European Psychiatry

www.cambridge.org/epa

Evaluating the performance of machine learning models for automatic diagnosis of patients with schizophrenia based on a single site dataset of 440 participants

Lung-Hao Lee[1,2,3], Chang-Hao Chen[1,3], Wan-Chen Chang[4,5,6], Po-Lei Lee[1,3], Kuo-Kai Shyu[1,3], Mu-Hong Chen[6,7], Ju-Wei Hsu[6,7], Ya-Mei Bai[6,7,8], Tung-Ping Su[7,8,9] and Pei-Chi Tu[5,6,7,10]*

[1]Department of Electrical Engineering, National Central University, Taoyuan City, Taiwan; [2]Department of Medical Humanities and Education, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan; [3]Pervasive Artificial Intelligence Research (PAIR) Labs, Hsinchu, Taiwan; [4]Department of Biomedical Engineering, National Yang Ming Chiao Tung University, Taipei, Taiwan; [5]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan; [6]Department of Psychiatry, Taipei Veterans General Hospital, Taipei, Taiwan; [7]Department of Psychiatry, Faculty of Medicine, National Yang-Ming Chiao Tung University, Taipei, Taiwan; [8]Institute of Brain Science, National Yang-Ming Chiao Tung University, Taipei, Taiwan; [9]Department of Psychiatry, Cheng Hsin General Hospital, Taipei, Taiwan and [10]Institute of Philosophy of Mind and Cognition, National Yang-Ming Chiao Tung University, Taipei, Taiwan

Research Article

Cite this article: Lee L-H, Chen C-H, Chang W-C, Lee P-L, Shyu K-K, Chen M-H, Hsu J-W, Bai Y-M, Su T-P, Tu P-C (2022). Evaluating the performance of machine learning models for automatic diagnosis of patients with schizophrenia based on a single site dataset of 440 participants. *European Psychiatry*, **65**(1), e1, 1–10
https://doi.org/10.1192/j.eurpsy.2021.2248

Received: 23 July 2021
Revised: 21 October 2021
Accepted: 24 October 2021

Keywords:
Automatic classification; functional connectivity; homogeneous; schizophrenic disorder; support vector machine; training sample size

Author for correspondence:
*Pei-Chi Tu,
E-mail: peichitu@gmail.com

Abstract

Background. Support vector machines (SVMs) based on brain-wise functional connectivity (FC) have been widely adopted for single-subject prediction of patients with schizophrenia, but most of them had small sample size. This study aimed to evaluate the performance of SVMs based on a large single-site dataset and investigate the effects of demographic homogeneity and training sample size on classification accuracy.

Methods. The resting functional Magnetic Resonance Imaging (fMRI) dataset comprised 220 patients with schizophrenia and 220 healthy controls. Brain-wise FCs was calculated for each participant and linear SVMs were developed for automatic classification of patients and controls. First, we evaluated the SVMs based on all participants and homogeneous subsamples of men, women, younger (18–30 years), and older (31–50 years) participants by 10-fold nested cross-validation. Then, we hold out a fixed test set of 40 participants (20 patients and 20 controls) and evaluated the SVMs based on incremental training sample sizes ($N = 40, 80, \ldots, 400$).

Results. We found that the SVMs based on all participants had accuracy of 85.05%. The SVMs based on male, female, young, and older participants yielded accuracy of 84.66, 81.56, 80.50, and 86.13%, respectively. Although the SVMs based on older subsamples had better performance than those based on all participants, they generalized poorly to younger participants (77.24%). For incremental training sizes, the classification accuracy increased stepwise from 72.6 to 83.3%, with >80% accuracy achieved with sample size >240.

Conclusions. The findings indicate that SVMs based on a large dataset yield high classification accuracy and establish models using a large sample size with heterogeneous properties are recommended for single subject prediction of schizophrenia.

© The Author(s), 2022. Published by Cambridge University Press on behalf of the European Psychiatric Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.

Introduction

Schizophrenia is a severe mental disorder accompanied by delusions, hallucinations, and cognitive impairment. It affects nearly 1% of the world's population and the biological underpinnings of schizophrenia have remained elusive despite decades of intensive research [1]. One important theory about its etiology is the dysconnectivity hypothesis, which proposes that the aberration of neural circuits during neural development plays a crucial role in the disease process [2]. The development of functional connectivity (FC) analysis [3,4] provides an optimal tool to test the hypothesis and have consistently identified FC abnormalities in widespread cortical and subcortical structures, including anterior cingulate cortex [5], thalamus [6,7], basal ganglion [8,9], and cerebellum [10] in patients with schizophrenia. With advancements in machine learning in medical imaging, researchers further explored the use of brain-wide FCs based on a specific anatomical or functional parcellation as features for single subject prediction of patients with schizophrenia [11]. Early studies based on a small sample size have reported classification performances of 93.2% (44 participants) [12] and 83% (56 participants) [13]. Several more recent studies have included larger samples. For example, Zhao et al. [14] included 283 participants (135 with schizophrenia and 148 healthy controls) and obtained an accuracy of 71% based on FC features, and Kalmady et al. [15] included

174 participants (81 with drug-naïve schizophrenia and 93 healthy controls) and reported an accuracy of 87% with ensemble learning. Lei et al. [16] evaluated five datasets of 112–192 participants and noted an average accuracy of 82.61% (77.1–87.3%). Together, these preliminary findings indicate that machine learning models are feasible for automated diagnosis of schizophrenia. However, the accuracy range varied substantially across these studies, and the relatively small sample size in many of them limited the application of the models in real-world clinical settings.

Sample size plays a key role in machine learning. A large single-site sample, which automatically covers more variation in disease features, is suggested to be helpful in building more robust classification models for real-world application than are other sample types [17–19]. However, several reviews of previous machine learning studies have observed a negative correlation between sample size and accuracy [18,20] with high accuracy predictions usually limited to studies with small samples [21]. One explanation is that sample size influences the trade-off between accuracy and generalizability [18]. Small, homogeneous samples are able to produce classification models with high accuracy, at the cost of low generalizability, whereas large, heterogeneous samples produce models with better generalizability at the cost of accuracy. However, recent simulation and empirical studies have highlighted the critical role of biased estimations in machine learning studies with small sample sizes. The high accuracy may have been obtained because of inherent large variance of performance in studies with small samples as well as publication bias in reporting significant effects [20] and biased validation processes with a limited sample size [22]. Notably, the popular K-fold cross-validation method produces strongly biased performance estimates with small samples because it does not ensure that the data used to validate the classifier are not part of the data used to train it [22]. Therefore, it was unclear whether high accuracy can be achieved for the identification of patients with schizophrenia based on a large heterogeneous sample with the current approach.

