# Alzheimer's Disease Brain Network Classification Using Improved Transfer Feature Learning with Joint Distribution Adaptation

Binglin Wang, Wei Li, Wenliang Fan, Xi Chen, Dongrui Wu, for the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—Alzheimer's disease significantly affects the quality of life of patients. This paper proposes an approach to identify Alzheimer's disease based on transfer learning using functional MRI images, which is especially useful when the training dataset is small. Transfer learning improves the performance of the classifier with the help of an auxiliary dataset, which may be obtained from a different population group and/or machine. First, we used the joint distribution adaptation method to project the source and target domain samples into a new feature space, and then we built a classifier that works well in both the source and target domains but emphasizes the target domain. In the classifier, we assigned larger weights to the target domain samples and minimized the weighted loss in classifying the samples in both domains. Experimental results verify the effectiveness of our proposed approach and, with the help of the auxiliary samples, the classification accuracy of our target dataset has been greatly improved.**

***Index Terms*—Alzheimer's Disease, Brain Network, Transfer Learning, Joint Distribution Adaptation.**

## I. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that occurs in the brain and significantly affects the patient's quality of life [1]. It is predicted that the number of AD patients will double in the next 20 years [2]. Diagnosing AD early and accurately can greatly benefit its treatment.

With the rapid advance in brain-analysis technology, complex network analysis has become a common method for studying the brain. Complex brain networks are constructed using several brain-imaging technologies, such as PET, EEG, and functional-MRI (fMRI). Betty et al. studied the changes in the brain network of AD patients [3]. Dosenbach et al. used complex networks to study the growth of the brain [4]. Many computer-assisted AD diagnosis techniques have also been proposed based on brain networks. For example, Zhang et al. used machine learning to distinguish AD patients using brain networks [5]. However, these algorithms usually need a large amount of training samples for reliable performance, but the collection of AD cases is difficult. Moreover, fMRI images may be collected in different locations on different machines, whose parameters and formats are not consistent.

The goal of transfer learning is to make use of information from a (large) auxiliary dataset to improve the learning on a (small) target dataset. It is a commonly used approach in machine learning when it is not possible to achieve a good performance in training a classifier based on a small dataset. Transfer learning has shown promising results in EEG-based brain–computer interfaces. For example, Wu et al. showed that transfer learning can effectively deal with individual differences and reduce the number of labeled subject-specific training samples [6]. However, there is little research on the applications of transfer learning to AD diagnosis based on fMRI data, which is more difficult and expensive to acquire than EEG data.

In this work, we consider the case where the target domain contains a very small number of samples and propose a transfer learning method for the fMRI images. We use the joint distribution adaptation (JDA) method to find a new feature space in which the source and target domain data are more consistent and design a classifier in that space by considering the information in both domains.

## II. Method

JDA reduces the differences between both the marginal and the conditional distributions between the two domains by projecting all samples onto a new lower-dimensional feature space [7].

The labeled source domain data $D_s = \{(x_1, y_1), \ldots, (x_{n_s}, y_{n_s})\} \in (X_s, Y_s)$ and target domain data $D_t = \{(x_{n_s+1}, y_{n_s+1}), \ldots, (x_{n_s+n_t}, y_{n_s+n_t})\} \in (X_t, Y_t)$ are given, where $n_s$ is the number of samples in the source domain and $n_t$ is the number of samples in the target domain; $X_s, X_t$ are multidimensional feature spaces and $Y_s$, $Y_t$ are label spaces. It is assumed that $X_s = X_t, Y_s = Y_t, P_s(X_s) \neq P_t(X_t)$, $Q_s(Y_s|X_s) \neq Q_t(Y_t|X_t)$, where $P(X)$ is the marginal probability distribution of $X$, and $f(x) = Q(Y|X)$ is the conditional probability distribution. JDA learns a feature representation in which the distribution differences are

B. Wang, W. Li, X. Chen, D. Wu are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Image Processing and Intelligent Control Key Laboratory of Education Ministry of China, Wuhan 430074, China (e-mail: binglin_wang@hust.edu.cn, liwei0828@mail.hust.edu.cn, chenxi@hust.edu.cn, drwu@hust.edu.cn ).

