Multi-biological classification for the diagnosis of schizophrenia
using multi-classifier, multi-feature selection and multi-cross
validation: An integrated machine learning framework study

Peng-fei Ke, BE 1, 3, 5; Dong-sheng Xiong, MS 1, 3, 5; Jia-hui Li, ME 1, 3, 5; Shi-jia Li, ME 1, 3, 5; Jie Song, MS 1, 3, 5; Xiao-yi Chen, BE 1, 3, 5; Jing Zhou, PhD 1, 3, 5; Gui-xiang Li, ME 4, 7; Jun Chen, ME 4, 7; Xiao-bo Li, PhD 8; Yu-ping Ning, PhD 2, 3; Feng-chun Wu, PhD 2, 3* and Kai Wu, PhD 1, 2, 3, 4, 5, 6, 7, 8, 9*

1 Department of Biomedical Engineering, School of Material Science and Engineering, South China University of Technology, Guangzhou 510006, China
2 The Affiliated Brain Hospital of Guangzhou Medical University, Guangzhou Huiai Hospital, Guangzhou 510370, 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
Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ, U.S.

Department of Nuclear Medicine and Radiology, Institute of Development, Aging and Cancer, Tohoku University, Sendai 980-8575, Japan

* Address for correspondence:

Kai Wu, PhD.
School of Material Science
South China University of Technology
Guangzhou, Guangdong 510006, China
Tel: 86-135-8037-8315
Email: kaiwu@scut.edu.cn

And

Fengchun Wu, Ph.D.
The Affiliated Brain Hospital of Guangzhou Medical University
Guangzhou Huiai Hospital
Guangzhou, Guangdong 510370, China
Tel: +86-020-8126-8212
Email: 13580380071@163.com
Abstract

Finding effective and objective biomarkers to inform the diagnosis of schizophrenia is of great importance yet remains challenging. However, there is relatively little work on multi-biological data for the diagnosis of schizophrenia. This was a cross-sectional study in which 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 showed that the performance of the classifier using multi-biological data was better than that of the classifiers using single biological data, with 91.7% accuracy and 96.5% AUC. The most discriminative features (top 5%) for the classification include gut microbiota features (*Lactobacillus*, *Haemophilus*, and *Prevotella*), blood features (superoxide dismutase, monocyte-lymphocyte ratio, and neutrophil), and electroencephalogram features (nodal local efficiency, nodal efficiency, and nodal shortest path length in the temporal and frontal-parietal areas). The proposed integrated framework may be help in understanding the pathophysiology of schizophrenia and developing biomarkers for schizophrenia using multi-biological data.
Introduction

Finding effective and objective biomarkers to inform the diagnosis of schizophrenia (SZ) is of great importance yet remains challenging [1, 2]. It is now broadly accepted that gut microbiota abundance, inflammation, immunological factors, and functional brain networks are altered in SZ [3-6], however, most of these alterations are observed at the group level with great variability among individuals with the same phenotypic diagnosis. Consequently, none so far has proven to have the ability to reliably aid in the differential diagnosis of SZ [1, 7]. Therefore, it is important to analyze how gut microbiota abundance, inflammatory factors, immunological factors and functional brain networks behave at an individual level; for example, this information could be used to better understand the pathology and identify objective biomarkers for the clinical diagnosis of SZ [8].

Recently, pattern recognition based on machine learning has gained increasing attention, which is well suited for the identification of subtle patterns of information in the data and, as a consequence, is useful to better predict diagnosis at an individual level [1, 9-11]. Using a variety of biological data, such as gut microbiota data, blood data, and electroencephalogram (EEG) data, along with machine learning techniques, hundreds of studies have been performed attempting to achieve accurate classification of patients with SZ [8, 12-14]. For instance, one study [15] 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. Another study [16] used Boruta variable selection to select the most discriminatory taxa and random forests to develop a classifier and predict SZ based on the important microbiota features. The receiver operating characteristic curve analysis
demonstrated that 12 significant microbiota biomarkers were capable of being used as diagnostic factors. A recent study \[8\] 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).

Despite these advances, previous discriminative studies of SZ have primarily focused on biomarkers extracted from one kind 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 might further improve the classification performance \[17-20\]. Given the complexity of psychiatric disorders, these findings have led to the hypothesis that combining multi-biological data will also improve our ability to understand SZ with promise for biomarker identification. However, this hypothesis lacks direct evidence, since only a few studies have used a combination of multi-biological data for a discriminative study of SZ.

