**Supplementary materials:**

## Participants

The initial sample contained 690 schizophrenia patients and 619 healthy subjects. The data have been partially described elsewhere (1). All the patients had a diagnosis of schizophrenia confirmed by trained psychiatrists using the Structured Clinical Interview for DSM-IV-TR Axis I Disorders (SCID-I, patient edition). The exclusion criteria were a history of somatic or neurologic disorders, serious medical illness, substance dependence, pregnancy, electroconvulsive therapy within the last six months, or a diagnosis of any other axis I disorder. All patients were receiving antipsychotic monotherapy, and none were taking antidepressants or mood stabilizers. Medication dosage was converted to chlorpromazine equivalents. Forty percent of our sample was experiencing their first episode, and the remaining patients had experienced a relapse of schizophrenia. The Positive and Negative Syndrome Scale (PANSS) was used to assess the positive, negative, and general psychopathology symptoms in the patients. All the recruited patients were in a state of acute psychosis and had a PANSS total scale higher than or equal to 60. The healthy controls, who had no current or previous axis I psychiatric disorders, were recruited from the local community near each site through advertisements. None of the HCs had any personal history of psychotic illness nor any family history of psychosis in their first-, second-, or third-degree relatives. All the participants were Han Chinese in origin, right-handed, and had no contraindications to MRI scanning. All images were carefully reviewed by four examiners, and those with artifacts such as motion, ghosting, low signal to noise ratio, or insufficient gray/white matter contrast were excluded. After extensive quality checking of the brain imaging data, 662 patients and 613 HCs were included in the analysis.

**T1-weighted scans processing**

All the T1-weighted images were processed using Statistical Parametric Mapping (SPM8, Wellcome Department of Imaging Neuroscience, London, UK; <http://www.fil.ion.ucl.ac.uk/spm/>) and Voxel-Based Morphometry toolbox version 8 (VBM8). Briefly, the images were bias-corrected and segmented into different tissues including GM, WM, and CSF images. The tissue images were then spatially normalized and resampled to a resolution of 1.5 × 1.5 × 1.5 mm3. To preserve the regional volumetric information, the segmented and normalized images were modulated by the Jacobian determinants of the transformation matrices derived from the normalization process (2). Finally, the modulated GM and WM partitions were smoothed with an 8 mm full width at half maximum Gaussian kernel for the subsequent statistical analysis. Smoothing enables the GM and WM volumes to be more normally distributed and compensates for inaccuracies in spatial normalization (3).

**Feature preprocessing**

First, each feature vector was *z* scored within each site. Since we used leave-one-set-out cross validation, all 8 sites were *z* scored within each site. *Z* score normalization can be used to reduce the range variations caused by scanners. For each feature vector *x*, the *z* score was calculated by:



whereis the mean and *s* is the SD of all the samples at each site (i.e., HC+SZ).

Second, we estimated the age and gender effects within the pooled HCs samples in the training set using a general linear model and applied the coefficients to the other participants in the training and test sets. Specifically, for the pooled HC samples in the training set we estimated β1 and β2 using the following formula:



where *Xage*, *Xgender* are the age and gender vectors for all the HC samples in the training set, and *Y* denotes the volumetric features, and ε denotes the error term.

The estimated β1 and β2 were then applied to all samples (HC+SZ) in the training and test sets separately to obtain the age and gender regressed training features and test features. This strategy has been demonstrated to be effective for removing the influence of demographic effects on the distinction between patients and controls (4). This was done voxel-wise.

Supplementary Fig. 8 shows the distribution of raw, *z* scored, and age and gender regressed data for the HCs at each site. Note that the age and gender regression was performed eight times because we performed eight leave-one-site-out experiments and calculated the total values of the overlapped and feature selected voxels for eight validations (i.e., 12666 voxels) to plot the distributions for the raw, *z* scored, and age and gender regressed data. The harmonization performance was evaluated using analyses of variance (ANOVAs), which were conducted using the Pingouin package (v.0.3.11, <https://pingouin-stats.org/index.html>). The violin plots were conducted using Python 3.8.5.

As shown in Supplementary Table 9, no differences were found for the main effects of sites after *z* score normalization and age and gender regression.