In the present study, we use a large single-site resting fMRI dataset of 220 patients with schizophrenia and 220 healthy controls to develop machine learning models for automatic identification of patients with schizophrenia based on brain-wide FCs and test the hypothesis that support vector machines (SVMs) based on larger, heterogeneous samples can also provide high classification accuracy. SVMs were adopted as machine learning models because they were most commonly used models in recent machine learning studies of psychiatric patients [14, 22–24] and showed superior performance than other traditional models [16]. Also, the sample size is too small for application of deep learning algorithms [25]. To the best our knowledge, this is the largest single site machine learning study of patients with schizophrenia based on brain-wise FCs to date. The data were collected using the same MRI machine and acquisition parameters from 2010 to 2019, thereby minimizing the confounding effects of medical center, MRI machine, and acquisition parameters. Given previous concerns of participant's homogeneity and training sample size on classification accuracy, we also investigated the effect of these two factors on model performance.

Materials and Methods

Participants

The resting fMRI data set included 220 patients with schizophrenia and 220 sex- and age-matched healthy controls. Their demographic characteristics are presented in Table 1. All patients were and recruited from outpatient and inpatient units of the Taipei Veterans General Hospital in Taiwan. Structured clinical interviews based on the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) [26] confirmed the diagnoses and the clinical status of schizophrenic patients was characterized using the Positive and Negative Syndrome Scale (PANSS) [27]. We excluded the participants with the following conditions: (a) substance abuse or dependency in the preceding 6 months; (b) a history of head injury that resulted in sustained loss of consciousness or cognitive sequelae; and (c) neurological illnesses or any other disorder that affects cerebral metabolism. Of these patients, there were seven

Table 1. Demographic and clinical features of the patients and controls in this study.

	SZ ($N = 220$)	HC ($N = 220$)	t/X^2	p
Sex (M/F)	120/100	110/110	0.91	0.39
Age (years)	31.7 ± 9.6	31.8 ± 9.7	−0.11	0.91
Education (years)	13.2 ± 2.9	14.6 ± 2.7	−5.34	<0.001
Age at onset	22.5 ± 6.8			
Length of illness	9.4 ± 8.1			
PANSS total	66.3 ± 15.9			
Positive subscale	15.1 ± 4.9			
Negative subscale	17.4 ± 5.3			
Psychopathology	33.8 ± 8.2			
Medication (% of total patients)				
Antipsychotics	94.1			
Antidepressant	28.6			
Mood stabilizers	41.8			

Abbreviations: F, female; HC, healthy control; M, male; PANSS, Positive and Negative Syndrome Scale for Schizophrenia; SZ, schizophrenia.

with comorbidity of other psychiatric disorders (detailed in Supplementary Table S1). The patients with schizophrenia were under stable treatments with various antipsychotics, antidepressants, and mood stabilizers before participating in the study.

Healthy controls were recruited through advertisements; they were screened by an experienced psychiatrist with the Mini International Neuropsychiatric Inventory Plus, and candidates with a possible major psychiatric illness were excluded. In addition, candidates with a history of first-degree relatives with axis-I disorders, including schizophrenia, major depressive disorder, and bipolar disorder, were excluded.

MRI image acquisition

MRI images were acquired using a 3.0 Tesla GE Discovery 750 whole-body high-speed imaging device with an eight-channel high-resolution brain coil. Head stabilization was achieved through cushioning, and all participants wore earplugs (29 dB rating) to attenuate noise. Automated shimming procedures were performed, and scout images were obtained. Resting-state functional images were collected using a gradient echo T2* weighted sequence (repetition time [TR]/echo time [TE]/flip angle = 2,500 ms/30 ms/90°). Forty-seven contiguous horizontal slices parallel to the intercommissural plane (voxel size: $3.5 \times 3.5 \times 3.5$ mm^3) were acquired and interleaved. These slices covered the cerebellum of each participant. During functional scanning, the participants were instructed to remain awake with their eyes open (each scan lasted 8 min and 24 s across 200 time points). In addition, a high-resolution structural image was acquired in the sagittal plane using a high-resolution sequence (TR = 2,530 ms, echo spacing = 7.25 ms, echo time TE = 3 ms, flip angle = 7°) and an isotropic 1-mm voxel (field of view: 256 × 256).

Quality control

Regarding head motion during image acquisition, we used the method of scrubbing within regression (spike regression) suggested by Yan et al. [28] to minimize the effect of head motion on FC measurement. This method identifies "bad" time points using a threshold of framewise displacement (FD) > 0.2 mm as well as one back and two forward neighbors [29]; each "bad" time point was modelled as a separate regressor in the regression models [30,31]. The detailed parameters of motion correction were also provided in Supplementary Table S2 and there was no significant difference between these two groups.

FC preprocessing

All preprocessing was performed using the Data Processing Assistant for Resting-State fMRI (http://www.restfmri.net), which is based on Statistical Parametric Mapping (http://www.fil.ion.ucl.ac.uk/spm) and the Resting-State fMRI Data Analysis Toolkit (http://www.restfmri.net). The functional scans received slice-timing correction, motion correction, and were normalized to a standard anatomical space (Montreal Neurological Institute). Additional preprocessing steps were used to prepare the data for FC analysis. These were as follows: (a) spatial smoothing using a Gaussian kernel (6-mm full width at half-maximum), (b) temporal filtering ($0.009\,\text{Hz} < f < 0.08\,\text{Hz}$), and (c) removal of spurious or nonspecific sources of variance through regression of the following variables. (a) Six head motion parameters and autoregressive models of motion: six head motion parameters, six head motion parameters one time point before, and the 12 corresponding squared items [32] (Friston 24-parameter model); (b) the mean whole-brain signal; (c) the mean signal within the lateral ventricles; and (d) the mean signal within a white matter mask. The regressors used in the method of scrubbing within regression were also included to minimize the effect of head motion on the measurement of FC. The regression of each of these signals was computed simultaneously, and the residual time course was then retained for the correlation analysis.

Calculation of brain-wise FCs

We chose three parcellations: the automated anatomical labeling atlas version 3 (AAL-3) [33], AAL-2 [34], and Shen's 268 parcellations [35], comprising 166, 120, and 268 regions of interest (ROIs), respectively (Figure 1). The mean time series were derived for each ROI by averaging the time course of all voxels within the ROI. Pearson's correlation coefficients for each pair of ROIs were calculated and z-transformed, yielding three FC matrices (166×166, 120×120, and 268×268) for each participant. By evaluating the model performance based on the three parcellations, we aimed to choose the one yielding best performance for later experiments. AAL-2 and AAL-3 were selected because the automated anatomical atlas [36] was widely used in neuroimaging research. Compared with AAL-2, AAL-3 had a more detailed parcellation of the thalamus (15 parts) and we would like to know if it was helpful for model performance. Shen 268 was selected because it was defined according neuroimaging-based parcellation algorithms based FC data and ever adopted in our previous machine learning study of patients with bipolar disorder [37].

