2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

DISCRIMINATING SCHIZOPHRENIA FROM NORMAL CONTROLS USING RESTING STATE FUNCTIONAL NETWORK CONNECTIVITY: A DEEP NEURAL NETWORK AND LAYER-WISE RELEVANCE PROPAGATION METHOD

Weizheng Yan[1], Sergey Plis[2], Vince D Calhoun[3], Shengfeng Liu[1], Rongtao Jiang[1], Tian-Zi Jiang[1,2], Jing Sui[1,2,3]**

1. Brainnetome center, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
2. University of Chinese Academy of Sciences, Beijing, China
3. The Mind Research Network/LBERI, Albuquerque, NM, USA

ABSTRACT

Deep learning has gained considerable attention in the scientific community, breaking benchmark records in many fields such as speech and visual recognition [1]. Motivated by extending advancement of deep learning approaches to brain imaging classification, we propose a framework, called "deep neural network (DNN)+ layer-wise relevance propagation (LRP)", to distinguish schizophrenia patients (SZ) from healthy controls (HCs) using functional network connectivity (FNC). 1100 Chinese subjects of 7 sites are included, each with a 50*50 FNC matrix resulted from group ICA on resting-state fMRI data. The proposed DNN+LRP not only improves classification accuracy significantly compare to four state-of-the-art classification methods (84% vs. less than 79%, 10 folds cross validation) but also enables identification of the most contributing FNC patterns related to SZ classification, which cannot be easily traced back by general DNN models. By conducting LRP, we identified the FNC patterns that exhibit the highest discriminative power in SZ classification. More importantly, when using leave-one-site-out cross validation (using 6 sites for training, 1 site for testing, 7 times in total), the cross-site classification accuracy reached 82%, suggesting high robustness and generalization performance of the proposed method, promising a wide utility in the community and great potentials for biomarker identification of brain disorders.

Index Terms— deep neural network, layer-wise relevance propagation, functional network connectivity, schizophrenia

1. INTRODUCTION

Resting-state fMRI (rs-fMRI) has been successfully employed to exploit neuronal underpinnings in neuropsychiatric disorders. A recent schizophrenia classification challenge demonstrated clearly, across a broad range of classification approaches, the value of resting state fMRI data in capturing useful information about schizophrenia [2]. Low signal-to-noise ratio, high dimension, and small sample size are still the main challenges in neuropsychiatric disorder diagnostics [3]. To overcome the difficulties, hundreds of machine learning methods have been carried out for dimension reduction and accurate classification of patients with heterogeneous mental and neurodegenerative disorders, including independent component analysis (ICA), support vector machine (SVM), and so on. It is worth mentioning that ICA is a method for recovering underlying signals from linear mixtures of these signals and draws upon higher-order signal statistics to determine a set of "components" that are maximally independent of each other. It is able to capture the complex nature of fMRI time courses while also producing consistent spatial components [4]. The resting state FNC features which are computed between each pair of selected independent components have been successfully exploited to automatically discriminate schizophrenia patients [5].

Deep learning methods have recently demonstrated unprecedented classification performance via a hierarchical representation of input data from research studies to industrial applications. More recently, the deep learning has shown its efficacy to neuroimaging data [6-8]. Nevertheless, because of the nested non-linear structure, it is not obvious which input dimensions are mainly responsible for a given prediction. LRP is a potential remedy for the lack of interpretability of DNNs that has limited their utility in clinical applications [9]. LRP explains individual classification decisions of a DNN by decomposing its output in terms of input variables.

The purpose of this study is two-fold. First, improving the classification accuracy of SZ patients vs. HC subjects by using modified DNN classifier. Second, extracting the highly discriminative FNCs with LRP. Therefore, we propose a

978-1-5090-6341-3/17/\$31.00 ©2017 IEEE

framework, called "DNN+LRP", for the discrimination of schizophrenia patients from healthy controls and explanation of individual network decisions. The framework integrates functional network connectivity patterns with deep learning methods. The philosophy is natural yet effective: (1) reducing the dimension of the fMRI data; (2) finding out the high-level and relevant feature representation. FNC based on ICA, as an efficient unsupervised and feature-selection method, can avoid the curse of dimensionality and improve the ratio of signal to noise. In addition, the DNN can derive hierarchical feature representations from the lower level features. What's more, the heat map derived by LRP can explain why the classifier reaches a certain decision in a single instance.