**Fisher score feature ranking**

The Fisher score is a criterion by which to quantify the discrimination between a feature and its clinical label (5). Given training instances *xi*, *i* = 1, …, *l*, the Fisher score of the *j*th feature is defined as:



where *n*+ and *n*- are the number of schizophrenia (positive) and healthy control (HC, negative) individuals, respectively; are the average of the *j*th features for the total sample, schizophrenia sample, and HC sample, respectively; is the *j*th feature of the *i*th schizophrenia/HC sample. The numerator denotes the inter-class variance, and the denominator denotes the sum of the variance within each class. The larger a feature’s Fisher score the more discriminative it is.

**DNN training setup**

The deep learning parameters momentum and learning rates were set to 0.9 and 0.1, respectively. ReLU was chosen as the activation function, and logistic regression was the classifier. The weights of each layer were initialized tothe standard Gaussian initialization of 0 mean and 0.01 standard deviation for weights and a constant initialization of 0 for biases. We optimized our algorithm for 150 epochs with a batch size of 50 and a momentum of 0.9. The learning rate started at 0.001 and was divided by 2 every 30 epochs. The weight decay term was fixed to 0.0001. We used a dropout with a fixed rate of 0.5 and batch normalization to enable the global statistics derived from all the training samples to be used to normalize every mini-batch of test data. The DNN models were trained by Pytorch (http://pytorch.org/). All the experiments were performed using a server with Intel(R) Xeon(R) CPU (3.60GHz), 32GB DDR3, and TITAN XP (Tesla) GPU (12G).

**Comparisons with SVM**

To compare SVM with DNN, support vector machines (SVMs) were performed in the automated classification procedure. SVM classifiers have been widely used in the field of neuroscience (6). SVMs select a small number of critical boundary samples from each class and build a linear discrimination function with the maximal margin. In the present study, SVM classifiers were implemented using the scikit-learn package in Python (7), which is based on the LIBSVM toolbox (8). The same feature selection method was performed to reduce the feature dimensions and avoid overfitting. We tested the performance of the Gaussian radial basis function (RBF) kernel and linear SVMs. The parameters *C* (a constant determining the trade-off between training error and model flatness) and ɣ (Gaussian kernel width) for RBF kernels and *C* for linear SVMs were optimized via cross-validation on the training data. We optimized parameters *C* and ɣ via cross validation on the training set for RBF and for linear SVMs via grid search (i.e., *C* = 2−5, 2−3, …, and 215 and γ = 2−15, 2−13, …, and 23 for RBF SVM, *C* = 2−5, 2−3, …, and 215 for linear SVM, respectively). We found the best performance was achieved when ɣ =2-15 and *C* = 23 using RBF kernels. We observed that a linear SVM when *C* = 2-5 had marginally inferior performance compared with the RBF kernel SVM (linear SVM BAC = 76.30% and AUC = 0.824 vs. RBF SVM BAC = 76.07% and AUC = 0.844). Therefore, we reported the results of the RBF kernel SVM in the manuscript.

The voxel probability maps of volumetric contributions to schizophrenia generated by SVM were based on the learned weight vector of a linear SVM. For a linear SVM with a -dimensional weight vector, the output for a -dimensional data can be written as , where and denote the weight and bias terms of SVM, respectively. A linear SVM creates a hyperplane that uses support vectors to maximize the distance between the two classes, which is determined by and . represents the vector which is orthogonal to the hyperplane, in which the absolute values of the coefficients in relation to each feature can then be used to determine feature contribution for the classification task. The final contribution of each feature was averaged across all the participants from the eight experiments and normalized between 0 and 1. A higher value indicates a greater discriminative ability for the classification of schizophrenia.

**Comparisons with CNN**

CNNs are special types of neural networks that require all the high-dimensional neuroimaging image data as the input. The convolutional operations in the CNNs enhance their ability to contextualize spatial information. We constructed 3D ResNet-18 models to test the performance of the CNNs.

**Volumetric abnormalities contributing to the DNN models**

To identify the consistent changes related to schizophrenia across the eight sites, voxel probability maps of the GM, WM, and CSF were generated with the value at each voxel indicating the contributions of the eight experiments to the DNN model. Because of the nested non-linear activation functions and numerous parameters, the interpretability of the DNN model was challenging. We used a layer-wise relevance propagation (LRP) (9) to back-propagate the final two-dimensional classification scores.