Machine learning model creation, training, and performance evaluations

SVM is a supervised learning model with an associated learning algorithm that analyzes data for classification [38]. The lower triangle elements of the FC matrix were congregated into a vector per subject and regarded as discriminative features to feed into the SVM for classifier training. The hyperparameters C = (1, 10, 100, 1000) and tolerance = (0.001, 0.01, 0.1, 1) of the SVM were optimized using grid search with cross-validation within the training set. To classify an FC matrix in the test set, its classification output was considered as true (positive) for schizophrenia if the probability of class 1 (i.e., diagnosed as schizophrenia) exceeded a predefined threshold (i.e., 0.5).

We used nested 10-fold cross-validation to evaluate SVMs with inner cross-validation for hyperparameter determination and outer cross-validation performance evaluations [39]. The entire dataset was divided into 10 folds that preserved the relative proportion of the 2 classes (i.e., schizophrenia and healthy controls) according to various experimental setups; nine folds were used as the training set, while the remaining fold was used as the test set. Each training set was used to perform inner cross-validation by dividing into 10 folds again, in which 9 folds were used to train the model and the remaining 1 fold was used for performance validation. This process was repeated 10 times until each of the 10 folds had served as the validation set for hyperparameter determination. The outer cross-validation process was also repeated 10 times until each of the 10 folds had served as the test set. The above process was repeated 10 times until each of the 10 folds had served as the test set. We repeated the experiment 100 times to avoid any bias introduced by random sampling in nested 10-fold cross-validation, and the mean ± standard deviation of the performance was reported.

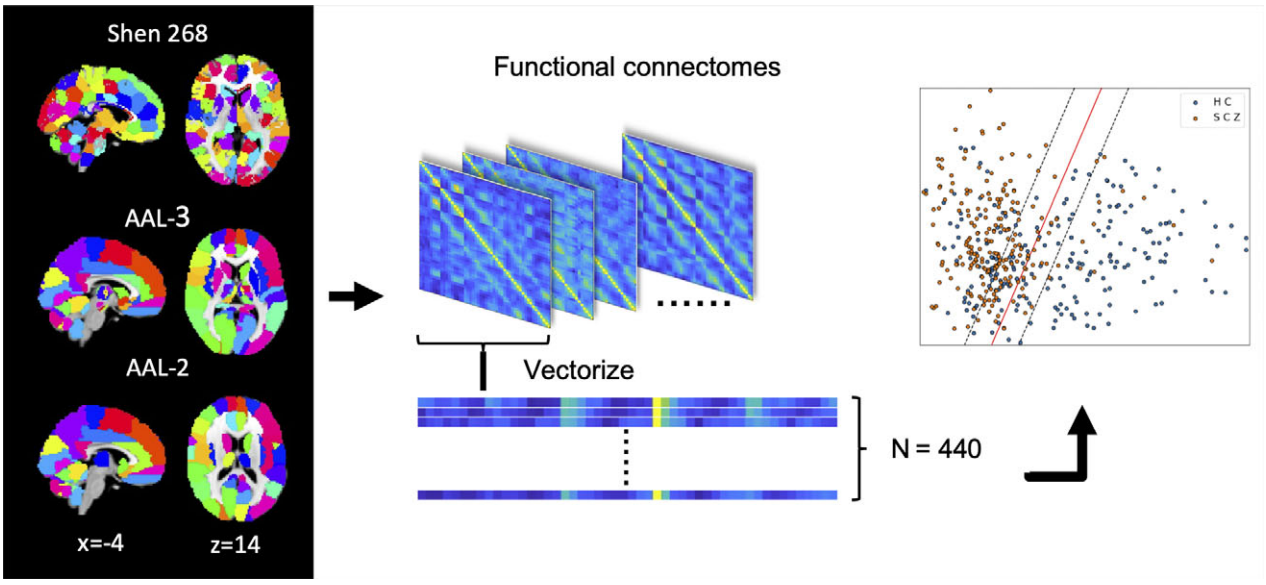

Figure 1. Automatic classifications of schizophrenic patients and healthy controls based on brain-wise functional connectivity. Brain-wise functional connectivity was calculated for each participant according to three different parcellations and linear support vector machines were developed and evaluated for performance. AAL-3 = the automated anatomical labeling atlas version 3; AAL-2 = the automated anatomical labeling atlas version 2.

The performances were evaluated by the following metrics: (a) accuracy: this is the fraction of predictions our machine learning model got right; (b) sensitivity (true positive rate): this refers to the proportion of testing instances who received a positive result out of those participants who actually have schizophrenia; (c) specificity (true negative rate): this refers to the proportion of testing instances who received a negative result out of those participants who do not actually have schizophrenia; (d) F1-score: it is the harmonic mean of precision and recall (sensitivity) that take both false positive and false negative into account; and (e) area under the curve (AUC): this provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0 (a model whose predications are totally wrong) to 1 (otherwise).

First, we evaluated the classification performance of SVMs based on brain-wide FC of three parcellations, and the one with the best performance was selected for later experiments.

Next, we investigated whether increasing the homogeneity of the demographic properties of sex and age improved SVM performance. We divided the whole sample by sex (235 men: 120 with schizophrenia and 115 healthy controls; 205 women: 100 with schizophrenia and 105 healthy controls) and age (212 younger adults: 18–30 years, 106 with schizophrenia and 106 healthy controls; 228 older adults: 31–50 years, 114 with schizophrenia and 114 healthy controls), and the SVMs based on the subsamples were evaluated using nested 10-fold cross-validation with 100 random sampling, and the mean ± standard deviation of the performance was reported. We also evaluated the generalizability of these SVMs to the participants with different demographic characteristics. The male- or female-specific SVMs, which were trained by only male or female participants, were used to classify the clinical status of the other subsamples with different sex, that is, female or male, respectively. In the similar way, we applied the SVMs trained by only younger participants to predict the clinical status of older adults and those SVMs trained with only older adults to predict participants with younger adults.

At last, we evaluated the effects of training sample size on SVM performance to understand what number of participants is necessary to have a robust machine learning model. We randomly selected a test set of 40 participants (20 with schizophrenia and 20 healthy controls) and fixed the same test set in each testing group for performance comparisons. For training sample size setups, we started with $N = 40$ (20 with schizophrenia and 20 healthy controls) randomly drawn from the other 400 participants and incrementally increased the 20 patients with schizophrenia and 20 healthy controls until the maximum training set size of $N = 400$ was reached. A model was built from the training set and tested on the test set repetitively until $N = 400$. We conducted 100 repetitions with different random samplings of participants, and the mean ± standard deviation of the performance was reported.

