Classification for psychiatric disorders including schizophrenia, bipolar disorder, and major depressive disorder using machine learning

Qingxia Yang a,*, Qiaowen Xing a, Qingfang Yang b, Yaguo Gong c,*

aDepartment of Bioinformatics, Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
bSecond Affiliated Hospital, Zhejiang Chinese Medical University, Hangzhou 310005, China
cSchool of Pharmacy, Macau University of Science and Technology, Macau

ABSTRACT

Schizophrenia (SCZ), bipolar disorder (BP), and major depressive disorder (MDD) are the most common psychiatric disorders. Because there were lots of overlaps among these disorders from genetic epidemiology and molecular genetics, it is hard to realize the diagnoses of these psychiatric disorders. Currently, plenty of studies have been conducted for contributing to the diagnoses of these diseases. However, constructing a classification model with superior performance for differentiating SCZ, BP, and MDD samples is still a great challenge. In this study, the transcriptomic data was applied for discovering key genes and constructing a classification model. In this dataset, there were 268 samples including four groups (67 SCZ patients, 40 BP patients, 57 MDD patients, and 104 healthy controls), which were applied for constructing a classification model. First, 269 probes of differentially expressed genes (DEGs) among four sample groups were identified by the feature selection method. Second, these DEGs were validated by the literature review including disease relevance with the psychiatric disorders of these DEGs, the hub genes in the PPI (protein–protein interaction) network, and GO (gene ontology) terms and pathways. Third, a classification model was constructed using the identified DEGs by machine learning method to classify different groups. The ROC (receiver operator characteristic) curve and AUC (area under the curve) value were used to assess the classification capacity of the model. In summary, this classification model might provide clues for the diagnoses of these psychiatric disorders.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In psychiatric disorders, schizophrenia (SCZ), bipolar disorder (BP), and major depressive disorder (MDD) are multigenic diseases with complex etiology [1]. These four psychiatric disorders are associated with high rates of morbidity, mortality, and suicide. There were evident differences among these psychiatric disorders. SCZ is a severe mental disorder and can cause delusions and hallucinations [2]. SCZ affects approximately 1% of the world’s population and generally appears in subjects aged 15 to 25 years [3]. BP is known as one disabilities worldwide and is characterized by a high suicide rate, sleep problems, and dysfunction of psychological traits [4]. BP is characterized by alternating episodes of mania interspersed with periods of depression [5]. MDD is the leading cause of disability resulting in the overall burden of disease. MDD is characterized by symptoms and causes emotional distress, functional impairment, and suicide [6].

There are many similar symptoms of these psychiatric disorders such as suicidal ideation, sleep disturbances, and cognitive deficits. The diagnostic boundaries among these psychiatric disorders remain difficult to define because of this similarity. Therefore, psychiatry is the last medicine area because the diagnosis only uses the symptoms due to a lack of biomarkers to assist the diagnosis [7]. Using these biomarkers, underlying molecular pathologies using biomarkers is necessary to address the burden of psychiatric diseases. For psychiatric disorders, developing more effective method for objective diagnoses has been a major international public health priority [8,9].

Identification of molecular measures (biomarkers) will provide insight into the biology underlying the shared symptoms and is beneficial to the diagnosis of psychiatric disorders [10]. To seek objective biomarkers, transcriptomic data has become a powerful...
technology for detecting gene expression [11]. Recently, there are plenty of studies exploring molecular biomarkers based on transcriptomics [12]. For instance, in the research of Lanz et al. [13], the STEP level is unchanged in the pre-frontal cortex and associative striatum of post-mortem human brain samples of SCZ, BP, and MDD subjects. As reported by Higgs et al. [14], the database including SCZ, BP, and MDD samples can offer an efficient tool for data mining, such as biomarkers elucidation for target discovery. However, a classification model based on machine learning is still highly necessary and beneficial to the diagnoses of psychiatric disorders.