Specifically, for the th network layer in a DNN that consisted of layers, we defined as the output of a neuron in layer , as the output of a neuron in the next layer . The can be calculated by

where is the network weight of layer , is a non-linear activation function. For each neuron at layer , we defined the propagation of relevance from layer to layer as

After defining the relevance between two adjacent network layers, we could back propagate the relevance layer by layer and finally get , representing the relevance between eachinput feature and the final prediction of the network.

By implementing a min-max normalization on the absolute value of the vector , which scales the values between 0 and 1, we could get , a vector with length M, which represents the contribution of the input features. M is the number of input features. The final contribution of each feature was averaged across all the participants from the eight experiments and normalized between 0 and 1. A higher value indicates a greater discriminative ability for the classification of schizophrenia.

**ROI-based contributions to the DNN and meta-analysis**

To identify the regions of interest (ROIs) that contribute to the DNN models, we calculated the mean probability values for each ROI in the Brainnetome Atlas (10). Similarly, the average *T* statistical values were also calculated for the ROIs. Supplementary Table 10 shows the ROIs with more than 150 voxels (~4ml in 3×3×3 mm3) in a probability map and a *T* map. We observed the thalamus contributed the most for the schizophrenia classification and in the statistical analysis. We observed the thalamus contributed the most for schizophrenia classification and statistical analysis, suggesting a crucial role for the thalamus in schizophrenia. In addition to the thalamus, the insula, precentral gyrus, inferior parietal lobule, middle frontal gyrus, orbital gyrus, inferior frontal gyrus, superior frontal gyrus, superior temporal gyrus, cingulate gyrus, middle temporal gyrus, and hippocampus have been identified in both of these methods. The basal ganglia and fusiform gyrus contributed to the DNN models, and the postcentral gyrus and lateral occipital cortex were detected in the meta-analysis, suggesting that the two approaches may complement each other.

**GAF analysis**

To control the confounding effect of the global assessment of functioning (GAF) scale, additional experiments were performed in which we included GAF as a covariate in a classification procedure. We found that the classification performance was comparable between the corrected and uncorrected GAF (Supplementary Table 4), indicating that what we found were schizophrenia-specific signatures in the brain volumes rather than a GAF-related brain signature and that a poor GAF is likely to be a result of impaired brain structures.

**References**

1. Li A, Zalesky A, Yue W, Howes O, Yan H, Liu Y, et al. A neuroimaging biomarker for striatal dysfunction in schizophrenia. Nat Med. 2020; 26(4): 558-65.

2. Good CD, Johnsrude IS, Ashburner J, Henson RN, Fristen K, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. In: 5th IEEE EMBS International Summer School on Biomedical Imaging: 16 IEEE, 2002.

3. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage. 2000; 11(6): 805-21.

4. Rozycki M, Satterthwaite TD, Koutsouleris N, Erus G, Doshi J, Wolf DH, et al. Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and Within Individuals. Schizophr Bull. 2017.

5. Chang Y-W, Lin C-J. Feature Ranking Using Linear SVM. In: JMLR Workshop and Conference Proceedings: Causation and Prediction Challenge: 53-64. IEEE Press, 2008.

6. Cui Y, Wen W, Lipnicki DM, Beg MF, Jin JS, Luo S, et al. Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach. NeuroImage. 2012; 59(2): 1209-17.

7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12(Oct): 2825-30.

8. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2(3): 27:1-:.

9. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015; 10(7): e0130140.

10. Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. Cereb Cortex. 2016; 26(8): 3508-26.

**Supplementary Fig. 1.** Flowchart of the classification procedure using T1-weighted volumetric features and leave-one-site-out validation.

Abbreviations: CSF, cerebrospinal ﬂuid; DNN, deep neural network; GM, gray matter; WM, white matter.



**Supplementary Fig. 2.** Statistical maps displaying gray matter volume reductions in schizophrenia patients compared with healthy controls at eight centers.

The color bar indicates *T* values corrected for false discovery rate with *p* < .05. Abbreviations: GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital

Siemens scanning site; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

****

**Supplementary Fig. 3.** Statistical maps displaying reduced (A) and increased (B) white matter volumes in schizophrenia patients compared with healthy controls at individual centers.

The color bar indicates *T* values corrected for false discovery rate with *p* < .05.

Abbreviations: GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; RWU, Renmin Hospital of Wuhan University.