Fig. 1 shows the framework corresponding technical details of the algorithm. The DNN classifier outperformed the conventional machine learning methods (i.e., SVM, Random Forests, AdaBoost classifier) and performed well while doing multi-sites prediction task, which suggests that the DNN-based method can lead to better feature learning. The biomarker subsets extracted by LRP are meaningful in the context of medical medicine.

2. MATERIALS AND METHODS

2.1. Overview

Fig. 1 presents an overview flow diagram of the analysis. First, the raw data are preprocessed following the standard procedure. Then the group ICA algorithm is applied to calculate group independent components (ICs). After removing ICs with artifacts, the functional network connectivity matrices are computed with the reserved ICs (Fig. 1a). Second, the FNC patterns are used as the input of the DNN classifier. The DNN classifier is trained and parameters are optimized using training and validation data from the cross-validation (CV) framework during the training phase (Fig. 1b). A classification of SZ patient and HC subject are performed for each individual in the test data during the test phase. Third, the explanations of how a DNN reaches a decision and the most informative FNC patterns are produced with layer-wise relevance propagation (LRP) (Fig. 1c).

2.2. Data description and preprocessing

The resting-state fMRI data are collected from 542 healthy and 558 schizophrenic patients. The details of the data are shown in Table I. The fMRI data are preprocessed using SPM8 software (http://www.fil.ion.ucl.ac.uk/spm/), motion corrected, spatially normalized into standard MNI space and slightly subsampled to voxel size $3 \times 3 \times 3$ mm^3, resulting in $53 \times 63 \times 46$ voxels. Then group ICA is performed on the preprocessed fMRI data using GIFT software (http://mialab.mrn.org/software/gift), fMRI images are decomposed via principal component analysis (PCA), with the first 100 components selected for dimension reduction. Then the Infomax algorithm is repeated 20 times using

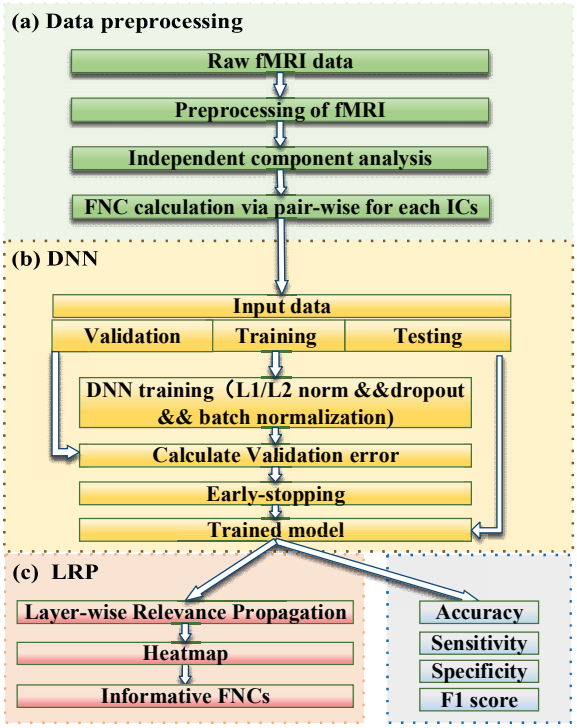

Fig. 1. Overview flow of the "DNN-LRP" framework

ICASSO to improve the reliability of the decomposition, resulting in 100 group independent components (ICs). 50 ICs are characterized as functional network connectivity (FNC) after removing ICs with artifacts [18]. The time courses (TCs) of 50 ICNs across the whole brain are post-processed by detrending, regressing out head motion, despiking and low-pass filtering (<0.15 Hz). Then the FNC matrices for all subjects are calculated as the Pearson's correlation between TCs of each pair of ICs. The magnitudes of FNC strength are used as the input of the DNN model.