In this work, one combined dataset including SCZ, BP, MDD, and healthy controls was obtained by integrating three transcriptomic studies. First, there were 268 samples in this dataset including 67 SCZ subjects, 40 BP subjects, 57 MDD subjects, and 104 healthy controls. The differentially expressed genes (DEGs) were discovered by the partial least squares-discriminant analysis (PLS-DA), and 269 probes of DEGs were identified for psychiatric disorders (SCZ, BP, MDD, and healthy controls). Second, these DEGs were validated by the literature review including disease relevance with the psychiatric disorders of these DEGs, the hub genes of the PPI (protein–protein interaction) network, GO (gene ontology) terms, and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. Third, a classification model was constructed applying machine learning method for classifying four groups based on the identified DEGs. Based on the independent set, the AUC (area under the curve) value and the ROC (receiver operator characteristic) curve were used for assessing the classification capacity of this model.

2. Materials and methods

2.1. Transcriptomic dataset for the psychiatric disorders

Based on popular databases including GEO (Gene Expression Omnibus) and SMRI (Stanley Medical Research Institute), the datasets of the prefrontal cortex from the Brodmann Area 9, 10, and 46 in the brain were collected by searching the keywords (schizophrenia, bipolar disorder, and major depressive disorder). As a result, three microarray datasets were used in this study, and each dataset included four sample groups (SCZ, BP, MDD, and healthy controls). As shown in Table 1, detailed information on these datasets was provided, such as dataset ID and the number of samples. The data analysis of the raw data for these datasets was performed using the R language. Herein, these three studies were integrated as a comprehensive dataset by matching the probe ID of the gene. After integration, the batch effects were removed for the comprehensive studies [15]. The combat function in the sva package was used to remove the batch effects for three different datasets [16]. This comprehensive dataset was used to identify the DEGs among different sample groups of SCZ, BP, MDD, and healthy controls.

2.2. Identifying DEGs for SCZ, BP, MDD, and healthy controls

To identify the DEGs among four sample groups, a popular feature selection algorithm, PLS-DA (partial least squares-discriminant analysis) [17] was applied in this study. PLS-DA was one of the most well-known machine learning methods as a useful feature selector [18]. Recently, PLS-DA was widely applied for identify features for omics data [19]. Because of substantial similarity, the identified DEGs were expected to classify SCZ, BP, MDD, and healthy subjects. PLS-DA can select differential features among multiple classes simultaneously. Herein, the DEGs of four sample groups (SCZ, BP, MDD, and healthy groups) were discovered by the PLS-DA model. The VIP (Variable Importance in the Projection > 2) value in the PLS-DA model was applied as the index for the DEGs. And the dysregulated genes among four groups were identified by the VIP value (>2) of the PLS-DA model [20].

2.3. Functional analysis for the DEGs identified in psychiatric disorders

The functional analysis for the DEGs identified in psychiatric disorders was conducted in this work. The analysis was conducted from three different perspectives, including (1) disease relevance of these DEGs, (2) disease relevance of the hub genes of the PPI network, and (3) disease relevance of the gene ontology terms and pathways. For these DEGs, the disease relevance with psychiatric disorders was surveyed by the literature review. A substantial percentage of the disease-related genes was expected for these psychiatric disorders. But a certain number of psychiatric disorder-unrelated genes was unavoidable because of the measurement variations. The disease relevance was represented by the percentage of disease-related genes among all DEGs.

To ensure the hub genes of psychiatric disorders, the STRING database [21] was used to construct protein–protein interaction (PPI) network. Using high confidence (0.7), the DEGs discovered in this study can be mapped into this PPI network. Cytoscape [22] was used for visualizing the interactions of genes in the PPI network. The hub genes were discovered from all genes with high interaction degrees (score > 10) for psychiatric disorders. The role of the hub genes in psychiatric disorders was confirmed using the literature review. Moreover, GSEA was used to conduct the enrichment of GO terms and KEGG pathways by the adjusted p-value (<0.05) [23]. The GO terms and KEGG pathways overrepresented were identified, and a comprehensive literature review was conducted to reveal the important role of these terms and pathways in psychiatric disorders.