In this study, an integrated framework of machine learning consisting of multi-biological data, multi-biological features, multi-classifiers, multi-feature selection algorithms and multi-cross validation methods (5M), was used to discriminate SZ patients from HCs. We hypothesized that combining multi-biological data may improve the classification performance compared to previously used techniques.
Materials and Methods

Participants

The final sample comprised 99 participants including 49 SZ patients and 50 HCs. The SZ patients were recruited from the Affiliated Brain Hospital of Guangzhou Medical University, Guangzhou and met the diagnostic criteria of 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. SZ patients 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 drug administration 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, 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 hours before the experiment and that they get enough sleep the night before.
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 with eyes closed was recorded from 16 scalp electrodes (i.e., Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, and T6) according to the International 10/20 System referenced to electrode Cz (UEA-B, symptom, CHINA), as shown in (eFig1, Supplementary). All electrode impedances were kept lower 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 (Math Works, Natick, MA). The preprocessing was divided into four steps: electrode positioning, filtering, elimination of bad signal segments and signal frequency band decomposition. To improve the quality of the signal, a bandpass filter of 0.1-45 Hz was performed. Then, the EEG signal was cut into several epochs of 2 s, and artifact noise, such as eye blinks, movement, was removed by technicians. Finally, the signal was divided into seven frequency bands 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).

Fecal Sample Collection and Preprocessing

Fresh fecal samples were collected from all subjects and then were stored at -80°C until DNA extraction. A total of 200 mg of each fecal sample was used for DNA extraction.
The DNA extraction method was consistent with our previously published report[4]. 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). Then, we used a pretrained Naïve Bayes classifier for taxonomic analysis, and this classifier was trained on the Greengenes database (Version 13.8). The raw sequence data reported in this article have been deposited in the 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 was drawn from control subjects and patients by simple venipuncture between 7.00 and 9.00 a.m., after overnight fasting and tobacco abstinence for more than 12 h. Blood biochemical indicators were detected by an automatic biochemical analyzer.

Multi-biological Features Extraction

EEG Features Extraction

In a brain functional network based on EEG signals, each channel can be considered as a node, and the correlation between two nodes can be regarded as the connecting edge. 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 Figure1.

Instantaneous phase $\phi(t)$ can be calculated from signal $x(t)$ by using the Hilbert transform:

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

Phase is computed by the following expression:
\[ \varphi_x(t) = \arctan \frac{\ddot{x}(t)}{x(t)} \]

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

\[ \varphi(t) = \varphi_x(t) - \varphi_y(t) \]

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

\[ R = \left| \frac{1}{N} \sum_{j=0}^{N-1} e^{i\varphi(j\Delta t)} \right| \]

Where \( i \) denotes the imaginary unit, \( N \) means the total number of samples, and \( \Delta t \) is the bespeak time between the successive samples from 1 to \( N-1 \).

In this study, the network threshold was determined by the method of network sparsity. According to the E-R random graph model, the sparsity of a fully connected network was not less than \( 2\ln N/N \), where \( N \) represented the number of nodes. In this study, the number of network nodes was 16, so the minimum connection sparsity was 34%. To ensure that the constructed network met the attributes of the small world, the maximum network sparsity was 73%. This study analyzed networks with a sparsity of 34% to 73%, and constructed different networks within the above range with 1% as the step size. To obtain the characteristic value, we selected the upper and lower thresholds to calculate the area under the curve of each attribute.

Global attributes, including the global clustering coefficient (aCp), shortest path length (aLp), global efficiency (aEg), local efficiency (aEloc) and nodal attributes, including the nodal clustering coefficient (aNCp), nodal shortest path length (aNLP), nodal efficiency (aNNe), nodal local efficiency (aNLe), and degree centrality (aDc) were computed as EEG features, with reference to the research of Rubinov M [21]. In this study, 64 global attributes and 640 nodal attributes were initially computed for analysis. Among them, any feature that was missing any participant was removed. This filtering resulted in 48 global attributes, and 526 node attributes.
that were selected for the final analysis. Therefore, in the selected set, there were no missing values and no imputation process was required.