**Supplementary Fig. 4.** Voxel probability maps of reliable gray matter volumetric contributions to schizophrenia using DNN (A) and SVM (B) classification approaches. A higher value indicates a greater discriminative ability for the classification of schizophrenia patients. We observed the regions identified by two approaches have similar patterns. Note that the values in A and B were not comparable because different methods were used to calculate the contributions of features.



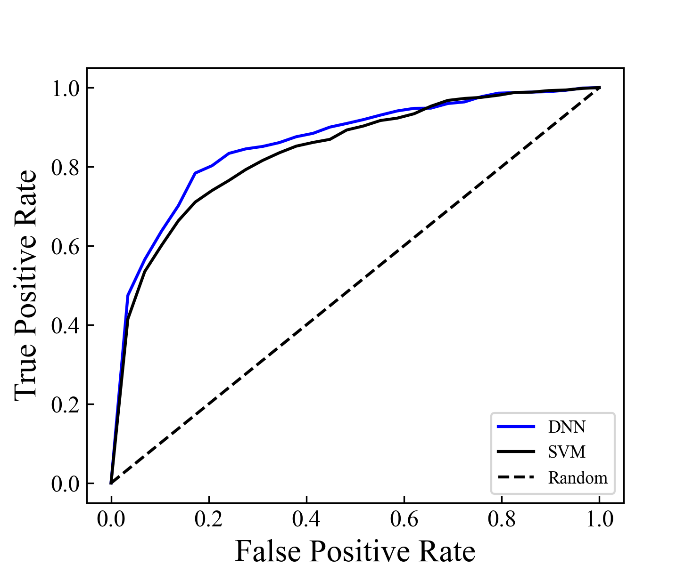
**Supplementary Fig. 5.** Voxel probability maps of reliable white matter (WM) and cerebrospinal ﬂuid (CSF) volumetric contributions to schizophrenia using eight classification experiments. A higher value indicates a greater discriminative ability for the classification of schizophrenia patients.



**Supplementary Fig. 6.** Voxel probability maps of reliable gray matter volumetric contributions to schizophrenia identified by relapse vs. healthy controls (A) and first episode vs. healthy controls (B) classification approaches. A higher value indicates a greater discriminative ability for the classification of schizophrenia patients.

****

**Supplementary Fig. 7.** Receiver operating characteristic **(**ROC) curves for schizophrenia patients vs. healthy controls classification using deep neural network (DNN) and support vector machine (SVM) methods. The classification was performed using the pooled eight sites data and a ten-fold cross validation.





**Supplementary Fig. 8.** Violin plots showing the distribution of raw (A), *z* scored (B), and age and gender regressed (C) data in healthy controls at each site. Boxes display the interval between the 25th and 75th percentiles (q1 and q3); white points indicate median values; red points indicate mean values; whiskers indicate the interval between q1−1.5×(q3−q1) and q3+1.5×(q3−q1), plots were smoothed for visualization with kernels by Scott's rule of thumb.

**Supplementary Table 1.** Demographic characteristics at each site a

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Patients with schizophrenia | | | Healthy controls | | |
| Site | Sample size | Gender  (M : F) | Age (years) b | GAF c | Gender  (M : F) | Age (years) b | GAF c |
| PKUH6 | 198 | 61 : 38 | 26.7 (6.4) | 43.1 (12.3) | 52 : 47 | 25.7 (5.4) | 90.1 (6.5) |
| HLG | 139 | 30 : 49 | 27.7 (6.9) | 65.9 (13.8) | 30 : 30 | 25.6 (5.5) | 99.3 (0.7) |
| XJ | 128 | 45 : 36 | 27.0 (5.8) | 51.6 (11.2) | 28 : 19 | 28.9 (5.2) | 94.6 (2.5) |
| HMS | 175 | 39 : 42 | 26.8 (5.4) | 47.9 (9.0) | 48 : 46 | 28.4 (6.4) | 94.0 (2.0) |
| GB | 193 | 69 : 30 | 27.3 (5.8) | 35.8 (13.7) | 53 : 41 | 25.9 (4.9) | 94.9 (4.6) |
| HMG | 133 | 33 : 23 | 29.6 (7.6) | 44.3 (10.9) | 38 : 39 | 31.3 (7.0) | 94.7 (1.7) |
| RWU | 170 | 36 : 46 | 26.3 (6.1) | 46.3 (8.1) | 45 : 43 | 24.8 (4.6) | 80.2 (2.3) |
| ZMD | 139 | 49 : 36 | 29.4 (7.5) | 43.6 (14.7) | 22 : 32 | 31.5 (5.9) | 93.1 (5.1) |
| Total | 1275 | 362 : 300 | 27.5 (6.5) | 46.7 (13.8) | 316 : 297 | 27.5 (6.1) | 92.0 (6.6) |

Abbreviations: GAF, Global Assessment of Functioning; GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital Siemens scanning site; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

a All sites are matched for gender and age.

b Values are means (SDs).

c Data were missing for 39 patients and 26 controls; all sites showed significant differences between patients and controls (*p* < .001); values are means (SDs).