Results

Demographic properties

The participants' demographic data are presented in Table 1. We controlled the age and sex distribution of each group to ensure a balanced study design. Differences in demographic characteristics among the two groups were examined using the chi-square test for categorical variables and the *t* test for continuous variables. The mean ages of the patients with schizophrenia and healthy controls were 31.7 and 31.8 years, respectively. No significant differences were noted in age and sex distribution between the two groups. However, patients with schizophrenia had significantly lower education illnesses than healthy controls.

The performance of SVMs based on three different parcellations

The detailed results of SVM performance are presented in Table 2. The mean accuracy of the SVMs based on AAL-3, AAL-2, and Shen's 268 parcellations were 85.05 ± 0.84%, 84.17 ± 0.88%, and 84.45 ± 0.89%, respectively. The SVMs based on AAL3 had slightly but significantly higher than those based on AAL-2 ($p < 0.01$) or Shen's 268 ($p < 0.01$). Therefore, brain-wide FC based on AAL-3 was adopted for later experiments.

Table 2. The performance of support vector machines based on different parcellations for automatic classifications of patients with schizophrenic disorder and healthy controls.

Different parcellations	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC (%)
AAL-3	85.05 ± 0.84	87.32 ± 1.01	82.78 ± 1.37	85.37 ± 0.79	92.28 ± 0.48
AAL-2	84.17 ± 0.88	85.74 ± 1.09	82.60 ± 1.42	84.38 ± 0.84	92.07 ± 0.50
Shen 268	84.45 ± 0.89	86.21 ± 1.22	82.68 ± 1.40	84.69 ± 0.89	91.97 ± 0.49

Abbreviations: AAL-2, the automated anatomical labeling atlas version 2; AAL-3, the automated anatomical labeling atlas version 3; AUC, area under curve; SVM, support vector machine.

Table 3. The performance of support vector machines based on different homogeneous subsamples for automatic classifications of patients with schizophrenic disorder and healthy controls.

Homogeneous subsamples	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC (%)
Men	84.66 ± 1.07	88.98 ± 1.12	80.34 ± 1.83	85.66 ± 0.97	92.11 ± 0.73
Women	81.56 ± 1.27	86.60 ± 1.72	76.51 ± 2.13	81.97 ± 1.26	90.13 ± 1.00
Younger adults	80.50 ± 1.38	83.71 ± 2.04	77.29 ± 2.14	81.06 ± 1.39	89.52 ± 0.97
Older adults	86.13 ± 0.87	91.51 ± 1.24	80.74 ± 1.36	86.91 ± 0.83	93.79 ± 0.69

Abbreviations: AUC, area under curve.

The effects of demographic homogeneity and training sample size on classification accuracy

The detailed demographic and clinical characteristics of subsamples according to sex or age range were provided in Supplementary Table S3. Sex-specific SVMs had accuracies of 84.66 ± 1.07% for men and 81.56 ± 1.27% for women (Table 3 and Figure 2a). We also evaluated the generalizability of the sex-specific SVMs to the participants of the other sex. The accuracy was 78.20% for predicting female participants by male-specific models and 81.33 ± 1.23% for predicting male participants by female-specific SVMs (Table 4). Thus, the female-specific SVMs, which yielded an accuracy of 81.56% for predicting female participants, generalized well to predict male participants with an accuracy of 81.33%, but not vice versa. Nevertheless, the sex-specific models had worse performances than the SVMs based on participants of both sexes.

Age-specific SVMs yielded accuracies of 80.50 ± 1.38% and 86.13 ± 0.87% for younger and older adults (Table 3 and Figure 2a), respectively. We also evaluated the generalizability of the age-specific SVMs to the participants of the other age range. The accuracies for predicting young participants by using the older adults-specific SVMs and vice versa were 77.24 ± 1.07% and 82.93 ± 1.04%, respectively (Table 4). The younger adults-specific SVMs, which yielded an accuracy of 80.5% for predicting younger participants, generalized well to predict older participants with an accuracy of 82.93%. In the contrary, the older adults-specific SVMs, which yielded an accuracy of 86.13% for predicting older participants, generalized poorly to predict clinical status of young participants with an accuracy of 77.24%.

The relationship between classification accuracy and training sample sizes was shown in Table 5 and Figure 2b. As training sample size increased from 40 to 400, the mean accuracy increased consistently from 72.61 to 83.32% and an average accuracy >81% was achieved after $N > 240$. According to the standard deviations of classification accuracy across 100 times of random sampling, the SVMs based on higher training sample sizes had lower variance in performance, suggesting a higher stability.

The FCs with greatest contributions to single subject classification

The identification of FCs contributing to differentiate patients from control subjects accurately provided a multivariate approach to identify biomarkers, which could lead to clinically useful tools for establishing both diagnosis and prognosis [40]. Therefore, we further analyzed the FCs contributing to classification performance. In each trained SVM, the absolute values of weights for each brain-wise FCs were regarded as feature importance and averaged across all SVMs based on AAL-3 with whole sample of $N = 440$. The top 20 FCs with the highest mean weights were listed in Table 6 and involved distributed cortical and subcortical structures (Figure 3). Among them, the thalamo-cerebellar FC had the highest mean weight and played the most important role in differentiating patients from controls.

Discussions

At present, psychiatric diagnoses are based largely on psychiatric interviews, and brain imaging does not play a vital role. However, the approach of combining imaging and machine learning is appealing and could be immensely useful if it is proven to be a robust means of establishing a psychiatric diagnosis. In this study, we used a large single-site dataset to build SVMs to classify patients with schizophrenia and healthy controls based on brain-wide FC, with an accuracy of 85%. In contrast to recent concerns about the biased estimations of classification performance in studies with small samples [23], the present results may provide a robust estimation of SVMs for automatic diagnosis of patients with schizophrenia based on brain-wise FCs. On the basis of our data, we recommend AAL-3 for the calculation of brain-wide FC because it yielded higher classification accuracy than AAL-2 and Shen's 268. Although the models using more homogenous subsamples of narrower age range (the older adult group) seemed to provide better classification accuracy than the overall model, they had poor generalization to other samples with different demographic properties.