**Gut Microbiota Features Extraction**

Through gene sequencing technology, 171 species of microbiota markers 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 Features Extraction**

The numbers of white blood cells (WBC), neutrophils (NEU), lymphocytes (LYM), platelets (PLT) and monocytes (MON) were recorded from complete blood counts after routine blood tests. Four indicators of blood 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 above five cell numbers. Moreover, the oxidative stress indicators, including superoxide dismutase (SOD), homocysteine and C-reactive protein (CRP), 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). Statistical tests were 2 tailed and included the two-sample t-test and \( \chi^2 \) test. Comparison of sex distribution between the two groups was performed using the \( \chi^2 \) test. Comparisons including age and education years between the two groups were performed using a two-sample t test. Unless otherwise specified, significance of all tests was set as \( p < 0.05 \), or FDR corrected \( p < 0.05 \).
Machine Learning

We developed an integrated framework of machine learning based 5M to discriminate SZ patients from HCs (Figure2). In brief, the framework involved 3 phases: data preparation phase, model training phase, and independent model testing phase.

Data Preparation Phase

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, 77 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 of 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.

Model Training Phase

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

Multi-feature Selection Given that some features are less effective, irrelevant, or redundant for classification, and too many features may cause “overfitting”, it is necessary to use effective feature selection methods to identify the discriminative features to 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.

**Multi-classifier** It is difficult to build a specific classification model directly based on multi-biological data due to its 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 find out the most suitable model and to evaluate classification performance based on single and combined biological features.

**Multi-cross Validation** To ensure that there was a large enough sample size to train the model and prevent overfitting caused by insufficient training, multi-cross validation methods were performed in the training set, including 10-fold, 5-fold, 3-fold and leave one out methods. Several combinations of the aforementioned procedures were investigated for optimized data analysis. PCA and RFE feature selection algorithms could not be carried out due to the small dimension of blood features. As a result, a total of 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 Phase**

In the third phase, we used an independent testing dataset to estimate the generalizability of 280 models arising from the second phase. To quantitatively estimate the performances of all the methods mentioned in this study, we utilized the metrics of accuracy, sensitivity and specificity.
Moreover, we plotted the receiver operating characteristic (ROC) curves and then calculated the area under the curve value (AUC) for each classification situation to examine the possibility of correctly discriminating SZ patients 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 by the original data. The statistical significance was set at $p < 0.05$. All the automatic classification work was performed on NEURO-LEARN (https://github.com/Raniac/NEURO-LEARN[22]), which is a solution for collaborative pattern analysis of neuroimaging data.
Result

Participants

The resulting data set comprised 99 participants, including 49 SZ patients (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). There was no significant statistical difference between the SZ patients group and the HCs group in either age ($t = 0.141, p = 0.888$) or sex ($t = 0.294, p = 0.769$). 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 total 280 models. The classification performance under the 10-fold cross validation method, 5-fold cross validation method, 3-fold cross validation method, and leave-one-out cross validation method (eTable 1, eTable 2, eTable 3, and eTable 4 in the Supplementary) were obtained. There was no significant difference among the results of multi-cross validation methods. Table 2 shows the classification performance of the model obtained using different input features under the 10-fold 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 single type of biological feature (Figure 4). In addition, we found that the blood features achieved the best classification 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, which may be due to the heterogeneity of biological data. The SVM, LR and RF classifiers without using any feature selection algorithm had better classification performance when using combined features, with AUCs over 90%.

**Discriminative Features**

In this subsection, the most informative features selected to differentiate the SZ patients 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 are 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 is the first discriminative study of SZ by combining multi-biological data of gut microbiota data, blood data, and EEG data. We developed an integrated framework of machine learning to discriminate SZ patients from HCs. The main findings of this study are as follows: 1) using the combination of three types of biological features as input features for the classification, the best performance was achieved, with 91.7% of accuracy, 91.7% of sensitivity, 91.7% of specificity, and 96.5% of AUC. 2) The most discriminative features (top 5%) include gut microbiota features (Lactobacillus, Haemophilus, and Prevotella), blood features (superoxide dismutase, monocyte-lymphocyte ratio, and neutrophil), 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 of SZ patients. 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 importantly, the integration of multi-logical data outperforms a single type of biological data, which is consistent with our findings[8]. 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[23]. In summary, based on the integrated framework of machine learning, the combination of multi-biological data substantially improves the classification performance for schizophrenia patients. 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\textsuperscript{[1]}.

