DISCRIMINATION OF ADHD CHILDREN BASED ON DEEP BAYESIAN NETWORK

A. Junyu Hao, B. Lianghua He

Tongji University, Department of Computer Science and Technology

Keywords: ADHD, Deep Learning, Bayesian Network, SVM, Deep Belief Network

ABSTRACT

Attention deficit hyperactivity disorder (ADHD) is a threat for the public health all the time, so the effective discrimination of it is significant and meaningful. In current research, Functional Magnetic Resonance Imaging (fMRI) data has become a popular tool for the analysis of ADHD. In this paper, we introduce the Deep Bayesian Network, a combination of Deep Belief Network and Bayesian Network, to classify the ADHD children from the normal. In Deep Bayesian Network, The Deep Belief Network is applied to normalize and reduce dimension of the fMRI data in every brodmann area. And the Bayesian Network is used to extract the feature of relationships between several well-performed brain areas by structure learning. According to the information of structure and probability in Bayesian Network, we predicted the subjects as control,combined ,inattentive or hyperactive using SVM classifier. The final results perform better than using single Deep Belief Network and the best results in ADHD-200 competition.

1. INTRODUCTION

Attention deficit hyperactivity disorder(ADHD) is among the most common psychiatric disorders of childhood that persists into adulthood in the majority of cases[1]. According to American Psychiatric Association's Diagnostic and Statistical Manual, the prevalence of ADHD in the whole world is approximately 5%, especially in the United States the prevalence among 8 to 15-years-olds reaches to 8.7% during the past years. Therefore, the methods of diagnosing the ADHD are in urgent need.

As an excellent method of measuring brain activation, fMRI signals[2] are studied to classify ADHD. Shengfu Liang et al.[3] utilizes the LDA classifier to discriminate ADHD by analyzing the rs-fMRI data[4], and the average accuracy of distinguishing normal and ADHD children reaches 80.08% through 50 times of 2-fold validation. Xunheng Wang[5] applies Kernel Principal Component Analysis (KPCA) method based on connectivity matrix of each functional meaningful brain region to find the abnormal pattern of ADHD. Then Support Vector Machine (SVM) as a classifier increases the accuracy rate to 81% using a leave-one-out cross validation. In the global ADHD-200 competition, Eloyan A et al.[6] has achieved relatively better scores by using rs-fMRI based on decomposition of CUR along with gradient boosting. In this paper, three datasets from ADHD-200 competition are applied to discriminate ADHD.

However, it will encounter a serious problem during analyzing the ADHD with fMRI data, which are the mass of data and data redundancy. Therefore, dimensionality reduction is essential. Although there are many dimensionality reduction methods proposed, but they only take mathematical requirements into consideration, rather than physical requirements. Luckily, Geoffrey E.Hinton et al.[7] derives a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory by using complementary prior. Because of the good performance in dimensionality reduction currently[8], Deep Learning has been applied for many areas including image processing[7][9], audio classification[10], natural language processing[11] and so on. Considering its powerful learning ability and advantage of dimensionality reduction and normalization, Deep Learning method will be used to analysis massive fMRI data to dig out the cognitive significance of brain in this paper.

Since the cause of ADHD is unknown, one should take into consideration the relationships between different areas, rather than analysis the brain area separately, which meets the requirements of Bayesian network. It is a graphical model that can encode probabilistic relationships among variables of interest. On the one hand, the model can be used to learn causal relationships and gain understanding about a problem domain. For example, in bioinformatics Bayesian network has been used for the interpretation and discovery of gene regulatory pathways[12]. On the other hand, it is an ideal representation for combining prior knowledge and data. Based on these, Bayesian network has been used in information retrieval[13], natural language processing[14] ,and for the analysis of a medical service's performance for management decisions[15]. In this paper, Bayesian network is applied to retrieval the information between different brain areas, which is shown by the edges in the graph of the Bayesian network.

