scientific reports

OPEN

An integrated machine learning framework for a discriminative analysis of schizophrenia using multi-biological data

Peng-fei Ke[1,3,5], Dong-sheng Xiong[1,3,5], Jia-hui Li[1,3,5], Zhi-lin Pan[1,3,5], Jing Zhou[1,3,5], Shi-jia Li[1,3,5], Jie Song[1,3,5], Xiao-yi Chen[1,3,5], Gui-xiang Li[4,7], Jun Chen[4,7], Xiao-bo Li[8], Yu-ping Ning[2,3], Feng-chun Wu[2,3]✉ & Kai Wu[1,2,3,4,5,6,7,9]✉

Finding effective and objective biomarkers to inform the diagnosis of schizophrenia is of great importance yet remains challenging. Relatively little work has been conducted on multi-biological data for the diagnosis of schizophrenia. In this cross-sectional study, we extracted multiple features from three types of biological data, including gut microbiota data, blood data, and electroencephalogram data. Then, an integrated framework of machine learning consisting of five classifiers, three feature selection algorithms, and four cross validation methods was used to discriminate patients with schizophrenia from healthy controls. Our results show that the support vector machine classifier without feature selection using the input features of multi-biological data achieved the best performance, with an accuracy of 91.7% and an AUC of 96.5% ($p < 0.05$). These results indicate that multi-biological data showed better discriminative capacity for patients with schizophrenia than single biological data. The top 5% discriminative features selected from the optimal model include the gut microbiota features (*Lactobacillus*, *Haemophilus*, and *Prevotella*), the blood features (superoxide dismutase level, monocyte-lymphocyte ratio, and neutrophil count), and the electroencephalogram features (nodal local efficiency, nodal efficiency, and nodal shortest path length in the temporal and frontal-parietal brain areas). The proposed integrated framework may be helpful for understanding the pathophysiology of schizophrenia and developing biomarkers for schizophrenia using multi-biological data.

Finding effective and objective biomarkers to inform the diagnosis of schizophrenia (SZ) is of great importance yet remains challenging[1,2]. Currently, increasing evidence has shown that the gut microbiome, blood and electroencephalogram (EEG) provide abundant clues for the diagnosis of SZ. Recently, several studies have indicated that patients with SZ show an altered gut microbiome composition[3–5], which is significantly associated with the severity of symptoms[3] and human brain structure and function[5]. Moreover, a large number of previous studies indicate alterations in both pro- and anti-inflammatory molecules in the central nervous system, which have also been detected in peripheral blood, and may correlate with SZ symptoms[6–8]. Furthermore, several EEG analyses indicate that patients with SZ show significant alterations in the power of various frequency bands, including the increases in delta and theta waves, the decreases in alpha waves and the increases in beta and gamma waves[9–12].

[1]Department of Biomedical Engineering, School of Material Science and Engineering, South China University of Technology, Guangzhou 510006, Guangdong, China. [2]The Affiliated Brain Hospital of Guangzhou Medical University, Guangzhou Huiai Hospital, Guangzhou 510370, Guangdong, China. [3]Guangdong Engineering Technology Research Center for Translational Medicine of Mental Disorders, Guangzhou 510370, China. [4]Guangdong Engineering Technology Research Center for Diagnosis and Rehabilitation of Dementia, Guangzhou 510500, China. [5]National Engineering Research Center for Tissue Restoration and Reconstruction, South China University of Technology, Guangzhou 510006, China. [6]Key Laboratory of Biomedical Engineering of Guangdong Province, South China University of Technology, Guangzhou 510006, China. [7]National Engineering Research Center for Healthcare Devices, Guangzhou 510500, China. [8]Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ, USA. [9]Department of Nuclear Medicine and Radiology, Institute of Development, Aging and Cancer, Tohoku University, Sendai 980-8575, Japan. ✉email: 13580380071@163.com; kaiwu@scut.edu.cn

However, most of these alterations are observed at the group level with substantial variability among individuals with the same phenotypic diagnosis. Consequently, none of these alterations has proven to have the ability to reliably aid in the differential diagnosis of SZ to date[1,13]. Therefore, studies analyzing how gut microbiota data, blood data and EEG data behave at an individual level are important; for example, this information could be used to better understand the pathology and identify objective biomarkers for the clinical diagnosis of SZ[14].

Recently, pattern recognition based on machine learning has attracted increasing attention, which is well suited for the identification of subtle patterns of information in the data and, consequently, is useful to better predict the diagnosis at an individual level[1,15–17]. Using a variety of biological data, such as gut microbiota data[4,18], blood data[14], and EEG data[12,19–22], along with machine learning techniques, hundreds of studies have been performed in an attempt to achieve the accurate classification of patients with SZ. For instance, a previous study[4] used Boruta variable selection to select the most discriminatory taxa and random forests methods to develop a classifier and predict SZ based on the important microbiota features. A receiver operating characteristic curve analysis revealed that 12 significant microbiota biomarkers were capable of being used as diagnostic factors. A more recent study[14] developed a probabilistic multi-domain data integration model consisting of immune and inflammatory biomarkers in peripheral blood and cognitive biomarkers using machine learning to discriminate patients with SZ from healthy controls (HCs). Another study[20] applied the 1-norm support vector machine (SVM) method based on EEG signals of 64 channels during a working memory task to classify patients with SZ versus healthy controls and an accuracy of 87% was achieved. Despite these advances, previous discriminative studies of SZ have primarily focused on biomarkers extracted from a single type of biological data, which only capture partial information about the human body and therefore influence the resulting classification performance. Currently, increasing evidence has shown that the combination of multimodal imaging data can further improve the classification performance[23–26].

In this study, we collected multi-biological data, including gut microbiota data, blood data and EEG data, from patients with SZ and HCs. An integrated framework of machine learning consisting of multi-biological data, multi-classifiers, multi-feature selection algorithms and multi-cross validation methods, was used to discriminate patients with SZ from HCs. Numerous previous studies have shown that: (1) combining multi-biological data provides more complementary information for discriminative analysis[14,24]; (2) multi-classifiers, multi-feature selection algorithms can better adapt to heterogeneous biological data[27,28]; (3) multi-cross validation methods can test the performance of models more credibly[21]. In this study, we proposed an integrated framework to improve the classification performance and the understanding of biomarker identification for SZ.

Materials and methods

Participants.

The final sample comprised 99 participants, including 49 patients with SZ and 50 HCs. Patients with SZ were recruited from the Affiliated Brain Hospital of Guangzhou Medical University, Guangzhou, and met the diagnostic criteria in the fourth edition of the Diagnostic and Statistical Manual of Mental Disorder-IV-Text Revision (DSM-IV-TR). The psychopathology and symptom severity of the patients were evaluated with the positive and negative syndrome scale (PANSS) and the psychiatric symptoms were steady for > 2 weeks; the PANSS evaluated the rate of change at ≤ 20% over 2 weeks and the total score on the PANSS was ≥ 30. Patients with SZ were excluded if they met any of the following criteria: (1) any other psychiatric axis I disorder meeting the DSM-IV criteria; (2) constipation, diarrhea, diabetes, hypertension, heart disease, thyroid diseases or any somatic diseases; (3) a history of epilepsy, with the exception of febrile convulsions; (4) a history of having received electroconvulsive therapy in the past 6 months; (5) lactating, pregnant, or planning to become pregnant; (6) alcohol dependence; or (7) noncompliant with drug treatment or a lack of legal guardians.