**Supplementary Table 2.** Clinical characteristics at each site

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | PANSS | | Age at onset of illness (years) | Duration of illness (month) | First episode | CPZ-eq at scan (mg/d)b |
| Positive | Negative |
| PKUH6  (SZ = 99) | 23.6 (4.5) | 18.2 (5.8) | 22.9 (5.9) | 51.0 (50.1) | 49 | 467.7 (215.5) |
| HLG  (SZ = 79) | 26.1 (3.0) | 16.6 (3.3) | 22.9 (6.3) | 64.2 (60.1) | 24 | NA |
| XJ  (SZ = 81) | 22.7 (4.8) | 22.4 (6.5) | 25.0 (6.0) | 26.1 (29.8) | 53 | 452.6 (128.7) |
| HMS  (SZ = 81) | 22.5 (2.8) | 19.1 (5.0) | 22.9 (5.0) | 41.6 (41.2)a | 34 | 339.5 (210.6) |
| GB  (SZ = 99) | 24.3 (4.1) | 22.5 (7.3) | 22.8 (5.9) | 56.5 (49.7) a | 29 | 391.7 (142.6) |
| HMG  (SZ = 56) | 24.6 (3.6) | 24.0 (5.9) | 24.9 (6.5) | 52.4 (68.7) | 22 | 351.8 (182.0) |
| RWU  (SZ = 82) | 23.8 (4.0) | 21.3 (5.9) a | 22.5 (5.8) a | 47.4 (48.3) a | 23 | 558.3 (153.9) |
| ZMD  (SZ = 85) | 24.7 (4.7) | 20.9 (5.7) | 24.7 (6.7) | 56.4 (54.3) | 35 | 600 (0) |
| Total  (SZ = 662) | 24.0 (4.2) | 20.5 (6.2) | 23.5 (6.1) | 49.6 (51.6) | 269 | 411 (204.3) |

a Missing data at the sites ranged from N = 1 to N = 7.

b Data were missing for 362 patients.

Abbreviations: GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital Siemens scanning site; NA, not available; PANSS, Positive and Negative Syndrome Scale; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

**Supplementary Table 3. Scanning sequences and parameters used at each center**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Center | PKUH6, HLG, XJ | HMS | GB | HMG | RWU | ZMD |
| Type of 3T MRI scanner | Vendor | Siemens | Siemens | Philips | General Electric | General Electric | General Electric |
| Model | Trio 3T | Verio 3T | Achieva 3T | Signa HDx 3T | Signa HDxt 3T | Signa HDxt 3T |
| MRI scan sequence and parameter | Sequence | MPRAGE | MPRAGE | 3D T1-TFE | BRAVO | BRAVO | BRAVO |
| TR (ms) | 2530 | 2530 | 8.18 | 8.06 | 7.79 | 6.78 |
| TE (ms) | 3.44 | 2.43 | 3.76 | 3.12 | 3.0 | 2.49 |
| TI (ms) | 1100 | 1100 | 1100 | 1100 | 1100 | 1100 |
| FA (°) | 7 | 7 | 7 | 7 | 7 | 7 |
| Orientation | Sagittal | Sagittal | Sagittal | Sagittal | Sagittal | Sagittal |
| Matrix size | 256\*256\*192 | 256\*256\*192 | 256\*256\*188 | 256\*256\*188 | 256\*256\*188 | 256\*256\*188 |
| Voxel size | 1\*1\*1 | 1\*1\*1 | 1\*1\*1 | 1\*1\*1 | 1\*1\*1 | 1\*1\*1 |

FA, flip angle; GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital Siemens scanning site; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; TE, echo time; TI, inversion time; TR, repetition time; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