Thanks to National Natural Sciences Foundation of China.(No.61272267, 61270220, 51075306,61273261)

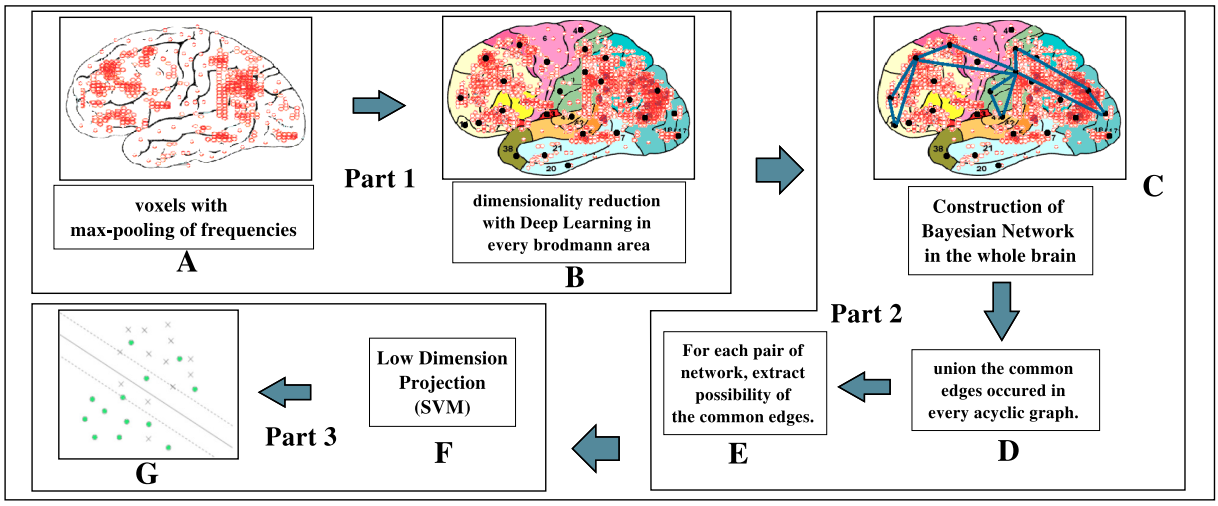


Fig. 1. The procedure of Deep Bayesian Network

Figure 1 shows the framework of proposed method. It can be seen that there are totally four steps. The first step is that pre-process the voxels with max-pooling of frequencies(**A** in Fig.1The procedure of Deep Bayesian Networkfigure.1). Then, Deep Belief Network is used to reduce dimensionality of data in every brodmann area(**B** in Fig.1The procedure of Deep Bayesian Networkfigure.1), which reconstructs the different number of voxels in every areas into the same number of features, and the Bayesian network can extract the relationships with the normalized fMRI data(**C-E** in Fig.1The procedure of Deep Bayesian Networkfigure.1). Finally, using SVM as a classifier to discriminate the ADHD children from normal(**F-G** in Fig.1The procedure of Deep Bayesian Networkfigure.1). The main contribution of our work is to propose a novel method which combine the dimension-reduced characteristic of Deep Belief Network with the global feature-extracted characteristic of Bayesian Network. Based on this two characteristics, Deep Belief Network can deal with the huge amount of datas, such as fMRI data. In addition, the features extracted are global, which can represent the whole brain.

The rest of this paper is organized as follows. Section 2 mainly describes the Deep Belief Network method, Bayesian network and the application of two graph model for ADHD data. Section 3 presents our experiments and results on ADHD dataset.

2. METHOD

Due to Deep Belief Network and Bayesian Network are two key parts of Deep Bayesian Network we proposed, we will introduce the two methods to you in brief.

2.1. Deep Belief Network

Deep Belief Network is made of a stack of restricted Boltzmann machines(RBM)[16] which is a two-layer, undirected, bipartite graphical model including visible units v and hidden units h.

The construction of Deep Belief Network includes two steps, which are pretraining and back propagation. Pretraining is a down-up unsupervised feature learning. It uses unlabelled data as the input of first RBM layer. After the training in the first layer, the output will act as the input of next layer. Pretraining RBM network make sure that the feature can be mapped into different feature space and keep more feature information at the same time. In each layer, the weights W and biases a, b of the RBM determine the energy of a joint configuration of the hidden and visible units $E(v, h)$,

$$E(v,h;\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} v_i h_j w_{ij} - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j \quad (1)$$

with model parameters $\theta = \{W, b, a\}$ and $v_i, h_j \in \{0, 1\}$,V, H represents the number of visible units and hidden units.

