x_j^t)\sim\mathcal{D}_t} w_j^t log[1 - \mathcal{D}(h_j^t)])$$

3. EXPERIMENTAL SETUP AND RESULT

3.1. Experimental Setup

We conduct our experiments in two settings according to train and validation data: (i)Train and Evaluate on the same site using 10-fold cross-validation. These recognition results are regarded as the upper-bound for the specific site. (ii) Train on one site and evaluate on the other, for the evaluation of the domain adaptability. The final metric used is the unweighted average recall (UAR).

3.1.1. Comparison Models

We first evaluate the vanilla SVM and DNN method without considering the domain discrepancy in this transfer learning setup. Then, we compare the proposed method with the following methods for evaluation of domain transferability:

- **DNN**: Vanilla DNN without consideration of domain shift problems between different data sites. Several hyperparameters are gird searched: layer size:[90-32], [90-32-16], batch size:16,32,64, dropout rate:.0,.2,.5 learning rate using Adam optimizer:0.001, 0.0007, 0.0001, and max epoch is 100.
- **DNN-Z**: A simple site-wise Z-score normalization is applied to extracted fMRI brain features as an straightforward naive domain adaptation (normalization) approach. Network parameters are the same as DNN.
- **DANN**: Domain adversarial neural network with the constraint on domain shift of feature space. Layer dimensions are searched among: [90-45], [90-45-16], and the rest of hyperparameters are alike DNN while max epoch is 700.
- **CDAN+E**: Conditional adversarial domain adaptation with entropy regularization illustrated in 2.2.3.

3.2. ADHD Recognition Results

Table 3 summarizes our ADHD recognition results in a cross-site transfer learning setup. Our proposed CDAN+E method achieves consistently better recognition rates across most of the pairing between sites. Several observations can be summarized. First, we find that 90ROI features generally reaches the highest UAR in contrast to other features sets when comparing within-site accuracy. This indicates that this representation contains the most discriminative information in classifying ADHD and could be robustly generalized to all sites. Hence, all the rest of the experiments are run 90ROI for further domain generalization. Second, we observe that NeuroIMAGE has the highest average UAR 0.739, while NYU performs the poorest with 0.581. As a source domain, NYU and Peking have higher accuracies of 0.661 and 0.662 each, while NeuroImage as a source domain has the lowest accuracy of 0.636. This phenomenon could probably be related to the sample size. Finally, it is interesting to see that some of our CDAN domain transfer results even surpass the recognition results which train and evaluate on a single domain. For example, when taking NYU as a source domain and NeuroIMAGE as a target domain, the UAR remarkably increases by 0.238. This may demonstrate CDAN-E not only learns a conditional distributional mapping between datasets but further additional variability that increases the robustness of the

Table 4. A summary of the T-test between any two different sites. The bold part refers to the larger percentage between males and females. *:P-value<0.05. **:P-value<0.001, I: combined, II:Hyperactive, III:Inattentive, gen: gender ratio

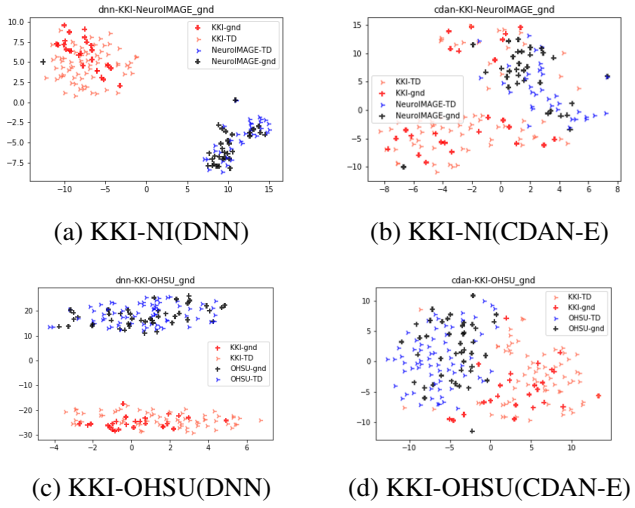
		K-O	K-P	K-NY	O-K	O-P	O-NY	P-O	P-K	P-NY	N-O	N-P	N-K
I	T	-	-5.93**	-9.26**	2.3*	2.82*	-14.7**	-	4.1**	-1.2	-	0.33	-2.06*
II	T	-0.77	-6.94**	-9.09**	2.45*	2.76*	-14.8**	-6.12**	3.84**	-1.38	-0.72	0.22	-1.85
III	T	-0.07	-3.98**	-9.15**	1.59	2.1*	-10.96**	-4.29**	4.44**	-1.12	-0.43	0.19	-2.78*
gen	M	54	34	32	46	42	16	39	27	40	39	43	23
(i)	F	62	51	57	47	32	37	60	29	46	42	41	32
gen	M	10	17	4	7	11	17	16	7	14	20	16	32
(ii)	F	8	32	5	5	24	24	15	16	11	21	27	32

model. In summary, the proposed CDAN-E could effectively improve the unsupervised transfer of ADHD recognition rate using fMRI by mitigating cross-site heterogeneous information, and we observe that the larger original data source dataset enhances the transferability.

