Manifold Learning Analysis for Allele-Skewed DNA Modification SNPs for Psychiatric Disorders

WENBIN LI¹,², DONGDONG LI², AND HENRY HAN³,⁴
¹Institute of Computational Science and Technology, Guangzhou University, Guangzhou 510006, China
²Department of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou 325000, China
³Department of Computer and Information Science, Fordham University, New York, NY 10023, USA
⁴School of Computer Science, Qinghai Normal University, Xining 810008, China

Corresponding authors: Wenbin Liu (wbliu6910@126.com) and Henry Han (xhan9@fordham.edu)

This work was supported in part by the National Science Foundation of China under Grant 61572367 and Grant 61573017, and in part by the Provincial Science Foundation of Qinghai, China, under Grant 2016ZJ904.

ABSTRACT Bipolar disorder (BPD) and schizophrenia (SCZ) are two severe worldwide psychiatric disorders. Identifying genetic components contributing to both disorders will provide meaningful insights into their pathogenesis and widely-existent misdiagnosis. In this study, we employ allele-skewed DNA modification (ASM-SNP) data to investigate the two psychiatric disorders via state-of-the-art manifold learning, data-driven feature selection, and novel pathway analysis. We propose a novel manifold learning analysis for ASM-SNP data of bipolar disorder and schizophrenia based on a data-driven feature selection algorithm: nonnegative singular value approximation (NSVA). Our results indicate that t-distributed stochastic neighbor embedding (t-SNE) outperforms its peers in distinguishing psychiatric disorder samples from normal ones in both visualization and phenotype classification. It achieves the best phenotype diagnosis results with the average AUC 0.95 by using only about 20% top-ranked SNPs. Furthermore, our results from manifold learning along with support vector machine analysis suggest that the possible non-separability of SCZ and BPD in genetics. We also validate that SCZ and BPD both share the same or similar genetic variations from pathway analysis. This study indicates the inevitable misdiagnosis issue between BPD and SCZ from a machine learning and systems biology approach. The result sheds light on the existing psychiatry research to reexamine the existing behavior-based classification for BPD and SCZ. To the best of our knowledge, this study is the first comprehensive investigation of BPD and SCZ in bioinformatics.

INDEX TERMS Manifold learning, bipolar disorder, schizophrenia, misdiagnosis, ASM-SNP, pathway.

I. INTRODUCTION

Schizophrenia (SCZ) and bipolar disorder (BPD), both of which are highly heritable, complex neuropsychiatric diseases that share some similar clinical symptoms [1], [2]. Schizophrenia is a chronic and severe mental health disorder characterized by hallucinations, delusions, and disorganized thinking [2]. Alternatively, bipolar disorder is a chronic mental illness that causes dramatic shifts in a person’s mood, energy, and ability to think clearly. People with bipolar disorder have high and low spirits, known as mania and depression. According to the National Alliance on Mental Illness (NAMI), every year, about 1% and 2.9% of Americans are diagnosed with schizophrenia and bipolar disorder respectively.

The underlying etiology of the neuropsychiatric disorders is not well understood or even unknown, but various research efforts have been made to decipher their genetic mechanisms from different standing points. Recent large-scale genome-wide association studies (GWAS) have revealed that psychiatric disorders share some common genetic risk loci [3]–[5]. For example, the Psychiatric Genomics Consortium GWAS identified 108 independently associated loci for schizophrenia, about 75% of which contain protein-coding genes that play essential roles in neurodevelopment and brain function [6]. It was found that most of these susceptibility variants located in noncoding genomic regions spanning multiple genes. Xiao et al. integrated transcriptomic data, epidemiological and bioinformatical methods, as well as in vitro experimental approaches to promote the translation of genetic discoveries to physiological mechanisms [7]. Furthermore, Smeland et al. observed substantial genetic enrichment on
associations with intelligence in both SCZ and BPD, indicating a polygenic overlap in their study [8].

DNA methylation plays a crucial role in many important genomic regulatory processes in many tissues, especially in the human brain [3, 9]. Different dysregulations in epigenetic mechanisms were reported to play a vital role in the pathogenesis of neurodevelopmental disorders such as schizophrenia [10]. Jaffe et al. identified DNA methylation changes in brain tissue were significantly enriched in genomic regions that conferred clinical risk for schizophrenia [11]; Hannon et al. showed more than 16,000 DNA methylation quantitative trait loci (mQTLs) and a substantial proportion of the genetic variants were related to schizophrenia risk in an extensive collection (n = 166) of human fetal brain samples [12].