2.4. Constructing classification model using Machine learning

It remains difficult to define the diagnostic boundaries among psychiatric disorders due to the similarity of symptoms. A classification model with superior performance is important for the diagnoses of SCZ, BP, and MDD samples. Therefore, the DEGs identified in this study were used to construct a model for classifying different groups of psychiatric disorders. A popular machine learning method, support vector machine (SVM), was a supervised technique and was applied for classification. Herein, a classification model applying SVM method was constructed based on the identified DEGs for SCZ, BP, and MDD groups. This classification model was validated using fivefold cross-validation. The AUC value and ROC curve of this model were used to assess the classification capacity. Using the comprehensive dataset, the fivefold cross-validation, the batch effects were removed for the comprehensive datasets [15]. The combat function in the sva package was used to remove the batch effects for three different datasets [16]. This comprehensive dataset was used to identify the DEGs among different sample groups of SCZ, BP, MDD, and healthy controls.
validation was applied in the classification model. The AUC value could quantify the classification capacity of the model to distinguish different classes. If the AUC value was 1, the classification capacity of the model to classify different groups was excellent enough. If the AUC value was 0, the classification capacity of the model was poor enough.

To validate the classification capacity for generalizing to other datasets, the independent set was applied in the constructed SVM model. In this model, the combined dataset (Table 1) was as the training set, and the independent sets (GSE127711 [24] and GSE38484 [25]) were as the test set due to a lack of associated datasets. In the independent discovery cohort of the first dataset (GSE127711), there were 124 SCZ patients, 260 BP patients, and 112 MDD patients in the blood samples. In the second dataset (GSE38484), there were 106 SCZ patients and 96 healthy subjects in human whole blood. To obtain all four groups, these two independent datasets were combined as a new independent set. In this dataset, there were 230 SCZ samples, 260 BP samples, 112 MDD samples, and 96 healthy samples. The gene expression of the comprehensive dataset for identifying DEGs was detected in the prefrontal cortex of the brain, and the gene expression of the independent set was detected in the blood samples. To generalize the model constructed in this study, the blood samples in the independent set were applied to measure the classification capacity.

### 3. Results and discussion

#### 3.1. Comprehensive dataset including SCZ, BP, MDD, and healthy groups

As shown in Fig. 1, the flowchart of this study included four parts: (1) the comprehensive transcriptomic dataset; (2) identification of DEGs by PLS-DA; (3) functional analysis; and (4) construction of the classification model. At the beginning of this study, three datasets (Table 1) were collected for the comprehensive transcriptomic dataset. One dataset (Stanley AltarC) was from the SMRI database [14] including 72 (21 SCZ, 11 BP, 11 MDD, and 33 healthy subjects) samples detected by the HG-U133A platform. For dataset GSE92538 [26] from the GEO database, 128 samples (31 SCZ, 12 BP, 29 MDD, and 56 healthy controls) were detected by HG-U133 Plus 2 platform. For dataset GSE39387 [13] from the GEO database, 68 samples (15 SCZ, 17 BP, 17 MDD, and 19 healthy controls) were detected by HG-U133 Plus 2 platform. After each dataset was processed and analyzed using the R language, the comprehensive dataset was combined by removing batch effects. In this comprehensive dataset by combining these three datasets, there were 22,277 probes of genes and 268 samples of prefrontal cortex including 67 SCZ patients, 40 BP patients, 57 MDD patients, and 104 healthy subjects.