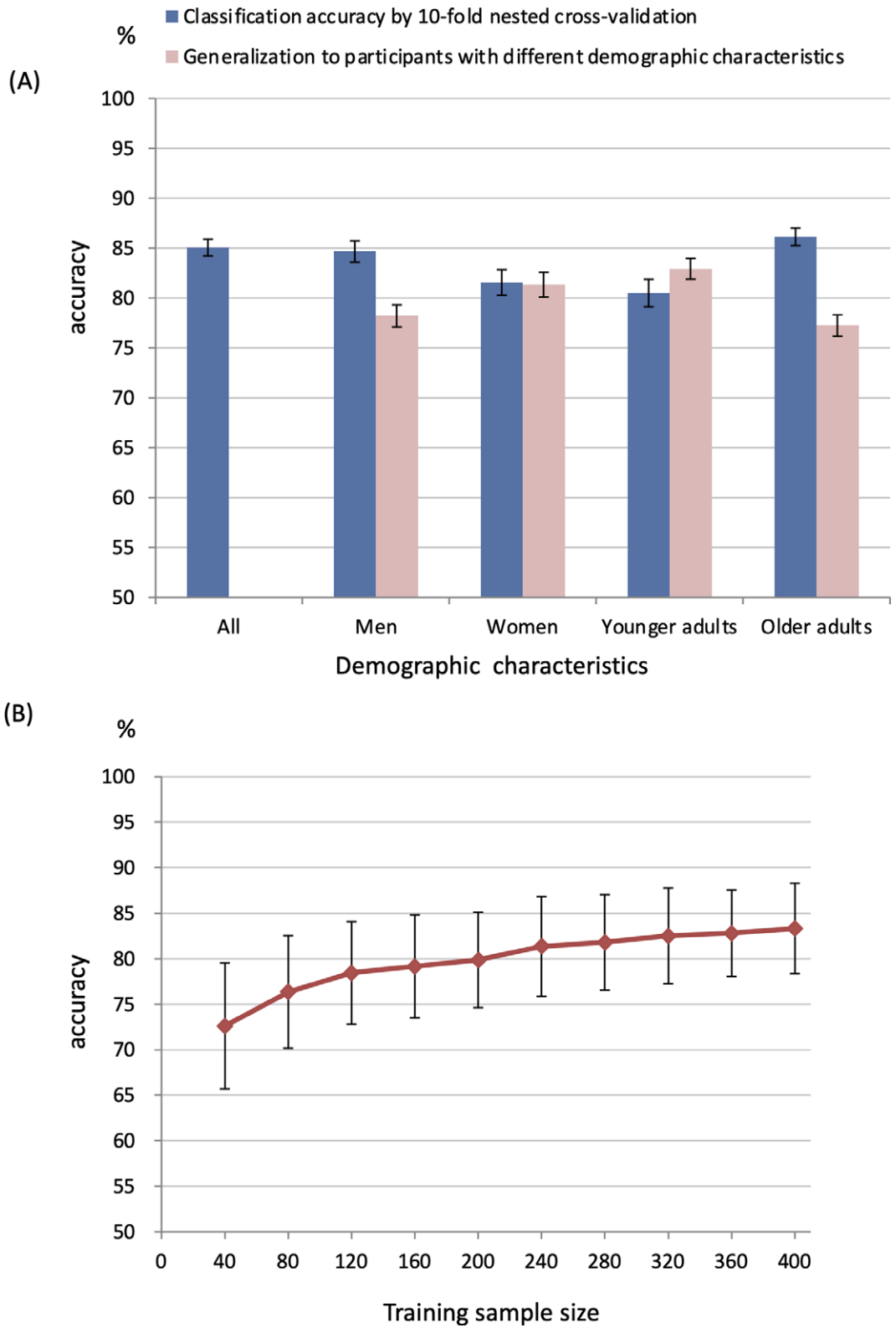

Figure 2. The effects of demographic homogeneity and training sample sizes on support vector machines (SVMs) performance. (a) The classification accuracy of SVMs based on all participants and those based on homogeneous subsamples of men, women, younger, and older participants were demonstrated. The SVMs based on homogeneous subsamples were also applied to the other participants with different demographic properties to understand their generalizability. (b) The classification accuracy of SVMs based on incremental training sample sizes improved consistently from 72.61 to 83.32% and >81% accuracy were achieved after training sample size >240.

Table 4. The performance of generalization of support vector machines to participants with different demographic characteristics for automatic classifications of patients with schizophrenic disorder and healthy controls.

Demographic characteristics	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC (%)
Men[a]	78.20 ± 1.10	84.01 ± 1.45	72.40 ± 1.64	78.84 ± 1.11	87.38 ± 0.66
Women[b]	81.33 ± 1.23	89.02 ± 0.95	73.64 ± 2.38	83.15 ± 0.98	90.65 ± 0.70
Younger adults[c]	82.93 ± 1.04	92.86 ± 1.29	73.01 ± 2.00	84.59 ± 0.95	92.00 ± 0.60
Older adults[d]	77.24 ± 1.07	74.03 ± 1.25	80.46 ± 1.80	76.17 ± 1.10	85.60 ± 0.69

Abbreviations: AUC, area under curve; SVMs, support vector machines.
[a]The classification performance of predicting female participants by male-specific SVMs.
[b]The classification performance of predicting male participants by female-specific SVMs.
[c]The classification performance of predicting older-adult participants by younger-adult-specific SVMs.
[d]The classification performance of predicting younger-adult participants by old-adult-specific SVMs.

Table 5. The performance of support vector machines based on different training sample size for automatic classifications of patients with schizophrenic disorder and healthy controls.

Training sample size	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC (%)
40	72.61 ± 6.92	78.29 ± 13.25	66.92 ± 13.53	73.52 ± 8.45	80.95 ± 7.14
80	76.36 ± 6.17	84.30 ± 8.07	68.42 ± 12.23	78.10 ± 5.40	84.90 ± 6.27
120	78.43 ± 5.63	84.72 ± 7.36	72.14 ± 10.84	79.73 ± 5.02	86.62 ± 5.50
160	79.16 ± 5.67	84.98 ± 7.77	73.33 ± 10.25	80.29 ± 5.24	87.66 ± 5.16
200	79.86 ± 5.24	85.43 ± 7.25	74.28 ± 9.09	80.90 ± 4.97	88.49 ± 4.81
240	81.36 ± 5.48	86.27 ± 6.61	76.46 ± 9.17	82.27 ± 5.06	89.43 ± 4.81
280	81.80 ± 5.25	86.32 ± 6.62	77.27 ± 8.16	82.58 ± 4.95	89.82 ± 4.78
320	82.50 ± 5.26	86.93 ± 6.90	78.08 ± 8.19	83.24 ± 4.98	90.24 ± 4.52
360	82.80 ± 4.75	86.81 ± 6.76	78.79 ± 7.61	83.46 ± 4.58	90.78 ± 4.09
400	83.32 ± 4.97	87.68 ± 6.30	78.95 ± 7.97	84.03 ± 4.66	91.07 ± 4.11

Abbreviations: AUC, area under curve.