W. Fan is with the Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China(fwl@hust.edu.cn)

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

explicitly reduced between 1) $P_s(X_s)$ and $P_t(X_t)$ 2) $Q_s(Y_s|X_s)$ and $Q_t(Y_t|X_t)$ [7].

Let $X = [x_1, x_2, \ldots, x_n]$ be the combined data matrix from both domains, where $n = n_s + n_t$. The JDA optimization problem is as follows:

$$\min_{A^T X H X^T A = I} \sum_{c=0}^{C} \mathrm{tr}(A^T X M_c X^T A) + a\|A\|_F^2 \quad (1)$$

where $H = I - \frac{1}{n}1$ is the centering matrix, **1** is the matrix of ones, $a$ is the regularization parameter, and $M_c$ is the maximum mean discrepancy matrix defined as

$$(M_c|c=0)_{ij} = \begin{cases} \frac{1}{n_s n_s}, x_i, x_j \in D_s \\ \frac{1}{n_t n_t}, x_i, x_j \in D_t \\ \frac{-1}{n_s n_t}, otherwise \end{cases} \quad (2)$$

$$(M_c|c=1,\ldots,C)_{ij} = \begin{cases} \frac{1}{n_s{}^{(c)} n_s{}^{(c)}}, \ x_i, x_j \in D_s{}^{(c)} \\ \frac{1}{n_t{}^{(c)} n_t{}^{(c)}}, \ x_i, x_j \in D_t{}^{(c)} \\ \frac{-1}{n_s{}^{(c)} n_t{}^{(c)}}, \begin{cases} x_i \in D_s{}^{(c)}, x_j \in D_t{}^{(c)} \\ x_j \in D_s{}^{(c)}, x_i \in D_t{}^{(c)} \end{cases} \\ 0 \quad , \ otherwise \end{cases} \quad (3)$$

in which $D_s{}^{(c)} = \{x_i : x_i \in D_s, y(x_i) = c\}$ is the set of all samples from the c-th class in the source domain and, $D_t{}^{(c)}$ is the set of all samples from the c-th class in the target domain, c=0 means $M_0$ is not related to the label. $y(x_i)$ is the true or predicted label of $x_i$.

When $c = 0$, the difference between the marginal probability distributions of the two domains is reduced. When $c = 1,2,\ldots,C$ the difference between the conditional probability distributions of the c-th class of the two domains is reduced. The optimal fitness matrix A can be found by solving (4) to obtain the eigenvectors corresponding to the k smallest eigenvalues [7].

$$(X \sum_{c=0}^{C} M_c X^T + aI)A = XHX^T A\phi \quad (4)$$

JDA maps the source and target domain samples to a more consistent new feature space, and then we build a classifier in this feature space. However, because the source domain has much more data than the target domain, whereas the classifier will be used in the target domain, we need to properly balance the two domains so that the target domain will not be overwhelmed by the source domain. Sometimes a single sample of the source domain contains less effective information than the target domain.

An intuitive representation of the algorithm is shown in Figure. 1. As the figure shows, the classification accuracy is low when we use only the target data and, with the help of the auxiliary samples, the classification boundaries are clearer. However, if the source domain sample distribution is slightly different from the target domain, as shown in Figures. 1c and 1d, then the target domain samples should be of greater importance.

The optimization problem of our classifier is

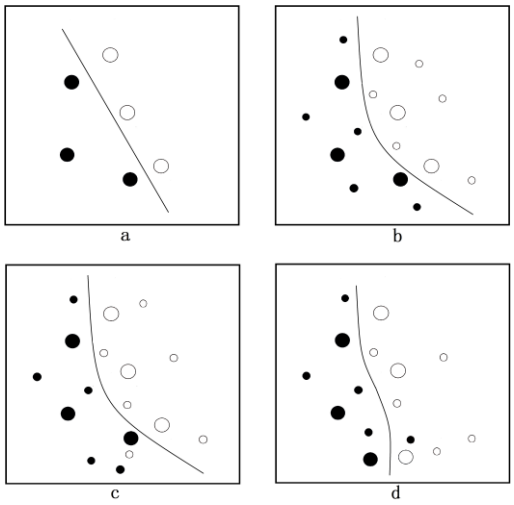


Figure 1. Intuitive representation of the improved JDA algorithm. Different colors represent different categories, the white balls represent the positive samples and the black balls are the negetive samples; different sizes represent different levels of importance, and the larger balls represent the target domain samples, while the smaller ones represent the source domain samples.

