EBioMedicine 30 (2018) 74–85

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI

Ling-Li Zeng [a,1], Huaning Wang [b,1], Panpan Hu [c,1], Bo Yang [a,1], Weidan Pu [d], Hui Shen [a], Xingui Chen [c], Zhening Liu [e], Hong Yin [f], Qingrong Tan [b,***], Kai Wang [c,**], Dewen Hu [a,*]

[a] College of Mechatronics and Automation, National University of Defense Technology, Changsha, China
[b] Department of Psychiatry, Xijing Hospital, Fourth Military Medical University, Xi'an, China
[c] Department of Neurology, The First Affiliated Hospital of Anhui Medical University, Hefei, China
[d] Medical Psychological Center, Second Xiangya Hospital, Central South University, Changsha, China
[e] Mental Health Institute, Second Xiangya Hospital, Central South University, Changsha, China
[f] Department of Radiology, Xijing Hospital, Fourth Military Medical University, Xi'an, China

ARTICLE INFO

Article history:
Received 24 January 2018
Received in revised form 6 March 2018
Accepted 16 March 2018
Available online 23 March 2018

Keywords:
Schizophrenia
Deep learning
Connectome
fMRI
Striatum
Cerebellum

ABSTRACT

Background: A lack of a sufficiently large sample at single sites causes poor generalizability in automatic diagnosis classification of heterogeneous psychiatric disorders such as schizophrenia based on brain imaging scans. Advanced deep learning methods may be capable of learning subtle hidden patterns from high dimensional imaging data, overcome potential site-related variation, and achieve reproducible cross-site classification. However, deep learning-based cross-site transfer classification, despite less imaging site-specificity and more generalizability of diagnostic models, has not been investigated in schizophrenia.

Methods: A large multi-site functional MRI sample ($n = 734$, including 357 schizophrenic patients from seven imaging resources) was collected, and a deep discriminant autoencoder network, aimed at learning imaging site-shared functional connectivity features, was developed to discriminate schizophrenic individuals from healthy controls.

Findings: Accuracies of approximately 85·0% and 81·0% were obtained in multi-site pooling classification and leave-site-out transfer classification, respectively. The learned functional connectivity features revealed dysregulation of the cortical-striatal-cerebellar circuit in schizophrenia, and the most discriminating functional connections were primarily located within and across the default, salience, and control networks.

Interpretation: The findings imply that dysfunctional integration of the cortical-striatal-cerebellar circuit across the default, salience, and control networks may play an important role in the "disconnectivity" model underlying the pathophysiology of schizophrenia. The proposed discriminant deep learning method may be capable of learning reliable connectome patterns and help in understanding the pathophysiology and achieving accurate prediction of schizophrenia across multiple independent imaging sites.

1. Introduction

Complex and heterogeneous symptoms with impairments in multiple cognitive domains, including perception, memory, attention, and executive function, and other negative symptoms pose a challenge to the objective diagnosis of schizophrenia based solely on clinical manifestations (APA, 2013). Searching for reliable biomarkers for the diagnosis and treatment of schizophrenia is clearly an international imperative.

The pathophysiology of schizophrenia has been proposed to be associated with the dysfunctional integration of distributed neuronal networks, giving rise to the concept of "widespread disconnectivity" in schizophrenia (Andreasen et al., 1999; Cheng et al., 2015b; Friston and Frith, 1995). Shen et al. (2010) used whole-brain functional connectivity MRI (fcMRI) pattern analysis to discriminate schizophrenic patients from healthy controls, demonstrating the potential of whole-brain fcMRI in the diagnosis of schizophrenia. Subsequently, a number of

* Correspondence to: D. Hu College of Mechatronics and Automation, National University of Defense Technology, 109 Deya Road, Changsha, Hunan 410073, China.
** Correspondence to: K. Wang Department of Neurology, The First Affiliated Hospital of Anhui Medical University, Hefei, Anhui 230022, China.
*** Correspondence to: Q. Tan Department of Psychiatry, Xijing Hospital, Fourth Military Medical University, Xi'an, Shaanxi 710032, China.
E-mail addresses: tanqingr@fmmu.edu.cn (Q. Tan), wangkai1964@126.com (K. Wang), dwhu@nudt.edu.cn (D. Hu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ebiom.2018.03.017

neuroimaging studies have shown progress in probing connectome-based biomarkers of schizophrenia (Arbabshirani et al., 2013; Cheng et al., 2015a; Kim et al., 2016; Mikolas et al., 2016) (see ref. (Arbabshirani et al., 2017) for a review). Cheng et al. (2015b) obtained an overall accuracy of 75.81% in the multi-site pooling classification based on voxel-based functional connectivity. Moreover, some previous studies have made an attempt at leave-site-out transfer classification of schizophrenia using MRI, yielding average accuracies of around 75.0% (Rozycki et al., 2017; Skåtun et al., 2017). In the leave-site-out transfer classification, a sample of a given imaging site was left for prediction and a sample of multiple independent sites was used for model training. Such transfer classification may be of more significance in clinical practice because when using multi-site data during model training, the final neuroimaging-based diagnostic classification models are much less imaging site-specific and should therefore be more generalizable. However, most studies suffered from either a small sample size or modest classification performance.

Recently, deep learning has attracted increasing attention in the field of machine learning and artificial intelligence and has been demonstrated to prodigiously improve learning performance in computer vision and image recognition (Lecun et al., 2015; Sun et al., 2013). Kim and colleagues used a deep neural network with weight sparsity control for whole-brain fcMRI classification of schizophrenia patients vs. healthy controls with a small sample size ($n = 100$) (Kim et al., 2016), illuminating the potential of deep learning in automatic diagnosis of clinical populations(Hazlett et al., 2017; Kawahara et al., 2017; Suk et al., 2013; Zhao et al., 2017). Furthermore, deep learning is capable of learning subtle hidden patterns from high dimensional neuroimaging data, perhaps providing cues for understanding the neural basis of neuropsychiatric disorders (Arbabshirani et al., 2017; Guo et al., 2017; Vieira et al., 2017). So far, the potential of deep learning of whole-brain fcMRI both in multi-site and cross-site classification of schizophrenia remains unknown.

In this study, we first collected the largest multi-site fcMRI raw dataset reported to date in the schizophrenia literature, including 1000+ participants (474 schizophrenic patients), and we developed a deep Discriminant Autoencoder Network with Sparsity constraint (DANS) for the automatic diagnosis of individuals with schizophrenia. In addition to multi-site pooling classification based on the pooled data of all sites, we conducted leave-site-out transfer classification, training learning models independently on imaging sites.

2. Materials and Methods

2.1. Participants

The dataset includes 474 patients with schizophrenia and 607 healthy controls from seven imaging resources. The first and second sub-datasets were both collected at Xijing Hospital in China (Xijing#1: 107 schizophrenic patients and 113 healthy controls; 49 of the patients were drug-naïve, whereas the remainder were receiving antipsychotic medications at the time of image acquisition; Xijing#2: 54 patients and 102 healthy controls); the third sub-dataset was from the First Affiliated Hospital of Anhui Medical University in China (AMU: 105 patients and 101 healthy controls); the fourth sub-dataset was collected at Second Xiangya Hospital in China (Xiangya: 56 schizophrenic patients and 42 healthy controls; 11 of the patients were drug-naïve, whereas the remainder were receiving antipsychotic medications); the fifth sub-dataset was from the Center for Biomedical Research Excellence (COBRE: 71 schizophrenic patients and 74 healthy controls; the patients were all receiving various antipsychotic medications, available at https://openfmri.org/); the sixth sub-dataset was from the University of California, Los Angles (UCLA: 58 schizophrenic patients and 134 healthy controls; stable medications were permitted for the patients, available at https://openfmri.org/) (Poldrack et al., 2016); and the last sub-dataset was from the Conte Center for the Neuroscience of Mental Disorders at Washington University School of Medicine in St. Louis (WUSTL: 23 schizophrenia patients and 41 healthy controls; 19 of the patients were receiving various antipsychotic medications; nine healthy controls with ages of <16 years were excluded, available at https://openfmri.org/).

All of the patients were evaluated by qualified psychiatrists using the Structured Clinical Interview for DSM-IV Axis I Disorders and the Patient Edition (First et al., 1996) and were required to meet the DSM-IV (APA, 2013) diagnostic criteria for schizophrenia. No patients had a history of neurological disorders, severe medical disorders, substance abuse, or electroconvulsive therapy. The symptoms severity of the patients was measured with the Positive and Negative Syndrome Scale (PANSS) assessment (Xijing#1, AMU, and Xiangya). All of the healthy controls, who had no relationship with the schizophrenic patients, were assessed in accordance with DSM-IV criteria, and none had acute physical illness, substance abuse or dependence, a history of head injury resulting in loss of consciousness, or major psychiatric or neurological disorders. All of the participants provided written informed consent, and this study was conducted according to the principles in the Declaration of Helsinki and was approved by the Ethics Committee or Institutional Review Boards of the respective hospitals or image centers.

2.2. Image Acquisition

2.2.1. Xijing#1

All data were collected on a 3·0-T Tim Trio scanner (Siemens, Erlangen, Germany) using a 12-channel phased-array head coil. Images were acquired using a gradient-echo echo-planar pulse sequence sensitive to blood oxygenation level-dependent (BOLD) contrast [repetition time (TR) = 2000 ms; echo time (TE) = 30 ms; flip angle (FA) = 90°; matrix = 64 × 64; field of view (FOV) = 220 mm; thickness = 4 mm; gap = 0·6 mm; slices = 33]. Each resting-state fMRI run lasted 8 min (240 TRs), and each subject underwent two runs. Subjects were instructed to stay awake, keep their eyes closed, and minimize head movement; no other task instruction was provided.