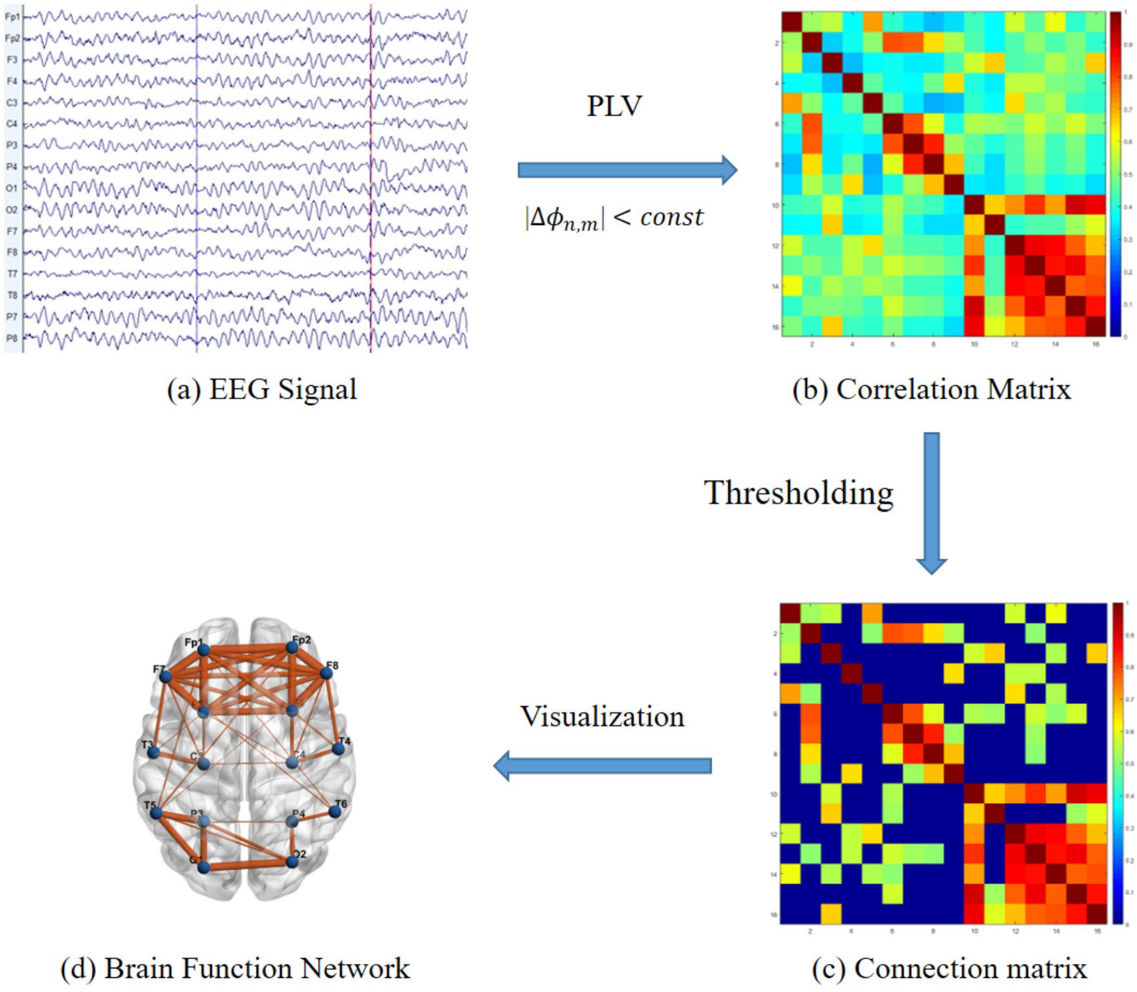
The HCs were solicited from the local community through advertisements and were screened for their family clinical history and a history of mental illness. All healthy subjects had no history of brain disease (such as pain, schizophrenia, concussion, brain trauma, etc.), ocular disease, treatment with psychotropic medication and drug abuse. In addition, the subjects were asked not to drink alcohol, tea, coffee or any other food or drugs that might excite the central nervous system within 48 h before the experiment and that they get enough sleep the night before the test.

The study protocol was approved by the ethics committees of the Affiliated Brain Hospital of Guangzhou Medical University, and written informed consent was obtained from each subject or their legal guardian prior to the study.

Multi-biological data acquisition and preprocessing.

EEG recording and preprocessing. Three minutes of resting EEG data were recorded from 16 scalp electrodes (i.e., Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, and T6) while the participant's eyes were closed according to the International 10/20 System and referenced to electrode Cz (UEA-B, symptom, China). All electrode impedances were maintained at less than 10 kΩ. Signals were amplified and digitized using a sampling rate of 1000 Hz and a 60-Hz low-pass filter during recording.

EEG preprocessing was conducted using MATLAB software (Math Works, Natick, MA). Preprocessing was divided into four steps: electrode positioning, filtering, elimination of bad signal segments and signal frequency band decomposition. A bandpass filter of 0.1–45 Hz was used to improve the quality of the signal. Then, the EEG signal was divided into several epochs of 2 s, and artifact noise, such as eye blinks and movement, was removed by technicians. Finally, the signal was divided into seven frequency subbands by a finite impulse response filter: delta band (1.5–4 Hz), theta band (4–8 Hz), alpha1 band (8–10 Hz), alpha2 band (10–13 Hz), beta1 band (13–20 Hz), beta2 band (20–30 Hz) and gamma band (30–45 Hz).



(a) EEG Signal

(b) Correlation Matrix

(c) Connection matrix

(d) Brain Function Network

Figure 1. Flow chart of the brain network construction of the EEG signal. *EEG* electroencephalogram, *PLV* phase locking value. Figure (**a**) was generated by an EEG processing tool of "EEGLAB" (Version 2019.0, https://sccn.ucsd.edu/eeglab/index.php), based on MATLAB (Version R2018a). Figure (**b**–**d**) were generated by a brain network visualization tool of "BrainNet Viewer" (Version1.62, https://www.nitrc.org/projects/bnv/), based on MATLAB (Version R2018a).

Fecal sample collection and preprocessing. Fresh fecal samples were collected from all subjects and then were stored at – 80 °C until DNA extraction. Two hundred milligrams of each fecal sample were used for DNA extraction.

The DNA extraction method was consistent with our previously published report[3]. Sequencing of the V4 region of the 16S rRNA gene was performed on the Illumina MiSeq platform. The row sequences were processed using QIIME2 (version 2018.6). Forward and reverse reads for each individual sample were demultiplexed, joined and quality filtered. We obtained a total of 4,561,105 joined sequences from these raw paired-end sequences, ranging from 15,449 to 95,651, and the average length of all joined sequences was approximately 251 bp. Then, the DADA2[29] algorithm was used for sequence quality control and feature table construction. After quality filtering, we obtained 4,148,451 high-quality reads, ranging from 13,581 to 90,203 and with a mean of 41,903.5 reads. Then, all the high-quality reads were clustered, 2031 features were obtained, and the frequency per feature ranged from 2 to 533,200, with an average of 3356.9. We used a pretrained naïve Bayes classifier for taxonomic annotation, and this classifier was trained on the Greengenes database (version 13.8). The raw sequence data reported in this article have been deposited in GenBank in the National Center for Biotechnology Information (NCBI) under accession numbers MT545156–MT547172, which are publicly accessible at https://www.ncbi.nlm.nih.gov.