More importantly, recent studies have suggested that allele-skewed DNA modification (ASM) could mediate or modulate phenotypic effects of DNA sequence variants. An allele-skewed DNA modification refers to various types of epigenetic cytosine marks in addition to methylation. Sarah et al. performed a comprehensive search of ASM effects in a total of 256 human post-mortem brain and sperm samples from individuals with significant psychosis and controls. They found that brain ASM-SNPs demonstrated a much stronger enrichment in a schizophrenia GWAS than in the other 17 large GWAS of non-psychiatric diseases and traits [12], [13].

The identified genetic variations from ASM-SNP data are crucial to revealing the underlying mechanisms of mental diseases. It will be desirable to analyze the impacts of the genetic variations by visualizing the distributions of phenotypes or sub-phenotypes (e.g., bipolar I and bipolar II) and via a comprehensive SNP analysis to dig knowledge and explore the genetic differences between schizophrenia and bipolar disorders besides investigating their meaningful phenotype clustering. It will directly answer the core questions in psychiatric disorder research: ‘how to distinguish schizophrenia (SCZ), bipolar disorder (BPD), or control samples?’; ‘which genes or pathways are associated with them?’ Effective answers to them will be essential to enhance diagnosis, intervention, and treatment of mental diseases.

However, there was almost no previous study on this important topic because of the lack of SNP data and effective analytics methods. Unlike other omics data, the specific SNP data of SCZ and BPD is relatively limited, not to mention ASM-SNP data. On the other hand, SNP data is typically high-dimensional data, where the number of variables (SNPs) is much greater than the number of observations (SCZ and BPD samples). Its feature selection is not well understood because it remains unclear how to select meaningful features among huge amounts of SNPs. Although previous studies have tackled SNP analysis related to other diseases, it is unknown how to effectively analyze SNP data on behalf of SCZ and BPD phenotype recognition and genetic etiology analysis [13].

In this study, we addressed the urgent challenges via state-of-the-art manifold learning and machine learning, novel SNP feature selection, and rigorous pathway analysis. To handle the SNP feature selection issue, we first employed nonnegative singular value approximation (NSVA), a data-driven feature selection algorithm, to identify the significant ASM-SNPs [14]. We found that the genes associated with the top 30% SNPs were significantly enriched in eight pathways involved in regulating nervous system functions. For example, Axon guidance and AMPK signaling pathways identified in this work were proved to be linked to SCZ and bipolar BPD [15].

We then employed different manifold learning methods along with their peers for the selected ASM-SNPs to investigate SCZ and BPD phenotype visualization and clustering under Gaussian mixture models [16]–[18]. Six manifold learning methods were investigated in this study. They included t-distributed stochastic neighbor embedding (t-SNE), locally linear embedding (LLE), isometric mapping (ISOMAP), kernel principal component analysis (KPCA), and multidimensional scaling (MDS) [19]–[28]. At the same time, we included other four peer methods: principal component analysis (PCA), sparse PCA, nonnegative matrix factorization (NMF), and random projection in experiments for comparisons [21], [25], [26].

Our results strongly showed that all the dimensional-reduction methods could achieve their best performance by only using the top-ranked ASM-SNPs under nonnegative singular value approximation (NSVA) [14]. The state-of-the-art manifold learning algorithm t-SNE achieved the best and most stable performance in phenotype clustering among all manifold learning and comparison methods. It suggested that SCZ and BPD phenotypes could be easily separated from controls, but there existed some high degree of misdiagnosis for SCZ and BPD. Our support vector machine (SVM) analysis also indicated that BPD and SCZ may not be separable from a machine learning perspective. Previous and recent clinical studies echoed the result and reported that BPD had a high misdiagnosis rate and commonly misdiagnosed as other mental disorders such as SCZ [29], [30]. The finding along with our pathway analysis indicated that SCZ and BPD could have the same root in genetics. They could be different variants from the same disease with shared phenotypes pathologically. As such, misdiagnosis between BPD and SCZ could be inevitable in terms of their genetic root. Furthermore, we demonstrated the advantage of t-SNE in phenotype diagnosis compared to other manifold learning peers by using different machine learning algorithms.

To the best of our knowledge, this is the first work that provides a novel and comprehensive ASM-SNP analysis for SCZ and BPD from machine learning and systems biology perspectives. Our study provided a novel genetic etiology analysis by finding pathways associated with SCZ and BPD via a data-driven feature selection. Our pathway analysis showed that genes associated with BPD and SCZ were significantly
enriched in eight pathways involved in brain neurotransmissions and relevant processes. The results not only echoed the state-of-art bioinformatics studies but also provided new insights on the misdiagnosis issue between BPD and SCZ.