2.2.2. Xijing#2

The MRI data were collected on a 3·0-T GE Signa scanner (GE Signa, Milwaukee, Wisconsin, USA). Echo-planar imaging (EPI) sequences were used to obtain BOLD-fMRI images [TR = 2000 ms; TE = 22·5 ms; FA = 30°; matrix = 64 × 64; FOV = 220 mm; thickness = 4 mm; gap = 0·6 mm; slices = 33]. Each subject underwent a resting-state run lasting 8 min (240 TRs). Subjects were instructed to stay awake, keep their eyes closed, and minimize head movement; no other task instruction was provided.

2.2.3. AMU

The MRI data were collected on a 3·0-T GE Signa scanner (GE Signa, Milwaukee, Wisconsin, USA). EPI sequences were used to obtain BOLD-fMRI images [TR = 2000 ms; TE = 22.5 ms; FA = 30°; matrix = 64 × 64; FOV = 220 mm; thickness = 4 mm; gap = 0·6 mm; slices = 33]. Each subject underwent a resting-state run lasting 8 min (240 TRs). Subjects were instructed to stay awake, keep their eyes closed, and minimize head movement; no other task instruction was provided.

2.2.4. Xiangya

The MR images were acquired on a 1·5-T GE Signa scanner (GE Signa, Milwaukee, Wisconsin, USA). Foam padding and earplugs were used to minimize head motion and reduce scanner noise. Participants were told to lie still, keep their eyes closed, and stay awake. EPI sequences were utilized to obtain functional images [TR = 2000 ms; TE = 40 ms, FA = 90°, matrix = 64 × 64; FOV = 240 mm; thickness = 5 mm; gap = 1 mm; slices = 20]. Each subject underwent a resting-state run lasting 6 min (180 TRs).

2.2.5. COBRE

The MR images were acquired on a 3·0-T Tim Trio scanner (Siemens, Erlangen, Germany) with single-shot full k-space EPI sequences [TR = 2000 ms; TE = 29 ms, FA = 75°, matrix = 64 × 64; FOV = 192 mm; thickness = 4 mm; no gap; slices = 32]. Each subject underwent a resting-state run, which lasted 5 min (150 TRs).

2.2.6. UCLA

All BOLD-fMRI data were collected on a 3·0-T Tim Trio scanner (Siemens, Erlangen, Germany). BOLD-fMRI scans were acquired using an asymmetrical spin-echo, echo-planar sequence (T2*) [TR = 2000 ms; TE = 30 ms; FA = 90°; matrix = 64 × 64; FOV = 192 mm; thickness = 4 mm; no gap; slices = 34]. Each subject underwent a resting-state run, which lasted 5 min (152 TRs). Subjects were instructed to rest quietly with their eyes closed.

2.2.7. WUSTL

All data were collected on a 3·0-T Tim Trio scanner (Siemens, Erlangen, Germany). BOLD-fMRI scans were acquired using an asymmetrical spin-echo, echo-planar sequence (T2*) [TR = 2500 ms; TE = 27 ms; FA = 90°; matrix = 64 × 64; FOV = 256 mm; thickness = 4 mm; no gap; slices = 33]. Each BOLD-fMRI run lasted 6 min and 50 s (164 TRs), and each subject underwent three working memory task runs.

2.3. Data Preprocessing

All of the BOLD-fMRI data were preprocessed by using the previously described procedures (Zeng et al., 2014a,b) with a statistical parametric mapping software package (SPM8, Welcome Department of Cognitive Neurology, Institute of Neurology, London, UK, http://www.fil.ion.ucl.ac.uk/spm). For each subject, the first five frames of the scanned data were discarded for magnetic saturation. The following steps were then performed: 1) slice timing correction; 2) motion correction; 3) normalization with an EPI template in the Montreal Neurological Institute atlas space (3-mm isotropic voxels); 4) spatial smoothing using a 6-mm full-width half-maximum Gaussian kernel; 5) linear detrend and band-pass temporal filtering (0·01–0·08 Hz); 6) regression of nuisance variables, including the six parameters obtained by rigid body head motion correction, ventricular and white matter signals, and their first temporal derivatives, quadratic terms, and squares of derivatives (32P) (Ciric et al., 2017; Satterthwaite et al., 2013); and 7) if frame-wise displacement (FD) at any point in time exceeded 0·3 mm, then that time point was scrubbed (Drysdale et al., 2017; Power et al., 2015).

2.4. Control of Motion Artifact

To control confounding effects of motion artifact, several strategies were conducted: First, 10 patients and 18 controls were excluded due to excessive head motion during scan acquisition (>2·5 mm translation and/or 2·5° rotation); second, frame-wise displacement (FD) and the temporal mean of the FD time series (mFD) was computed for each run (Power et al., 2012), and we excluded 93 patients and 60 controls with high levels of gross motion (mFD > 0·3 mm) (Satterthwaite et al., 2013); third, a confound regression strategy of nuisance variables as the aforementioned was used. Fourth, if FD at any point in time exceeded 0·3 mm, then that time point was scrubbed (Drysdale et al., 2017; Power et al., 2015). After volume censoring, we excluded 14 patients and 143 controls with <100 time points. The results of the analysis of controls for motion-related artifact can be seen in Fig. S1, illuminating the sufficiency of the control of motion artifact.

After stringent control of motion artifact (117 patients and 221 controls were excluded totally) and balancing for age and gender between the patient and control groups (We first calculated the Chi-square value and sorted the subjects by age for each group. Then the subjects with minimal/maximal ages and a certain gender were removed to guarantee that *P*-values were >0·10 both in the Chi-square test of gender and in the two-sample *t*-test of age), 357 patients and 377 controls were finally retained for further analysis. The participants' demographic and clinical characteristics are summarized in Table 1, and the patients and controls were matched for each site ($P > 0{\cdot}10$, Pearson Chi-square test/ two-sample *t*-test). In addition, the entire patient and control groups were also well matched in gender (patients vs. controls: 144/213 vs. 161/216 females/males, $P = 0{\cdot}515$, Pearson Chi-square test), age (patients vs. controls: $28{\cdot}49 \pm 8{\cdot}58$ vs. $29{\cdot}32 \pm 9{\cdot}07$ years, $P = 0{\cdot}202$, two-sample *t*-test), mFD before and after motion scrubbing (patients vs. controls: $0{\cdot}129 \pm 0{\cdot}063$ vs. $0{\cdot}128 \pm 0{\cdot}060$ mm, $P = 0{\cdot}886$, and $0{\cdot}103 \pm 0{\cdot}040$ vs. $0{\cdot}107 \pm 0{\cdot}040$ mm, $P = 0{\cdot}108$, two-sample *t*-test), and data loss (patients vs. controls: $6{\cdot}44 \pm 7{\cdot}96\%$ vs. $5{\cdot}94 \pm 3{\cdot}95\%$, $P = 0{\cdot}381$, two-sample *t*-test).

2.5. Functional Connectivity Measure

To ensure the optimal use of the wealth of information present in fcMRI scans, we used multi-atlas based whole-brain fcMRI in the multivariate pattern analysis, which measures functional connectivity of the same image in different spaces of multiple atlases (Min et al., 2014). The first one includes 176 regions of interest (ROIs) based on the 17-functional network parcellation of the human brain (Buckner et al., 2011; Choi et al., 2012; Yeo et al., 2011). The second one includes 160 ROIs from several meta-analyses of fMRI activation studies (Dosenbach et al., 2010). All of the 160 ROIs were modeled as 6-mm radius spheres. The last one is the automated anatomical labeling (AAL) atlas, which included 116 regions (Tzourio-Mazoyer et al., 2002).

For each brain atlas, we evaluated functional connectivity between each pair of regional averaged time courses using the Pearson correlation coefficient. Then, all of the correlation coefficients were converted to *z*-scores by applying Fisher *r*-to-*z* transformation. Thus, three correlation matrices were obtained for each subject. To remove potential site-related variation in functional connectivity measures, *Z*-standardization was used to normalize the functional correlation matrices (Fig. S2) (Yan et al., 2013). Then, the standardized correlation matrices were used as classification features in further analysis.

2.6. Discriminant Autoencoder Network with Sparsity Constraint (DANS)

In the current study, we developed a deep DANS neural network aimed at learning imaging site-shared rather than site-specific fcMRI features. In contrast to the conventional sparse autoencoder network (SAN) (Kim et al., 2016), we introduced an optimized discriminant item based on correlation function in the cost function at the pre-training stage to generate discriminating fcMRI features for binary classification. In the conventional autoencoder neural network, a least-square cost function is defined in the pre-training of weight matrices: $\min_{w,\ w_d} \|\boldsymbol{O}-\boldsymbol{X}\|_F$, where $\boldsymbol{w}$ denotes weight matrix for linear dimensionality reduction, $\boldsymbol{w_d}$ denotes weight matrix for linear reconstruction, $\boldsymbol{O} = \boldsymbol{g}(\boldsymbol{w_d h} + \boldsymbol{q})$ denotes reconstructed output, $\boldsymbol{h} = \boldsymbol{f}(\boldsymbol{wx} + \boldsymbol{p})$ denotes hidden layer output, $\boldsymbol{g}(\cdot)$ and $\boldsymbol{h}(\cdot)$ denote nonlinear projections, $\boldsymbol{p}$ and $\boldsymbol{q}$ denote offsets, and $\|\cdot\|_F$ denotes L_F-norm. For the multi-layer autoencoder network, let $\boldsymbol{h}(s-1)$ denote the input of *S*th layer; then the hidden layer output is $\boldsymbol{h}(s) = \boldsymbol{f}(\boldsymbol{w}(s)\boldsymbol{h}(s-1) + \boldsymbol{p}(s))$, the reconstructed output is $\boldsymbol{O}(s) = \boldsymbol{g}(\boldsymbol{w_d}(s)\boldsymbol{h}(s) + \boldsymbol{q}(s))$, and the cost function is $\min_{w(s),\ w_d(s)} \boldsymbol{J_1}(s) = \|\boldsymbol{O}(s)-\boldsymbol{h}(s-1)\|_F$. Because the previous studies demonstrate that classification performance may benefit from the control of the sparsity of layer weights(Kim et al., 2016), a sparsity constraint $c_1\|\boldsymbol{w}(s)\|_1 + c_2\|\boldsymbol{w}(s)\|_2^2$ was also included in the cost function to minimize the risk of overfitting and further to improve the generalizability of the discriminant deep learning framework, where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote L_1-norm and L_2-norm, respectively, and c_1 and c_2 denote the regularization coefficients selected in the ranges of $[10^{-5},\ 10^{-4}, 10^{-3},\ 10^{-2}, 10^{-1}]$. To reduce the computation burden of the estimate of hyperparameters

Table 1
Summary of the demographic information and clinical characteristics in the present study.