$$f = argmin \sum_{i=1}^{n_s} w_{s,i} l(f(x_i), y_i) + w_t \sum_{i=n_s+1}^{n_s+n_{t,l}} w_{t,i} l(f(x_i), y_i) + \sigma\|f\|_K^2 \quad (5)$$

where $n_{t,l}$ and $n_{t,u}$ are the number of labeled and unlabeled samples in the target domain, respectively; $l(\cdot)$ is the loss function; σ is a regularization parameter; $w_t$ is the overall weight for all target domain samples; $w_{s,i}$ and $w_{t,i}$ are the weights for the i-th sample in the source and target domain, respectively, whose purpose is to balance the positive and negative samples.

There are three terms in (5), the 1-st term minimizes the weighted loss on fitting the samples in the source domain; the 2-nd term minimizes the weighted loss of the target domain; and the 3-rd term minimizes the structural risk of the model. The solution of (5) is [8] [9]

$$f(x) = \sum_{i=1}^{n_s+n_t} \alpha_i K(x_i, x) + b = \alpha^T K(X, x) + b \quad (6)$$

where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_{n_s+n_{t,l}+n_{t,u}}]^T$, $K \in R^{(n_s+n_{t,l}+n_{t,u})\times(n_s+n_{t,l}+n_{t,u})}$ is the kernel matrix with $K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$; so we can convert the optimization problem to [**Error! Bookmark not defined.**]:

$$\alpha = \underset{\substack{\alpha \in \mathbb{R}^{n_s+n_{t,l}+n_{t,u}} \\ \xi \in \mathbb{R}^{n_s+n_{t,l}}}}{argmin} \sum_{i=1}^{n_s+n_{t,l}} E_{ii}\xi_i + \sigma\alpha^T K\alpha$$

$$s.t. \quad y_i\left[\sum_{i=1}^{n_s+n_{t,l}+n_{t,u}} \alpha_i K(x_i, x_j) + b\right] \geq 1 - \xi_i \quad (7)$$

$$\xi_i \geq 0, \quad i = 1,2,\ldots,n_s + n_{t,l}$$

where $E_{ii} = diag(w_{s,1}, \ldots, w_{s,n_s}, w_t w_{t,n_s+1}, \ldots, w_t w_{t,n_s+n_{t,l}})$.

The pseudo code of the overall algorithm is summarized in Algorithm 1.

**Algorithm 1:** the improved JDA algorithm

**Input:**

$n_s$ labeled source domain samples: $\{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i=1}}^{\mathrm{n_s}}$;

$n_{t,l}$ labeled target domain samples: $\{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i=n_s+1}}^{\mathrm{n_s+n_{t,l}}}$,

$n_{t,u}$ unlabeled target domain samples: $\{\mathrm{x_i}\}_{\mathrm{i=n_s+n_{t,l}+1}}^{\mathrm{n_s+n_{t,l}+n_{t,u}}}$,

Subspace dimension: $k$;

Regularization parameters: $\lambda$;

Maximum iterations: $T$

**Output:**

labels: $\{\widehat{y_i} := \mathrm{f}(\mathrm{A^T x_i})\}_{\mathrm{i=n_s+n_{t,l}+1}}^{\mathrm{n_s+n_{t,l}+n_{t,u}}}$