II. MATERIALS AND METHODS

In this section, we introduced our proposed methods as well as related data acquisition and preprocessing. The proposed methods mainly consist of a data-driven feature selection algorithm for ASM-SNPs, manifold learning for BPD and SCZ visualization and clustering, downstream machine learning, and relevant pathway analysis. Figure 1 summarizes the workflow of our study.

First, raw ASM-SNPs were filtered through a quality-control procedure. The statistically significant ASM-SNPs were selected according to its false discovery ratios (FDR) from the ANOVA-test. Second, we proposed a data-driven feature selection method: nonnegative singular value approximation (NSVA) to glean the significant ASM-SNPs. Third, we exploited manifold learning along with its peers to visualize the ASM-SNP data and conduct phenotype clustering under the Gaussian mixture model (GMM). Furthermore, we employed state-of-the-art machine learning to discriminate SCZ and BPD phenotypes using the low-dimension embedding SNP data under t-SNE, which is the best-performed manifold learning model in phenotype clustering [20], [22]. Finally, we conducted a pathway enrichment analysis of genes associated with the top 30% ASM-SNPs to identify pathways related to SCZ and BPD. To decipher misdiagnosis between SCZ and BPD, we employed support vector machines (SVM) to validate the non-separability of BPD and SCZ by using important ASM-SNPs selected by nonnegative singular value approximation.

A. DATA ACQUISITION AND PREPROCESSING

We downloaded the raw SNP dataset GSE71443 from the NCBI GEO database [13]. It included 74 control subjects, 65 bipolar disorder (BPD), and 64 schizophrenia (SCZ) individuals from the Harvard Brain Tissue Resource Center and the Stanley Medical Research Institute [13], [31]. Most of the individuals were of European origin: 74 in the control group, 65 in the BPD group, and 64 in the SCZ group respectively.

To obtain the significant differential SNP loci, we first filtered those SNPs with missing annotations, not on autosomes or sex chromosomes, or diverged from Hardy-Weinberg equilibrium (HWE) with \( \rho < 10^{-10} \). Finally, we had a total of 627,693 SNPs left [13].

As the SNPs with FDR < 0.01 were generally associated with mental disease [13], we employed ANOVA to screen statistically significant SNPs under the cutoff. Then we removed those SNPs in the linkage disequilibrium (LD) with one another (\( r^2 > 0.25 \)). Finally, we obtained a dataset with a total of 1,345 ASM-SNPs in the BPD group, 3,051 in the SCZ group, and 1,457 in the control group respectively. In other words, our ASM-SNP dataset after preprocessing had a total of 5,843 SNPs (features) across 74 control, 65 BPD, and 64 SCZ samples.

B. NONNEGATIVE SINGULAR VALUE APPROXIMATION (NSVA): A DATA-DRIVEN FEATURE SELECTION

The SNP data after preprocessing is still high-dimensional data, in which the number of features is much greater than the number of observations (samples). How to select meaningful SNPs for the upstream analysis remained a challenge. Traditional feature selection methods assumed SNP data subject to a specific probability distribution and employed some corresponding statistical tests to identify meaningful features (SNPs) according to their \( p \)-values. The model-driven approaches suffered from ad-hoc results, because SNP data after preprocessing may not follow the assumed distribution even if their original data followed it. Thus, a data-driven feature selection method especially needed to perform meaningful feature selection for the dataset after preprocessing.

In this paper, we employed nonnegative singular value approximation (NSVA), a data-driven feature selection algorithm to identify meaningful SNPs from the SNP data after preprocessing [14]. It did not assume input data following any data distribution. Instead, it conducted feature selection by ranking the importance of each SNP by taking advantage of the nonnegativity of input SNP data. We describe the algorithm briefly as follows. More details about this method and its related proof can be found in Han’s work [14].

Given a nonnegative matrix with \( n \) features and \( m \) samples: \( X \in \mathbb{R}^{n \times m} \geq 0 \), it can be decomposed by singular value decomposition(SVD) as:

\[
X = \sum_{k=1}^{r} u_k s_k v_k^T
\]

where \( s_k \) represents the \( k^{th} \) singular values, \( v_k \in \mathbb{R}^m \), \( u_k \in \mathbb{R}^n \), and \( r = \min(n, m) \) is the rank of \( X \). Then, we have the following results:

1. Both \( u_1 \) and \( v_1 \) have only nonnegative entries, i.e., \( u_1 \geq 0 \) and \( v_1 \geq 0 \).
2. The vectors \( v_1 \) and \( u_1 \) contain at least one negative entry when \( j, k \geq 2 \).
3. Matrix \( X \) has the following first-level singular value approximation:

\[
X \sim u_1 s_1 v_1^T = \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ij} v_{ij}
\]

Since the nonnegativity of the \( u_1 \) and \( v_1 \), the \( i^{th} \) row of the matrix \( X \), which is an ASM-SNP in our context, can be approximated as \( X_i^T \sim s_1 u_1 v_1^T \). Thus, the contribution score of the \( i^{th} \) feature to all samples is defined as \( u_{ij} \sim (1/s_1) X_i^T v_1 \), where \( v_1 \) is the first eigenvector of \( X^T X \). Therefore, it can be used to determine the degree of importance of the \( i^{th} \) feature because it measures its projection on the largest singular value direction. Similarly, \( u_1 \) can indicate the importance of each feature. To rank the importance of each feature, all the entries in the nonnegative vector \( u_1 \) are sorted in a descending order. The order indicates the importance degree of the corresponding features.
FIGURE 1. The workflow of proposed ASM-SNP analysis.
The data-driven feature selection is good for high-dimensional omics data feature selection because the data variation explanation ratio of the first singular-value direction: $\rho = s_1 / \sum_{i=1} s_i$ is generally $\geq 50\%$. The SNP dataset after preprocessing also shares the property.

C. MANIFOLD

A manifold can be viewed as a set of data points that locally behave like the Euclidean space. Given a point $x_i = \{x_{i1}, x_{i2}, \cdots, x_{in}\}^T$ on a manifold, there exists a neighborhood $S$ consisting of a number of points close to $x_i$ under a distance metric. The neighbor $S$ behaves like the Euclidean space $\mathbb{R}^m$. The ‘behaves like’ mathematically means there is a 1-1 mapping available between $S$ and $\mathbb{R}^m$, i.e., $S$ is homeomorphic to the Euclidean space.

A line is a typical one-dimensional manifold. All points on the line locally behave like the one-dimensional Euclidean space. On the other hand, the ASM-SNP dataset in our context can be simply viewed as a dataset on a high-dimensional manifold because each point (an SNP) has $m > 3$ entries ($m = 203$ in our context).

D. MANIFOLD LEARNING

Manifold learning aims to find another ‘manifold’ mostly in a low-dimensional space to represent the original manifold in the high-dimensional space by preserving distance metrics at least locally in mapping. Unlike PCA conducting a linear dimension reduction by mapping data into a low-dimensional subspace spanned by maximum variance directions, manifold learning conducts nonlinear dimension reduction by preserving distance metrics between data points according to different standards. Such a distance-preservation mechanism enables manifold learning to maintain data locality well and unveil latent structures of data compared to PCA. The data locality refers to the intrinsic local geometric structure of data, which is not maintained by the classic decomposition-based linear dimension reduction method like PCA. PCA actually suffers from a holistic learning mechanism and lacks the full preservation of distance metrics besides missing the purely additive property [32].

Manifold learning (e.g., t-SNE) has demonstrated a better nonlinear feature extraction by maintaining good data locality for high-dimensional data [22], [23], [32]. Most manifold learning methods employ different nearest-neighbor learning approaches to embed a high-dimensional manifold to a low-dimensional manifold to capture intrinsic nonlinear data structures. To some degree, manifold learning has been an indispensable tool for high-dimensional data analysis.

ASM-SNP data in this study is typical high-dimensional data [33]. It is reasonable to exploit manifold learning to investigate SCZ and BPD phenotype clustering and analyze their relationships in the low-dimensional space generated from different manifold learning. It may contribute to explaining the widely-existing misdiagnosis between the two psychiatric disorders from a machine learning standing point [29], [30]. We briefly introduce major manifold learning algorithms in the following sections. The details about the other peer methods such as PCA, NMF, sparse PCA, and random projection can be found in [16], [34]–[40].

E. t-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (t-SNE)

Given a dataset $X = \{x_1, x_2, \cdots, x_n\}$ in a high-dimensional manifold: $x_i \in \mathbb{R}^N$, t-SNE aims to embed it into a low-dimensional manifold by computing a corresponding a low-dimensional dataset $Y = \{y_1, y_2, \cdots, y_n\}$, $y_i \in \mathbb{R}^l$, $l \ll N$. To achieve the embedding, t-SNE minimizes the Kullback-Leibler (K-L) distance between distributions $P$ and $Q$ [20]–[22].