Sites	Group	Age (yrs)	Gender (F/M)	Pre-scrubbing mFD (mm)	Post-scrubbing mFD (mm)	Data loss	Positive scale	Negative scale	Total scale	Duration of ill (yrs)
Xijing#1	Patient	26·40 ± 6·69	45/58	0·121 ± 0·057	0·099 ± 0·029	0·056 ± 0·082	21·70 ± 7·43	21·34 ± 8·76	87·44 ± 29·60	2·00 ± 2·57
	Control	27·79 ± 5·19	29/46	0·128 ± 0·055	0·106 ± 0·033	0·058 ± 0·077				
	P-value	0·136[a]	0·502[b]	0·462[a]/0·292[c]	0·102[a]/0·132[c]	0·872[a]				
Xijing#2	Patient	27·15 ± 7·81	21/26	0·096 ± 0·052	0·077 ± 0·030	0·036 ± 0·048				
	Control	27·67 ± 8·42	26/34	0·093 ± 0·045	0·080 ± 0·029	0·031 ± 0·056				
	P-value	0·745[a]	0·889[b]	0·682[a]/0·978[c]	0·638[a]/0·378[c]	0·637[a]				
AMU	Patient	30·73 ± 8·42	44/44	0·112 ± 0·056	0·090 ± 0·037	0·043 ± 0·055	11·79 ± 3·49	10·91 ± 2·98	65·80 ± 6·90	5·67 ± 5·77
	Control	31·04 ± 10·97	53/41	0·100 ± 0·047	0·089 ± 0·034	0·031 ± 0·058				
	P-value	0·829[a]	0·388[b]	0·121[a]/0·176[c]	0·905[a]/0·988[c]	0·155[a]				
Xiangya	Patient	24·50 ± 6·00	17/25	0·107 ± 0·039	0·086 ± 0·025	0·055 ± 0·050			83·34 ± 14·64	
	Control	24·87 ± 4·85	11/19	0·094 ± 0·038	0·078 ± 0·021	0·037 ± 0·049				
	P-value	0·783[a]	0·744[b]	0·174[a]/0·162[c]	0·166[a]/0·261[c]	0·134[a]				
COBRE	Patient	32·36 ± 13·62	4/21	0·200 ± 0·055	0·151 ± 0·032	0·129 ± 0·096				
	Control	33·34 ± 10·60	14/33	0·181 ± 0·050	0·147 ± 0·029	0·114 ± 0·089				
	P-value	0·737[a]	0·198[b]	0·137[a]/0·202[c]	0·547[a]/0·662[c]	0·520[a]				
UCLA	Patient	35·15 ± 9.03	8/26	0·193 ± 0·048	0·157 ± 0·028	0·130 ± 0·098				
	Control	31·88 ± 9·41	18/32	0·188 ± 0·046	0·149 ± 0·018	0·108 ± 0·086				
	P-value	0·116[a]	0·255[b]	0·639[a]/0·594[c]	0·141[a]/0·118[c]	0·299[a]				
WUSTL	Patient	24·29 ± 3·62	5/13	0·170 ± 0·070	0·122 ± 0·039	0·102 ± 0·110				
	Control	23·04 ± 3·52	10/11	0·146 ± 0·036	0·123 ± 0·025	0·065 ± 0·061				
	P-value	0·283[a]	0·204[b]	0·164[a]/0·155[c]	0·937[a]/0·767[c]	0·192[a]				

[a] Two-sample *t*-test.
[b] Pearson Chi-square test.
[c] Wilcoxon rank sum test.

(regularization coefficients), we fixed the c_1 and c_2 for all hidden layers. In addition to L_1– and L_2-norm regularization, a dropout strategy was also employed to prevent overfitting (Li et al., 2014), where 50% of nodes and respective connections were temporarily removed to extract different sets of features that could independently produce a useful output.

In this study, we additionally introduced an optimized discriminant item based on a correlation function, i.e. $\sum_{i,\,j=1}^{i=N_j,\,j=C}\rho_i^j(s)$, in the cost function at the pre-training stage to generate discriminating fcMRI features for binary classification ($C = 2$). $\rho_i^j(s)$ denotes the correlation function between the feature vector of Sample i of Class C_j and the averaged feature vector of samples of other classes, and $\rho_i^j(s) = \widetilde{\boldsymbol{m}}_j(s)^T\boldsymbol{h}_i^j(s)/\sqrt{\|\widetilde{\boldsymbol{m}}_j(s)\|^2\|\boldsymbol{h}_i^j(s)\|^2}$, where $\boldsymbol{h}_i^j(s)$ denotes the feature vector of Sample i of Class C_j, $\widetilde{\boldsymbol{m}}_j(s) = \frac{1}{N-N_j}\sum_{i,k=1(k\neq j)}^{i=N_k,\,k=C}h(s)$ denotes the averaged feature vector of samples of other classes, and C, N, N_j, N_k denote the number of classes ($C = 2$ here), sample size of training dataset, sample sizes of Class C_j and Class C_k, respectively. An autoencoder neural network always minimizes the cost function using the initial input and final output. Because the models were trained in a layer-wise way here, the error of between input and output of a given hidden layer is equivalent to the error between initial input and final output of the network. Thus, the cost function of the layer-wise training model could be as follows:

$$\min_{\boldsymbol{w}(s),\boldsymbol{w}_d(s)}\mathbf{J}(s) = \mathbf{J}_1(s) + c_1\|\boldsymbol{w}(s)\|_1 + c_2\|\boldsymbol{w}(s)\|_2^2 + c_3\sum_{i,j=1}^{i=N_j,\,j=C}\rho_i^j(s)$$

where c_3 denotes the regularization coefficient selected in the ranges of $[0{\cdot}1, 0{\cdot}5, 1, 5, 10]$. The aim of including the optimized discriminant item was to speed up the convergence and improve classification performance by learning imaging site-shared features rather than site-specific features.

In the back propagation of the conventional autoencoder neural network, the gradients $\frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{w}(s)}$, $\frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{p}(s)}$, $\frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{w}_d(s)}$, $\frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{q}(s)}$, $\frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{h}(s)}$, $\frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{O}(s)}$ can be calculated according to the computational formula on http://ufldl.stanford.edu/wiki/index.php/ Backpropagation_Algorithm. After introducing a sparsity constraint $c_1\|\boldsymbol{w}(s)\|_1 + c_2\|\boldsymbol{w}(s)\|_2^2$ and an optimized discriminant item $\sum_{i,\,j=1}^{i=N_j,\,j=C}\rho_i^j(s)$ in the cost function:

$$\min_{\boldsymbol{w}(s),\boldsymbol{w}_d(s)}\mathbf{J}(s) = \boldsymbol{J}_1(s) + c_1\|\boldsymbol{w}(s)\|_1 + c_2\|\boldsymbol{w}(s)\|_2^2 + c_3\sum_{i,j=1}^{i=N_j,\,j=C}\rho_i^j(s)\,,$$

where $\boldsymbol{J}_1(s) = \|\boldsymbol{O}(s) - \boldsymbol{h}(s-1)\|_F$, $\boldsymbol{h}(s) = \boldsymbol{f}(\boldsymbol{w}(s)\boldsymbol{h}(s-1) + \boldsymbol{p}(s))$, $\boldsymbol{O}(s) = \boldsymbol{g}(\boldsymbol{w_d}(s)\boldsymbol{h}(s) + \boldsymbol{q}(s))$, and $\rho_i^j(s) = \widetilde{\boldsymbol{m}}_j(s)^T\boldsymbol{h}_i^j(s)/\sqrt{\|\widetilde{\boldsymbol{m}}_j(s)\|^2\|\boldsymbol{h}_i^j(s)\|^2}$, where $\widetilde{\boldsymbol{m}}_j(s) = \frac{1}{N-N_j}\sum_{i,k=1(k\neq j)}^{i=N_k,\,k=C}h(s)$; excepting $\frac{\partial\mathbf{J}(s)}{\partial\boldsymbol{w}_d(s)}$, $\frac{\partial\mathbf{J}(s)}{\partial\boldsymbol{q}(s)}$, $\frac{\partial\mathbf{J}(s)}{\partial\boldsymbol{O}(s)}$, the computational formula of the other gradients need to be modified:

(1) The weight matrix for linear dimensionality reduction $\boldsymbol{w}(s)$:

$$\frac{\partial\mathbf{J}(s)}{\partial\boldsymbol{w}(s)} = \frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{w}(s)} + c_1\ \mathrm{sgn}(\boldsymbol{w}(s)) + 2c_2\boldsymbol{w}(s),$$

(2) The hidden layer output $\boldsymbol{h}(s)$:

$$\frac{\partial\mathbf{J}(s)}{\partial\boldsymbol{h}(s)} = \frac{\partial\mathbf{J}_1(s)}{\partial\boldsymbol{h}(s)} + \frac{c_3\widetilde{\boldsymbol{m}}_j(s)}{\sqrt{\|\widetilde{\boldsymbol{m}}_j(s)\|^2\left\|\boldsymbol{h}_i^j(s)\right\|^2}} + \sum_{k=1,k\neq j}^{C}\frac{1}{N-N_k}\sum_{i=1}^{N_k}\frac{c_3\boldsymbol{h}_i^k(s)}{\sqrt{\|\widetilde{\boldsymbol{m}}_k(s)\|^2\left\|\boldsymbol{h}_i^k(s)\right\|^2}}$$
$$-\rho\left(\boldsymbol{h}_i^j(s),\widetilde{\boldsymbol{m}}_j(s)\right)\frac{c_3\boldsymbol{h}_i^j(s)}{\left\|\boldsymbol{h}_i^j(s)\right\|^2} - \sum_{k=1,k\neq j}^{C}\frac{1}{N-N_k}\sum_{i=1}^{N_k}\rho(\boldsymbol{h}_i^k(s),\widetilde{\boldsymbol{m}}_k(s))\frac{c_3\widetilde{\boldsymbol{m}}_k(s)}{\left\|\widetilde{\boldsymbol{m}}_k(s)\right\|^2},$$

(3) The offset in the back propagation $\boldsymbol{p}(s)$:

$$\frac{\partial \mathbf{J}(s)}{\partial \boldsymbol{p}(s)} = \sum \left(\frac{\partial \mathbf{J}(s)}{\partial \boldsymbol{h}(s)} \right) \cdot \mathbf{f}'(\boldsymbol{w}(s)\boldsymbol{h}(s-\mathbf{1}) + \boldsymbol{p}(s)).$$

In the current study, the learning rate was fixed at 10^{-3}, the dropout was fixed at 0·5, and the updates of weight matrices were stopped if the epoch number reached 100.