Blood collection and preprocessing. Three milliliters of blood were collected from control subjects and patients by simple venipuncture between 7.00 and 9.00 a.m., after an overnight fasting and tobacco abstinence for more than 12 h. Blood biochemical indicators were detected with an automatic biochemical analyzer.

Multi-biological feature extraction.

EEG feature extraction. In this study, we used the phase-locked value (PLV) method to quantify the functional connectivity (FC) between any two channels of EEG signals, as shown in Fig. 1.

The instantaneous phase $\varnothing(t)$ was calculated from the signal $x(t)$ by using the Hilbert transform:

$$\tilde{x}(t) = h(t) * x(t) = \int_{-\infty}^{+\infty} x(\tau)h(t-\tau)d\tau = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau}d\tau$$

The phase was computed using the following expression:

$$\varnothing_x(t) = \arctan\frac{\tilde{x}(t)}{x(t)}$$

Phase synchronization is defined as the locking of phases of two oscillators:

$$\varnothing(t) = \varnothing_x(t) - \varnothing_y(t)$$

The phase-locking value (PLV) is defined as:

$$\mathrm{R} = \left|\frac{1}{N}\sum_{j=0}^{N-1} e^{i\varnothing(j\Delta t)}\right|$$

where i denotes the imaginary unit, N indicates the total number of samples, and Δt is the bespeak time between the successive samples j from 1 to N – 1.

In this study, a cost threshold strategy was used to analyze global and nodal attributes of the functional brain network (FBN). The cost threshold should be greater than $2 * \ln(\mathrm{N})/N$, where N represented the number of nodes, to ensure that the small-world properties of FBNs were estimable[30]. Moreover, the resulting brain networks should have sparse properties and distinguishable properties compared to the degree-matched random networks. Thus, we selected the small-world regime as a range of cost thresholds ($34\% \le \text{cost} \le 73\%,\ \text{step} = 1\%$). The area under the curve for each attribute was then calculated across the range of cost thresholds and used in a subsequent analysis. Here, all the global and nodal attributes were calculated using the toolbox of BCT[31]. Global attributes include the global clustering coefficient (aCp), shortest path length (aLp), global efficiency (aEg), local efficiency (aEloc), aGamma, aLambda, and aSigma. Nodal attributes include the clustering coefficient (aNCp), nodal shortest path length (aNLp), nodal efficiency (aNe), nodal local efficiency (aNLe), and degree centrality (aDc). In this study, 56 global attributes of an FBN and 640 nodal attributes of 16 nodes were computed from the whole band and seven frequency subbands. Importantly, any features with missing values for any participant were removed. Finally, 48 global attributes and 526 node attributes were used for the subsequent analysis.

Gut microbiota feature extraction. Through gene sequencing technology, microbiota markers from 171 species were obtained from all subjects. Among them, any microbiota marker that was missing in more than 85% of the participants was removed. Ninety-four microbiota markers were removed, and 77 gut microbiota markers were selected for the final analysis.

Blood feature extraction. The white blood cells (WBC) count, neutrophils (NEU) count, lymphocytes (LYM) count, platelets (PLT) count and monocytes (MON) count were recorded from complete blood counts after routine blood tests. Four blood indicators inflammation and immunity, including the neutrophil–lymphocyte ratio (NLR), platelet–lymphocyte ratio (PLR), monocyte–lymphocyte ratio (MLR) and systemic immune inflammation index (SIII), were calculated based on the numbers of the five cell types described above. Moreover, the oxidative stress indicators, including superoxide dismutase (SOD), homocysteine and C-reactive protein (CRP) levels, were also detected in the collected serum. In conclusion, we collected a total of 12 blood features for the final analysis.

Statistical analysis. Statistical analyses were conducted using SPSS software version 22 (IBM). The comparison of the sex distribution between the two groups was performed using the χ^2 test. Comparisons including age and education years between the two groups were performed using a two-tailed two-sample t test. Unless specified otherwise, the significance of all tests was set to $p < 0.05$, or FDR-corrected $p < 0.05$.

Machine learning. We developed an integrated framework of machine learning to discriminate patients with SZ from HCs (Fig. 2). Briefly, the framework involved three phases: the data preparation, model training, and independent model testing.

Data preparation. Data preparation included feature extraction and subject grouping. We extracted three types of biological features from fecal data, blood data, and EEG data, namely, gut microbiota features, blood features, and EEG features, respectively. For the final analysis, Seventy-seven gut microbiota features, 12 blood features, and 574 EEG features were selected for the final analysis. Three types of biological features were used as input features for machine learning, either individually or in combination, to form four input feature sets. At this stage, we randomly split the set of participants into two groups, a training dataset and an independent testing dataset, at a ratio of 3:1. The training dataset was used to train the model parameters, and the independent testing dataset was used to evaluate the performance of the trained model.

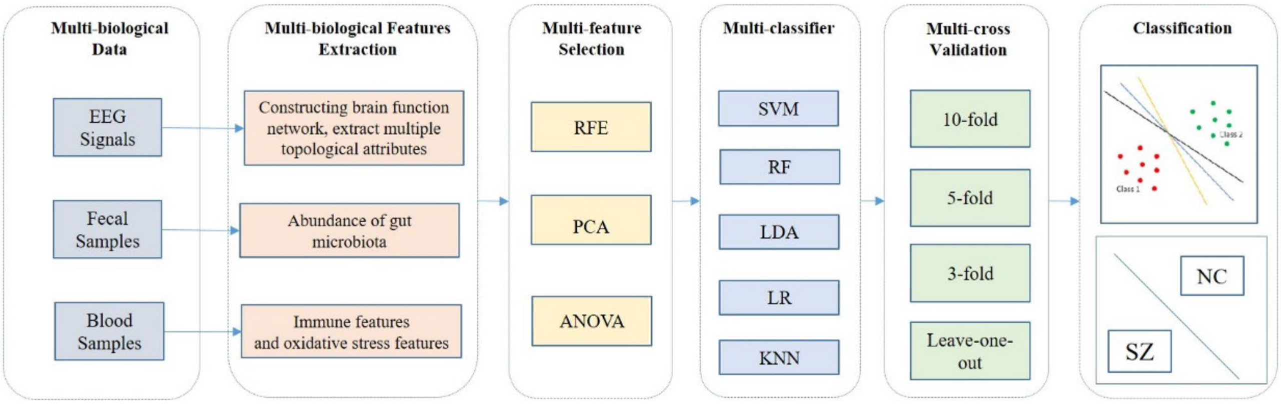


Figure 2. Overview of the proposed integrated machine learning framework for classifying schizophrenia. The proposed integrated machine learning framework for classifying schizophrenia consists of 5 M-methods. (**a**) Multi-biological data were collected from all subjects, including electroencephalogram (EEG) data, fecal data and blood data. (**b**) Multi-biological features were extracted from multi-biological data. (**c**) Multi-feature selection algorithms were used to eliminate redundant features, including recursive feature elimination (RFE), principal component analysis (PCA), and analysis of variance (ANOVA) (**d**) Multi-classifier were used to match heterogeneous biological features including support vector machine (SVM), random forest (RF), linear discriminant analysis (LDA), logistic regression (LR), and k-nearest neighbor (KNN) methods. (**e**) Multi-cross validation methods including tenfold, fivefold, threefold, and leave-one-out methods, were used to evaluate the performance of the trained model.