The Gaussian distribution $P$ models the pairwise similarity between points in the original high-dimensional manifold. The similarity between $x_i$ and $x_j$ is represented as the joint probability $p_{ij} = \frac{p_{ii}p_{jj}}{2\pi\sigma_i^2}$. The joint probability is defined as the normalized sum of two conditional probabilities $p_{ii}$ and $p_{jj}$. The high probabilities correspond to similar points that is control, BPD, or SCZ sample in our context. The conditional probability $p_{ij}$ models the standardized similarity between $x_i$ and its neighbor $x_j$:

$$p_{ij} = \frac{e^{-\frac{1}{2\sigma_i^2}\|x_i - x_j\|^2}}{\sum_{k \neq i} e^{-\frac{1}{2\sigma_i^2}\|x_i - x_k\|^2}}$$

where $\sigma_i^2$ is the Gaussian kernel variance of $x_i$. It is set to control the perplexity of the distribution $P : H (P) = 2^{-\sum_{ij}\log_2 p_{ij}}$. A low perplexity indicates a probability distribution can predict the distance relationships between data points well. The perplexity controls the size of neighborhoods of t-SNE and $H (P) = 20$ means all but about 20 nearest neighbors of point $x_i$ will have $p_{ij} \sim 0$.

Similarly, the similarity $Q$ measures the similarity between two corresponding points $y_i$ and $y_j$ in the low-dimensional embedding. A normalized Students t-distribution with one degree of freedom is used to calculate the joint probability $q_{ij}$.

$$q_{ij} = \left(1 + \frac{\|y_i - y_j\|^2}{Z}\right)^{-1}$$

where $Z = \sum_{k \neq l} \left(1 + \frac{\|y_k - y_l\|^2}{Z}\right)^{-1}$. The exact locations of the low dimensional embedding points $y_i$, $i = 1, 2, \cdots, n$ are learned by minimizing the K-L distance between the two joint distributions:

$$KL (P||Q) = \sum_{i=1}^n \sum_{j=1, j \neq i}^n p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The objective function is non-convex. It is minimized by a general gradient descending way with the gradient:

$$\frac{d\varphi}{dy_i} = 4 \sum_{i \neq j} (p_{ij} - q_{ij}) q_{ij} Z (y_i - y_j)$$
t-SNE has demonstrated its built-in advantages in capturing intrinsic nonlinear data structures and maintaining good data locality. Also, for this reason, t-SNE is reported to be more robust with respect to the presence of outliers and noise [41]. The complexity of the t-SNE is \( O(n^3) \) for given \( n \) points in most implementations though some t-SNE variants (e.g., Barns-Hut-SNE) only have a \( O(n \log n) \) complexity under different acceleration techniques [22], [42].

**F. LOCALLY LINEAR EMBEDDING (LLE)**

LLE claims to be able to generate highly nonlinear embeddings for high-dimensional data while keeping the underlying neighborhood structure than PCA and MDS besides avoiding the local minima issue in optimization [23], [24]. The principle of LLE is that a point in the high dimensional manifold can be represented linearly by its adjacent points. LLE reconstructs each point using its nearest neighbors by solving the following optimization problem. The solution to the problem is a weight matrix \( W = (w_{ij}) \) that represents the intrinsic neighborhood geometric structures.

Given a dataset \( X = \{x_1, x_2, \cdots, x_n\} \) in a high-dimensional manifold: \( x_i \in \mathbb{R}^N \), the \( k \) nearest neighbors of \( x_i \) are determined by minimizing the cost function:

\[
\varphi (w) = \sum_{i=1}^{n} \left\| x_i - \sum_{j=q(i)} w_{ij} x_j \right\|^2
\]  

where \( q(i) \) is an index set of the \( k \) nearest neighbors of \( x_i \), \( x_j \) is the nearest neighbor of \( x_i \), and \( w_{ij} \) is a linear weight assigned to \( x_i \) and \( x_j \). If \( x_i \) is not in the neighborhood of \( x_j \), \( w_{ij} = 0 \). The row sum of the weight matrix is required to be 1: \( \sum_j w_{ij} = 1 \).

LLE seeks a corresponding low-dimensional embedding \( Y = \{y_1, y_2, \cdots, y_n\} \), \( y_i \in \mathbb{R}^l \), \( l < N \), to maintain the 'local linear relations' by minimizing another loss function:

\[
\varphi (Y) = \sum_{i=1}^{n} \left\| y_i - \sum_{j=p(i)} w_{ij} y_j \right\|^2
\]  

where \( p(i) \) is the index set of the \( k \) nearest neighbors of \( y_i \). The point \( y_i \) is the embedding of \( x_i \) and \( y_j \) is one of the closest neighbors of \( y_i \). The optimal embedding finding is equivalent to solving a sparse eigenvalue problem [23].