2.3. DNN training with L1/L2 norm regularization and parameter optimization

The DNN model consists of one input layer, multiple hidden layers, and one output layer. The target values of the two output nodes in the output layer are assigned as $[1,0]^T$ and $[0,1]^T$ for the input pattern from HC and the SZ group. The high-dimensional characterization of an individual may cause overfitting on the training dataset. To reduce the susceptibility of overfitting, we limit the complexity by controlling sparsity of the parameters with $L1/L2$ norm regularization. The loss function, $L(W)$ of the DNN for the supervised fine-tuning step is defined using the mean squared error (MSE), $L1/L2$ terms as follows:

$$L(\boldsymbol{W}) = \frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{y}^{\{n\}} - \boldsymbol{t}^{\{n\}}\|^2 + \sum_{l=1}^{L}\beta_l\|\boldsymbol{W}^{\{l\}}\| + \frac{1}{2}\sum_{l=1}^{L}\gamma_l\|\boldsymbol{W}^{\{l\}}\|^2 \quad (1)$$

Table I Characteristics of the participants in this study

Group	Healthy	Schizophrenia	P
Number	542	558	NA
Age ($mean \pm sd$)	28.0±7.2	27.6±7.1	0.06
Gender (male/%)	276/51%	292/52%	1.96

Where $\boldsymbol{y}^{\{n\}}$ is the output vector for the n_{th} subject, n in the training set, $\boldsymbol{t}^{\{n\}}$ is the target output values for the n_{th} subject. β and γ are the $L1/L2$ norm regularization parameters, respectively. $\boldsymbol{W}^l$ is the DNN weights matrix in the l_{th} hidden layer. N is the number of subjects in training data, L is the number of hidden layers in the DNN. The adjustable parameters are jointly optimized through minimizing the misclassification error over the training set [2].

The stochastic gradient-based optimization algorithm we used to minimize the loss is *Adam*, a method for efficient stochastic optimization that only requires first-order gradients with little memory requirement [10]. The method combines the advantages of two recently popular methods: *RMSProp*, which works well in online and non-stationary settings, and *AdaGrad*, which works well with sparse gradients. Stochastic regularization methods, such as *dropout*, are also effective ways to prevent overfitting and often used in practice due to their simplicity. The neurons which are dropped out do not contribute to the forward pass and do not participate in backpropagation. *Batch normalization* can also aid generalization [11]. In this practice, in addition to $L1/L2$ norm regularization, we also applied both *dropout* and *batch normalization* after applying the *ReLU* nonlinearity in all hidden layers.

2.4. Multi-sites prediction

By now, a large amount of data from multi-sites studies has been available for developing, training and evaluating automated classifiers. However, their translation to the clinical remains challenging in part due to their limited generalizability across different datasets [12]. Generally speaking, machine learning methods built on a relatively low number of subjects suffer from poor generalizability [13]. Combining datasets, even if acquired on different scanners and from different centers, can theoretically improve predictive models. Therefore, the limitations of small sample sizes can be overcome by combining multi-sites data, and unbalanced datasets can benefit from the more balanced distribution that derives from merging multiple datasets [14]. Furthermore, the robustness and generalization performance of the machine learning method has important effects on the classification results.

To reduce the susceptibility to overfitting, we limit the complexity of DNN classifier with some tricks, including $L1/L2$ *norm regularization*, *dropout* and *batch normalization*. To test multi-site prediction performance of the DNN model, we elects to use a leave-one-site-out cross validation method, which allows one center left for testing and the remaining centers used for model creation. The procedure proceeds until each center had performed once in the test set. The average of all these datasets is what generated our multi-site predictive values.

2.5. Layer-wise Relevance Propagation for Deep Neural Network Architectures

A deep neural network is a feed-forward graph of elementary computational units. Each of them realizing a simple function of type [15].

$$x_j^{l+1} = \max\left(0, \sum_i x_i^{\{l\}} w_{ij}^{\{l,l+1\}} + b_j^{l+1}\right) \quad (2)$$

Where j indexes a neuron at a particular layer $l+1$, where Σ runs over all lower-layer neurons connected to neuron i, and where $w_{ij}^{\{l,l+1\}}$, b_j^{l+1} are parameters specific to pairs of adjacent neurons and learned from the data.