**Supplementary Table 4**. Classification performance using a combination of gray matter, white matter, and CSF volumes after correcting for age, gender, and GAF with deep neural networks.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Site | BAC (%) | Sensitivity (%) | Specificity (%) | AUC |
| PKUH6 | 80.30 | 73.74 | 86.87 | 0.860 |
| HLG | 83.44 | 83.54 | 83.33 | 0.862 |
| XJ | 77.57 | 67.90 | 87.23 | 0.801 |
| HMS | 77.65 | 69.14 | 86.17 | 0.809 |
| GB | 81.47 | 76.77 | 86.17 | 0.874 |
| HMG | 82.71 | 87.50 | 77.92 | 0.861 |
| RWU | 83.43 | 80.49 | 86.36 | 0.881 |
| ZMD | 80.32 | 84.71 | 75.93 | 0.838 |
| Average | 80.86 | 77.97 | 83.75 | 0.848 |

Abbreviations: AUC, area under the receiver operating characteristic curve; BAC, balanced accuracy; GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital Siemens scanning site; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

**Supplementary Table 5**. Classification performance using deep neural networks for relapse vs. healthy controls and first episode patients vs. healthy controls.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Relapse vs healthy controls | | | | First episode vs healthy controls | | | |
| Site | BAC  (%) | Sensitivity  (%) | Specificity  (%) | AUC | BAC  (%) | Sensitivity  (%) | Specificity  (%) | AUC |
| PKUH6 | 85.43 | 84.00 | 86.87 | 0.900 | 76.62 | 71.43 | 81.82 | 0.773 |
| HLG | 86.59 | 78.18 | 95.00 | 0.880 | 76.25 | 79.17 | 73.33 | 0.822 |
| XJ | 80.05 | 75.00 | 85.11 | 0.818 | 74.63 | 64.15 | 85.11 | 0.778 |
| HMS | 80.32 | 78.72 | 81.91 | 0.831 | 69.90 | 82.35 | 57.45 | 0.689 |
| GB | 83.81 | 85.71 | 81.91 | 0.910 | 80.74 | 82.76 | 78.72 | 0.832 |
| HMG | 82.30 | 97.06 | 67.53 | 0.849 | 85.06 | 90.91 | 79.22 | 0.909 |
| RWU | 85.28 | 83.05 | 87.50 | 0.869 | 74.18 | 60.87 | 87.50 | 0.696 |
| ZMD | 81.67 | 80.00 | 83.33 | 0.822 | 79.81 | 80.00 | 79.63 | 0.826 |
| Average | **83.18** | **82.72** | **83.65** | **0.860** | 77.15 | 76.45 | 77.85 | 0.791 |

Abbreviations: AUC, area under the receiver operating characteristic curve; BAC, balanced accuracy; GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital Siemens scanning site; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

**Supplementary Table 6.** Classification performancefor schizophrenia patients vs. healthy controls using deep neural network (DNN) and support vector machine (SVM) methods. The classification was performed using the pooled eight sites data and a ten-fold cross validation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC |
| SVM | 78.98 | 76.36 | 81.61 | 0.842 |
| DNN | 82.73 | 80.17 | 85.28 | 0.865 |

**Supplementary Table 7.** Classification performance using only gray matter (GM) and the combination of gray matter and white matter (GM + WM) volumetric features with deep neural networks.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Site | GM | | | | GM + WM | | | |
| BAC (%) | Sensitivity (%) | Specificity (%) | AUC | BAC (%) | Sensitivity (%) | Specificity (%) | AUC |
| PKUH6 | 75.76 | 71.72 | 79.80 | 0.808 | 78.79 | 80.81 | 76.77 | 0.845 |
| HLG | 77.97 | 75.95 | 80.00 | 0.835 | 79.24 | 78.48 | 80.00 | 0.851 |
| XJ | 74.55 | 70.37 | 78.72 | 0.796 | 76.78 | 64.20 | 89.36 | 0.790 |
| HMS | 69.04 | 75.31 | 62.77 | 0.736 | 73.12 | 72.84 | 73.40 | 0.776 |
| GB | 81.87 | 81.82 | 81.91 | 0.874 | 82.90 | 82.83 | 82.98 | 0.874 |
| HMG | 83.04 | 80.36 | 85.71 | 0.846 | 80.76 | 87.50 | 74.03 | 0.874 |
| RWU | 78.51 | 69.51 | 87.50 | 0.865 | 79.05 | 85.37 | 72.73 | 0.844 |
| ZMD | 75.11 | 70.59 | 79.63 | 0.801 | 74.86 | 68.24 | 81.48 | 0.803 |
| Average | 76.98 | 74.45 | 79.51 | 0.820 | 78.19 | 77.53 | 78.84 | 0.832 |