The fine tuning of the current DANS model is the same as the conventional autoencoder neural network. First, we used the weight matrix in the pre-training as the initials. Then the template $\boldsymbol{M}(s)$ with sparse weights can be generated with a threshold of $e(s) > 0$ (10^{-3} was selected here):

$$\boldsymbol{M}_{ij}(s) = \begin{cases} 1 & |\boldsymbol{w}_{ij}(s)| \geq e(s) \\ 0 & |\boldsymbol{w}_{ij}(s)| < e(s) \end{cases}$$

In the fine tuning, the $\boldsymbol{M}(s)$ could not be modified, and we just updated the corresponding weights if $\boldsymbol{M}_{ij}(s) = 1$, so that there was no change in the distribution of sparse weights.

Linear support vector machines (SVMs) were used in the classifier layer of the DANS neural network. Thus, the entire procedure of the DANS framework included three steps: (1) layer-wise pre-training to generate initial weight matrices using a training dataset, (2) fine tuning to determine hyperparameters with a softmax layer using a validation dataset, and (3) testing to generate classification results with linear SVMs using a testing dataset. The flowchart of the DANS framework can be seen in Fig. 1.

The DANS networks were constructed with three hidden layers and 100 nodes in each layer, which were optimized using the training and validation datasets (Fig. S3). A conventional SAN network without the discriminant item was also constructed with three hidden layers and 100 nodes in each layer in the multi-site pooling classification. The model training and validation experiments revealed that the discriminant item in the cost function could speed up the convergence of the deep learning network and improve classification performance, as shown in Fig. S4.

2.7. Multivariate Pattern Classification

In this study, we conducted two types of classification: (1) *k*-fold multi-site pooling classification, in which all seven datasets were pooled together, and then *k*-fold cross-validation strategies were used to evaluate the classification performance; and (2) leave-site-out transfer classification, in which the sample of a given imaging site was left for testing, and the sample of other sites was used for training. These two types of classification were independent from each other.

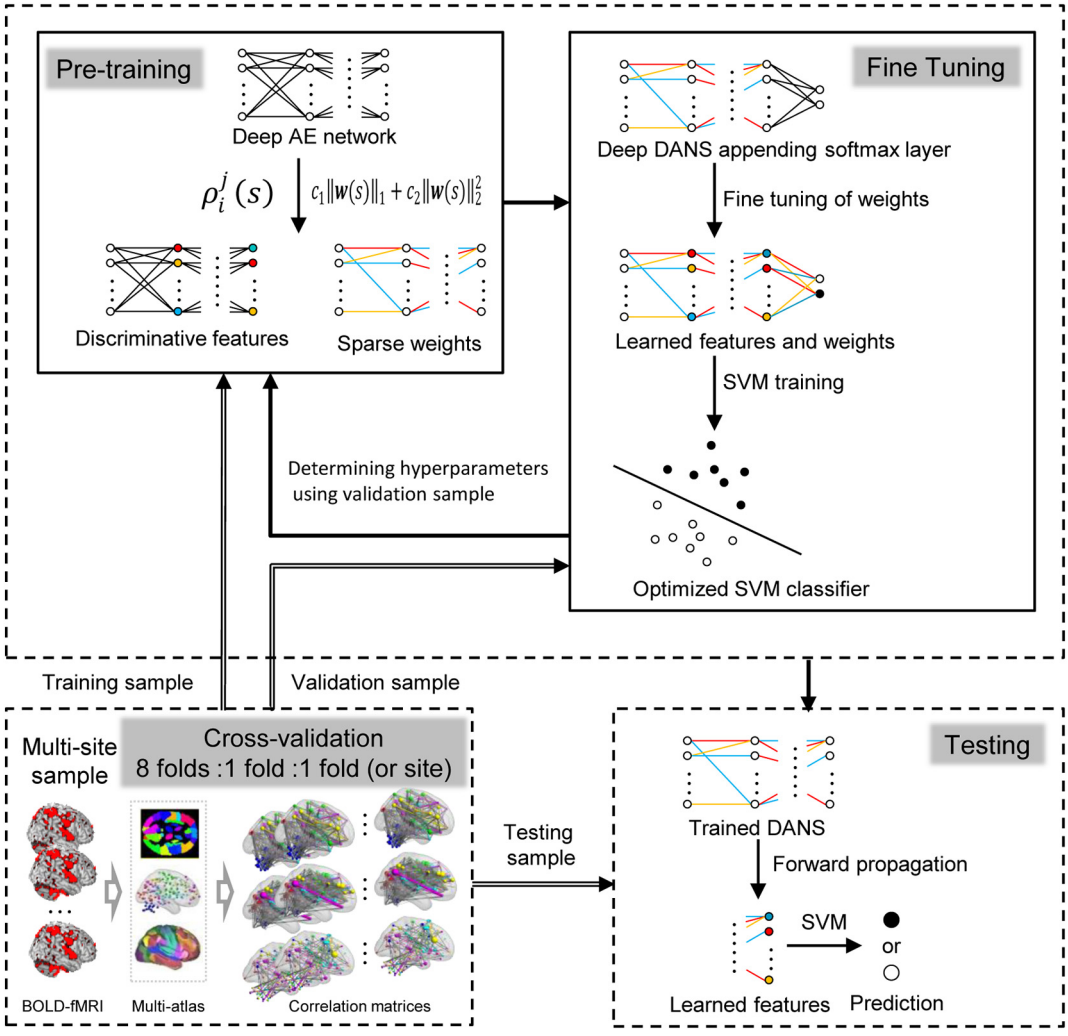

Fig. 1. The flowchart of deep discriminant autoencoder neural network with sparsity constraint (DANS) in distinguishing schizophrenic patients from healthy controls. The entire procedure of the DANS framework includes three steps: (1) layer-wise pre-training to generate initial weight matrices using training dataset (eight folds), (2) fine tuning to determine hyperparameters with a softmax layer using validation dataset (one fold), and (3) testing to generate classification results with linear support vector machines using testing dataset (one fold or site).

In addition to the proposed deep learning method, the canonical linear SVM and linear discriminant analysis (LDA) with feature selection were used in this study. Because initially reducing the number of features accelerates computation and diminishes noise (Dosenbach et al., 2010), a hybrid feature reduction strategy was adopted in the study. First, a univariate feature selection with Kendall tau rank correlation coefficient was used to eliminate half of the correlation features (Shen et al., 2010; Zeng et al., 2012). Then, multivariate recursive feature elimination (RFE) was used in combination with SVM or LDA (RFE-SVM, RFE-LDA) (Martino et al., 2008). It should be noted that feature selection was done on the training dataset, embedded in cross-validation cycles. All the classifiers are supervised, so that the models were trained to differentiate patients from controls rather than differentiate different sites, and the final accuracy is due to the disorder itself rather than site difference.

In the multi-site pooling classification, we used a k-fold ($k = 10$) cross-validation strategy to estimate the generalization ability of the classifiers. In the DANS and SAN classification frameworks, we used eight folds as the training dataset, one fold as the validation dataset, and one fold as the testing dataset. In the RFE-SVM and RFE-LDA classification frameworks, we used nine folds as the training dataset and one fold as the testing dataset, and the training dataset was further partitioned in nine splits both in the RFE-SVM and RFE-LDA nested cycles (Martino et al., 2008). The multi-site pooling classification with a five-fold cross-validation strategy can be seen in the Supplementary data.

In the leave-site-out transfer classification, we used the sample of a given imaging site as the testing dataset and the sample of other sites as the training dataset. The training dataset was further randomly partitioned into nine folds, and eight folds were used for training and one fold for validation in the DANS and SAN classification frameworks. The training dataset was also partitioned in nine splits both in the RFE-SVM and RFE-LDA nested cycles (Martino et al., 2008).

The performance of a classifier can be quantified using the accuracy, sensitivity, and specificity based on the results of cross-validation (k-fold or leave-site-out). Note that the sensitivity represents the proportion of patients correctly predicted, whereas the specificity represents the proportion of controls correctly predicted. The overall proportion of samples correctly predicted was evaluated by the accuracy (generalization rate). The full k-fold cross-validation procedure was repeated ten times to generate the means and standard deviations of accuracy, sensitivity, and specificity. We used two-sample/paired t-test and Kolmogorov-Smirnov test (a nonparametric test) to compare classification performance between different algorithms.