We also found that classification accuracy increased with incremental increases in training sample size from 40 to 400, with an accuracy of >81% achieved with $N > 240$. These findings suggest that establishing an SVM based on a large single-site dataset covering varied demographics and disease features may be optimal for the automatic diagnosis of schizophrenia.

Our model had a mean accuracy of 85%, which is slightly better than those reported in recent machine-learning studies based on brain-wide FC: 82.4% [15], 81.74% [16], and 82.61% [41]. Notably, the performance of these SVMs was highly consistent—between 80 and 85%—suggesting that brain-wide FC is a reliable feature for automatic classification of patients with schizophrenic disorder. The reported high accuracy (>90%) in early studies with small samples may have been due to high variability and over-optimistic estimation of accuracy during cross-validation within a small sample. A recent study systemically investigated the issues with structural MRIs of 1,868 patients with major depressive disorder and healthy controls from the international Predictive Analytic Competition [23]. They mimicked the process by which researchers would draw samples of various sizes (N = 4–150) and concluded that a strong risk of misestimation and an accuracy of up to 95% can be observed with sample sizes of 20, mainly due to accuracy overestimation during cross-validation. They recommended using sufficiently large test sets to offset the performance misestimation.

Studies have rarely explored sex- and age-specific machine learning models. One diffusion spectrum imaging study that used a diagnostic index based on whole-brain patterns of altered white matter tract integrity did separate models by sex [42]. The overall prediction accuracy was approximately 84% for men, 82% for women, and 76% for men and women together. The results implied that sex has a significant effect on structural connectivity patterns, and it may be helpful to establish different models for male and female participants to improve prediction performance. In our study, sex-specific SVMs performed worse than those based on both sexes. By contrast, the older adult-specific SVMs had slightly better performance than the SVMs based on all ages, but with poor generalization to younger participants. Therefore, it may be practical to establish SVMs based on participants covering various demographic properties in the clinical setting.

We noted that a higher sample size provided better performance and improved the reliability of the SVMs by decreasing performance variance. Our findings are consistent with a previous simulation study suggesting that a larger sample size may improve model stability [20]. Several studies have also explored the relationship between training sample sizes and classification accuracy, but the results have exhibited some disagreement. A study trained SVMs based on structural MRI features and demonstrated a consistent increase of classification accuracy to approximately 70% with increases in sample size (N = 10, 20, 30, …, 220), and the accuracy appeared not to have reached its maximum. Another resting fMRI study used intersubject correlation in functional connectome as to classify patients with schizophrenia and reported higher performance associated with larger training samples [43]. By contrast, one study evaluated SVMs based on structural MRI to classify patients with major depressive disorder with variable training set

Table 6. The functional connectivity features with greatest contributions to single subject classification of patients with schizophrenia.

Rank	Structure 1	Structure 2	Mean weight
1	Thalamus, medial geniculate (R)	Cerebellum, Lobule VIIB (R)	0.4751
2	Substantia nigra, pars compacta (L)	Inferior parietal gyrus (R)	0.4586
3	Insula (R)	Anterior orbital gyrus (R)	0.4487
4	Thalamus, medial geniculate nucleus (R)	Inferior frontal gyrus, triangular part (R)	0.4446
5	Paracentral lobule (L)	Supplementary motor area (L)	0.4219
6	Red nucleus (R)	Precentral gyrus (L)	0.4186
7	Superior temporal gyrus (R)	Gyrus rectus (R)	0.4186
8	Cerebellum, Lobule VIIB (R)	Superior occipital gyrus (R)	0.4047
9	Substantia nigra, pars reticulate (R)	Insula (L)	0.3951
10	Temporal pole; superior temporal gyrus (R)	Paracentral lobule (R)	0.3917
11	Heschl gyrus (R)	Middle cingulate and paracingulate gyri (R)	0.3869
12	Cerebellum, Crus1 (R)	Superior frontal gyrus, medial orbital (R)	0.3868
13	Thalamus, medial geniculate (R)	Fusiform gyrus (L)	0.3857
14	Paracentral lobule (L)	Rolandic operculum (R)	0.3847
15	Anterior cingulate cortex, supracallosal (L)	Medial orbital gyrus (R)	0.3789
16	Parahippocampal gyrus (L)	Posterior cingulate gyrus (L)	0.3771
17	Thalamus, mediodorsal lateral parvocellular (L)	Thalamus, anteroventral nucleus (R)	0.3761
18	Thalamus, medial geniculate (R)	Rolandic operculum (L)	0.3753
19	Superior temporal gyrus (R)	Superior frontal gyrus, medial orbital (L)	0.3706
20	Thalamus, pulvinar lateral (L)	Amygdala (R)	0.3697

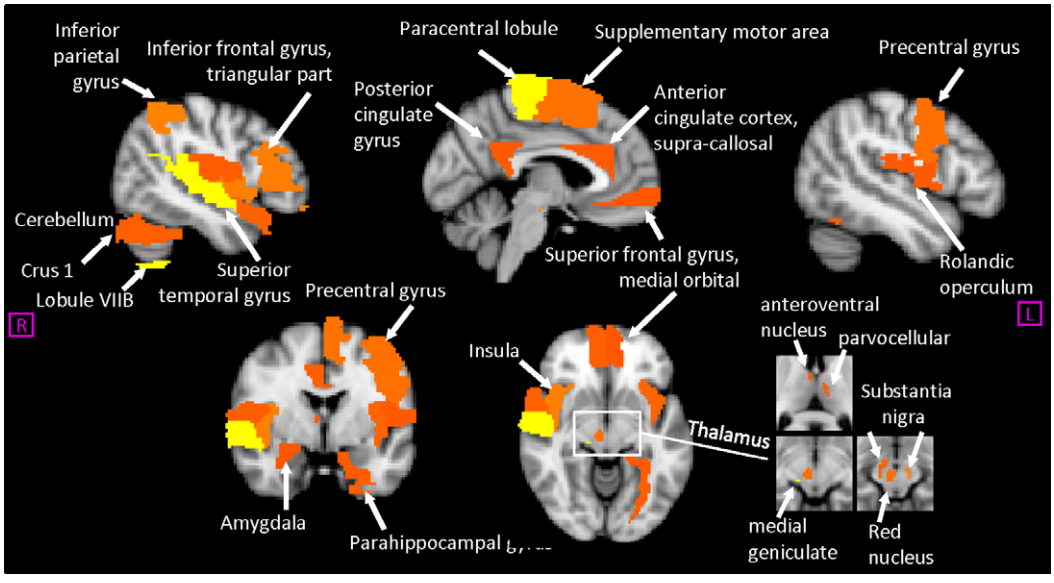

Figure 3. The cortical and subcortical structures involved in the functional connectivities with greatest contributions to single subject classification of patients with schizophrenia.

size $N = 5$–150 and reported no performance improvement for $N > 30$. Thus, the relationship between classification performance and training sample sizes may depend on the features (structural or functional) and complexities of algorithms, and a higher training sample may not generally lead to better performance. Our findings indicate that the performance continued to improve at $N = 400$; we therefore suggest increasing the sample size of the dataset even further with the current models.