Because of the nested non-linear structure, it is not obvious which input dimensions are mainly responsible for a given prediction. Layer-wise relevance propagation is a method to compare scores for image pixels and image regions denoting to the prediction of the classifier for one particular test image. The extreme specificity of the LRP-derived heat maps can open up new avenues for investigating neural activity underlying complex perception of decision-related processes. The spatiotemporal heat maps represent a new quality of explanatory resolution that allows us to explain why the classifier reaches a certain decision in a single instance. A modularity analysis incorporating the learned features of the DNN can reveal how FNC patterns are decomposed in each of the hidden layers to better discriminate SZ patients from HCs.

A deep network derives its complexity from the interconnection of a large number of these elementary units, and from the availability of an efficient algorithm for learning the model (error backpropagation). The output of a deep neural network is obtained by evaluating these neurons in a feed-forward pass. Conversely, the same graph can be used to redistribute the relevance at the output of the network onto pixel-wise relevance scores $R_i^{\{l\}}$ [16], by using a local redistribution rule:

$$R_i^{\{l\}} = \sum_j \frac{z_{ij}}{\sum_i z_{ij}} R_j^{\{l+1\}}, z_{ij} = x_i^{\{l\}} w_{ij}^{\{l,l+1\}} \quad (3)$$

Where i indexes a node at a particular layer l and Σ runs over all upper-layer neurons to which neuron i contributes.

For a more theoretical view of LRP, we refer the reader to [16, 17], where the author shows a close connection between LRP and a deep Taylor decomposition. An implementation of LRP can be found and downloaded from www.heatmapping.org.

3. RESULTS AND DISCUSSION

3.1. DNN implementation and Comparison with typical classification methods

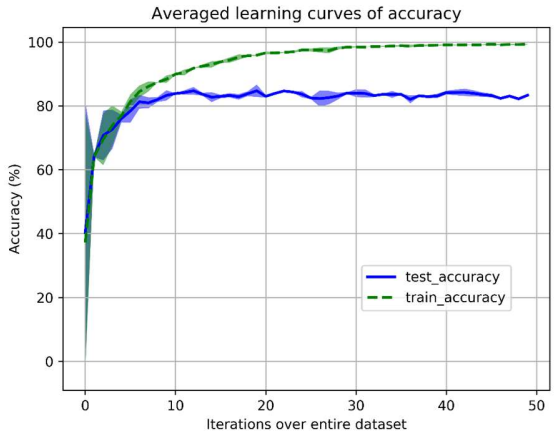

Fig. 2. The learning curves of accuracy from the training and testing data.

The DNN model with multiple hidden layers are trained using a standard error back propagation algorithm using batches of 16 randomly drawn training samples. To simplify our experiments, we did not use any unsupervised pre-training even though we expect that it will help. The weights of each layer are initialized randomly with zero-mean and unit-variance. What's more, the weights are controlled with $L1/L2$ norm regularization to further improve the classification performance [2]. We tested 1-5 hidden layers and the results showed that using 3 hidden layers could obtain the best performance. After performing grid search runs, we obtain the best architecture of the DNN model, which consisted of 3 hidden layers and the number of hidden units in each hidden layer was $[50,50,50]$. The constants β, γ, $dropout$ rate are hyper-parameters whose values are determined using a validation set. The β is fixed to 1 and the γ is fixed to 10. The learning rate is initially set as 0.001. To overcome the overfitting problem, the $dropout$ rate was fixed to 0.5. The learning curves of accuracy from the training and test data are shown in Fig. 2.