#### 3.2. DEGs identified for psychiatric disorders using the comprehensive dataset

DEGs were discovered by the PLS-DA method to classify different groups of psychiatric disorders simultaneously based on the comprehensive transcriptomic data. Using the cutoff of VIP value ($\geq 2$) of the PLS-DA model, there were 269 probes of DEGs identified in this study (as shown in Supplementary Figure S1). As shown in Table 2, detailed information on the top 20 DEGs with the highest VIP values was provided. The dysregulated information of all DEGs between two groups (including between SCZ and BP, between SCZ and MDD, as well as between BP and MDD) was shown in Supplementary Table S1. As demonstrated in Fig. 2, the boxplots were applied to visualize and compare the differential expression of the top 9 DEGs among four groups directly. For example, the gene expression of NEK1 with the highest VIP value ($\text{VIP} = 3.00$) has a strong association with a chromosome 4 genetic locus identified as significantly associated with SCZ [27]. It showed an increase in NEK1 after antidepressant treatment in responders [28]. Moreover, it was reported that the expression of CDC42BPA with the second highest VIP value ($\text{VIP} = 2.95$) differed significantly among SCZ, BP, and controls [29].

#### 3.3. Functional analysis for DEGs identified among multiple psychiatric disorders

The functional analysis of the DEGs was performed from three different perspectives including (1) disease relevance for the DEGs,
(2) disease relevance for the hub genes of the PPI network, and (3) disease relevance for the enriched GO terms and KEGG pathways.

(1) Disease relevance with psychiatric disorders for the DEGs.

To evaluate the disease relevance of the DEGs discovered among multiple disorders, the top 20 DEGs among different groups were surveyed by the comprehensive literature review. The disease relevance between each DEG and psychiatric disorders (SCZ, BP, MDD, or cognition) was described in Supplementary Table S2. A great disease relevance (90 %) of the top 20 DEGs was verified. For these DEGs, it was reported that DACH1 was a transcription factor acting as a neurogenic cell-fate determining factor [30]. The mutations of TANC2 were associated with both pediatric neurodevelopmental and adult neuropsychiatric disease [31]. ANKRDL1 was a nuclear coregulator in the developing brain, which determined precursor proliferation, neurogenesis, and neuronal positioning [32]. It was reported that KMT2A, NFA5T, SCAF11, and GSK3B were upregulated in neurons of BP [33]. Several genetic variants of RORA were associated with BP [34], and the polymorphisms of RORA were associated with risk for various forms of psychopathology including BP and MDD [35]. PHACTR1 showed the association with SCZ in the combined analysis and the locus was located in an SCZ linkage region [36]. ZEB1 was an element of a common pathway involved in SCZ [37]. EHBP1 was down-regulated in the medial prefrontal cortex of adult SHANK3-overexpressing mice, and variants of SHANK3 were causally associated with numerous neuropsychiatric and neuropsychiatric disorders including BP and SCZ [38]. CSRNP2, CAMK2G, PIK3R1, and PRPF40A were identified as the hub genes in the anterior cingulate cortex regions of the brain for MDD [49]. Therefore, these hub genes discovered using the PPI network had an important role in SCZ, BP, and MDD, which showed the reliability of the DEGs discovered in this work.

(2) Disease relevance with psychiatric disorders for the hub genes in the PPI network.

STRING database was used for constructing the PPI network [42], and the hub genes were discovered based on the CytoHubba [43] of Cytoscape [22]. As shown in Fig. S3, the hub network for all DEGs was constructed. The degree of nodes in this PPI network was shown in Supplementary Table S3. As shown in Fig. 3B, the top 13 nodes with the highest score (≥10) of the network using the MCC algorithm on CytoHubba were marked with red and yellow colors. The intersection between the top 13 genes by the MCC algorithm (as shown in Fig. 3C) and the top 20 genes with high degree (≥10) in this PPI network was regarded as the hub genes. There were 9 hub genes including ESF1, PAK1IP1, SF3B1, RBM25, KRAS, SRRM2, CAMK2G, PIK3R1, and PRPF40A. As shown in Fig. 4, the 9 hub genes were validated to confirm the differential expression among four groups (SCZ group, BP group, MDD group, and healthy group) using the boxplots. From these boxplots, there were significant changes for these DEGs.