**Begin:**

a. Calculate the MMD matrix $M_0$ by (2); set $\{M_c := 0\}_{c=1}^{C}$;

b. Solve (4), and choose the eigenvectors corresponding to the k smallest eigenvalues as A, so $Z := A^T X$;

c. Calculate the kernel matrix by $\{(A^T x_i, y_i)\}_{i=1}^{n_s+n_{t,l}}$, and calculate $\alpha, b$ by (7);

d. Calculate $\{f(x_i)\}_{i=n_s+n_{t,l}+1}^{n_s+n_{t,l}+n_{t,u}}$ by (6), and get $\{\widehat{y_i}\}_{i=n_s+n_{t,l}+1}^{n_s+n_{t,l}+n_{t,u}}$,

e. Update the MMD matrix $\{M_c\}_{c=1}^{C}$ by (3);

f. If $\{\widehat{y_i}\}_{i=n_s+n_{t,l}+1}^{n_s+n_{t,l}+n_{t,u}}$ no longer changes or get maximum iterations T, return; else go to Step b;

**End**

## III. EXPERIMENTS

### A. Datasets

We have 292 resting fMRI cases in the source domain and 26 cases in the target domain. Source data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. For up-to-date information, see www.adni-info.org. The target data were obtained from the Tongji Hospital, Huazhong University of Science and Technology.

The ADNI dataset is a widely used public dataset, it has more samples than Tongji Hospital. Table 1 presents the summary of these two datasets. We have 117 samples of patients with Alzheimer's Disease (AD) and 175 normally control (NC) samples in the ADNI dataset, 12 AD samples and 14 NC samples in the Tongji dataset. Where Mini–Mental State Examination (MMSE) is a questionnaire to measure cognitive impairment, and healthy people usually have large MMSE values.

TABLE I. SUMMARY OF THE ADNI AND TONGJI DATASETS.

| Dataset | Information | AD | NC |
|---|---|---|---|
| ADNI | Number (M/F) | 117 (56/61) | 175 (98/77) |
| | MMSE | 21.3 ± 3.5 | 28.9 ± 1.5 |
| | Age | 74.6 ± 7.5 | 75.5 ± 6.1 |
| Tongji | Number (M/F) | 12 (6/6) | 14 (8/6) |
| | MMSE | 16.7 ± 3.0 | 28.5 ± 1.2 |
| | Age | 65.7 ± 11.9 | 65.7 ± 7.5 |

### B. Data Preprocessing

We used the data processing assistant for resting-state fMRI (DPARSF) [10] to preprocess all samples. We removed the first ten volumes of each time series, corrected the slice time, realigned the volumes for head motion correction, and normalized them to the EPI template. We excluded samples with significant head motions. Subsequently, the images were spatially smoothed, and the linear trends of the time courses were removed. Next, we applied a [0.01, 0.08] Hz band-pass filter to the time courses of each voxel, and the global mean signal, white matter signal, and cerebrospinal fluid sign were regressed out. Finally, we extracted the time series of 90 ROIs of each sample following the AAL template [11].

### C. Feature Extraction

The Pearson correlation coefficient can be used to measure the linear relationship between variables, which is often used in the construction of brain networks [12]. As the brain network is a $90 \times 90$ symmetric matrix, we used the 4005 unique terms as our features. Given that the number of features is very large, it is necessary to select the most discriminative features. For the sequence of each feature of the training sample, we calculated its Kendall correlation coefficient with the sequence of labels and selected the features with large correlation coefficient as the inputs to the classifier.

### D. Experimental Results

We divided the samples into three parts: the entire ADNI dataset as the labeled source domain data $S$, some samples from the Tongji dataset as the unlabeled target domain samples $T_u$, and the remaining Tongji dataset samples as the labeled target domain samples $T_l$. Three methods were compared:

A. $T_l$ for training and $\mathrm{T_u}$ for the test, i.e., we only used the Tongji dataset for training and test;

B. $T_l$ combined with $S$ for training and $\mathrm{T_u}$ for the test, i.e., we simply combined the samples from different domains for training; and

C. $T_l$ and $S$ for training and $T_u$ or the test, by applying our proposed algorithm.

Classification accuracy was calculated by

$$\mathrm{Accuracy} = (\mathrm{x}: \mathrm{x} \in T_u \wedge \hat{y}(\mathrm{x}) = \mathrm{y}(\mathrm{x}))/(\mathrm{x}: \mathrm{x} \in T_u) \quad (7)$$

The relationship between the classification accuracy and the number of features is shown in Figure. 2.