To investigate the effectiveness of the proposed method, we compared the proposed method with several state-of-art methods. To avoid possible bias caused by certain partitioning of training and test samples, 10-folds cross validation which is known to be an unbiased estimator of the generalization performance of a classifier is performed. The data collected from each hospital were also split into 10 folds to balance the data. FNC features are used as the input of different machine learning methods. Table II shows the classification results achieved by four methods, including SVMRFE[19], Random Forest[20], AdaBoost classifier[21]. We also reshaped the 1225 features into 35*35 features to test the performance of convolutional neural network (CNN). The values in Table II are the averaged results of the 10-fold cross validation. We also compared the performance of DNN with other models one by one according to the cross-validation

Table II Performance comparison of several methods

Methods	ACC(%)	SEN(%)	SPEC(%)	F
DNN(proposed)	**84.75**	**86.68**	**82.79**	**0.85**
SVMRFE [19]	77.09	76.36	77.85	0.77
RandomForests[20]	76.81	77.05	76.81	0.77
AdaBoost	70.98	71.31	70.63	0.71
CNN	78.63	75.24	82.21	0.77

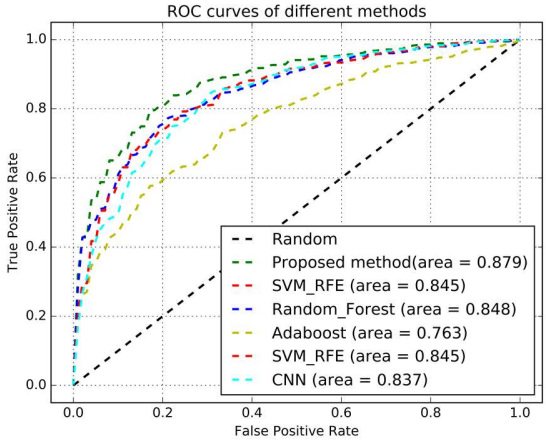

Fig. 3. Receive operating characteristic curves of different methods.

results. The result show that the performance of DNN is significantly better than other methods (P<0.01).

The classification accuracy, along with the sensitivity and specificity rates are then computed based on how many correct predictions are made with all the folds summed up against the sample size. In addition, we plot the ROC curves of these four methods in Fig. 3. In statistics, a receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The top left corner of the ROC plot is the "ideal" point - a false positive rate of zero and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better. As can be seen from Table 2 and Fig. 3, for the binary classification problem, the proposed DNN method outperforms SVMRFE, Random Forests, AdaBoost and CNN in terms of classification accuracy, sensitivity and AUC measures.

TensorFlow (https://www.tensorflow.org) was used with the above parameters to train the DNN model. The DNN classifier with three hidden layers were trained with a system consists of Intel(R) Xeon(R) CPU (3.60GHz), 32GB DDR3, and TITAN X (Pascal) GPU (12G). As for a model's complexity, the DNN method requires more computational time and resources than the competing methods. However, the computational burden of our method is mostly involved in the computation during a training phase, which can be performed offline. In other words, the high computational burden or complexity affects only the training step, while the required computation for testing is only matrix-vector multiplication and simple nonlinear function operations.

3.2. Cross-site prediction

Table III Results of the classification models

Test Site	M/F	ACC	SEN	SPEC	F score
Site 1	98/93	80%	75%	85%	0.78
Site 2	60/83	80%	77%	85%	0.79
Site 3	102/81	84%	80%	87%	0.82
Site 4	69/49	81%	90%	75%	0.85
Site 5	55/90	79%	78%	80%	0.78
Site 6	89/82	82%	82%	83%	0.82
Site 7	69/80	85%	86%	83%	0.85
Total	542/558	82%	81%	83%	0.81

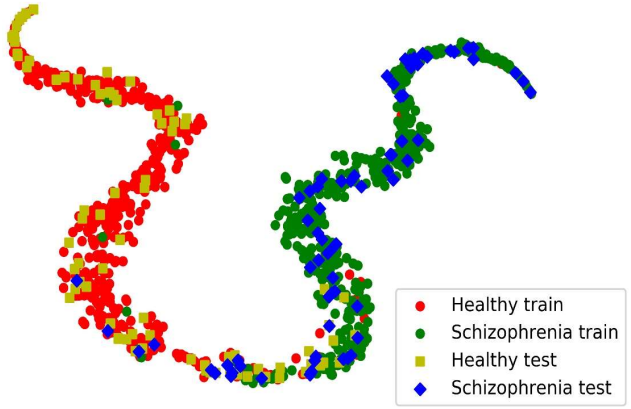

Fig. 4. Visualization of 1100 subjects of the last hidden layer by t-SNE. The color differentiates the class (patients and controls) and the training (Site 1-6: 951 subjects) from testing (Site 7: 149 subjects) data.