To date, although there are numerous discriminative studies of SZ using either data of blood-based\textsuperscript{[24-26]}, or neuroimaging data \textsuperscript{[20 27-30]}, few studies have investigated the potential of biomarkers for the diagnosis of SZ using gut microbiota data. In our study, among the top 34 features shown in Table 3, gut microbiota features accounted for a large proportion, indicating that they played an important role in classifying patients with SZ from HCs. An increasing amount of evidence suggests that the gut microbiota bidirectionally communicates with the central nervous system through the microbiome-gut-brain axis (MGBA), thereby influencing brain function and behavior\textsuperscript{[31 32]}. Recently, a few studies have focused on the role of the MGBA in SZ and revealed several alterations in the gut microbiota in SZ patients\textsuperscript{[33-36]}. These reports of altered gut microbionas are consistent with those identified in this study for which the most informative features of gut microbiota include \textit{Lactobacillus}, \textit{Haemophilus}, \textit{Collinsella}, \textit{Clostridium}, and \textit{Prevotella}. Furthermore, Yuan et al.\textsuperscript{[37]} have shown that changes in the gut microbiota and its metabolites may cause neuronal damage. \textit{Lactobacillus} can stimulate TNF production, based on this, \textit{Lactobacillus} may induce changes in inflammatory factors that induce SZ \textsuperscript{[38]}. 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) \textsuperscript{[39]}. \textit{Clostridium} is the main source of propionate in the gut, indicating that \textit{Clostridium} may influence the BBB and act on the brain by regulating SCFAs. In addition, \textit{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\textsuperscript{[40]}, 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 classification included superoxide dismutase (SOD), monocyte-lymphocyte ratio (MLR), monocyte (MON), neutrophil (NEU), C-reactive protein (CRP), white blood cell (WBC) and neutrophil-lymphocyte ratio (NLR), which is consistent with previous studies using conventional univariate statistical analysis. Numerous studies and evidence suggest that the oxidative stress contributes to the pathogenesis of SZ, and abnormalities in antioxidant enzymes, including SOD activities, are frequently found in patients diagnosed with SZ\textsuperscript{[41-43]}. A previous study\textsuperscript{[44]} indicated that SOD activities remained lower in patients with SZ and may be an important indirect biomarker of oxidative stress in SZ. The present findings provide additional evidence of increased oxidative stress in SZ. Blood inflammatory and immune system abnormalities in patients with SZ have been widely reported, which lead to an increase in various inflammatory markers. NEU count was demonstrated to be increased in chronic SZ patients\textsuperscript{[45]}. Increased MON count has been reported in chronic SZ patients as well\textsuperscript{[46, 47]}. Furthermore, a moderately increased CRP in SZ patients compared to HCs has been observed\textsuperscript{[48-50]}. It has been indicated that subjects with SZ have significantly elevated WBC. MLR and NLR have recently been used as indicators of inflammation, and predictors of cardiovascular disease, the leading cause of death in SZ. A recent meta-analysis investigated there was a significant increase in NLR in
patients with SZ \cite{51}. Elevated MLR and NLR have been observed in SZ, suggesting an increased inflammatory response in SZ \cite{45}. Our experimental results are consistent with these studies.

Table 3 shows that the EEG features with heavy weight primarily derive from the delta and alpha2 frequency bands, and partly from the beta and gamma frequency bands. Previous investigators found increases in delta and theta waves, decreases in alpha waves and increases in beta and gamma waves in schizophrenic individuals \cite{8, 12, 29, 30, 52, 53}. 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 \cite{29, 30, 54}. 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 SZ. EEG studies have shown that in the resting state, small world attributes in SZ patients were disrupted, with a lower clustering coefficient and a longer shortest path length \cite{55}. In addition, global and local efficiency in SZ patients is lower compared to healthy people \cite{56}. 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 SZ \cite{57}. The frontal and temporal lobes are primarily associated with higher cognitive functions, among which the temporal lobe was associated with hearing and language functions, which has been confirmed by MRI studies \cite{58}. These results are in accordance with previous structural and functional neurological findings.