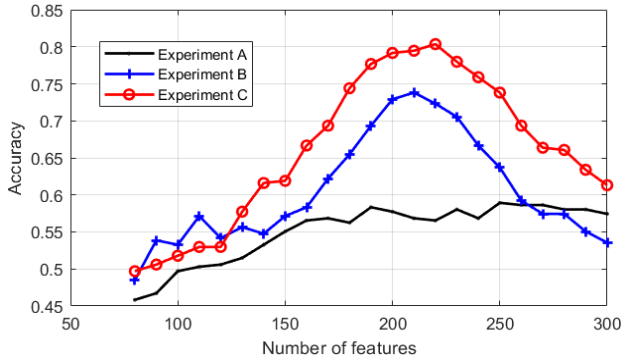


Figure 2. Results with our proposed algorithm compared with control groups.

In Experiment A, which used only the small Tongji dataset, the maximum testing classification accuracy was approximately 58%, close to a random guess. In Experiment B, we mixed the Tongji dataset with the ADNI dataset in training, and the maximum testing classification accuracy was approximately 75%, much better than Experiment A. This implied that, although the ADNI and Tongji datasets were from different populations and different fMRI scanners, they still shared similarities, so one dataset can help with the classification of the other. In Experiment C, the proposed improved JDA algorithm was used to transfer knowledge from the ADNI to the Tongji dataset, and a maximum classification accuracy of 80% was achieved, a 22% improvement over Experiment A, and a 7% improvement over Experiment B.

As we can see, when the number of features is approximately 220, the performance of the classifier is the best, and its accuracy is 80.4%. The mixing matrix for the classifier when using 220 features is presented in Table 2.

To visualize the distribution of samples in different datasets, we used t-SNE to reduce the features of the ADNI and Tongji datasets to two dimensions. The results are shown in Figure. 3. Before JDA (Figure. 3a), we can observe that the data distributions of the two domains are different. The Tongji samples are mostly distributed on the edge of the ADNI samples. As most machine learning algorithms assume that the

TABLE II. MIXING MATRIX FOR THE CLASSIFIER WHEN USING 220 FEATURES, WHERE ROWS REPRESENT THE TRUE LABEL AND COLUMNS REPRESENT THE RESULT OF THE CLASSIFIER.

| | **AD** | **NC** |
|---|---|---|
| **AD** | TP: 0.7083 | FN：0.2917 |
| **NC** | FP：0.1012 | TN：0.8988 |

training and testing datasets have the same distribution [13], such an inconsistency usually results in poor generalization performance. After transfer learning (Figure. 3b), we can observe that the difference between the source and target domain distributions is significantly reduced. As a result, the classification performance of Experiment C was improved, a 22% improvement over Experiment A, and a 7% improvement over Experiment B.

In summary, these experimental results indicate that the improved JDA algorithm can effectively extract relevant model information from the source domain data and assist the classification of the small dataset in the target domain. So, our approach can significantly improve the classification accuracy of the target domain samples.

## IV. CONCLUSION

This work proposes a transfer learning approach to classify AD based on fMRI images. Particularly, it considers the problem in which we have a large number of auxiliary samples in the source domain and a very small number of samples in the target domain. These two domains have different distributions. Our approach first mapped the samples in the two domains onto a more consistent feature space and then assigned larger weights to the target domain samples and minimized the weighted loss in classifying the samples in both domains. This ensured that the designed classifier worked well in both the source and the target domains but emphasized the target domain. Experimental results show that our approach can significantly improve the classification accuracy, which may help in the development of a computer-assisted AD diagnosis system.