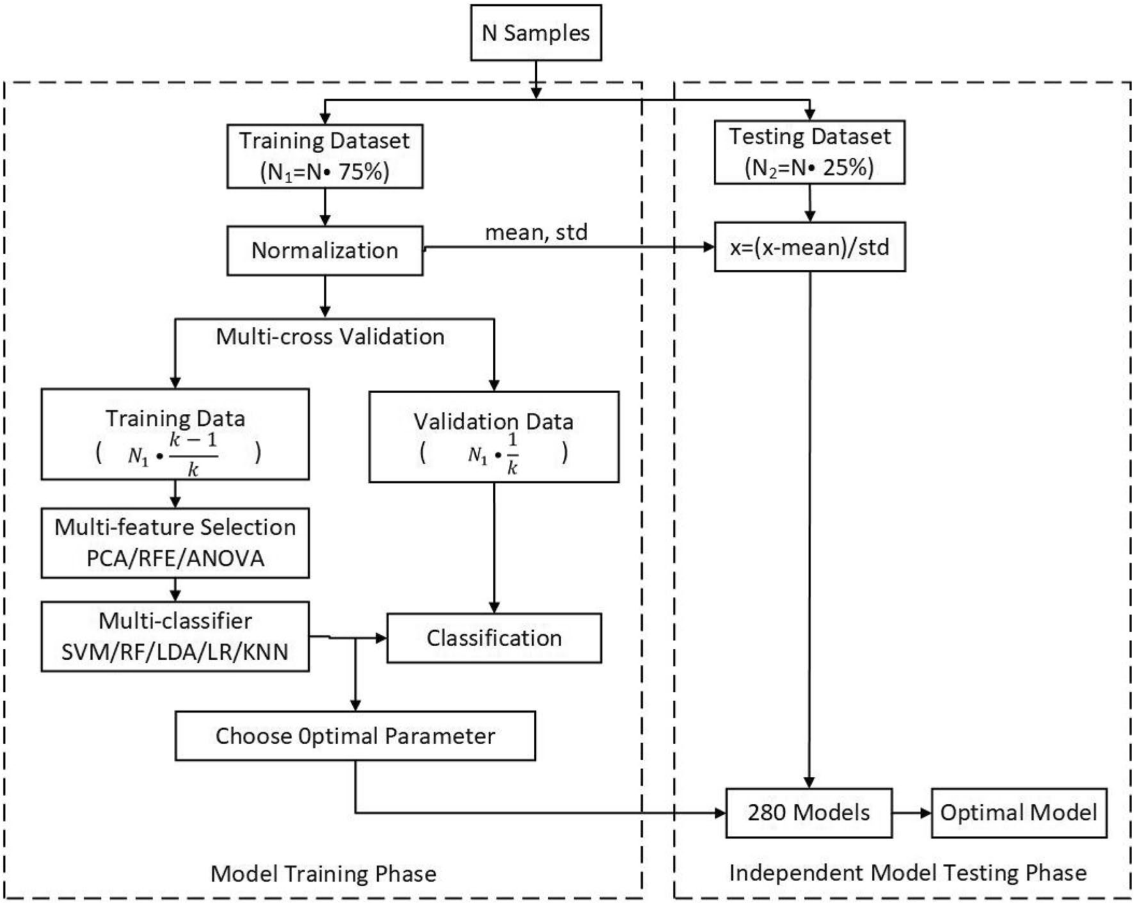

N Samples
Training Dataset (N_1=N• 75%)
Testing Dataset (N_2=N• 25%)
Normalization
mean, std
x=(x-mean)/std
Multi-cross Validation
Training Data ($N_1 \bullet \frac{k-1}{k}$)
Validation Data ($N_1 \bullet \frac{1}{k}$)
Multi-feature Selection PCA/RFE/ANOVA
Multi-classifier SVM/RF/LDA/LR/KNN
Classification
Choose Optimal Parameter
280 Models
Optimal Model
Model Training Phase
Independent Model Testing Phase

Figure 3. Flowchart of the machine learning classification method.

Characteristic	HCs (n = 50)	SZs (n = 49)	*p* value
Age, mean (SD) (years)	41.7 (13.1)	42.1 (12.5)	0.89
Sex, No (%)			
Male	23 (46.0)	24 (49.0)	0.77
Female	27 (54.0)	25 (51.0)	
Education years, mean (SD) (years)	14.2 (3.6)	11.6 (3.4)	< 0.001
PANSS, mean (SD)	NA	58.84 (17.50)	NA

Table 1. Demographic and clinical characteristics used in the analysis. *HCs* healthy controls, *SZs* patients with schizophrenia, *NA* not applicable.

Model training. The specific details of the model training phase and independent model testing phase are shown in Fig. 3. The model training procedures included three steps: multi-feature selection algorithms, multi-classifier, and multi-cross validation methods.

Because some features are less effective, irrelevant, or redundant for classification, and too many features may cause "overfitting", effective feature selection methods can be used to identify the discriminative features and facilitate disease classification and interpretation. Three feature selection algorithms were used on each classifier, including principal component analysis (PCA), recursive feature elimination (RFE) and analysis of variance (ANOVA), to observe the classification effect of the classifier.

A specific classification model directly based on multi-biological data is difficult to build due to heterogeneity. Therefore, the use of several machine learning methods to construct different classification models is meaningful. In this study, we used five different popular classifiers including support vector machine (SVM), random forest (RF), linear discriminant analysis (LDA), logistic regression (LR) and k-nearest neighbor (KNN), to determine the most suitable model and to evaluate classification performance based on single and combined biological features.

Multi-cross validation methods were used to analyze the training set, including tenfold, fivefold, threefold and leave-one-out methods, and to ensure that the sample size was sufficiently large to train the model and prevent overfitting caused by insufficient training. Several combinations of the aforementioned procedures were investigated for optimized data analysis. PCA and RFE feature selection algorithms were unable to be used due to the small dimension of blood features. As a result, 280 models were obtained based on four input feature sets, five classifiers, three feature selection algorithms and four cross validation methods. Model training in the second phase was performed with their application restricted to the training data set.

Independent model testing. In the third phase, we used an independent testing dataset to estimate the generalizability of 280 models arising from the second phase. We utilized the metrics of accuracy, sensitivity and specificity to quantitatively estimate the performance of all the methods mentioned in this study. Moreover, we plotted receiver operating characteristic (ROC) curves and then calculated the area under the curve (AUC) for each classification situation to examine the possibility of correctly discriminating patients with SZ and HCs.