The choice of brain parcellations has been rather arbitrary in previous machine learning studies using the brain-wide FC as features. AAL-3 is a recently announced brain parcellation [33]. Compared with AAL-2, AAL-3 has 26 new regions, a new subdivision of the thalamus into 15 parts, and subdivision of the anterior cingulate cortex into subgenual, pregenual, and supracallosal parts. Given the critical role of the thalamocortical FC in schizophrenic disorder [6,7,44], finer parcellations of the thalamus

in AAL-3 may have contributed to its higher performance in our study. Nevertheless, the SVMs based on the three parcellations all had high accuracy, thus supporting the reliability of the models.

Our study had several limitations. First, all our patient groups received treatment with various antipsychotics, so the performance of our models on drug-naïve or first-present patients remains unclear. While diagnostic interviews had the most critical values in first-presentation patients, the factor may limit their clinical applications. Secondly, our machine learning models adopted the features of brain-wise FCs and was limited to only one modality. Previous studies suggested that multi-modal techniques may provide superior performance [24,45], and it should be explored about the performance of SVMs using multi-modal features in the large single site dataset. Finally, our dataset was limited to a single site, precluding the cross-site generalization of our models. Models based on single-site datasets have a much lower performance in cross-site generalization [46,47], likely due to various confounding factors such as different MRI machines, acquisition parameters, and diagnostic processes. Future studies should explore the performance of SVMs using multimodal features in large single-site and multi-site datasets.

Conclusions

In this study, SVMs trained on brain-wide FC retrieved from a large single-site dataset of patients with schizophrenia and healthy controls provided a classification accuracy of 85.05%. The results provided support for the diagnostic values of brain-wise FCs in patients with schizophrenia with the largest single site sample size to date. The feature importance analysis found that the thalamo-cerebellar FC played the most important role in differentiating patients from controls and might serve as potential neural biomarker for patients with schizophrenia. AAL-3 was recommended for brain-wise FC constructions. The use of more homogenous participants with the same sex or age range did not provide better performance and establishing SVMs with a large sample size with heterogeneous properties is a recommend for their applications in single subject prediction of patients with schizophrenia.

Supplementary Materials. To view supplementary material for this article, please visit http://dx.doi.org/10.1192/j.eurpsy.2021.2248.

Acknowledgments. We gratefully thank all the participants who took part in this research and all the research assistants and staff who facilitated their involvement.

Data Availability Statement. The data that support the findings of this study are available from the authors.

Restrictions in relation to potentially person identifiable information apply.

Financial Support. The study was supported by grants from Taipei Veterans General Hospital (V99C1-040, V101C1-159, V104C-039, V105C-119, V106C-091, and V107C-100), Taiwan Ministry of Science and Technology (NSC 99-2628-B-010-021-MY2, MOST 105-2314-B-075-056-MY2, MOST 103-2314-B-075-065-MY2, and MOST 109-2314-B-075-062) and the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3 and MOST 108-2634-F-008-003- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

Authorship contributions. Conceptualization: L.-H.L. and P.-C.T.; Formal analysis: L.-H.L., C.-H.C., W.-C.C., and P.-C.T.; Funding acquisition: L.-H.L., P.-L.L., K.-K.S., J.-W.H., Y.-M.B., and P.-C.T.; Investigation: L.-H.L., C.-H.C., W-C.C., M.-H.C., J.-W.H., Y.-M.B., T.-P.S., and P.-C.T.; Methodology: L.-H.L., C.-H.C., W.-C.C., M.-H.C., and P.-C.T.; Supervision: L.-H.L., P.-L.L., K.-K.S., Y.-M.B., and T.-P.S.; Validation: L.-H.L., C.-H.C., W.-C.C., and P.-C.T.; Project administration: C.-H.C., W.-C.C., and P.-C.T.; Resources: C.-H.C., W.-C.C., P.-L.L., M.-H.C., Y.-M.B., and P.-C.T.; Software: C.-H.C. and W.-C.C.; Data curation: C.-H.C., K.-K.S., J.-W.H., Y.-M.B., and P.-C.T.; Visualization: W.-C.C. and T.-P.S.; Writing – original draft: L.-H.L. and P.-C.T.; Writing – review & editing: L.-H.L., W.-C.C., and P.-C.T.

Conflicts of Interest. The authors declare that they have no conflict of interest.