The second step is back propagation, which is a top-down supervised learning. It uses the labelled data to fine-tune the whole network, which spreads the error message Δw_{ij} to every level of RBM from the top to down.

$$\Delta w_{ij} = \epsilon(< v_i, h_j >_{data} - < v_i, h_j >_{recon}) \quad (2)$$

Where ϵ is the learning rate and the angle brackets manifests the expectations relative to the distribution specified in the subscript. This procedure is viewed as a initialization of parameter weights W, which overcome the shortcomings of

local optimum and long train time due to random initialization of parameter weights. compared with RBM, Convolutional Netural Network(CNN) is more popular in field of image recognition[17][18]. Considering the specificity of fMRI, which is a 4D time series and huge amount of data, we apply a generalized technology deep learning based on RBM other than Convolutional Netural Network(CNN) and autocode.

2.2. Bayesian Network

A Bayesian network $B =< N, A, \Theta >$ is a directed acyclic graph (DAG) $< N, A >$ with a conditional probability table (CPT) for each node, collectively represented by Θ. Each node $n \in N$ represents a domain variable x_i, and each arc $a \in A$ between nodes represents a probabilistic dependency $p(x_i|x_j)$. The joint probability $p(X)$ can be calculated as follows:

$$p(X) = \prod_{i=1}^{m} p(x_i|\pi_i)) \tag{3}$$

where π_i denotes the parent $i = 1, \cdots, m$ nodes of x_i. There are two major tasks in learning a BN: structure learning and parameters learning. Currently, there are three methods to learn the structure of Bayesian network, which are method based on searching and scoring, method based on testing independence and the mix of two. The third method performs well, especially for Max-Min Hill Climbing(MMHC) algorithm. This paper mainly apply Max-Min Hill Climbing (MMHC) algorithm to learn structure efficiently, which can reduce the searching spaces and improve the learning efficiency of the whole algorithm.

2.3. The Construction of Deep Bayesian Network

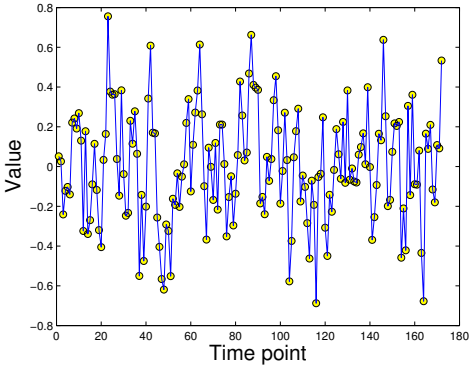
2.3.1. Original Idea

For all our experiments we used the preprocessed rs-fMRI data released for the competition. All the fMRI data volumes are of size $49 \times 58 \times 47$ voxels, but the number of samples across time varies among the data capturing centers. For example, there are 172 samples in NYU and 148 samples in KKI. For this huge 4-D fMRI data, large number of calculation is absolutely necessary, which will cause intolerable time consuming. Fortunately, Brodmann divided the cerebral cortex into 52 areas according the brain's cognitive function[19]. In fact, only 41 brodmann area are used considering some brodmann area possess few voxels. Based on this, we do some research on each brodmann area in order to reduce dimensionality the fMRI data. What is worth mentioning is that variable number in each brodmann area will be normalize to the same size, which achieves the effect of normalization. Deep Belief Network is exact suitable to this. But the performance in single brodmann area is unsteady. some of brodmann areas perform well, such as the areas of 10, 18, 30 etc and some of brodmann areas perform bad, such as the areas of 22, 26, 36, 38 etc. In addition, due to the huge fMRI data, we cann't deal with whole brain data with Deep Belief Network exactly. So we select the performed-well brodmann areas to construct Bayesian Network, to explore the differences in the whole brain between ADHD and normal person.