A permutation test was applied to evaluate the statistical significance of the classification results. In our analysis, we disrupted the labels of all samples 1000 times, and the p value was computed as the proportion of accuracies that were no less than the accuracy obtained with the original data. The statistical significance was set to $p < 0.05$. All automatic classification work was performed using NEURO-LEARN (https://github.com/Raniac/NEURO-LEARN[32]), which is a solution for collaborative pattern analysis of neuroimaging data.

Results

Participants.

The resulting data set comprised 99 participants, including 49 patients with SZ (mean [SD] age, 42.06 [12.48] years; 24 [49.0%] males) and 50 HCs (mean [SD] age, 41.70 [13.07] years; 23 [46.0%] males). Significant differences in either age ($t = 0.141$, $p = 0.888$) and sex ($t = 0.294$, $p = 0.769$) were not observed between the patients with SZ and HC group. See Table 1 for a detailed description of other characteristics.

Classification results and analysis.

We used an independent testing dataset to estimate the generalizability of the 280 models. The classification performance of the tenfold cross validation method, fivefold cross validation method, threefold cross validation method, and leave-one-out cross validation method (eTables 1–Table 4 in the "Supplementary S1") was obtained. No significant differences were observed among the results of multi-cross validation methods. Table 2 shows the classification performance of the model obtained using different input features with tenfold cross validation methods. The optimal classification performance was achieved when multi-biological features were combined as input features, with 91.7% accuracy, 91.7% sensitivity, 91.7% specificity, and 96.5% AUC. The performance of the classifier based on multi-biological features was better than that of the classifiers using a single type of biological feature (Fig. 4). In addition, the blood features achieved the best classification performance when using a single type of biological feature, with an accuracy of 83.3% and an AUC of 87.5%. When gut microbiota features, blood features, and EEG features were used as input feature sets alone, the classifiers and feature selection algorithms of the optimal model were inconsistent, potentially due to the heterogeneity of biological data. The SVM, LR and RF classifiers without using any feature selection algorithm displayed better classification performance when using combined features, with AUCs greater than 90%.

Input feature	Feature Selection Method	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p* value[a]
Gut microbiota features (n = 77)	RFE	RF	70.8	58.3	83.3	0.80	0.03
Blood features (n = 12)	None[b]	KNN	83.3	83.3	83.3	0.88	0.010
EEG features (n = 574)	RFE	RF	79.2	83.3	75.0	0.90	0.010
Combined features (n = 663)	None	SVM	91.7	91.7	91.7	0.97	0.010

Table 2. Classification performance of the optimal model including different input features using the integrated machine learning framework (tenfold). *AUC* area under the receiver operating characteristic curve, *RFE* recursive feature elimination, *KNN* k-nearest neighbor, *LR* logistic regression, *RF* random forest, *SVM* support vector machine, *EEG* electroencephalogram. [a]The statistical significance of the permutation test was set to $p < 0.05$. [b]None means no feature selection algorithm was used.

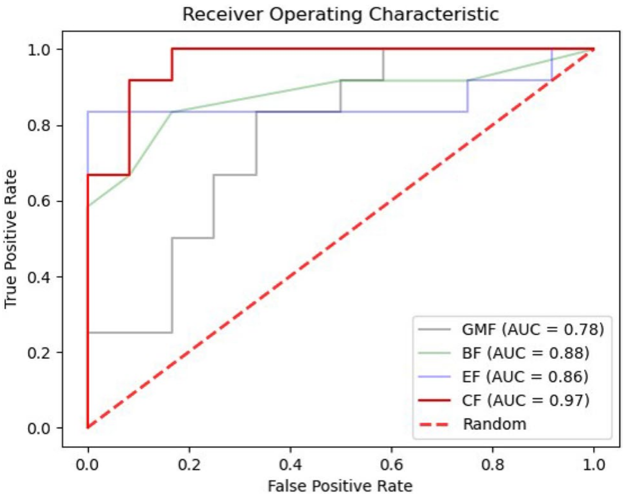


Figure 4. Areas under the receiver operating characteristic curves (AUC) for the best model comparing the gut microbiota features, blood features, electroencephalogram features and the combination of GMV, BF and EF as the input for machine learning. Each curve in the figure represents the ROC curve of the best model using different input features. *GMF* gut microbiota features, *BF* blood features, *EF* electroencephalogram features, *CF* combined features. This figure was generated by "Visual Studio Code" (Version 1.56, https://code.visualstudio.com/).

Discriminative features. In this subsection, the most informative features selected to differentiate the patients with SZ from HCs are reported. We discuss the most discriminative features from the optimal model that were generated when combined features were used. For quantitative analysis, the top 34 (5% of the total number of features) commonly selected features are summarized in Table 3, which shows the top 34 features for classification listed in descending order of their weights, including 14 gut microbiota features, 8 blood features, and 12 EEG features.

Discussion

To the best of our knowledge, this discriminative study of SZ is the first to combine multi-biological data of gut microbiota data, blood data, and EEG data. We developed an integrated framework of machine learning to discriminate patients with SZ from HCs. The main findings of this study are described below. (1) Using a combination of three types of biological features as input features for the classification, the best performance was achieved, with an accuracy of 91.7%, a sensitivity of 91.7%, a specificity of 91.7%, and an AUC of 96.5%. (2) the most discriminative features (top 5%) included gut microbiota features (*Lactobacillus*, *Haemophilus*, and *Prevotella)*, blood features (superoxide dismutase level, monocyte-lymphocyte ratio, and neutrophil count), and EEG features (nodal local efficiency, nodal efficiency, and nodal shortest path length in the temporal and frontal-parietal areas).

In this study, we developed an integrated framework of machine learning using a combination of multi-biological data, which is a promising direction for the identification of biomarkers for the diagnosis, prognosis, and treatment patients with SZ. The comparison of classification performance with existing studies is listed in Table 4. A recent study indicated that the diagnosis of SZ can be predicted with possible clinical utility by a computational machine learning algorithm using the combination of blood and cognitive biomarkers; more

Number	Feature name	Feature type	Number	Feature name	Feature type
1	SOD	Blood	18	*Undefined*[b]	GM
2	MLR	Blood	19	*Anaerostipes*	GM
3	*Lactobacillus*	GM	20	PLT	Blood
4	MON	Blood	21	alpha2_aNLe_P4	EEG
5	*Haemophilus*	GM	22	*Dialister*	GM
6	*Prevotella*	GM	23	beta1_aLambda	EEG
7	NEU	Blood	24	*Slackia*	GM
8	CRP	Blood	25	*Undefined*	GM
9	*Megamonas*	GM	26	*Odoribacter*	GM
10	theta_aNLe_T6[a]	EEG	27	*Ruminococcus*[c]	GM
11	theta_aNe_T6	EEG	28	theta_aDc_FP1	EEG
12	theta__aNCp_T6	EEG	29	alpha2__aNCp_P4	EEG
13	WBC	Blood	30	beta2_aNLe_FP2	EEG
14	NLR	Blood	31	beta2_aDc_O2	EEG
15	*Collinsella*	GM	32	*Gemmiger*	GM
16	gamma_aDc_F7	EEG	33	alpha2_aNLe_T4	EEG
17	*Clostridium*	GM	34	alpha2__aNCp_T4	EEG