Reference

[1] Dhindsa RS, Goldstein DB. From genetics to physiology at last. Nature. 2016;530(7589):162–3.
[2] Andreasen NC, Paradiso S, O'Leary DS. "Cognitive dysmetria" as an integrative theory of schizophrenia: a dysfunction in cortical-subcortical-cerebellar circuitry? Schizophr Bull. 1998;24(2):203–18.
[3] Biswal B, Yetkin FZ, Haughton VM, Hyde JS. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med. 1995;34(4):537–41.
[4] Greicius MD, Krasnow B, Reiss AL, Menon V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc Natl Acad Sci U S A. 2003;100(1):253–8.
[5] Tu P, Buckner RL, Zollei L, Dyckman KA, Goff DC, Manoach DS. Reduced functional connectivity in a right-hemisphere network for volitional ocular motor control in schizophrenia. Brain. 2010;133(Pt 2):625–37.
[6] Tu PC, Lee YC, Chen YS, Hsu JW, Li CT, Su TP. Network-specific cortico-thalamic dysconnection in schizophrenia revealed by intrinsic functional connectivity analyses. Schizophr Res. 2015;166(1–3):137–43.
[7] Anticevic A, Haut K, Murray JD, Repovs G, Yang GJ, Diehl C, et al. Association of thalamic dysconnectivity and conversion to psychosis in youth and young adults at elevated clinical risk. JAMA Psychiatry. 2015;72(9):882–91.
[8] Tu PC, Hsieh JC, Li CT, Bai YM, Su TP. Cortico-striatal disconnection within the cingulo-opercular network in schizophrenia revealed by intrinsic functional connectivity analysis: a resting fMRI study. Neuroimage. 2012;59(1):238–47.
[9] Karcher NR, Rogers BP, Woodward ND. Functional connectivity of the striatum in schizophrenia and psychotic bipolar disorder. Biol Psychiatry Cogn Neurosci Neuroimaging. 2019;4(11):956–65.
[10] Chen YL, Tu PC, Lee YC, Chen YS, Li CT, Su TP. Resting-state fMRI mapping of cerebellar functional dysconnections involving multiple large-scale networks in patients with schizophrenia. Schizophr Res. 2013;149(1–3):26–34.
[11] Du Y, Fu Z, Calhoun VD. Classification and prediction of brain disorders using functional connectivity: promising but challenging. Front Neurosci. 2018;12:525.
[12] Tang Y, Wang L, Cao F, Tan L. Identify schizophrenia using resting-state functional connectivity: an exploratory research and analysis. Biomed Eng Online. 2012;11:50.
[13] Arbabshirani MR, Kiehl KA, Pearlson GD, Calhoun VD. Classification of schizophrenia patients based on resting-state functional network connectivity. Front Neurosci. 2013;7:133.
[14] Zhao W, Guo S, Linli Z, Yang AC, Lin C-P, Tsai S-J. Functional, anatomical, and morphological networks highlight the role of basal Ganglia–Thalamus–Cortex circuits in schizophrenia. Schizophr Bull. 2020;46:422–31.
[15] Kalmady SV, Greiner R, Agrawal R, Shivakumar V, Narayanaswamy JC, Brown MRG, et al. Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. NPJ Schizophr. 2019;5(1):2.
[16] Lei D, Pinaya WHL, van Amelsvoort T, Marcelis M, Donohoe G, Mothersill DO, et al. Detecting schizophrenia at the level of the individual: relative diagnostic value of whole-brain images, connectome-wide functional connectivity and graph-based metrics. Psychol Med. 2020;50(11):1852–61.

[17] Krystal JH, State MW. Psychiatric disorders: diagnosis to therapy. Cell. 2014;157(1):201–14.
[18] Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front Psychiatry. 2016;7:50.
[19] Rashid B, Calhoun V. Towards a brain-based predictome of mental illness. Hum Brain Mapp. 2020;41(12):3468–535.
[20] Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage. 2018;180:68–77.
[21] Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage. 2017; 145(Pt B):137–65.
[22] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019;14(11): e0224365.
[23] Flint C, Cearns M, Opel N, Redlich R, Mehler DMA, Emden D, et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. Neuropsychopharmacology. 2021;46(8): 1510–7.
[24] Lei D, Pinaya WHL, Young J, van Amelsvoort T, Marcelis M, Donohoe G, et al. Integrating machining learning and multimodal neuroimaging to detect schizophrenia at the level of the individual. Hum Brain Mapp. 2020; 41(5):1119–35.
[25] Cearns M, Hahn T, Baune BT. Recommendations and future directions for supervised machine learning in psychiatry. Transl Psychiatry. 2019; 9(1):271.
[26] First M, Spitzer R, Gibbon M, Williams J. Structured clinical interview for DSM-IV Axis I disorders, research version, patient edition with psychotic screen (SCID-I/P W/PSY SCREEN). New York: Biometrics Research, New York State Psychiatric Institute; 1997.
[27] Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1987;13(2):261–76.
[28] Yan C-G, Cheung B, Kelly C, Colcombe S, Craddock RC, Di Martino A, et al. A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. Neuroimage. 2013; 76:183–201.
[29] Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Steps toward optimizing motion artifact removal in functional connectivity MRI; a reply to Carp. Neuroimage. 2013;76:439–41. doi:10.1016/j.neuroimage.2012.03.017.
[30] Lemieux L, Salek-Haddadi A, Lund TE, Laufs H, Carmichael D. Modelling large motion events in fMRI studies of patients with epilepsy. Magn Reson Imaging. 2007;25(6):894–901.
[31] Satterthwaite TD, Elliott MA, Gerraty RT, Ruparel K, Loughead J, Calkins ME, et al. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage. 2013;64: 240–56. doi:10.1016/j.neuroimage.2012.08.052.
[32] Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R. Movement-related effects in fMRI time-series. Magn Reson Med. 1996;35(3):346–55.
[33] Rolls ET, Huang C-C, Lin C-P, Feng J, Joliot M. Automated anatomical labelling atlas 3. Neuroimage. 2020;206:116189.
[34] Rolls ET, Joliot M, Tzourio-Mazoyer N. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. Neuroimage. 2015;122:1–5.
[35] Shen X, Tokoglu F, Papademetris X, Constable RT. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage. 2013;82:403–15.
[36] Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002;15(1):273–89.
[37] Chen YL, Tu PC, Huang TH, Bai YM, Su TP, Chen MH, et al. Using minimal-redundant and maximal-relevant whole-brain functional connectivity to classify bipolar disorder. Front Neurosci. 2020;14:563368.
[38] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):Article no. 27.
[39] Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6(1):10.
[40] Gutiérrez-Gómez L, Vohryzek J, Chiêm B, Baumann PS, Conus P, Cuenod KD, et al. Stable biomarker identification for predicting schizophrenia in the human connectome. Neuroimage Clin. 2020;27:102316.
[41] Cui LB, Liu L, Wang HN, Wang LX, Guo F, Xi YB, et al. Disease definition for schizophrenia by functional connectivity using radiomics strategy. Schizophr Bull. 2018;44(5):1053–9.
[42] Chen YJ, Liu CM, Hsu YC, Lo YC, Hwang TJ, Hwu HG, et al. Individualized prediction of schizophrenia based on the whole-brain pattern of altered white matter tract integrity. Hum Brain Mapp. 2018;39(1):575–87.
[43] Ji GJ, Chen X, Bai T, Wang L, Wei Q, Gao Y, et al. Classification of schizophrenia by intersubject correlation in functional connectome. Hum Brain Mapp. 2019;40(8):2347–57.
[44] Woodward ND, Karbasforoushan H, Heckers S. Thalamocortical dysconnectivity in schizophrenia. Am J Psychiatry. 2012;169(10):1092–9.
[45] Lin X, Li W, Dong G, Wang Q, Sun H, Shi J, et al. Characteristics of multimodal brain connectomics in patients with schizophrenia and the unaffected first-degree relatives. Front Cell Dev Biol. 2021;9:631864.
[46] Cai X-L, Xie D-J, Madsen KH, Wang Y-M, Bögemann SA, Cheung EFC, et al. Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data. Hum Brain Mapp. 2020; 41(1):172–84.
[47] Orban P, Dansereau C, Desbois L, Mongeau-Pérusse V, Giguère C, Nguyen H, et al. Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. Schizophr Res. 2018;192:167–71.