As shown in Supplementary Table S4, a great disease relevance (78 %) for the hub genes was discovered between the hub genes and psychiatric disorders by a literature review. It was reported that SF3B1 was associated with SCZ and neurodevelopmental disorders in the largest SCZ genome-wide association study [44]. KRAS mutations were associated with depression severity and higher rates of probable depression in patients with metastatic colorectal cancer [45]. A mechanistic pathway involving CAMK2G was reported in stress and the trauma-related manifestation of anxiety and depression across species [46]. The interaction effects of the polymorphisms in hsa-miR-219, CAMK2G, GRIN2B, and GRIN3A might confer susceptibility to SCZ in the Chinese Han population [47]. PIK3R1 was the shared susceptibility gene for SCZ and BP, which might be a potential diagnostic biomarker for BP [48]. PIK3R1 and PRPF40A were identified as the hub genes in the anterior cingulate cortex regions of the brain for MDD [49]. Therefore, these hub genes discovered using the PPI network had an important role in SCZ, BP, and MDD, which showed the reliability of the DEGs discovered in this work.

(3) Disease relevance with psychiatric disorders for the Enriched GO terms and KEGG pathways.

Moreover, 33 KEGG pathways have been enriched using the DEGs discovered in this study (as shown in Fig. S5 and Supplementary Table S5), including regulation of actin cytoskeleton, neurotrophin signaling pathway, focal adhesion, calcium signaling pathway, and insulin signaling pathway. The regulation of the actin cytoskeleton was likely to be shared between SCZ and BP [50]. Rare variants in the neurotrophin signaling pathway were implicated in SCZ risk [51]. The evidence for altered motility and focal adhesion dynamics was consistent with dysregulated gene expression in the FAK signaling pathway. Alterations in cell adhesion dynamics and cell motility can affect the trajectory of brain development in SCZ [52]. A detailed characterization of the risk loci showed that cal-

Table 2
Detailed information on the top 20 DEGs identified by the PLS-DA (partial least squares discriminant analysis) method with the cutoff of Variable Importance in the Projection (VIP > 2). SCZ: schizophrenia, BP: bipolar disorder, and MDD: major depressive disorder.

| Order | Probe ID | Entrez ID | Symbol | VIP | Up-or Down -Regulated |
|-------|----------|-----------|--------|-----|-----------------------|
|       |          |           |        |     | SCZ vs BP | SCZ vs MDD | BP vs MDD |
| 1     | 213328_at | 4750      | NEK1   | 3.00| Down      | Down       | Up        |
| 2     | 214664_at | 8476      | CDC42BP1| 2.95| Down      | Down       | Up        |
| 3     | 205472_s_at| 1602      | DACH1  | 2.92| Down      | Down       | Down      |
| 4     | 208425_s_at| 26.115    | TANC2  | 2.84| Down      | Down       | Down      |
| 5     | 202905_s_at| 4683      | NFB    | 2.84| Down      | Down       | Up        |
| 6     | 208993_s_at| 9360      | PFPG   | 2.84| Down      | Down       | Down      |
| 7     | 219437_s_at| 25.123    | ANKRDL1| 2.78| Down      | Down       | Up        |
| 8     | 212079_s_at| 4297      | KMT2A  | 2.78| Down      | Down       | Down      |
| 9     | 208003_s_at| 10.725    | NFA5T  | 2.76| Down      | Down       | Up        |
| 10    | 213830_s_at| 9169      | SCAF11 | 2.75| Down      | Down       | Up        |
| 11    | 210479_s_at| 6095      | RORA   | 2.74| Down      | Down       | Up        |
| 12    | 213638_at | 221.692   | PHACTR1| 2.74| Down      | Down       | Down      |
| 13    | 212758_s_at| 6935      | ZEB1   | 2.74| Down      | Down       | Up        |
| 14    | 212650_at | 23.301    | EHB1   | 2.74| Down      | Down       | Down      |
| 15    | 220462_at | 80.034    | CSRNP3 | 2.72| Down      | Down       | Down      |
| 16    | 202040_s_at| 5927      | KDM5A  | 2.72| Down      | Down       | Up        |
| 17    | 209945_s_at| 2932      | GSK3B  | 2.72| Down      | Down       | Up        |
| 18    | 201996_s_at| 23.013    | SPEN   | 2.71| Down      | Down       | Down      |
| 19    | 220940_at | 57.730    | ANKRDL36B| 2.67| Down      | Down       | Down      |
| 20    | 209376_s_at| 9169      | SCAF11 | 2.67| Down      | Down       | Up        |
Calcium signaling pathway genes might play pivotal roles in SCZ [53], and the downregulated signaling pathways in depression mice included the calcium signaling pathway [54]. It was suggested that there existed abnormalities of the insulin signaling pathway in SCZ and that antipsychotic drug effects on this pathway were therapeutic in SCZ [55].