Table 3. Top 34 features (5%) showing the most discriminative biomarkers for multi-biological predictions. The top 34 features are listed in the descending order of their weights. *GM* gut microbiota, *EEG* electroencephalogram, *SOD* superoxide dismutase, *MLR* monocyte–lymphocyte ratio, *MON* monocyte, *NEU* neutrophil, *CRP* C-reactive protein, *WBC* white blood cell, *NLR* neutrophil–lymphocyte ratio, *PLT* platelet, *aNLe* nodal local efficiency, *aNe* nodal efficiency, *aNCp* nodal clustering coefficient, *aDc* degree centrality. [a]The EEG features are represented as a_b_c, where a represents the frequency band, b represents brain network attributes, and c represents the electrode channel. [b]Undefined *Lachnospiraceae*. [c]Undefined *Ruminococcaceae.*

References	Sample size	Input feature	Feature selection method	Classifier	Cross validation method	Performance
Shen et al.[4]	SZ = 64 HC = 53	Gut microbiota	Boruta variable selection	RF	None	AUC = 0.837
Brisa et al.[14]	SZ = 58 HC = 123	Blood and cognitive	PLS-DA	LDA	Tenfold	Accuracy = 0.86 AUC = 0.89
Jason et al.[20]	SZ = 40 HC = 12	EEG	None	SVM	None	Accuracy = 0.87 Sensitivity = 0.90 Specificity = 0.77
Sai Krishna Tikka et al.[21]	SZ = 38 HC = 20	EEG	None	SVM	Hold-out	Accuracy = 0.79 Sensitivity = 0.92 Specificity = 0.50 AUC = 0.71
Our best	SZ = 49 HC = 50	Gut microbiota Blood EEG	None	SVM	Tenfold	Accuracy = 0.92 Sensitivity = 0.92 Specificity = 0.92 AUC = 0.97

Table 4. Comparison of classification performance with existing research. *RF* random forest, *PLS-DA* partial least squares discriminant analysis, *LDA* linear discriminant analysis, *EEG* electroencephalogram, *SVM* support vector machine, *AUC* area under the receiver operating characteristic curve.

importantly, the integration of multi-biological data outperforms a single type of biological data[14], consistent with our findings. Interestingly, an early SVM-based prediction of the later development of SZ in a familial high-risk cohort is possible and can be improved by combining schizotypal and neurocognitive features with neuroanatomical variables[33]. In summary, based on the integrated framework of machine learning, the combination of multi-biological data substantially improves the classification performance for patients with SZ. Our results revealed that the features from multiple biological datasets provided complementary information and can help to develop effective and objective biomarkers for the clinical diagnosis of SZ[1].

To date, although numerous discriminative studies of SZ have used either data of blood-based[6,34,35], or neuroimaging data[9,10,26,36,37], few studies have investigated the potential of biomarkers for the diagnosis of SZ using gut microbiota data. Based on accumulating evidence, the gut microbiota bidirectionally communicates with the central nervous system through the microbiome-gut-brain axis (MGBA), thereby influencing brain function and behavior[38,39]. Recently, a few studies have focused on the role of the MGBA in SZ and revealed several alterations in the gut microbiota in patients with SZ[4,40–42]. These reports of an altered gut microbiotas are consistent with the finding from study, for which the most informative features of the gut microbiota include *Lactobacillus*, *Haemophilus*, *Collinsella*, *Clostridium,* and *Prevotella*. Furthermore, Yuan et al.[42] have shown that changes in the gut microbiota and its metabolites may cause neuronal damage. *Lactobacillus* stimulate TNF production; therefore, *Lactobacillus* may induce changes in inflammatory factors that induce SZ[43]. On the other hand, short-chain fatty acids (SCFAs), the primary bacterial metabolites produced, can enter the central nervous system through the blood–brain barrier (BBB)[44]. *Clostridium* is the main source of propionate in the gut, indicating that *Clostridium* may influence the BBB and act on the brain by regulating SCFAs. In addition, *Collinsella* has been shown to produce the proinflammatory cytokine IL-17a and to alter intestinal permeability by promoting the release of neurotransmitters produced by gut microbiota[45], thereby acting on the central nervous system. Above all, these investigations suggested that the gut microbiota may affect the central nervous system by acting on several pathways, providing a physiological basis for validating the use of the gut microbiota as a biomarker in the classification of the two groups.

Among the blood features we extracted, those that contributed the most to the classification included SOD level, MLR, MON count, NEU count, CRP level, WBC cunt and NLR, consistent with previous studies using conventional univariate statistical analysis. Numerous studies and increasing evidence suggest that the oxidative stress contributes to the pathogenesis of SZ, and abnormalities in antioxidant enzymes, including SOD activity, are frequently observed in patients diagnosed with SZ[46–48]. A previous study[49] indicated that SOD activity remained lower in patients with SZ and may be an important indirect biomarker of oxidative stress in individuals with SZ. The present findings provide additional evidence of increased oxidative stress in patients with SZ. Blood inflammatory and immune system abnormalities in patients with SZ have been widely reported, which lead to an increase in levels of inflammatory markers. The NEU count was reported to be increased in patients with chronic SZ[50]. An increased MON count has also been reported in patients with chronic SZ[51,52]. Furthermore, a moderately increased CRP level in patients with SZ compared to HCs has been observed[53–55]. Subjects with SZ have significantly elevated WBC counts. The MLR and NLR have recently been used as indicators of inflammation, and predictors of cardiovascular disease, the leading cause of death in patients with SZ. A recent meta-analysis revealed a significant increase in the NLR in patients with SZ[56]. Elevated MLR and NLR have been observed in patients with SZ, suggesting an increased inflammatory response in individuals with SZ[50]. Our experimental results are consistent with these studies.

Table 3 shows that the EEG features with heavy weight are primarily derived from the delta and alpha2 frequency bands and partly from the beta and gamma frequency bands. Previous investigators observed increases in delta and theta waves, decreases in alpha waves and increases in beta and gamma waves in individuals with SZ[9,10,12,14,57,58]. Moreover, the most prominent change was in the spectral power of the delta wave, which may support the development of a biological marker for diagnosing patients with SZ[9,10,59]. In addition, among these EEG features, node attributes including nodal local efficiency (aNLe), nodal efficiency (aNe), nodal clustering coefficient (aNCp), and degree centrality (aDc), contributed most to classifying patients with SZ. EEG studies have shown a disruption in the small-world attributes of patients with SZ in the resting state, with a lower clustering coefficient and a longer shortest path length[60]. In addition, global and local efficiency are lower in patients with SZ than that in healthy people[61]. The most discriminative EEG features in Table 3 are primarily concentrated in the temporal lobe and partly in the frontal lobe. Abnormalities in temporal and frontal lobe function and structure have been widely reported in patients with SZ[62] The frontal and temporal lobes are primarily associated with higher cognitive functions, among which the temporal lobe is associated with hearing and language functions, which have been confirmed by MRI studies[63]. These results are consistent with previous structural and functional neurological findings.

Limitations

The present study has several limitations. First, since this study employed a cross-sectional design, we cannot infer causality. Some evidence suggests that immune-inflammatory markers are altered from the beginning of SZ, and researchers have broadly accepted that inflammation plays a causal role in SZ. However, from a diagnostic perspective, this finding is irrelevant. A specific marker must only discriminate between two conditions, regardless of whether it is a cause, consequence, or correlate of the pathophysiological process. Second, a significant difference in education years was observed between the two participant groups, although the results remained unchanged when this factor was included as a covariate. Third, the sample size was moderate. A larger independent sample is essential to examine the reproducibility of our findings.