4. ANALYSIS

To realize the effect of our adversarial domain adaptation learning, we first visualize the data distribution between vanilla DNN method and CDAN+E using t-SNE(see Fig 2.) with setting perplexity to 30. We could immediately see that after applying adversarial losses, two clusters are merged into one cluster, indicating that the features of two sites are close in the original dimension after applying CDAN+E. Furthermore, there is a clear clustering among data that has similar ADHD label, which indicates that not only domain discrepancy on feature space is mitigated but also clinical labels are successfully transferred.

Second, we split the participants into two groups:(i)(R&R): correctly predicted in both DNN and CDAN (ii)(W&R): wrongly predicted in DNN but corrected after applying CDAN. Then the two-sided Student's t-test was performed on ADHD index, Hyperactive and Inattentive between (i) and (ii) with pairs of two sites. We immediately notice that there is a significant statistical correlation among each site with t-value ranges from -2.06 to -14.8 with respect to all phenotypes. This indicates that before domain adaptation, the vanilla DNN model fails to identify severe ADHD patients, while our CDAN model can successfully identify those that mitigates this issue through domain invariant constraint. Moreover, through comparison of the gender ratio ($=\frac{Amountofmale/femaleofthecase}{Totalamountofmale/femaleofthetargetset}$), wee see that 10 out of 12 female gender-ratio surpass the male with a large margin in case(i), while the difference between 2 ratios is decreased in case(ii). This shows that it is easier to directly transfer female samples in this dataset between sites while male samples are more vulnerable to site effects. Our proposed CDAN alleviates partially this problem and reaches a similar performance between gender.



(a) KKI-NI(DNN) (b) KKI-NI(CDAN-E)

(c) KKI-OHSU(DNN) (d) KKI-OHSU(CDAN-E)

Fig. 2. The t-SNE plot of the feature extracted by DNN an CDAN+E. Select K-NI and K-O to illustrate here. (red plus: objects from source sample set, tomato tri: TD objects from source sample set, black plus: objects with ADHD from target sample set, blue tri: TD objects from target sample set)

5. CONCLUSION

In this work, we develop an approach in tackling the challenges of the domain discrepancy in rs-fMRI data caused by different collection sites. To enhance the robustness on automatic ADHD recognition in a cross-site setting, we introduce the conditional adversarial domain adaptation techniques to mitigate this domain shift problem. The experiments show that our proposed method achieves an improvement in cross-site ADHD classification, and further visualization indicates that our method does mitigate the site-mismatch problem through pulling data distribution closer, while statistical analysis reveals that male patients could be more severely affected by collection sites. To our best knowledge, this is one of the first work on ADHD recognition under cross-site prediction scenarios. There are multiple future directions. An immediate one would be verifying results on other large-scale cross-site brain imaging datasets. Second, how to generalize toward

multiple unspecific sites simultaneously would be another challenge. By better understanding of potential latent variability of multi-site data analysis would help in advancing a many human-centered computational research in clinical applications [16].

6. REFERENCES

[1] Alaa M Hamed, Aaron J Kauer, and Hanna E Stevens, "Why the diagnosis of attention deficit hyperactivity disorder matters," *Frontiers in psychiatry*, vol. 6, pp. 168, 2015.

[2] Irene M Loe and Heidi M Feldman, "Academic and educational outcomes of children with adhd," *Journal of pediatric psychology*, vol. 32, no. 6, pp. 643–654, 2007.

[3] SUBCOMMITTEE ON ATTENTION-DEFICIT et al., "Adhd: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents," *Pediatrics*, pp. peds–2011, 2011.

[4] Stephen V Faraone and Kevin M Antshel, "Diagnosing and treating attention-deficit/hyperactivity disorder in adults," *World Psychiatry*, vol. 7, no. 3, pp. 131–136, 2008.

[5] Qingjiu Cao, Yufeng Zang, Li Sun, Manqiu Sui, Xiangyu Long, Qihong Zou, and Yufeng Wang, "Abnormal neural activity in children with attention deficit hyperactivity disorder: a resting-state functional magnetic resonance imaging study," *Neuroreport*, vol. 17, no. 10, pp. 1033–1036, 2006.

[6] Chao-Zhe Zhu, Yu-Feng Zang, Qing-Jiu Cao, Chao-Gan Yan, Yong He, Tian-Zi Jiang, Man-Qiu Sui, and Yu-Feng Wang, "Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder," *Neuroimage*, vol. 40, no. 1, pp. 110–120, 2008.

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

[8] Ling-Li Zeng, Huaning Wang, Panpan Hu, Bo Yang, Weidan Pu, Hui Shen, Xingui Chen, Zhening Liu, Hong Yin, Qingrong Tan, et al., "Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri," *EBioMedicine*, vol. 30, pp. 74–85, 2018.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[10] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.

[11] Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock, "The neuro bureau adhd-200 preprocessed repository," *Neuroimage*, vol. 144, pp. 275–286, 2017.

[12] Timothy E Wilens, Joseph Biederman, Stephen V Faraone, MaryKate Martelon, Diana Westerberg, and Thomas J Spencer, "Presenting adhd symptoms, subtypes, and comorbid disorders in clinically referred adults with adhd," *The Journal of clinical psychiatry*, vol. 70, no. 11, pp. 1557, 2009.

[13] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[15] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, 2018, pp. 1640–1650.

[16] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.