As shown in Fig. 5B and Supplementary Table S6, the enrichment analysis of GO terms was performed to discover the biological processes (BP) terms. For instance, adult neurogenesis concerning regulatory signaling molecules would be helpful to identify how abnormalities might contribute to the pathophysiology of SCZ [56], and altered adult neurogenesis was postulated as an aetiological mechanism for BP [57]. As demonstrated in Fig. 5C and Supplementary Table S7, the molecular functions (MF) terms were enriched using DEGs. Such as, the Alu element in the RNA binding motif protein (RBMX2) was found to be linked to BP [58]. As demonstrated in Fig. 5D and Supplementary Table S8, a lot of key cell components (CC) terms were enriched using the DEGs in this study. And a growing body of evidence connected a dysfunctional microtubule cytoskeleton with neuropsychiatric ill-
nesses [59]. Using the literature review, the GO terms and KEGG pathways enriched were validated that they played an important role in the development of psychiatric disorders.

3.4. Constructing the classification model for multiple psychiatric disorders

As one of the supervised machine learning algorithms, SVM can be used to construct a classification model. The classification of SVM can be applied for two or more classes using the e1071 package. A single SVM does binary classification and can classify samples between two classes. SVM can be applied for classifying multiple groups using the One-to-Rest approach. To classify multiple classes, each binary classifier is set to per each class. In this approach, the classifier can use \( m \) SVM models and each model will predict membership in one of the \( m \) classes. In this study, the SVM method was used to construct a model for classifying multiple groups (SCZ, BP, MDD, and healthy controls). The classification model was constructed for classifying samples of SCZ, BP, MDD, and healthy groups based on the DEGs identified by the PLS-DA method using the comprehensive dataset (Table 1). Because it was hard to obtain good performance when using all genes due to the interference of the irrelevant genes, these DEGs differential among four groups were applied for constructing well-performed classification model. In the multi-class classification models, there were four SVM models for SCZ, BP, MDD, and healthy groups. And the total model was obtained using the micro value of all SVM models. For the combined dataset (Table 1), the performance of the classification model was assessed by 5-fold cross-validation. The AUC value and ROC curve were used to assess the classification

Fig. 3. (A) The PPI network was constructed using DEGs (differentially expressed genes) among schizophrenia, bipolar disorder, major depressive disorder, and healthy controls. (B) The top 13 hub nodes with the highest MCC (score \( \geq 10 \)) in the network were marked with red and yellow colors using the MCC algorithm on CytoHubba. (C) The scores for the top 13 hub nodes were ranked by the MCC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
performance. Based on the comprehensive dataset, the AUC values and ROC curves are interpreted for SCZ groups (Fig. 6A), BP groups (Fig. 6B), MDD groups (Fig. 6C), healthy groups (Fig. 6D), and (Fig. 6E) total micro value for all groups using the 5-fold cross-validation. Overall, the AUC value of 5-fold cross-validation for four groups was 0.94 in the SVM model using the comprehensive dataset.