To ensure the optimal use of the wealth of information, we used multi-atlas fusions in the classification at the feature and label levels, respectively. In the feature-level fusion, the functional connectivity features of the three atlases were merged together and were considered as the inputs of a given classifier. In the label-level fusion, three classifiers were trained based on the functional connectivity features of the three atlases, respectively, and then we applied a standard majority-voting scheme to resolve disagreements in the prediction outputs of the three classifiers.

2.8. Estimation of the Discriminative Power of Functional Connectivity

The discriminative power of the functional correlations for the first hidden layer of the DANS network in the ten-fold multi-site pooling classification was calculated using the following steps: First, for each node, the weight vectors were averaged for ten iterations in a full ten-fold cross-validation. Second, the nodal weight vectors were converted to binary ones by a threshold, with which the top 1% of the elements with the largest absolute values were retained. Then, a synthesized weight vector of the first layer was obtained by averaging the weight vectors of all nodes of this layer. Because the full ten-fold cross-validation was repeated ten times, the final weight vector was obtained

Table 2
The results of ten-fold multi-site pooling classification.

Altas	RFE-SVM (%)			RFE-LDA (%)			SAN (%)			DANS (%)		
	SS	SC	ACC	SS	SC	ACC	SS	SC	ACC	SS	SC	ACC
AAL116	77·2 ± 1·0	80·4 ± 1·2	78·8 ± 0·8	76·8 ± 0·8	82·3 ± 1·6	79·7 ± 1·0	81·2 ± 1·2	84·3 ± 1·5	82·8 ± 0·7	82·0 ± 1·1	85·5 ± 0·7	83·8 ± 0·5
ROI160	73·3 ± 0·9	75·1 ± 1·9	74·2 ± 0·9	73·4 ± 0·9	75·5 ± 1·4	74·5 ± 1·0	77·2 ± 1·4	78·4 ± 1·1	78·1 ± 0·8	78·9 ± 1·7	80·4 ± 1·3	79·6 ± 0·9
Parcel176	77·6 ± 1·1	79·7 ± 1·1	78·6 ± 0·8	77·9 ± 1·2	81·1 ± 1·2	79·5 ± 0·8	79·3 ± 1·5	81·8 ± 1·7	80·6 ± 1·1	81·0 ± 1·0	84·2 ± 1·2	82·7 ± 0·9
Feature-fusion	79·3 ± 0·9	82·8 ± 1·4	81·1 ± 0·7	79·2 ± 0·9	83·9 ± 1·5	81·6 ± 0·9	82·2 ± 1·2	84·5 ± 0·8	83·4 ± 0·9	83·2 ± 1·0	85·4 ± 0·9	84·3 ± 0·7
Label-fusion	78·7 ± 1·0	82·7 ± 1·4	80·7 ± 1·0	78·5 ± 1·2	83·2 ± 1·1	80·9 ± 0·9	81·6 ± 1·6	85·1 ± 1·2	83·4 ± 1·1	83·1 ± 2·0	86·8 ± 2·1	85·0 ± 1·2

RFE, recursive feature elimination; SVM, support vector machine; LDA, linear discriminant analysis; SAN, sparse autoencoder network; DANS, discriminant autoencoder network with sparsity constraint; SS, sensitivity; SC, specificity; ACC, accuracy.

by averaging the ten synthesized ones, which indicates the discriminative power of the functional correlations in this layer. For the second and third hidden layers, because they used the output of the former hidden layer as an input, and the nodes always worked at an approximate linear interval, the weight vectors could be multiplied forward approximately. Region weights, representing the relative contributions in the classification of schizophrenia, were denoted by the sum of the weights of the relevant functional correlations. In this case, the final weights of the functional correlations and regions could be obtained. *Z*-standardization was used to normalize the connectivity and region weights. The weights of functional correlations and regions in leave-site-out transfer classification could be calculated in a similar way.

We grouped the brain regions of the 17-network parcellation into six functional subsystems for visualization: visual (VN), somatomotor (SMN), ventral attention/salience (vATN), dorsal attention (dATN), frontoparietal control (FPN), and default networks (DN). The brain regions of the AAL template and 160 ROIs were also labeled with the six networks according to their locations.

3. Results

3.1. Ten-Fold Multi-Site Pooling Classification

Accuracies of 83·8 ± 0·5%, 79·6 ± 0·9%, and 82·7 ± 0·9% were obtained by using the DANS method with the AAL template, 160 ROIs, and 17-network parcellation, respectively, which were significantly higher than those obtained by using the RFE-SVM and RFE-LDA classifiers ($P < 0·001$, two-sample *t*-test and Kolmogorov-Smirnov test, Table 2 and Fig. 2A), and the SAN network ($P < 0·05$, two-sample *t*-test and Kolmogorov-Smirnov test, Table 2 and Fig. 2A). When fusing multiple atlases at the feature and label levels, accuracies of 84·3 ± 0·7% and 85·0 ± 1·2% were obtained by using the DANS method, respectively (Table 2 and Fig. 2A), which were significantly higher than the accuracies obtained by using the RFE-SVM and RFE-LDA classifiers ($P < 0·001$, two-sample *t*-test and Kolmogorov-Smirnov test, Table 2 and Fig. 2A), and the SAN network ($P < 0·05$, two-sample *t*-test and Kolmogorov-Smirnov test, Table 2 and Fig. 2A). In addition, we also conducted five-fold multi-site pooling classification, and the accuracies were slightly lower than those of ten-folds (Table S1).

3.2. Leave-Site-Out Transfer Classification

The results of single-atlas based transfer classification can be seen in Table S2–S4. The accuracies of 79·8 ± 4·2%, 77·2 ± 4·9%, and 78·4 ± 4·8% were obtained by using the DANS method with the AAL template, 160 ROIs, and 17-network parcellation, respectively, which were significantly higher than the accuracies obtained by using the RFE-SVM and RFE-LDA classifiers ($P < 0·05$, paired *t*-test or Kolmogorov-Smirnov test, Fig. 2B). When fusing multiple atlases at the feature and label levels, accuracies of 80·4 ± 4·4% and 81·0 ± 4·9% were obtained by using the DANS method, which were significantly higher than the accuracies obtained by using the RFE-SVM and RFE-LDA classifiers ($P < 0·05$, paired *t*-test, Tables 3, 4 and Fig. 2B). Noted that one sample used data collected at a 1·5-T scanner (Xiangya), and another used functional rather than resting-state data (WUSTL). Given that the rest of the sites used resting-state data obtained at 3·0-T scanners, we conducted additional analysis that leaves out both the Xiangya and WUSTL data, and the full results of leaving-two sites-out cross-validation based on multi-atlas fusion can be seen in the Supplementary data (Table S5 & S6), which reveals that similar results could be obtained.

3.3. Most Discriminating Functional Connectivity

We first analyzed the most discriminating functional connectivity based on the 17-network parcellation in the multi-site pooling classification. The fcMRI features learned by the nodes in the first hidden layer were analyzed (Fig. S5). A significant observation was that cortical-striatal-cerebellar circuit exhibited great weights. Some nodes learned cortical-striatal functional connectivity features, some nodes learned striatal-cerebellar functional connectivity features, and some nodes learned direct cortical-cerebellar functional connectivity features (Fig. 3).

It was observed that the most discriminating regions include the left caudate, posterior cingulate cortex (PCC), bilateral temporo-parietal cortices, and right inferior parietal lobule (IPL) within the DN, bilateral prefrontal cortex (PFCl), dorsal prefrontal cortex (PFCd), and left temporal cortex within the FPN, left PFCd, right pre-central cortex (PrC), medial frontal cortex (FrMed), right ventral prefrontal cortex (PFCv), and bilateral putamen within the vATN, left frontal eye field (FEF) within the dATN, and central sulcus and putamen within the SMN

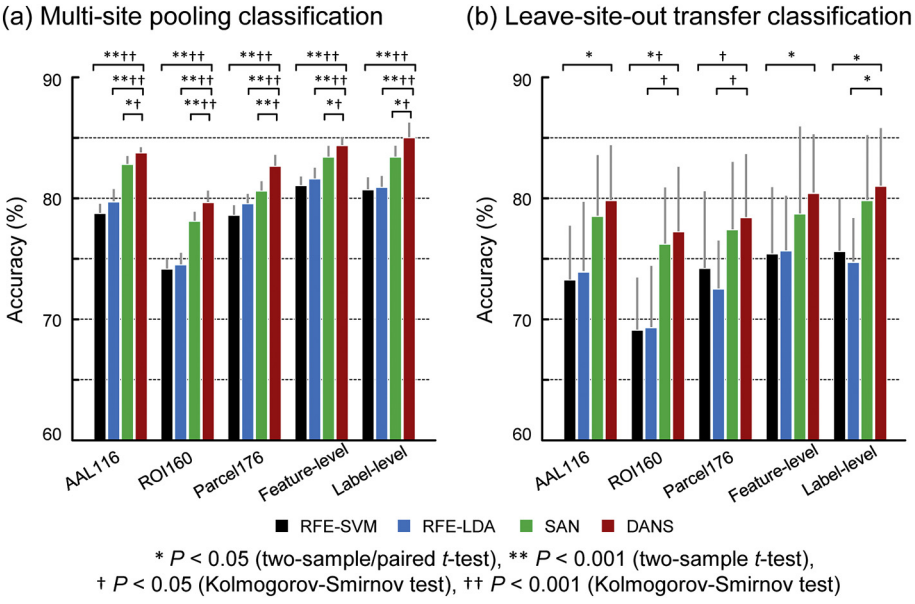

Fig. 2. The results of ten-fold multi-site pooling classification (A) and leave-site-out transfer classification (B). RFE, recursive feature elimination; SVM, support vector machine; LDA, linear discriminant analysis; SAN, sparse autoencoder network; DANS, discriminant autoencoder network with sparsity constraint.

Table 3
The results of leave-site-out transfer classification based on multi-atlas fusion at the feature level.