Limitations

The present study has several limitations. First, since this is a cross-sectional study, we cannot
infer causality. There is evidence that immuno-inflammatory markers are altered from the beginning of SZ, and it is broadly accepted that inflammation plays a causal role in SZ. However, from a diagnostic perspective, this is irrelevant. A given marker has only to discriminate between two conditions, regardless of whether it is a cause, consequence, or correlate of the pathophysiological process. Second, there was a significant difference between the two participant groups in terms of education years, although the results remained unchanged when taking it into consideration 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 NCs, which substantially improved the classification performance. Our results demonstrate that features from multiple biological data provide complementary information that aids in providing effective and objective biomarkers to inform the clinical diagnosis of schizophrenia and our framework is effective in conveying comprehensive and complementary information for the purpose of classification.
References


Genome medicine, 8(1), 43 (2016).


Acknowledgment

This work was supported by the National Key Research and Development Program of China (2020YFC2004300, 2020YFC2004301, 2019YFC0118800, 2019YFC0118802, 2019YFC0118804, 2019YFC0118805), the National Natural Science Foundation of China (31771074, 81802230), the Guangdong Key Project in “Development of new tools for diagnosis and treatment of Autism” (2018B030335001), and the Science and Technology Program of Guangzhou (201704020168, 201704020113, 201807010064, 201803010100, 201903010032).

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Author Contributions

P.K. wrote the main manuscript text. P.K., K.W., F.W., and D.X. designed the study. P.K., K.W., F.W., D.X., and J.L. conducted experiments and analyzed the data. J.S., S.L. and X.C. edited 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.

Declaration of Interest

The authors declare no conflict of interest.
Figure 1. Flow Chart of Brain Network Construction of EEG Signal.

Abbreviations: EEG, Electroencephalogram; PLV, Phase locking value.
Figure 2. Overview of the Proposed Integrated Machine Learning Framework for Schizophrenia Classification.

The proposed integrated machine learning framework for schizophrenia classification consists of 5M methods. (a) Multi-biological data of all subjects were collected 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) (e) Multi-cross validation methods including 10-fold, 5-fold, 3-fold, and leave-one-out were used to evaluate the performance of the trained model.
Figure 3. The Flowchart of Classification Method by Machine Learning.
Figure 4. Areas Under the Receiver Operating Characteristic Curves (AUC) of the best model Comparing the Gut Microbiota Features, Blood Features, Electroencephalogram Features and the Combination of GMV, BF and EF as the Input of Machine Learning.

Each curve in the figure represents the roc curve of the best model using different input features.

Abbreviations: GMF, gut microbiota features; BF, blood features; EF, electroencephalogram features; CF, combined features.
Table 1. Demographic and clinical characteristics used in the analysis.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>HCs (n=50)</th>
<th>SZs (n=49)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD), y</td>
<td>41.7 (13.1)</td>
<td>42.1 (12.5)</td>
<td>0.89</td>
</tr>
<tr>
<td>Sex, No (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>23 (46.0)</td>
<td>24 (49.0)</td>
<td>0.77</td>
</tr>
<tr>
<td>Female</td>
<td>27 (54.0)</td>
<td>25 (51.0)</td>
<td></td>
</tr>
<tr>
<td>Education years, mean (SD), y</td>
<td>14.2 (3.6)</td>
<td>11.6 (3.4)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>PANSS, mean (SD)</td>
<td>NA</td>
<td>58.84 (17.50)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Abbreviations: HCs, healthy controls; SZs, schizophrenia patients; NA, not applicable.
Table 2. Classification Performance of the optimal model under different input features using the integrated machine learning framework (10-fold).

<table>
<thead>
<tr>
<th>Input feature</th>
<th>Feature Selection method</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>AUC</th>
<th>p value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gut microbiota features</td>
<td>RFE</td>
<td>RF</td>
<td>70.8</td>
<td>58.3</td>
<td>83.3</td>
<td>0.80</td>
<td>.03</td>
</tr>
<tr>
<td>Blood features (n=12)</td>
<td>None&lt;sup&gt;b&lt;/sup&gt;</td>
<td>KNN</td>
<td>83.3</td>
<td>83.3</td>
<td>83.3</td>
<td>0.88</td>
<td>.010</td>
</tr>
<tr>
<td>EEG features (n=574)</td>
<td>RFE</td>
<td>RF</td>
<td>79.2</td>
<td>83.3</td>
<td>75.0</td>
<td>0.90</td>
<td>.010</td>
</tr>
<tr>
<td>Combined features (n=663)</td>
<td>None</td>
<td>SVM</td>
<td>91.7</td>
<td>91.7</td>
<td>91.7</td>
<td>0.97</td>
<td>.010</td>
</tr>
</tbody>
</table>

Abbreviations: 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;

<sup>a</sup>  The statistical significance of permutation test is set at p < 0.05.