In this study, the independent dataset was applied to generalize the constructed SVM model. The combined dataset (Table 1) was regarded as the training set, and the independent set by combining GSE127711 and GSE38484 was regarded as the test set. In this classification, the micro value was calculated for four groups of psychiatric disorders (SCZ, BP, MDD, and healthy controls). The AUC value of the independent set was 0.71 in the classification model. As shown in Fig. 6F, the AUC value and ROC curve were used to assess the performance of model using the independent set (Table 1). From the results, the classification performance is only good (AUC > 0.7) for classifying four groups simultaneously. The genes of the training set were detected in the prefrontal cortex, and the genes of the independent set were detected in the blood samples.

Fig. 4. The boxplots of the 9 hub genes of the PPI network using the intersection between the genes with the highest degree (score ≥ 10) of the PPI network and the top 13 hub nodes ranked by the MCC (score ≥ 10) in the network using the MCC algorithm on CytoHubba software. The blue, red, green, and purple indicated the SCZ (schizophrenia), BP (bipolar disorder), MDD (major depressive disorder), and CTRL (healthy controls), respectively. Statistically significant differences in cortical thickness: *p < 0.05, **p < 0.001, ***p < 0.0001. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Because of the differences in the data type between the training set and test set, it is very difficult to obtain superior performance for the classification capacity by the independent test. In the future, the classification model with superior performance can be developed using other machine learning methods, which will be helpful for the diagnoses of psychiatric disorders.
4. Conclusions

In this work, a combined dataset comprising 67 SCZ patients, 40 BP patients, 57 MDD patients, and 104 healthy controls was collected. First, 269 probes of DEGs were discovered based on the PLS-DA method to classify the samples into four groups. Second, these DEGs were validated by the literature review including disease relevance with the psychiatric disorders of these DEGs, the hub genes of the PPI network, and enriched GO terms and KEGG pathways. Third, a classification model was constructed by machine learning method using the DEGs identified in four groups. By ROC curve and AUC value, a strong capacity to classify samples among multiple groups was demonstrated. Moreover, the constructed SVM model was generalized using the independent set. In sum, the classification model constructed might provide clues for the diagnoses of these psychiatric disorders.

CRediT authorship contribution statement

Qingxia Yang: Conceptualization, Methodology, Software, Writing – original draft. Qiaowen Xing: Visualization, Investigation. Qingfang Yang: Software, Validation. Yaguo Gong: Data curation, Writing – review & editing.
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (62201289), the National Natural Science Foundation of Jiangsu (BK20210597), and the NUPITSF (Grant No. NY220169).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.09.014.

References:

[1] McGuinness AJ, Davis JA, Dawson SL, Loughman A, Collier F, O’Hely M, et al. A systematic review of gut microbiota composition in observational studies of major depressive disorder, bipolar disorder and schizophrenia. Mol Psychiatry 2022;27(4):1920–35.
[2] Jauhar S, Johnston M, McKenna PJ. Schizophrenia. Lancet 2022;399(10323):473–86.
[3] Bighelli I, Rodolico A, Garcia-Mieres H, Pitschel-Walz G, Hansen WP, Schneider-Thoma J, et al. Psychosocial and psychological interventions for relapse prevention in schizophrenia: a systematic review and network meta-analysis. Lancet Psychiatry 2021;8(11):999–80.
[4] Ruderfer DM, Fanous AH, Ripke S, McQuillin A, Amdur RL, et al. Polygenic associations between NR1D1, RORA and RORB genes and bipolar disorder. J Psychiatr Res 2013;47(1):325–33.
[5] Le Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 2014;15(1):S11.
[6] Guo H, Bettella E, Marcogliese PC, Zhao R, Andrews JC, Nowakowski TJ, et al. Integrative brain transcriptome-wide isoform-level dysregulation in ASD, schizophrenia and bipolar disorder. Mol Psychiatry 2021;26(1):186–202.
[7] Konopaske GT, Balu DT, Presti KT, Chan G, Benes FM, Coyle JT. Dysbindin-1 contributes to prefrontal cortical dendritic arbor pathology in schizophrenia. Schizophrenia Res 2018;201:270–7.
[8] Gallagher D, Voronova A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Construction of the classification model using key genes and chemometric strategies. J Agric Food Chem 2019;67(4):1302–11.
[9] Bighelli I, Rodolico A, Garcia-Mieres H, Pitschel-Walz G, Hansen WP, Schneider-Thoma J, et al. Psychosocial and psychological interventions for relapse prevention in schizophrenia: a systematic review and network meta-analysis. Lancet Psychiatry 2021;8(11):999–80.
[10] Ruderfer DM, Fanous AH, Ripke S, McQuillin A, Amdur RL, et al. Polygenic associations between NR1D1, RORA and RORB genes and bipolar disorder. J Psychiatr Res 2013;47(1):325–33.
[11] Konopaske GT, Balu DT, Presti KT, Chan G, Benes FM, Coyle JT. Dysbindin-1 contributes to prefrontal cortical dendritic arbor pathology in schizophrenia. Schizophrenia Res 2018;201:270–7.
Wingo AP, Velasco ER, Florido A, Lori A, Choi DC, Jovanovic T, et al. Expression of the PPM1F gene is regulated by stress and associated with anxiety and depression. Biol Psychiatry 2018;83(3):284–95.

Zhang Y, Fan M, Wang Q, He G, Fu Y, Li H, et al. Polymorphisms in microRNA genes and genes involving in NMDAR signaling and schizophrenia: a case-control study in Chinese Han population. Sci Rep 2015;5:12984.

Huang J, Chen Z, Zhu L, Wu X, Guo X, Yang J, et al. Phosphoinositide-3-kinase regulatory subunit 1 gene polymorphisms are associated with schizophrenia and bipolar disorder in the Han Chinese population. Metab Brain Dis 2020;35(5):785–92.

Wei Y, Qi K, Yu Y, Lu W, Xu W, Yang C, et al. Analysis of differentially expressed genes in the dentate gyrus and anterior cingulate cortex in a mouse model of depression. Biomed Res Int 2021;2021:5013565.

Zhao Z, Xu J, Chen J, Kim S, Reimers M, Bacanu SA, et al. Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. Mol Psychiatry 2015;20(5):563–72.

Kranz TM, Goetz RR, Walsh-Messinger J, Goetz D, Antonius D, Dolgalev I, et al. Rare variants in the neurotrophin signaling pathway implicated in schizophrenia risk. Schizophr Res 2015;168(1–2):421–8.

Fan Y, Abrahamsen G, Mills R, Calderon CC, Tee JY, Leyton L, et al. Focal adhesion dynamics are altered in schizophrenia. Biol Psychiatry 2013;74(6):418–26.

Xie Y, Huang D, Wei L, Luo XJ. Further evidence for the genetic association between CACNA1A and schizophrenia. Hereditas 2018;155:16.

Huang J, Chen Z, Zhu L, Wu X, Guo X, Yang J, et al. Phosphoinositide-3-kinase regulatory subunit 1 gene polymorphisms are associated with schizophrenia and bipolar disorder in the Han Chinese population. Metab Brain Dis 2020;35(5):785–92.

Wei Y, Qi K, Yu Y, Lu W, Xu W, Yang C, et al. Analysis of differentially expressed genes in the dentate gyrus and anterior cingulate cortex in a mouse model of depression. Biomed Res Int 2021;2021:5013565.

Zhao Z, Xu J, Chen J, Kim S, Reimers M, Bacanu SA, et al. Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. Mol Psychiatry 2015;20(5):563–72.

Kranz TM, Goetz RR, Walsh-Messinger J, Goetz D, Antonius D, Dolgalev I, et al. Rare variants in the neurotrophin signaling pathway implicated in schizophrenia risk. Schizophr Res 2015;168(1–2):421–8.