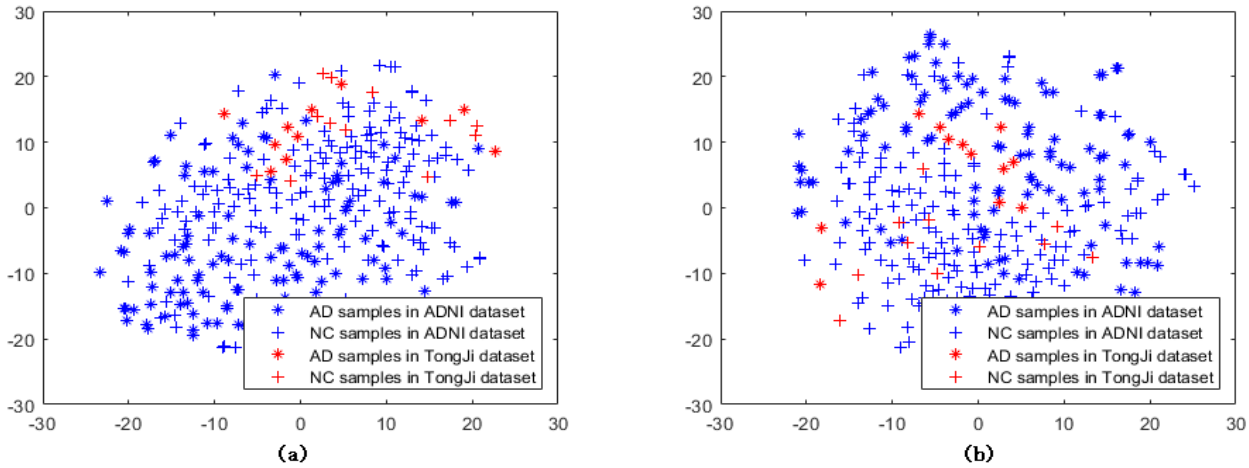


Figure 3. Distributions of ADNI and Tongji datasets. (a) Distributions before JDA. (b) Distributions after JDA.

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (61473131, 60905024).

Data collection and sharing for this project in the source domain was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

## REFERENCES

[1] T. Arendt, "Synaptic degeneration in Alzheimer's disease". Acta neuropathologica, vol 118, no 1, pp. 167-179, 2009.

[2] R. Brookmeyer, E. Johnson, K. Ziegler-Graham and H. Arrighi, "Forecasting the global burden of Alzheimer's disease", Alzheimer's & Dementia, vol. 3, no. 3, pp. 186-191, 2007.

[3] B. Tijms, A. Wink, W. de Haan, W. van der Flier, C. Stam, P. Scheltens, and F. Barkhof, "Alzheimer's disease: connecting findings from graph theoretical studies of brain networks", Neurobiology of Aging, vol. 34, no. 8, pp. 2023-2036, 2013.

[4] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems", Nature Reviews Neuroscience, vol. 10, no. 3, pp. 186-198, 2009.

[5] C Wee, P Yap, D Zhang, K Denny, J Browndyke, G Potter, K Welsh-Bohmer, L Wang, and D Shen, "Identification of MCI individuals using structural and functional connectivity networks", NeuroImage, vol. 59, no. 3, pp. 2045-2056, 2012.

[6] D. Wu, V. Lawhern, W and B. Lance, "Reducing offline BCI calibration effort using weighted adaptation regularization with source domain selection". IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2016: 3209-3216.

[7] M. Long, J. Wang, G. Ding, J. Sun, and P. Yu, "Transfer feature learning with joint distribution adaptation". IEEE International Conference on Computer Vision. IEEE, 2013: 2200-2207.

[8] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples". Journal of Machine Learning Research, 2006, 7(1):2399-2434.

[9] M. Long, J. Wang, G. Ding, S. Pan and P. Yu, "Adaptation Regularization: A General Framework for Transfer Learning", IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 5, pp. 1076-1089, 2014.

[10] http://rfmri.org/DPARSF.

[11] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. 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.

[12] M. Lynall, D. Bassett, and R. Kerwin. "Functional Connectivity and Brain Networks in Schizophrenia", Journal of Neuroscience, vol. 30, no. 28, pp.9477-9487,2010

[13] S. Pan and Q. Yang, "A Survey on Transfer Learning", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.