Site	RFE-SVM (%)			RFE-LDA (%)			SAN (%)			DANS (%)		
	SS	SC	ACC	SS	SC	ACC	SS	SC	ACC	SS	SC	ACC
Xijing#1	71·8	90·7	79·8	61·2	90·7	73·6	68·9	93·3	79·2	70·9	93·3	80·3
Xijing#2	78·7	68·3	72·9	74·5	75·0	74·8	76·6	78·8	77·6	78·7	78·3	78·5
AMU	83·0	80·9	81·9	88·6	78·7	83·5	88·6	88·3	88·5	89·8	88·3	89·1
Xiangya	88·1	56·7	75·0	88·1	56·7	75·0	95·2	56·7	79·2	100·0	60·0	83·3
COBRE	80·0	57·4	65·3	84·0	61·7	69·4	76·0	59·6	65·3	72·0	74·5	73·6
UCLA	64·7	88·0	78·6	58·8	84·0	73·8	79·4	88·0	84·5	76·5	80·0	78·6
WUSTL	66·7	81·0	74·4	61·1	95·2	79·5	66·7	85·7	76·9	61·1	95·2	79·5
Mean ± STD	76·1 ± 8·7	74·7 ± 14·0	75·4 ± 5·5	73·8 ± 13·4	77·4 ± 14·3	75·7 ± 4·5	78·8 ± 10·2	78·6 ± 14·7	78·7 ± 7·2	78·4 ± 11·9	81·4 ± 11·3	80·4 ± 4·4

RFE, recursive feature elimination; SVM, support vector machine; LDA, linear discriminant analysis; SAN, sparse autoencoder network; DANS, discriminant autoencoder network with sparsity constraint; SS, sensitivity; SC, specificity; ACC, accuracy.

Table 4
The results of leave-site-out transfer classification based on multi-atlas fusion at the label level.

Site	RFE-SVM (%)			RFE-LDA (%)			SAN (%)			DANS (%)		
	SS	SC	ACC	SS	SC	ACC	SS	SC	ACC	SS	SC	ACC
Xijing#1	60·2	96·0	75·3	59·2	92·0	73·0	70·9	92·0	79·9	73·8	93·3	82·0
Xijing#2	76·6	65·0	70·1	72·3	73·3	72·9	83·0	73·3	77·6	80·9	78·3	79·4
AMU	81·8	79·8	80·8	85·2	77·7	81·3	92·1	86·2	89·0	92·1	88·3	90·1
Xiangya	90·5	63·3	79·2	90·5	60·0	77·8	97·6	60·0	81·9	97·6	66·7	84·7
COBRE	92·0	57·4	69·4	80·0	66·0	70·8	72·0	70·2	70·8	68·0	76·6	73·6
UCLA	64·7	86·0	77·4	58·8	82·0	72·6	67·7	88·0	79·8	58·8	90·0	77·4
WUSTL	72·2	81·0	76·9	61·1	85·7	74·4	66·7	90·5	79·5	55·6	100·0	79·5
Mean ± STD	76·9 ± 12·2	75·5 ± 13·9	75·6 ± 4·3	72·5 ± 13·1	76·7 ± 11·2	74·7 ± 3·6	78·5 ± 12·4	80·0 ± 12·2	79·8 ± 5·4	75·2 ± 14·8	84·7 ± 10·6	81·0 ± 4·9

RFE, recursive feature elimination; SVM, support vector machine; LDA, linear discriminant analysis; SAN, sparse autoencoder network; DANS, discriminant autoencoder network with sparsity constraint; SS, sensitivity; SC, specificity; ACC, accuracy.

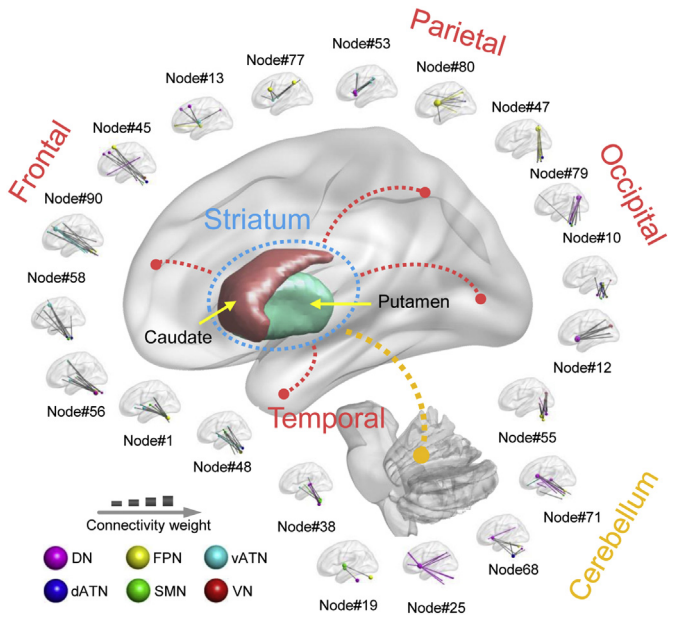

Fig. 3. Cortical-striatal-cerebellar functional connectivity features exhibited great weights in the classification of schizophrenia. The lines representing the discriminating functional connections are scaled with their discriminative power. DN, default network; FPN, frontoparietal control network; vATN, ventral attention network; dATN, dorsal attention network; SMN, somatomotor network; VN, visual network.

(Fig. 4A). The right PFCl, left caudate, left temporo-parietal cortex, and right FrMed exhibited the greatest region weights. Summing up, the most discriminating regions were primarily located within the DN, FPN, and vATN (Fig. 4B), and similar results could be obtained for the AAL template and 160 ROIs (Fig. S6). Furthermore, the region weights obtained in the leave-site-out transfer classification were quite similar with those in the multi-site pooling classification ($R = 0{\cdot}88, 0{\cdot}86$, and $0{\cdot}79$ for the AAL template, 160 ROIs, and 17-network parcellation, respectively, $P < 0{\cdot}001$, Fig. 4D).

We divided the most discriminating connectivity features of the 17-network parcellation into intra- and inter-network groups, as shown in Fig. 4C. The most discriminating intra-network functional correlations were primarily located within the DN, FPN, and vATN, whereas the most discriminating inter-network functional correlations between the DN and FPN (16·4%) and between the DN and vATN (13·2%) exhibited the highest percentages. Similar results were obtained for the AAL template and 160 ROIs (Fig. S7). The connectivity weights were highly correlated across the folds or sites ($P < 0{\cdot}001$, Fig. S8), and the average connectivity weights obtained in the leave-site-out transfer classification were quite similar with the average connectivity weights obtained in the multi-site pooling classification ($R = 0{\cdot}76, 0{\cdot}72$, and $0{\cdot}65$ for the AAL template, 160 ROIs, and 17-network parcellation, respectively, $P < 0{\cdot}001$, Fig. 4D). The top 1% of the most discriminating functional correlations of the three atlases can be seen in Fig. S9.

4. Discussion

We have developed a deep DANS network with multi-atlas fcMRI to discriminate schizophrenic patients from healthy controls in a large multi-site sample. The accuracies of approximately 85·0% and 81·0% were obtained in the multi-site pooling classification and leave-site-out transfer classification between the patients and controls, respectively, suggesting the potential of discriminant deep learning of multi-atlas fcMRI in searching biomarkers to achieve clinical diagnosis of schizophrenia across multiple independent imaging sites. In addition, the results revealed dysregulation of the cortical-striatal-cerebellar circuit in schizophrenia, and the most discriminating functional correlations were primarily located within and across the DN, FPN, and vATN, perhaps implying the potential roles of these subsystems in the "disconnectivity" model underlying the pathophysiology of schizophrenia.

As known, the current clinical diagnosis of schizophrenia is based solely on clinical manifestations. In recent years, a number of previous studies attempted to find stable neuroimaging-based biomarkers and design neuroimaging-based diagnostic tools with the claim that heterogeneous psychiatric disorders can be diagnosed robustly, accurately and rapidly in an automatic fashion. So far, it has been a big challenge to find reliable neuroimaging-based biomarkers for the diagnostic classification of schizophrenic individuals. Compared with previous multi-site studies (Cheng et al., 2015b; Rozycki et al., 2017; Skåtun et al., 2017), the current study developed a discriminant deep learning method, yielding an improvement (>5.0%) of average accuracy both in the multi-site pooling classification and leave-site-out transfer classification. From this view of point, the present study may mark an important breakthrough by enhancing the capabilities of psychiatrists by bringing deep learning to the task of diagnosing schizophrenia across sites. The promising classification results may derive from the following aspects: First, the powerful deep DANS network was able to improve the binary-class feature learning in the large multi-site sample, ensuring that the learned features are most discriminating between the two populations and are independent of the choice of imaging sites. We additionally tested the conventional SAN network (without the discriminant item in the cost function) and obtained significantly lower accuracies in the multi-site pooling classification (Fig. 2A), which suggests that the optimized discriminant item in the cost function at the pre-training stage helped improve classification performance. Though the improvement of the accuracies of the DANS was not significant in the leave-site-out transfer classification relative to the conventional SAN, but the DANS network obtained higher accuracies in most sites. Second, increasing the training sample size may improve classification performance of the deep DANS network. For example, the accuracies of ten-folds were higher than those of five-folds in multi-site pooling classification (Table 2 and S1). Third, multiple atlases enriched the discriminating fcMRI features (Fig. 2). The accuracies obtained from multi-atlas fusion were higher than those from a single atlas, suggesting that classification performance could be improved if more subtle brain atlases were included. In this study, promising accuracies were also obtained in the leave-site-out transfer classification. Such experiments may be of significance in clinical practice because the results indicate the independence of the classification models with respect to imaging sites. Perhaps because the classification model learned no knowledge from the testing site, the accuracies of leave-site-out transfer classification were relatively lower than the accuracies of multi-site pooling classification, and some sites had different sensitivity/specificity. However, the current results suggest that the proposed discriminant deep learning method with fcMRI may provide a promising investigative tool for diagnostic classification of individuals with schizophrenia across independent imaging sites, and may be the first step to build neuroimaging-based discriminative models to predict onset of schizophrenia in a high-risk sample or to differentiate schizophrenia from other disorders with clinical overlap.