LLE keeps local data geometric structures in the low-dimensional embedding by maintaining the local linear relations in each neighbor. It makes a high-dimensional manifold locally isometric to its low-dimensional embedding. However, it may miss or distort the whole global data geometric structure in the embedding space, when the neighborhood size selection in LLE is inappropriate or input data contains noise. For this reason, LLE may show some level of instability in dimension reduction for high-dimensional SNP data.

**G. MULTIDIMENSIONAL SCALING (MDS)**

Classic MDS aims to preserve the Euclidean distance between samples in the low dimensional space as consistent as possible in the high-dimensional space under a cost function. There are different variants of MDS (e.g., Local MDS) with quite different characteristics. We focus on the classic one in our context [25]–[27]. Compared with LLE and PCA, it demonstrates a high complexity and slow convergence [24]. MDS uses the similarity between the paired samples on behalf of the Euclidean distance to construct an appropriate low-dimensional space.

Given input dataset \( X = \{x_1, x_2, \cdots, x_n\}, x_i \in \mathbb{R}^N \), \( D = (d_{ij}) \), \( d_{ij} \in \mathbb{R}^{N \times N} \) is the pairwise similarity matrix, which \( d_{ij} \) represents the distance between samples \( x_i \) and \( x_j \). The distance metric by default is the Euclidean distance though other metrics can also be used. The low-dimensional reduced embedding \( Y = \{y_1, y_2, \cdots, y_n\}, y_i \in \mathbb{R}^l \), \( l < N \), can be obtained by minimizing the following loss function:

\[
\varphi (Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| d_{ij} - d_{ij}^{(y)} \right\|^2
\]  

where \( d_{ij}^{(y)} \) is the Euclidean distance between embedding samples \( y_i \) and \( y_j \).

The optimization problem can be solved by conducting singular value decomposition \( XX^T \) by assuming \( X \) is a standardized dataset. Since the classic MDS simply seeks to calculate the embedding data by minimizing the difference between the two similarity matrices, it may not capture data locality as well as t-SNE because does not use the nearest neighborhood search mechanism to maintain intrinsic data geometry locally [41].

**H. ISOMETRIC FEATURE MAPPING (ISOMAP)**

The ISOMAP algorithm aims to keep the geodesic distances between points under embedding. The geodesic distance is the shortest path between two points on a curved surface in the manifold. Compared with LLE, it can keep the whole global data geometric structure better, but it suffers from high complexities and can be sensitive to noise data [24], [28].

ISOMAP first employs the nearest neighbor search \( G \) to build a neighborhood graph that is defined in the neighborhood of each point of the input dataset \( X \) [28]. The neighborhood is determined by an Euclidean distance threshold \( \varepsilon \) so that all points in the neighborhood has a distance less than the threshold \( d(x_i, y_j) \leq \varepsilon \).

Then, it calculates the pairwise geodesic distance of the point \( x_i \) and \( x_j \) in a graph \( G \) by finding the shortest path via the Dijkstra algorithm to obtain the similarity matrix \( D(G) = (d_{ij})^G \) in the geodesic space. Finally, it employs the classical MDS algorithm to construct the embedding data \( Y \) by using \( (d_{ij})^G \) to substitute \( d_{ij} \) in equation (9).

It can somewhat be viewed as a more generalized MDS that can extract more nonlinear data characteristics in embedding because it applies MDS locally in the geodesic space of the original high-dimensional manifold [26]. The similarity matrix in the geodesic space can be more representative than that in the Euclidean space. However, ISOMAP is sensitive to the neighborhood size and noise. If the neighborhood size was too large, ISOMAP would suffer from high complexity, because of the Dijkstra algorithm \( O(n^3) \) [28].
I. KERNEL PRINCIPAL COMPONENT ANALYSIS (KPCA)

Kernel PCA (KPCA) is an extension of PCA in the feature space by using kernel tricks [19]. It formalizes the steps of PCA in the form of inner products and employs kernel tricks to calculate the data projections on the kernel PCs. Compared to other manifold learning algorithms, it seeks an ‘embedding’ of the original space into the feature space, which is a high-dimensional or infinite-dimensional manifold, to retrieve latent data characteristics that cannot be discovered in the original space. Unlike PCA, the real kernel PCs cannot be found explicitly, because the covariance matrix in the feature space is unknown due to the unknown mapping function that maps input data into the feature space. KPCA has the following three steps.