Conclusions

In conclusion, we developed an integrated framework of machine learning and used the combination of multi-biological data to discriminate patients with SZ from HCs, which substantially improved the classification performance. Based on our results, features from multiple biological datasets provide complementary information that aids in providing effective and objective biomarkers to inform the clinical diagnosis of SZ, and our framework is effective at conveying comprehensive and complementary information for the purpose of classification.

Received: 29 January 2021; Accepted: 30 June 2021
Published online: 19 July 2021

References

1. Fernandes, B. S. *et al.* The new field of "precision psychiatry". *BMC Med.* **15**(1), 80 (2017).
2. McCutcheon, R. A., Reis Marques, T. & Howes, O. D. Schizophrenia—An overview. *JAMA Psychiat.* **77**(2), 201–210 (2020).
3. Li, S. *et al.* Altered gut microbiota associated with symptom severity in schizophrenia. *PeerJ* **8**, e9574 (2020).

4. Shen, Y. *et al.* Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia: A cross-sectional study. *Schizophr. Res.* **197**, 470–477 (2018).
5. Li, S., *et al.* The gut microbiome is associated with brain structure and function in schizophrenia. *Sci. Rep.* **11**, 9743 (2021).
6. Chan, M. K. *et al.* Applications of blood-based protein biomarker strategies in the study of psychiatric disorders. *Prog. Neurobiol.* **122**, 45–72 (2014).
7. Tomasik, J., Rahmoune, H., Guest, P., Bahn, S. Neuroimmune biomarkers in schizophrenia. *Schizophr. Res.* **176**, 3–13 (2014).
8. Colpo, G. D., Leboyer, M., Dantzer, R., Trivedi, M. H. & Teixeira, A. L. Immune-based strategies for mood disorders: facts and challenges. *Expert Rev. Neurother.* **18**(2), 139–152 (2018).
9. Alfimova, M. V. & Uvarova, L. G. Changes in EEG spectral power on perception of neutral and emotional words in patients with schizophrenia, their relatives, and healthy subjects from the general population. *Neurosci. Behav. Physiol.* **38**(5), 533–540 (2008).
10. Boutros, N. N. *et al.* The status of spectral EEG abnormality as a diagnostic test for schizophrenia. *Schizophr. Res.* **99**(1–3), 225–237 (2008).
11. Gong, Q. & He, Y. Depression, neuroimaging and connectomics: A selective overview. *Biol. Psychiatry* **77**(3), 223–235 (2015).
12. Kim, J. W. *et al.* Diagnostic utility of quantitative EEG in un-medicated schizophrenia. *Neurosci. Lett.* **589**, 126–131 (2015).
13. Stephan, K. E. *et al.* Charting the landscape of priority problems in psychiatry, part 1: Classification and diagnosis. *Lancet Psychiatry* **3**(1), 77–83 (2016).
14. Fernandes, B. S. *et al.* Precision psychiatry with immunological and cognitive biomarkers: A multi-domain prediction for the diagnosis of bipolar disorder or schizophrenia using machine learning. *Transl. Psychiatry* **10**(1), 162 (2020).
15. Deo, R. C. Machine learning in medicine. *Circulation* **132**(20), 1920–1930 (2015).
16. Chen, Z. *et al.* Detecting abnormal brain regions in schizophrenia using structural MRI via machine learning. *Comput. Intell. Neurosci.* **2020**, 1–13 (2020).
17. Ji, D. *et al.* Machine learning of discriminative gate locations for clinical diagnosis. *Cytometry A* **97**(3), 296–307 (2020).
18. He, Y. *et al.* Gut microbiome and magnetic resonance spectroscopy study of subjects at ultra-high risk for psychosis may support the membrane hypothesis. *Eur. Psychiatry* **53**, 37–45 (2018).
19. Phang, C. R., Noman, F., Hussain, H., Ting, C. M. & Ombao, H. A multi-domain connectome convolutional neural network for identifying schizophrenia from EEG connectivity patterns. *IEEE J. Biomed. Health Inform.* **24**(5), 1333–1343 (2020).
20. Johannesen, J. K., Bi, J., Jiang, R., Kenney, J. G. & Chen, C. A. Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatr. Electrophysiol.* **2**, 3 (2016).
21. Tikka, S. K. *et al.* Artificial intelligence-based classification of schizophrenia: A high density electroencephalographic and support vector machine study. *Indian J. Psychiatry* **62**, 273 (2020).
22. Luo, Y., *et al*. Biomarkers for prediction of schizophrenia: Insights from resting-state EEG microstates. *IEEE Access* **8**, 213078–213093 (2020).
23. Sharpee, T. O. *et al.* 25th annual computational neuroscience meeting: CNS-2016. *BMC Neurosci.* **17**(Suppl 1), 54 (2016).
24. Zhuang, H. *et al.* Multimodal classification of drug-naïve first-episode schizophrenia combining anatomical, diffusion and resting state functional resonance imaging. *Neurosci. Lett.* **705**, 87–93 (2019).
25. Sui, J. *et al.* Combination of resting state fMRI, DTI, and sMRI data to discriminate schizophrenia by N-way MCCA + jICA. *Front. Hum. Neurosci.* **7**, 235 (2013).
26. Li, X. *et al.* Altered topological characteristics of morphological brain network relate to language impairment in high genetic risk subjects and schizophrenia patients. *Schizophr. Res.* **208**, 338–343 (2019).
27. Dai, Z. *et al.* Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). *Neuroimage* **59**(3), 2187–2195 (2012).
28. Donnelly-Kehoe, P. A., Pascariello, G. O. & Gómez, J. C. Looking for Alzheimer's Disease morphometric signatures using machine learning techniques. *J. Neurosci. Methods* **302**, 24–34 (2018).
29. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**(7), 581–583 (2016).
30. Erdös, P. & Rényi, A. On random graphs I. *Publ. Math. Debrecen* **6**, 290–297 (1959).
31. Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage* **52**(3), 1059–1069 (2010).
32. Lei, B., *et al*. NEURO-LEARN: A solution for collaborative pattern analysis of neuroimaging data. *Neuroinformatics* **19**, 79–91 (2020).
33. Zarogianni, E., Storkey, A. J., Johnstone, E. C., Owens, D. G. & Lawrie, S. M. Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophr. Res.* **181**, 6–12 (2017).
34. Schwarz, E. *et al.* Identification of a blood-based biological signature in subjects with psychiatric disorders prior to clinical manifestation. *World J. Biol. Psychiatry* **13**(8), 627–632 (2012).
35. Fernandes, B. *et al.* Serum thiobarbituric acid reactive substances in bipolar disorder during mania and schizophrenia: A biomarker with possible diagnostic implications. In: *4th biennial meeting of international society for bipolar disorders* (2010).
36. Wu, F. *et al.* Structural and functional brain abnormalities in drug-naive, first-episode, and chronic patients with schizophrenia: A multimodal MRI study. *Neuropsychiatr. Dis. Treat* **14**, 2889–2904 (2018).
37. Lu, X. B. *et al.* Analysis of first-episode and chronic schizophrenia using multi-modal magnetic resonance imaging. *Eur. Rev. Med. Pharmacol. Sci.* **22**(19), 6422–6435 (2018).
38. Cryan, J. F. & Dinan, T. G. Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* **13**(10), 701–712 (2012).
39. Sampson, T. R. *et al.* Gut microbiota regulate motor deficits and neuroinflammation in a model of parkinson's disease. *Cell* **167**(6), 1469–1480.e1412 (2016).
40. Schwarz, E. *et al.* Analysis of microbiota in first episode psychosis identifies preliminary associations with symptom severity and treatment response. *Schizophr. Res.* **192**, 398–403 (2018).
41. Zhu, F. *et al.* Transplantation of microbiota from drug-free patients with schizophrenia causes schizophrenia-like abnormal behaviors and dysregulated kynurenine metabolism in mice. *Mol. Psychiatry* **25**(11), 2905–2918 (2020).
42. Yuan, X., Kang, Y., Zhuo, C., Huang, X. F. & Song, X. The gut microbiota promotes the pathogenesis of schizophrenia via multiple pathways. *Biochem. Biophys. Res. Commun.* **512**(2), 373–380 (2019).
43. Rocha-Ramírez, L. M. *et al.* Probiotic lactobacillus strains stimulate the inflammatory response and activate human macrophages. *J. Immunol. Res.* https://doi.org/10.1155/2017/4607491 (2017).
44. De Vadder, F. *et al.* Microbiota-generated metabolites promote metabolic benefits via gut-brain neural circuits. *Cell* **156**(1), 84–96 (2014).
45. Chen, J. *et al.* An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med.* **8**(1), 43 (2016).
46. Ben Othmen, L. *et al.* Altered antioxidant defense system in clinically stable patients with schizophrenia and their unaffected siblings. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **32**(1), 155–159 (2008).
47. Zhang, X. Y. *et al.* Antioxidant enzymes and lipid peroxidation in different forms of schizophrenia treated with typical and atypical antipsychotics. *Schizophr. Res.* **81**(2–3), 291–300 (2006).
48. Zhang, X. Y. *et al.* Disrupted antioxidant enzyme activity and elevated lipid peroxidation products in schizophrenic patients with tardive dyskinesia. *J. Clin. Psychiatry* **68**(5), 754–760 (2007).