Abbreviations: AUC, area under the receiver operating characteristic curve; GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital Siemens scanning site; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

**Supplementary Table 8.** Classification performance using a combination of gray matter, white matter, and cerebrospinal fluid volumetric features with a convolutional neural network (CNN) for schizophrenia patients vs. healthy controls. We used 3D ResNet-18 as the CNN model for comparison.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Site | CNN | | | |
| BAC (%) | Sensitivity (%) | Specificity (%) | AUC |
| PKUH6 | 76.91 | 86.15 | 67.67 | 0.821 |
| HLG | 76.23 | 78.26 | 74.19 | 0.828 |
| XJ | 63.05 | 55.88 | 70.21 | 0.716 |
| HMS | 66.93 | 70.93 | 62.92 | 0.725 |
| GB | 75.51 | 82.81 | 68.22 | 0.862 |
| HMG | 74.33 | 89.36 | 59.3 | 0.796 |
| RWU | 70.85 | 77.27 | 64.42 | 0.817 |
| ZMD | 68.69 | 69.23 | 68.14 | 0.758 |
| Average | 71.56 | 76.24 | 66.89 | 0.790 |

Abbreviations: AUC, area under the receiver operating characteristic curve; BAC, balanced accuracy; GB, Guangzhou Brain Hospital; HLG, Beijing Huilongguan Hospital; HMG, Henan Mental Hospital GE scanning site; HMS, Henan Mental Hospital Siemens scanning site; PKUH6, Peking University Six Hospital; RWU, Renmin Hospital of Wuhan University; XJ, Xijing Hospital; ZMD, Zhumadian Psychiatric Hospital.

**Supplementary Table 9.** Analysis of variance of main effects of sites for raw, *z* scored and age and gender regressed data.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | ANOVA | |
| *F* | *p* |
|  | Raw | 29.079 | < .001 |
| Stage 1 | *Z* score normalization | 1.834 | .163 |
| Stage 2 | Age and gender regression | 1.825 | .167 |

**Supplementary Table 10.** ROIs that contributed to the classification of schizophrenia based on the contribution of voxel probability maps to the DNN models and the contributions of statistical results in the meta-analysis.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DNN | | | | Meta-analysis | | | |
| Ranking | ROIs | Contributions | Voxel number | Ranking | ROIs | Statistical T | Voxel number |
| **1** | **Thalamus** | **0.77** | 356 | **1** | **Thalamus** | **-8.02** | 296 |
| 2 | Basal ganglia | 0.69 | 274 | 2 | Superior temporal gyrus | -7.90 | 1196 |
| 3 | Insula | 0.67 | 543 | 3 | Insula | -7.20 | 568 |
| 4 | Precentral gyrus | 0.66 | 156 | 4 | Middle temporal gyrus | -7.04 | 440 |
| 5 | Inferior parietal lobule | 0.65 | 167 | 5 | Orbital gyrus | -6.93 | 1060 |
| 6 | Middle frontal gyrus | 0.63 | 220 | 6 | Precentral gyrus | -6.91 | 557 |
| 7 | Orbital gyrus | 0.62 | 728 | 7 | Inferior frontal gyrus | -6.88 | 717 |
| 8 | Inferior frontal gyrus | 0.62 | 244 | 8 | Hippocampus | -6.61 | 169 |
| 9 | Superior frontal gyrus | 0.62 | 255 | 9 | Inferior parietal lobule | -6.54 | 417 |
| 10 | Superior temporal gyrus | 0.61 | 493 | 10 | Postcentral gyrus | -6.52 | 385 |
| 11 | Cingulate gyrus | 0.60 | 603 | 11 | Cingulate gyrus | -6.44 | 773 |
| 12 | Fusiform gyrus | 0.59 | 217 | 12 | Middle frontal gyrus | -6.21 | 708 |
| 13 | Middle temporal gyrus | 0.58 | 237 | 13 | Superior frontal gyrus | -5.93 | 550 |
| 14 | Hippocampus | 0.55 | 254 | 14 | Lateral occipital cortex | -5.85 | 162 |