The first step is to compute a kernel matrix \( K = (k_{ij}) \), \( k \left( x_i, x_j \right) = (\varphi(x_i), \varphi(x_j)) \) for normalized input data \( X = \{ x_1, x_2, \ldots, x_n \} \). The kernel function \( k(x, y) \) is generally a nonlinear kernel (e.g., Gaussian kernel) and the function \( \varphi \) maps input data to its feature space. The second step is to compute the eigenvalue and eigenvectors of the kernel matrix \( K: K\alpha = \lambda\alpha \), which is the equivalent eigenvalue problem of PCA in the feature space. The third step is to project data in the feature space onto kernel PCs.

Kernel PCA is quite sensitive to noise data because noise can be amplified in the feature space and make extracted data characteristics meaningless [19], [41], [43]. It does not preserve the intrinsic geometric structures well as other manifold learning methods though it preserves distance metrics in mapping.

J. PHENOTYPE CLUSTERING UNDER THE GAUSSIAN MIXTURE MODELS (GMM)

As the control, BPD, and SCZ samples were normally distributed under the whole population, we employed Gaussian mixture models (GMM) to cluster them under manifold learning and its peers to examine phenotype distribution in different embedding spaces. GMM assumed the data points to be clustered as Gaussian distribution sampling points. The parameters of the Gaussian distribution were estimated by sampling points using a maximum likelihood estimation method [18]. It estimated the probability of each component generated by the input data as follows:

\[
\gamma(i, k) = \frac{\pi_k N(x_i/\mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i/\mu_j, \sigma_j)} \tag{10}
\]

where \( \gamma(i, k) \) represented the probability of the sample \( x_i \) generated by the \( k \)th component. The parameter \( \pi_k \) indicated the probability of \( k \)th component selected under the normal distribution \( N(\mu_k, \sigma_k) \). The parameters were estimated in the fly of computing until the GMM clustering converged.

We used the silhouette coefficient to evaluate the phenotype clustering of control, SCZ and BPD samples from GMM. The silhouette coefficient measures the difference between the intra-cluster distance and the inter-cluster distance for clusters [44]. It is defined as

\[
SilC = \frac{b - a}{\max\{a, b\}} \tag{11}
\]

where \( a \) is the average distance within the cluster, and \( b \) is the closest average distance between different phenotype groups. The silhouette coefficient ranges from \(-1\) to \(1\). The larger the silhouette coefficient implies that the closer the homogeneous samples are and the farther the heterogeneous samples are. That is, a good phenotype clustering should result in a large silhouette coefficient and vice versa. In our implementation, we employ the Euclidean distance to calculate all silhouette coefficients.

III. RESULTS

In this section, we examined the phenotype clustering, visualization and classification of control, BPD and SCZ samples based on the selected ASM-SNPs under different manifold learning methods and their peers. The methods included t-SNE, LLE, ISOMAP, MDS, kernel PCA, PCA, sparse PCA, NMF, and random projection. Finally, we conducted a pathway enrichment analysis of genes corresponding to the 30% top-ranked ASM-SNPs selected by nonnegative singular value approximation (NSVA) to validate their pathological functional associations with BPD and SCZ [14].

The results from phenotype clustering and visualization suggested that the misdiagnosis between SCZ and BPD could be inevitable in the clinic. Our pathway analysis showed they shared almost the same genetic variations. The same genetic roots caused them to share overlapped or the same symptoms. BPD and SCZ symptoms are likely different expressions of the same genetic disease for different individuals. The traditional psychiatric categorization can be blind to the same genetic root. On the other hand, the SCZ and BPD samples demonstrated a strong non-separable property in visualization under machine learning. The non-separable property of their SNP data echoed that the two diseases could be the same psychiatric disease. Furthermore, t-SNE distinguished itself in clustering and visualization among all dimension reduction methods for its good locality unveiling in the low-dimensional space.

A. ASM-SNP PHENOTYPE CLUSTERING ANALYSIS

We used the Gaussian mixture models (GMM) to conduct phenotype clustering by using the two-dimensional projection data from manifold learning and its peers [18]. We selected the top 10%, 20%, ..., 90%, and 100% ASM-SNPs by using NSVA to evaluate the effectiveness of the dimension reduction methods in capturing built-in data structures and latent data characteristics. Figure 2 illustrated the corresponding average silhouette coefficient consisting of top-ranked ASM-SNPs on behalf of the percentage.

We uniformly set the neighbor size as 5 in the LLE and ISOMAP in the experiment. Accordingly, we set the perplexity value in t-SNE as 5. We chose the sparsity in sparse PCA as 0.8 to pursue a high-level sparseness for each PC.
The coefficient of the regularization in NMF was set as $\alpha = 1.0$. Also, we employed a cosine kernel in kernel PCA.