In the classification of schizophrenic patients, the majority of discriminating functional correlations were related to the DN, FPN, and vATN. Schizophrenic patients consistently display deficits in a multitude of cognitive domains(Sheffield and Barch, 2016).The activity within the DN has been linked to task-related and spontaneous internally-guided processes spanning autobiographical memory retrieval and mentalizing (Addis et al., 2007; Saxe and Kanwisher, 2003). The DN anomaly has been consistently demonstrated in schizophrenia in previous studies (Bluhm et al., 2007; Camchong et al., 2011; Garrity et al., 2007; Wang et al., 2015; Whitfield-Gabrieli et al., 2009). In the current study, the abnormal functional connectivity related to DN regions including the left caudate, PCC, and bilateral temporal lobes may be associated with episodic memory deficits in patients with schizophrenia and also auditory

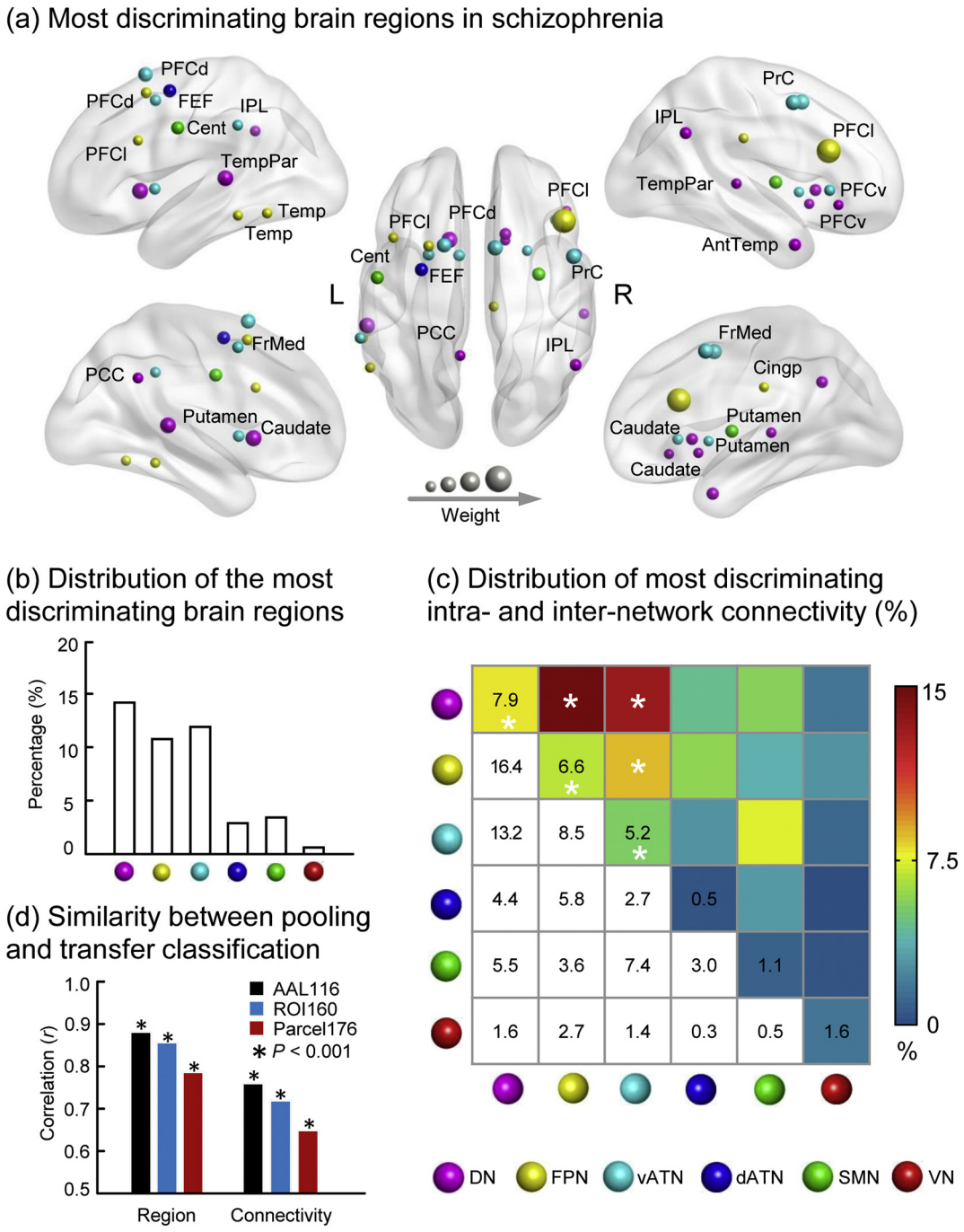

Fig. 4. Most discriminating brain regions and functional connectivity in schizophrenia. (A) Most discriminating brain regions in schizophrenia. (B) Percentages of the most discriminating brain regions of each network. (C) Percentages of most discriminating functional connectivity within and between brain networks. (D) The region weights ($R = 0{\cdot}88$, $0{\cdot}86$, and $0{\cdot}79$ for the AAL template, 160 ROIs, and 17-network parcellation, respectively, $P < 0{\cdot}001$) and connectivity weights ($R = 0{\cdot}76$, $0{\cdot}72$, and $0{\cdot}65$ for the AAL template, 160 ROIs, and 17-network parcellation, respectively, $P < 0{\cdot}001$) obtained in the leave-site-out transfer classification were quite similar with those in the ten-fold multi-site pooling classification. Regions are color-coded by category and are scaled with their discriminative power. DN, default network; FPN, frontoparietal control network; vATN, ventral attention network; dATN, dorsal attention network; SMN, somatomotor network; VN, visual network; L/R, left/right.

hallucinations (Mondino et al., 2016; Vercammen et al., 2010; Wang et al., 2014b). Poorer working memory ability has also been consistently observed in schizophrenia and is often attributed to abnormal functioning of the FPN, especially PFCl-related connectivity alterations (Eryilmaz et al., 2016; Nielsen et al., 2017; Wu et al., 2017). Recent evidence reveals that deficient DN suppression and altered dynamics between DN and FPN are also involved in the neuropathology of working memory deficit in schizophrenia (Pu et al., 2016; Whitfield-Gabrieli et al., 2009). In addition, it was observed that the functional circuit across the putamen, dorsal and medial frontal areas (especially the PFCd and FrMed) within the salience network (vATN) was altered in schizophrenia. The finding of an abnormal salience network could be explained by the dopamine hypothesis in schizophrenia to some extent (Hietala et al., 1995; Reith et al., 1994), i.e., dysregulated dopamine transmission may cause improper assignment of salience to irrelevant stimulus externally or internally, leading to delusional thought and hallucinations, respectively (Braff, 1993; Cohen and Servan-Schreiber, 1992; Kapur, 2003). Thus, salience-related functional connectivity changes together with

the DN and FPN alterations that underlie impaired attention may represent important factors in the positive symptoms of schizophrenia (Camchong et al., 2011; He et al., 2013; Unschuld et al., 2014).

Another significant observation was that the cortical-striatal-cerebellar circuit exhibited great weights. It was notable that the cortical-striatal, striatal-cerebellar, and cortical-cerebellar functional connectivity features were learned by the DANS network, implying dysregulation of the cortical-striatal-cerebellar circuit in schizophrenia. Yoon et al. (2013) observed a link between impaired prefrontal-basal ganglia functional connectivity and the level of psychosis, and Sarpal et al. (2015) found a negative relationship between reduction in psychosis and functional connectivity of striatal regions. In addition, altered connections between the frontal regions and caudate were associated with executive functioning impairments in schizophrenia (Morey et al., 2005; Repovs et al., 2011; Tu et al., 2012; Yoon et al., 2013). Previous studies have also demonstrated an important role of the cerebellum in schizophrenia (Shen et al., 2010; Shinn et al., 2015; Wang et al., 2014a; Yu et al., 2013a; Yu et al., 2013b). In the current study, the deep learning method discovered an integrated cortical-striatal-cerebellar circuit based on multi-site fMRI data, which suggests that dysregulation of the circuit may be a common pathway linking the pathogenesis of cognitive deficits and psychosis in schizophrenia.

The current study posed several limitations. First, accurate classification may benefit from homogeneous datasets by standardizing MRI scanners and scanning parameters. Second, this study is limited by potential confounding effects from medication and possible long duration of illness. Due to a limited sample size and a lack of individualized clinical information including medication and illness duration for some resources, it is important to examine the influence of the two conditions in the future. Third, the fusion of multi-modal neuroimaging evidence such as structural abnormality is necessary as a synthesized biomarker for more reliable clinical diagnosis of this complex disorder (Sui et al., 2015). Fourth, validation experiments are essential if the transfer classification model is applied to a clinical population at a new imaging site. Fifth, data preprocessing may be critical for multivariate pattern analysis, and the impacts of each preprocessing step for the final performance in the deep learning of brain imaging may need to be investigated. In summary, the machine learning-based diagnostic classification of multiple neuropsychiatric disorders with a large multi-site imaging dataset is of great significance for clinical practice and may be an important future direction.

Funding

National Science Foundation of China (61722313, 61503397, 61420106001, 61773391, 31571149, and 81571309), Fok Ying Tung Education Foundation (161057), National Clinical Research Center on Mental Disorders (2015BAI13B02), and Key Research and Development Program of Shaanxi Province (2017ZDXM-SF-047).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None.

Author Contributions

LLZ and DH designed research, HW, PH, XC, ZL, HY, KW and QT performed research; LLZ, PH, and HS contributed new reagents/analytic tools; LLZ, HW, and BY analyzed data, and LLZ, HW, BY, WP, HS, and DH wrote the initial draft, and all authors contributed to, read, and approved the final manuscript.