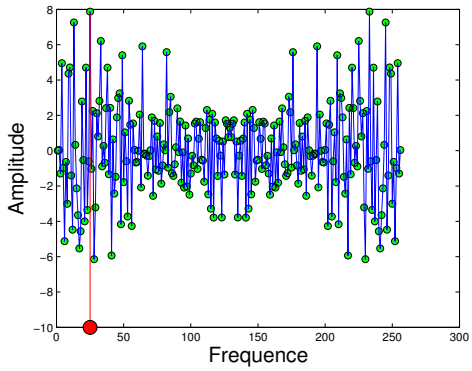
2.3.2. Deep Feature Extraction

In this paper, the original fMRI data D is a 4-D time series, here we unroll it as this $D = [D_1, \cdots, D_{49\times58\times47}], D_i = [d_1, \cdots, d_n], d_i \in R$, n is decided by the data capturing centers. The data D_i is produced by each voxel, which is show in Figure 2(a)Subfigure 2(a)subfigure.2.1. The frequency information is quite significant[20], which can be used as the characteristics of each voxel by the study in our research group. So we select the frequence with the max amplitude as the feature of each voxel by Fast Fourier Transform, which is show in Figure 2(b)Subfigure 2(b)subfigure.2.2.

$$x_i = max(fft(D_i)) \tag{4}$$



(a) Value information in each voxel



(b) Frequence information in each voxel

Fig. 2. Preprocess in each voxel

After that, we divide the dealt data $X = [x_1, \cdots, x_{49\times58\times47}]$ into 41 subsets $X = [X_{m_1}, \cdots, X_{m_41}]$

according to Brodmann area which is the input of Deep Belief Network.

To construct the Deep Belief Network, the forward calculation and backward propagation are needed in every layer.

$$[\theta, X] = RBM(X) \quad (5)$$

$$\theta = backprop(\theta, X) \quad (6)$$

The pseudo-code for application of Deep Belief Network on ADHD data can be seen as following Algorithm 1Deep Feature Extractionalgorithm.1 .

Algorithm 1 Deep Feature Extraction

Input: original fMRI data D

1: Get frequencies with max amplitude in every voxel x_i

$$x_i \leftarrow max(fft(D_i))$$

2: Pre-process frequency information: $X \leftarrow prepro(X_f)$

3: **for** $k \leftarrow 1, 3$ **do**

4: train the weights $\theta = [w, a, b]$ of single RBM

$$[\theta, X] \leftarrow RBM(X)$$

5: **end for**

6: Adjust weights with Backpropagation

$$\theta \leftarrow backprop(\theta, X)$$

7: Reduce dimensionality: $Y \leftarrow DBN(\theta, X)$

Output: dimensionality-reduced data Y

2.3.3. Structure Learning of Deep Bayesian Network

Due to the reason of ADHD is unknown, it is meaningful to analysis the dimensionality-reduced data Y and sample label L in the whole brain with Bayesian Network. First, filter the data Y to wipe out the noise. Second, the MMHC algorithm can speed up the structure of Bayesian network, but it needs the limited parent nodes of each node. So the dependent nodes I are computed by conditional independence testing in filtered data Y_f.

$$I = Indep(Y_f) \quad (7)$$

Third, we union the limited parent nodes U and use it to learn the structure of all data including training data and testing data. By this, we can get a DAG G and a table P including the information of probability of edge shown in equation below.

$$[G, P] = BN(U, Y_f) \quad (8)$$

Forth, the probability of each edge in Bayesian network will be extracted out and viewed as feature F of ADHD children and normal children.

Finally, SVM will be acted as classifier to train the training samples and classify the testing samples, which will get the classified label of testing sample C. We choose to use the SVM classifiers for the following reasons. First, the SVM can classify the data points from two classes, which are not easily separable in the feature space, by using a kernel trick to project the data points into a hyperspace where the separation is easy. Second, the SVM regresses the feature space without over fitting on the data by allowing miss classification with a penalty. The accuracy T can be computed by predicted labels L and true labels Y. The pseudo-code for application of BN on ADHD data can be seen as following Algorithm 2Structure Learning of Deep Bayesian Networkalgorithm.2.