The multi-site sample included brain structural MRI scans from 542 HCs and 558 SZs, acquired in seven different hospitals and on different MRI scanners. To test whether the model we proposed is robust enough, we perform the cross-site validation by leaving out one site for testing each time and using the other six sites to train the model. This procedure goes through all seven sites, i.e. a leave-one-site-out cross-validation (See Table Ⅲ). Finally, we averaged results of seven sites prediction as the final cross-site prediction accuracy. The result indicates that the DNN model we proposed is generalizable enough to do multi-site prediction task. To visualize the performance of DNN classifier, we visualized the last hidden layer of DNN with tSNE method [22]. 6 sites (951 subjects) are used as training data and 1site(149 subjects) as testing data. The result is also displayed on a 2D map which indicates that the proposed DNN model can successfully distill details and pull the classes apart.

3.3. Feature selection with LRP

Most of the functional connectivity features are redundant to classification and only a small number of them are significantly relevant to the behavioral symbols of schizophrenia. We are interested in which features contributed most to group discrimination. The goal is achieved by exploiting the quantitative advantage of the LRP model, whereby features with the greatest absolute weight value could be factored out.

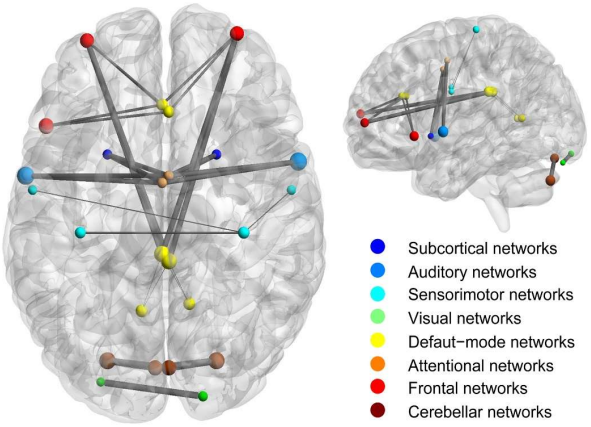

Fig. 5. Visualization of the top 30 significant features pertaining to SZ and HC classification

In the experiments, the procedure of extracting the most informative features is as follow: Firstly, each subject in test data set is used as the input of the trained DNN model. The outputs of the test subjects are figured out with forwarding propagation algorithm. According to the formula(3), for each subject, the DNN output value (softmax result) is used as the input of the LRP model to project the heat map. To eliminate the irrelevant features and select the most discriminative features, firstly, we selected the top 200 features of each subject according to its heat map value. After that, the frequency of each feature is counted. Finally, the top30 features are selected according to their frequency. Moreover, the selected features are visualized with BrainnetViewer (https://www.nitrc.org/projects/bnv/).

With the novel feature selection strategy, some functional connectivity between certain brain regions of the frontal network and subcortical network are found to exhibit the highest discriminative power. The regions have long been demonstrated important for execution, decision-making, and working memory, which are key components of evaluating the cognitive deficit.

4. CONCLUSION

In the present study, the DNN classifier model trained with $L1/L2$ norm regularization (*dropout* and *batch normalization*) demonstrated the feasibility of the DNN classifier toward the automated diagnosis of SZ patients by using resting-state FNC patterns as input patterns. The proposed DNN classifier significantly improved performance for the diagnosis of SZ patients from HC subjects relative to several conventional machine learning methods. The cross-site prediction

accuracy suggests high robustness and generalization performance of the proposed method. Moreover, to the best of our knowledge, this is the first attempt of combining DNN with LRP for brain disease classification using resting state fMRI data. We have provided a showcase of how LRP can add an explanatory layer to the highly effective technique of DNN in the fMRI domain. Our results show that LRP provided highly detailed accounts of relevant information in high-dimensional fMRI data that may be useful in analysis scenarios where single trials need to be considered individually.