49. Raffa, M. *et al.* Reduced antioxidant defense systems in schizophrenia and bipolar I disorder. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **39**(2), 371–375 (2012).
50. Özdin, S., Sarisoy, G. & Böke, Ö. A comparison of the neutrophil-lymphocyte, platelet-lymphocyte and monocyte-lymphocyte ratios in schizophrenia and bipolar disorder patients—A retrospective file review. *Nord J. Psychiatry* **71**(7), 509–512 (2017).
51. Wilke, I. *et al.* Investigations of cytokine production in whole blood cultures of paranoid and residual schizophrenic patients. *Eur. Arch Psychiatry Clin. Neurosci.* **246**(5), 279–284 (1996).
52. Zorrilla, E. P., Cannon, T. D., Gur, R. E. & Kessler, J. Leukocytes and organ-nonspecific autoantibodies in schizophrenics and their siblings: Markers of vulnerability or disease?. *Biol. Psychiatry* **40**(9), 825–833 (1996).
53. Miller, B. J., Buckley, P., Seabolt, W., Mellor, A. & Kirkpatrick, B. Meta-analysis of cytokine alterations in schizophrenia: Clinical status and antipsychotic effects. *Biol. Psychiatry* **70**(7), 663–671 (2011).
54. Miller, B. J., Gassama, B., Sebastian, D., Buckley, P. & Mellor, A. Meta-analysis of lymphocytes in schizophrenia: Clinical status and antipsychotic effects. *Biol. Psychiatry* **73**(10), 993–999 (2013).
55. Miller, B. J., Culpepper, N. & Rapaport, M. H. C-reactive protein levels in schizophrenia: A review and meta-analysis. *Clin. Schizophr. Relat. Psychoses* **7**(4), 223–230 (2014).
56. Karageorgiou, V., Milas, G. P. & Michopoulos, I. Neutrophil-to-lymphocyte ratio in schizophrenia: A systematic review and meta-analysis. *Schizophr. Res.* **206**, 4–12 (2019).
57. Miyauchi, T. *et al.* Computerized EEG in schizophrenic patients. *Biol. Psychiatry* **28**(6), 488–494 (1990).
58. Sponheim, S. R., Clementz, B. A., Iacono, W. G. & Beiser, M. Clinical and biological concomitants of resting state EEG power abnormalities in schizophrenia. *Biol. Psychiatry* **48**(11), 1088–1097 (2000).
59. Sponheim, S. R., Iacono, W. G., Thuras, P. D., Nugent, S. M. & Beiser, M. Sensitivity and specificity of select biological indices in characterizing psychotic patients and their relatives. *Schizophr. Res.* **63**(1–2), 27–38 (2003).
60. Jamal, W., Das, S. & Maharatna, K. Existence of millisecond-order stable states in time-varying phase synchronization measure in EEG signals. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2013**, 2539–2542 (2013).
61. Chen, J. *et al.* Variability in resting state network and functional network connectivity associated with schizophrenia genetic risk: A pilot study. *Front. Neurosci.* **12**, 114 (2018).
62. van den Heuvel, M. P. & Fornito, A. Brain networks in schizophrenia. *Neuropsychol. Rev.* **24**(1), 32–48 (2014).
63. Modinos, G. *et al.* Neuroanatomy of auditory verbal hallucinations in schizophrenia: A quantitative meta-analysis of voxel-based morphometry studies. *Cortex* **49**(4), 1046–1055 (2013).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2020YFC2004300, 2020YFC2004301, 2019YFC0118800, 2019YFC0118802, 2019YFC0118804, and 2019YFC0118805), the National Natural Science Foundation of China (31771074, 81802230, and 32000980), the Key Research and Development Program of Guangdong (2018B030335001, 2020B0101130020, and 2020B0404010002), the Guangdong Basic and Applied Basic Research Foundation Outstanding Youth Project (2021B1515020064), the Guangdong Basic and Applied Basic Research Foundation (2019A1515110427), the Key Platform and Scientific Research Project of Guangdong Provincial Education Department (2018KTSCX246), the Science and Technology Program of Guangzhou (201807010064, 201803010100, 201903010032, and 202103000032), the Key Laboratory Program of Guangdong Provincial Education Department (2020KSYS001), and the Scientific Research Project of Traditional Chinese Medicine of Guangdong (20211306). All procedures performed in studies involving human participants were conducted in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Author contributions

P.K. wrote the main text of the manuscript. P.K., K.W., F.W., and D.X. designed the study. P.K., K.W., F.W., D.X., Z.P., and J.L. conducted experiments and analyzed the data. J.S., S.L. and X.C. edited the manuscript. Z.P. and J.Z. revised the manuscript. K.W. and F.W. obtained funding. J.Z., G.L., J.C., Y.N. and X.L. provided technical and material support. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-94007-9.

Correspondence and requests for materials should be addressed to F.W. or K.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.