Algorithm 2 Structure Learning of Deep Bayesian Network

Input: dimensionality-reduced data Y, labels of data L

1: Filter the data: $X_f \leftarrow filter(Y)$

2: Test Conditional independence

$$I \leftarrow Indep(X_f)$$

3: union the parent's limitation: $U \leftarrow union(I)$

4: Learning structure of Bayesian Network

$$[G, P] \leftarrow BN(U, Y_f)$$

5: Extract feature: $F \leftarrow extract(G, P)$

6: Classify the features: $C \leftarrow svm(F, L)$

7: Computer accuracy: $T \leftarrow com(L, C)$

Output: accuracy of discrimination T

3. EXPERIMENTS AND RESULTS

3.1. Data

The data, provided by Neuro Bureau for the ADHD 200 competition, is used for our study. There are eight different centers contributing to the compilation of the whole data set, which makes it diverse as well as complex. In total it consists of 776 training and 197 test subjects. Besides the basic fMRI data, different phenotypic information, such as age, gender, handedness, IQ, is also provided for each subject.

The experimental validations of our proposed method are performed on the training and test data sets of 3 of the data centers - New York University(NYU), Peking University(Peking-1) and Kennedy Krieger Institute(KKI). Consider Table 1 for an overview of the data userd in our study. For NYU, the training subjects are 216, and testing subjects are 41; for Peking-1 dataset, the training subjects are 85, and testing subjects are 50; and for KKI the training subjects and testing subjects are 83 and 11 respectively.

In compliance with the Health Insurance Portability and Accountability Act(HIPAA) privacy rules, all data used for the experiments of this article are fully anonymized.

Table 1. Demographic Information of three Datasets

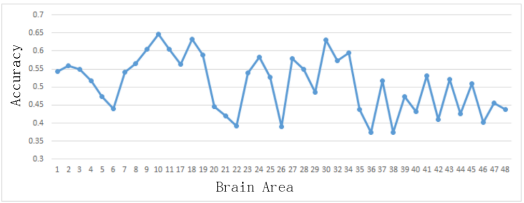
type	NYU		Peking-1		KKI	
	train 216	test 41	train 85	test 50	train 83	test 11
control	98	12	61	27	61	8
combined	73	22	7	9	16	3
inattentive	2	0	0	1	5	0
hyperactive	43	7	17	13	1	0

3.2. The Experiment of Parameter Design

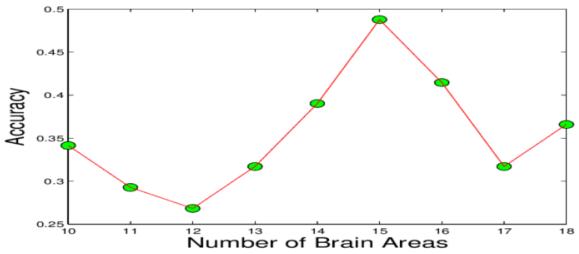
To construct a Deep Bayesian Network, we must set up the number of network nodes, which is equal to selected brodmann areas. If the value of number is small, Bayesian network will not extract the enough information of relationships between brain areas. If the value of number is too large, the probability that the unrelated brain areas will be chose as a node will get higher, which has a side effect on the final classification. In addition, with the increase of the variables, the search space of network structure will present exponentially. So the experiment of parameter designing is significant.

3.2.1. Parameter Design

As NYU dataset in the ADHD-200 competition achieved the lowest discrimination results, Deep Belief Network is particularly tested on the NYU dataset of 41 areas and softmax as a classifier generates the accuracies of different brain areas,which is convenient for the choice of node in the Bayesian network. The results of 41 regions are shown in Figure 3(a)Subfigure 3(a)subfigure.3.1.