<sup>b</sup> None means no feature selection algorithm is used.
Table 3. Top 34 features (5%) showing the most discriminative biomarkers for multi-biological prediction\textsuperscript{a}.

<table>
<thead>
<tr>
<th>Number</th>
<th>Feature name</th>
<th>Feature type</th>
<th>Number</th>
<th>Feature name</th>
<th>Feature type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SOD</td>
<td>Blood</td>
<td>18</td>
<td>Undefined\textsuperscript{c}</td>
<td>GM</td>
</tr>
<tr>
<td>2</td>
<td>MLR</td>
<td>Blood</td>
<td>19</td>
<td>Anaerostipes</td>
<td>GM</td>
</tr>
<tr>
<td>3</td>
<td>Lactobacillus</td>
<td>GM</td>
<td>20</td>
<td>PLT</td>
<td>Blood</td>
</tr>
<tr>
<td>4</td>
<td>MON</td>
<td>Blood</td>
<td>21</td>
<td>alpha2_aNLe_P4</td>
<td>EEG</td>
</tr>
<tr>
<td>5</td>
<td>Haemophilus</td>
<td>GM</td>
<td>22</td>
<td>Dialister</td>
<td>GM</td>
</tr>
<tr>
<td>6</td>
<td>Prevotella</td>
<td>GM</td>
<td>23</td>
<td>beta1_aLambda</td>
<td>EEG</td>
</tr>
<tr>
<td>7</td>
<td>NEU</td>
<td>Blood</td>
<td>24</td>
<td>Slackia</td>
<td>GM</td>
</tr>
<tr>
<td>8</td>
<td>CRP</td>
<td>Blood</td>
<td>25</td>
<td>Undefined.</td>
<td>GM</td>
</tr>
<tr>
<td>9</td>
<td>Megamonas</td>
<td>GM</td>
<td>26</td>
<td>Odoribacter</td>
<td>GM</td>
</tr>
<tr>
<td>10</td>
<td>theta_aNLe_T6\textsuperscript{b}</td>
<td>EEG</td>
<td>27</td>
<td>Ruminococcus</td>
<td>GM</td>
</tr>
<tr>
<td>11</td>
<td>theta_aNe_T6</td>
<td>EEG</td>
<td>28</td>
<td>theta_aDc_FP1</td>
<td>EEG</td>
</tr>
<tr>
<td>12</td>
<td>theta_aNCp_T6</td>
<td>EEG</td>
<td>29</td>
<td>alpha2_aNCp_P4</td>
<td>EEG</td>
</tr>
<tr>
<td>13</td>
<td>WBC</td>
<td>Blood</td>
<td>30</td>
<td>beta2_aNLe_FP2</td>
<td>EEG</td>
</tr>
<tr>
<td>14</td>
<td>NLR</td>
<td>Blood</td>
<td>31</td>
<td>beta2_aDc_O2</td>
<td>EEG</td>
</tr>
<tr>
<td>15</td>
<td>Collinsella</td>
<td>GM</td>
<td>32</td>
<td>Gemmiger</td>
<td>GM</td>
</tr>
<tr>
<td>16</td>
<td>gamma_aDc_F7</td>
<td>EEG</td>
<td>33</td>
<td>alpha2_aNLe_T4</td>
<td>EEG</td>
</tr>
<tr>
<td>17</td>
<td>Clostridium</td>
<td>GM</td>
<td>34</td>
<td>alpha2_aNCp_T4</td>
<td>EEG</td>
</tr>
</tbody>
</table>

Abbreviation: 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.

\textsuperscript{a} The top 34 features are listed in the descending order of their weights.

\textsuperscript{b} The EEG features are represented as a\_b\_c, where a represents frequency band, b represents brain network attributes, and c represents electrode channel.

\textsuperscript{c} Undefined \textit{Lachnospiraceae}.

\textsuperscript{d} Undefined \textit{Ruminococcaceae}.