Acknowledgments

This work was supported by the National Science Foundation of China (61722313, 61503397, 61420106001, 61773391, and 81571309), the Fok Ying Tung Education Foundation (161057), the National Clinical Research Center on Mental Disorders (2015BAI13B02), and the Key Research and Development Program of Shaanxi Province (2017ZDXM-SF-047).

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2018.03.017.

References

Addis, D.R., et al., 2007. Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. Neuropsychologia 45, 1363–1377.
Andreasen, N.C., et al., 1999. Defining the phenotype of schizophrenia: cognitive dysmetria and its neural mechanisms. Biol. Psychiatry 46, 908–920.
APA, 2013. Diagnostic and Statistical Manual of Mental Disorders. Fifth edition. American Psychiatric Press, Washington, DC.
Arbabshirani, M.R., et al., 2013. Classification of schizophrenia patients based on resting-state functional network connectivity. Front. Neurosci. 7. https://doi.org/10.3389/fnins.2013.00133.
Arbabshirani, M.R., et al., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145, 137–165.
Bluhm, R.L., et al., 2007. Spontaneous low-frequency fluctuations in the BOLD signal in schizophrenic patients: anomalies in the default network. Schizophr. Bull. 33, 1004–1012.
Braff, D.L., 1993. Information processing and attention dysfunctions in schizophrenia. Schizophr. Bull. 19, 233–259.
Buckner, R.L., et al., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 2322–2345.
Camchong, J., et al., 2011. Altered functional and anatomical connectivity in schizophrenia. Schizophr. Bull. 37, 640–650.
Cheng, H., et al., 2015a. Nodal centrality of functional network in the differentiation of schizophrenia. Schizophr. Res. 168, 345–352.
Cheng, W., et al., 2015b. Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry. NPJ Schizophrenia 1, 15016.
Choi, E.Y., et al., 2012. The organization of the human striatum estimated by intrinsic functional connectivity. J. Neurophysiol. 108, 2242–2263.
Ciric, R., et al., 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. NeuroImage 154, 174–187.
Cohen, J.D., Servan-Schreiber, D., 1992. Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. Psychol. Rev. 99, 45–77.
Dosenbach, N.U.F., et al., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361.
Drysdale, A.T., et al., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23, 28–38.
Eryilmaz, H., et al., 2016. Disrupted working memory circuitry in schizophrenia: disentangling fMRI markers of Core pathology vs other aspects of impaired performance. Neuropsychopharmacology 41, 2411–2420.
First, M.B., et al., 1996. Structured clinical interview for DSM-IV axis I disorder-patients edition (SCID-I/P). Biometrics Research Department. New York State Psychiatric Institute, New York.
Friston, K.J., Frith, C.D., 1995. Schizophrenia: a disconnection syndrome? Clin. Neurosci. 3, 89–97.
Garrity, A.G., et al., 2007. Aberrant "default mode" functional connectivity in schizophrenia. Am. J. Psychiatry 164, 450–457.
Guo, X.Y., et al., 2017. Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. Front. Neurosci. 11. https://doi.org/10.3389/fnins.2017.00460.
Hazlett, H.C., et al., 2017. Early brain development in infants at high risk for autism spectrum disorder. Nature 542, 348–351.
He, Z., et al., 2013. Aberrant intrinsic brain activity and cognitive deficit in first-episode treatment-naive patients with schizophrenia. Psychol. Med. 43, 769–780.
Hietala, J., et al., 1995. Presynaptic dopamine function in striatum of neuroleptic-naive schizophrenic patients. Lancet 346, 1130–1131.
Kapur, S., 2003. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. Am. J. Psychiatry 160, 13–23.
Kawahara, J., et al., 2017. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage 146, 1038–1049.
Kim, J., et al., 2016. 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 124, 127–146.
Lecun, Y., et al., 2015. Deep learning. Nature 521, 436–444.
Li, F., et al., 2014. Robust deep learning for improved classification of AD/MCI patients. International Workshop on Machine Learning in Medical Imaging, pp. 240–247.
Martino, F.D., et al., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58.
Mikolas, P., et al., 2016. Connectivity of the anterior insula differentiates participants with first-episode schizophrenia spectrum disorders from controls: a machine-learning study. Psychol. Med. 46, 2695–2704.
Min, R., et al., 2014. Multi-atlas based representations for Alzheimer's disease diagnosis. Hum. Brain Mapp. 35, 5052–5070.
Mondino, M., et al., 2016. Effects of Fronto-temporal transcranial direct current stimulation on auditory verbal hallucinations and resting-state functional connectivity of the left Temporo-parietal junction in patients with schizophrenia. Schizophr. Bull. 42, 318–326.
Morey, R.A., et al., 2005. Imaging Frontostriatal function in ultra-high-risk, early, and chronic schizophrenia during executive processing. Arch. Gen. Psychiatry 62, 254–262.
Nielsen, J.D., et al., 2017. Working memory modulation of Frontoparietal network connectivity in first-episode schizophrenia. Cereb. Cortex 27, 3832–3841.
Poldrack, R.A., et al., 2016. A phenome-wide examination of neural and cognitive function. Sci. Data 3, 160110.
Power, J.D., et al., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage 59, 2142–2154.
Power, J.D., et al., 2015. Recent progress and outstanding issues in motion correction in resting state fMRI. NeuroImage 105, 536–551.
Pu, W., et al., 2016. Failed cooperative, but not competitive, interaction between large-scale brain networks impairs working memory in schizophrenia. Psychol. Med. 46, 1211–1224.
Reith, J., et al., 1994. Elevated dopa decarboxylase activity in living brain of patients with psychosis. Proc. Natl. Acad. Sci. U. S. A. 91, 11651–11654.
Repovs, G., et al., 2011. Brain network connectivity in individuals with schizophrenia and their siblings. Biol. Psychiatry 69, 967–973.
Rozycki, M., et al., 2017. Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. Schizophr. Bull. https://doi.org/10.1093/schbul/sbx137.
Sarpal, D.K., et al., 2015. Antipsychotic treatment and functional connectivity of the striatum in first-episode schizophrenia. JAMA Psychiatry 72, 5–13.
Satterthwaite, T.D., et al., 2013. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. NeuroImage 64, 240–256.
Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". NeuroImage 19, 1835–1842.
Sheffield, J.M., Barch, D.M., 2016. Cognition and resting-state functional connectivity in schizophrenia. Neurosci. Biobehav. Rev. 61, 108–120.
Shen, H., et al., 2010. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. NeuroImage 49, 3110–3121.
Shinn, A.K., et al., 2015. Aberrant cerebellar connectivity in motor and association networks in schizophrenia. Front. Hum. Neurosci. 9:134. https://doi.org/10.3389/fnhum.2015.00134.
Skåtun, K.C., et al., 2017. Consistent functional connectivity alterations in schizophrenia Spectrum disorder: a multisite study. Schizophr. Bull. 43, 914–924.
Sui, J., et al., 2015. In search of multimodal neuroimaging biomarkers of cognitive deficits in schzophrenia. Biol. Psychiatry 78, 794–804.
Suk, H.-I., et al., 2013. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859.
Sun, Y., et al., 2013. Hybrid deep learning for face verification. IEEE Trans. Pattern Anal. Mach. Intell. 38, 1997–2009.
Tu, P.C., et al., 2012. Cortico-striatal disconnection within the cingulo-opercular network in schizophrenia revealed by intrinsic functional connectivity analysis: a resting fMRI study. NeuroImage 59, 238–247.
Tzourio-Mazoyer, N., et al., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289.
Unschuld, P.G., et al., 2014. Prefrontal brain network connectivity indicates degree of both schizophrenia risk and cognitive dysfunction. Schizophr. Bull. 40, 653–664.
Vercammen, A., et al., 2010. Auditory hallucinations in schizophrenia are associated with reduced functional connectivity of the Temporo-parietal area. Biol. Psychiatry 67, 912–918.
Vieira, S., et al., 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci. Biobehav. Rev. 74, 58–75.
Wang, L., et al., 2014a. Disruptive changes of cerebellar functional connectivity with the default mode network in schizophrenia. Schizophr. Res. 160, 67–72.
Wang, X., et al., 2014b. Disrupted resting-state functional connectivity in minimally treated chronic schizophrenia. Schizophr. Res. 156, 150–156.
Wang, H., et al., 2015. Evidence of a dissociation pattern in default mode subnetwork functional connectivity in schizophrenia. Sci. Rep. 5, 14655.
Whitfield-Gabrieli, S., et al., 2009. Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. Proc. Natl. Acad. Sci. U. S. A. 106, 1279–1284.
Wu, X.J., et al., 2017. Functional network connectivity alterations in schizophrenia and depression. Psychiatry Res. Neuroimaging 263, 113–120.
Yan, C.-G., et al., 2013. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. NeuroImage 80, 246–262.
Yeo, B.T.T., et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165.
Yoon, J.H., et al., 2013. Impaired prefrontal-basal ganglia functional connectivity and substantia Nigra hyperactivity in schizophrenia. Biol. Psychiatry 74, 122–129.
Yu, Y., et al., 2013a. Convergent and divergent functional connectivity patterns in schizophrenia and depression. PLoS One 8, e68250.
Yu, Y., et al., 2013b. Functional connectivity-based signatures of schizophrenia revealed by multiclass pattern analysis of resting-state fMRI from schizophrenic patients and their healthy siblings. Biomed. Eng. Online 12, 10.
Zeng, L.-L., et al., 2012. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain 135, 1498–1507.
Zeng, L.-L., et al., 2014a. Unsupervised classification of major depression using functional connectivity MRI. Hum. Brain Mapp. 35, 1630–1641.
Zeng, L.-L., et al., 2014b. Neurobiological basis of head motion in brain imaging. Proc. Natl. Acad. Sci. U. S. A. 111, 6058–6062.
Zhao, Y., et al., 2017. Constructing fine-granularity functional brain network atlases via deep convolutional autoencoder. Med. Image Anal. 42, 200–211.