(a) Accuracy of NYU in 41 Areas



(b) Accuracy in Different Number of Brain Areas

Fig. 3. Parameter Design in NYU dataset

From Figure 3(a)Subfigure 3(a)subfigure.3.1, we can see that different brain areas have different performance of discrimination. The areas of 10, 18, 30, 9, 11, 19, 34, 32, 8, 17, 2, 28, 3, 1, 7, 23, 25, 41 perform well. According to brodmann definition, it is clear that prefrontal cortex(9,10,25), visual cortex(8,17,18,19), somatosensory cortex(1,2,3,7) and cingulate cortex(23,30,32) is related to the ADHD closely. Therefore, they are selected as the input of Bayesian Network. The Figure 3(b)Subfigure 3(b)subfigure.3.2 shows that the accuracy of discriminate fluctuate a little from 10 to 18, but it reaches the peak when the number of brain areas is equal 15. So we will choose 15 brain areas in the front to construct the Bayesian network. Besides, using the relationships between brain areas to discriminate ADHD is better than using information in single brain area and the best result in ADHD-200 competition.

3.3. Performances on NYU, Peking-1 and KKI dataset

The same experiments are excuted on the NYU, Peking-1 and KKI dataset. The results released by ADHD-200 competition are 35.19% for NYU, 51.05% for Peking-1 and 61.90% for KKI respectively. The prediction accuracies of Attributed graph distance[21] (AGD) are none, 58.82%,54.55%. The Deep Bayesian Network gains a higher prediction accuracies than Deep Belief Network single, which are 48.78% for NYU, 54.00% for Peking-1 and 72.72% for KKI. The details show in Figure 4Performance in Different Datasetsfigure.4.

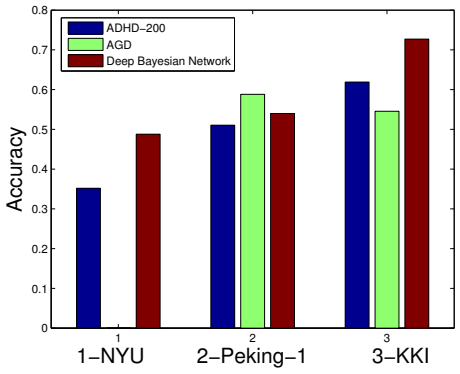


Fig. 4. Performance in Different Datasets

From this chart, we can see that Deep Bayesian Network improves the prediction accuracies in these three datasets compared with the results of ADHD-200 competition. Besides, the increase of accuracy is the highest in the NYU dataset than other two datasets. The different number of training samples have a big effect on the accuracy. This effect is shown in Figure 5Performance in Different Number of training samplesfigure.5.

To verify generality of our method, we make another exper-

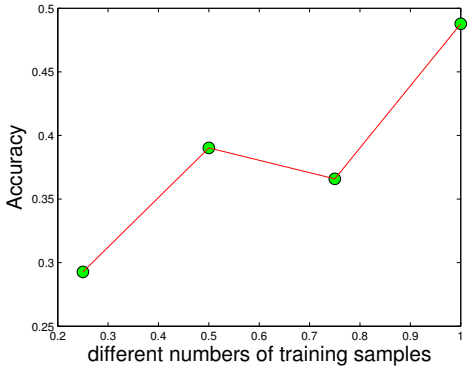


Fig. 5. Performance in Different Number of training samples

iment, which is described as follow. First, we mix the training dataset and the testing dataset in the ADHD-200 competition. Second, we randomly select same number of training dataset in the competition. Third, the left dataset will be acted as testing dataset to check the performance by Deep Bayesian Network method. we repeat the three steps above for 100 times, the detail information is show in Tab 2. Considering that discriminating the ADHD is important and meaningful, we take the prediction accuracy of the method along with the specificity and sensitivity values into consideration.

Table 2. The Detail information in Different Datasets

Dataset	Accuracy	Specificity	Sensitivity
NYU	64.7	68.8	43.9
Peking-1	66.3	87.7	22.9
KKI	59.0	83.0	55.6

Although the data in Tab 2 have no comparison with the results in the ADHD-200 competition, the generality of our method is good.