5. ACKNOWLEDGEMENT

This work was supported by the National Key Basic research and Development Program (973, No. 2011CB707800), the National High-Tech Development Plan (863, No. 2015AA020513), "100 Talents Plan" of Chinese Academy of Sciences, the Chinese National Science Foundation Number 81471367, the Strategic Priority Research Program of the Chinese Academy of Sciences (grant No. XDB02060005). The authors report no biomedical financial interests or potential conflicts of interest.

6. REFERENCES

[1] W. H. L. P. Sandra Vieira, Andrea Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience and Biobehavioral Reviews*, 2017.

[2] J. Kim, V. D. Calhoun, E. Shim, and J. H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia," *Neuroimage,* vol. 124, no. Pt A, pp. 127-46, Jan 01, 2015.

[3] R. N. Boubela, K. Kalcher, W. Huf, E. M. Seidel, B. Derntl, L. Pezawas, C. Nasel, and E. Moser, "fMRI measurements of amygdala activation are confounded by stimulus correlated signal fluctuation in nearby veins draining distant brain regions," *Sci Rep,* vol. 5, pp. 10499, May 21, 2015.

[4] G. H. Turner, and D. B. Twieg, "Study of temporal stationarity and spatial consistency of fMRI noise using independent component analysis," *IEEE Transactions on Medical Imaging,* vol. 24, no. 6, pp. 712-718, 2005.

[5] M. R. Arbabshirani, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun, "Classification of schizophrenia patients based on resting-state functional network connectivity," *Frontiers in Neuroscience,* vol. 7, pp. 133, 07/30, 2013.

[6] H. I. Suk, S. W. Lee, D. Shen, and I. Alzheimer's Disease Neuroimaging, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *Neuroimage,* vol. 101, pp. 569-82, Nov 01, 2014.

[7] H. I. Suk, C. Y. Wee, S. W. Lee, and D. Shen, "State-space model with deep learning for functional dynamics estimation in resting-state fMRI," *Neuroimage,* vol. 129, pp. 292-307, Apr 01, 2016.

[8] H. G. Heather Cody Hazlett, Brent C. Munsell, Dennis W. Shaw, Lonnie Zwaigenbaum, "Early brain development in infants at high risk for autism spectrum disorder," *Nature,* vol. 542, pp. 3, 2017.

[9] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience and Biobehavioral Reviews,* vol. 74, pp. 58-75, Mar 2017.

[10] D. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems.* pp. 1097-1105.

[12] C. Wachinger, M. Reuter, A. D. N. Initia, and A. I. B. Life, "Domain adaptation for Alzheimer's disease diagnostics," *Neuroimage,* vol. 139, pp. 470-479, Oct 1, 2016.

[13] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics,* vol. 12, no. 2, pp. 229-44, Apr 2014.

[14] M. Nieuwenhuis, H. G. Schnack, N. E. van Haren, M. S. Schaufelberger, and P. Dazzan, "Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients," *Neuroimage,* vol. 145, no. Pt B, pp. 246-253, Jan 15, 2017.

[15] G. E. Hinton, and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, no. 5786, pp. 504-507, Jul 28, 2006.

[16] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Muller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One,* vol. 10, no. 7, pp. e0130140, 2015.

[17] I. Sturm, S. Lapuschkin, W. Samek, and K. R. Muller, "Interpretable deep neural networks for single-trial EEG classification," *J Neurosci Methods,* vol. 274, pp. 141-145, Dec 01, 2016.

[18] E. A. Allen, E. B. Erhardt, E. Damaraju, W. Gruner, J. M. Segall, R. F. Silva, G. D. Pearlson, J. P. Phillips, J. R. Sadek, M. Stevens, U. Teuscher, R. J. Thoma, and V. D. Calhoun, "A baseline for the multivariate comparison of resting-state networks," *Front Syst Neurosci,* vol. 5, pp. 2, 2011.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning,* vol. 46, no. 1, pp. 389-422, 2002.

[20] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[21] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Statistics and Its Interface,* vol. 2, no. 3, pp. 349-360, 2009.

[22] G. H. Laurens van der Maaten, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 9(Nov), 2008.