4. CONCLUSION

In this paper, a novel method called Deep Bayesian Network, is proposed to classify fMRI ADHD image data. Because of the combination Deep Belief network and Bayeisan network, Deep Bayesian Network can compute relationships among brodmann brain areas more effectively. A series of experimental results also prove that Deep Bayesian network improves the classification performatnce of ADHD greatly comparing with the ADHD-200 competition results.

In the future, we could find more effective method to choose the different brain areas as the input of Bayesian Network. In addition, the distance between different brain areas is different, so they have different influence over each other. we could take the area's neighborhood into consideration manually to strengthen the structure learning.

5. ACKNOWLEDGMENT

This work was supported by National Natural Sciences Foundation of China (No.61272267, 61270220, 51075306, 61273261), Program for New Century Excellent Talents in University (NCET-11-0381), Fundamental Research Funds for the Central Universities, State Key Laboratory of Software Engineering.

6. REFERENCES

[1] S. J. Kooij, S. Bejerot, A. Blackwell, H. Caci, M. Casas-Brugué, P. J. Carpentier, D. Edvinsson, J. Fayyad, K. Foeken, M. Fitzgerald, *et al.*, "European consensus statement on diagnosis and treatment of adult adhd: The european network adult adhd," *BMC psychiatry*, vol. 10, no. 1, p. 67, 2010.

[2] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional magnetic resonance imaging*, vol. 1. Sinauer Associates Sunderland, MA, 2004.

[3] S.-F. Liang, T.-H. Hsieh, P.-T. Chen, M.-L. Wu, C.-C. Kung, C.-Y. Lin, and F.-Z. Shaw, "Differentiation between resting-state fmri data from adhd and normal subjects: Based on functional connectivity and machine learning," in *Fuzzy Theory and it's Applications (iFUZZY), 2012 International Conference on*, pp. 294–298, IEEE, 2012.

[4] M. P. Milham, "Open neuroscience solutions for the connectome-wide association era," *Neuron*, vol. 73, no. 2, pp. 214–218, 2012.

[5] X. Wang, Y. Jiao, and Z. Lu, "Discriminative analysis of resting-state brain functional connectivity patterns of attention-deficit hyperactivity disorder using kernel principal component analysis," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, vol. 3.

[6] A. Eloyan, J. Muschelli, M. B. Nebel, H. Liu, F. Han, T. Zhao, A. D. Barber, S. Joel, J. J. Pekar, S. H. Mostofsky, *et al.*, "Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging," *Frontiers in systems neuroscience*, vol. 6, 2012.

[7] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

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

[9] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," *Communications of the ACM*, vol. 54, no. 10, pp. 95–103, 2011.

[10] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, pp. 1096–1104, 2009.

[11] R. Sarikaya, G. E. Hinton, and A. Deoras, "Application of deep belief networks for natural language understanding.," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 4, pp. 778–784, 2014.

[12] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[13] R. Baeza-Yates, B. Ribeiro-Neto, *et al.*, *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[14] W. W. Chapman, M. Fizman, B. E. Chapman, and P. J. Haug, "A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia," *Journal of biomedical informatics*, vol. 34, no. 1, pp. 4–14, 2001.

[15] S. Acid, L. M. de Campos, J. M. Fernández-Luna, S. Rodrıguez, J. Marı?a Rodrı?guez, and J. Luis Salcedo, "A comparison of learning algorithms for bayesian networks: a case study based on data from an emergency medical service," *Artificial intelligence in medicine*, vol. 30, no. 3, pp. 215–232, 2004.

[16] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[18] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 98–113, 1997.

[19] K. Brodmann and L. J. Garey, *Brodmann's: Localisation in the Cerebral Cortex*. Springer Science & Business Media, 2007.

[20] A. M. Smith, B. K. Lewis, U. E. Ruttimann, Q. Y. Frank, T. M. Sinnwell, Y. Yang, J. H. Duyn, and J. A. Frank, "Investigation of low frequency drift in fmri signal," *Neuroimage*, vol. 9, no. 5, pp. 526–533, 1999.

[21] S. M. Dey S, Rao A R, *Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects*, vol. 8: 64. Frontiers in Neural Circuits, 2014.