According to the silhouette coefficient plot, it was easy to find that t-SNE had the best and most stable phenotype clustering performance with respect all datasets among all methods. The other methods all demonstrated large oscillations of their silhouette coefficients no matter they were manifold learning methods or decomposition-based methods (e.g., NMF). This probably lied in that t-SNE had a built-in advantage in revealing local data characteristics and maintaining the original intrinsic global data structure than the other. The normalized Gaussian and t-distribution-based similarity calculations as well as K-L distance-based embedding computing not only maintained the geometric structure of data well in embedding but also made it robust to data with noise.

The MDS and random projection methods both had the worst performance in clustering. It suggested that both of them could be sensitive to the noise of the SNP data. The MDS’ Euclidean distance preservation mechanism under noise could result in a low-quality low-dimension data projection and poor clustering. Similarly, the possible noise in the SNP data could bring more ad-hoc and lower the quality of the embedding in random projection for its ‘randomness’.

The clustering performance of LLE fluctuated with the percentage of selected SNPs. This indicated that LLE might not be feasible for ASM-SNP data because it lacks global data structure keeping in dimension reduction. LLE made a high-dimensional manifold locally isometric to its embedding by maintaining the linear-combination relationship between points in each neighborhood. Such a local linear relationship keeping mechanism let LLE miss to capture the global data geometric structure in the embedding. When the neighborhood size in LLE was ‘unmatched’ to input data or input data with noise (e.g., SNP data), it would show a high-degree oscillation in the clustering. For the same reason, LLE performed well when more SNPs were selected because more features contributed to the generalization of the local linear relationship keeping.

On the other hand, the clustering performance of sparse PCA also demonstrated large oscillations. It probably lied in the unpredictable impacts of sparseness added in the dimension reduction on clustering. However, sparse PCA obviously outperformed PCA, KPCA, MDS, NMF, and random projection in clustering. ISOMAP generally outperformed LLE when $<50\%$ SNPs were selected in the embedding. It was probably because of its good global data geometric capturing mechanism.

Manifold learning showed advantages in phenotype clustering because the top methods were mostly from manifold learning. t-SNE, LLE, sparse PCA, and ISOMAP demonstrated better performance than PCA, KPCA, MDS, NMF, and random projection in SNP phenotype clustering. It was noted that t-SNE outperformed LLE, sparse PCA and ISOMAP for its built-in advantages in unveiling local data locality while maintaining the global data structure. The average silhouette coefficients of phenotype clustering under t-SNE attained its best value when using only about 40% or fewer SNPs. The best-performed t-SNE achieved a decent phenotype clustering by only using the least SNPs. It implied that SCZ and BPD phenotype clustering would need the group of most important SNPs. Alternatively, it demonstrated the effectiveness of nonnegative singular value approximation (NSVA) in SNP feature selection.
B. ASM-SNP DATA PHENOTYPE VISUALIZATION

Figure 3. presented the visualization of 74 control, 65 BPD, and 64 SCZ samples by each dimension reduction method at its largest silhouette coefficient. t-SNE achieved the most meaningful visualization compared to its peers.

Figure 3(a) showed almost all control samples under t-SNE were clustered in four well-shaped control groups in the embedding space. One control group had the majority of control samples. The rest of the controls formed their sub-clusters. It might suggest that control samples from heterogeneous control collections in data acquisition.

However, the tightly-clustered control groups had clear boundaries with those of BPD and SCZ samples rather than mixed with them. The BPD and SCZ samples were wired together as several independent groups. In other words, health and non-health subjects (SCZ and BPD) both had their local regions in visualization.

The SCZ and BPD samples mixed-together in t-SNE visualization supported the fact that there widely existed misdiagnosis between BPD and SCZ in the clinic because BPD shared similar or same psychotic symptoms with and SCZ. For example, some BPD people also experienced similar psychotic symptoms as SCZ people. Also, the bipolar type schizoaffective disorder was already counted as a type of SCZ [30].

On the other hand, LLE, ISOMAP, sparse PCA, PCA, and NMF can capture a portion of control subjects in visualization. However, unlike the phenotype distribution visualization in t-SNE, control and non-control samples (BPD and SCZ) did not have their local regions, i.e., they all mixed or wired together. For example, ISOMAP (see Figure 3(d)) even unveiled global data structures of the AMS-SNP data but failed to unveil their local regions in clustering. The visualizations under LLE, sparse PCA, PCA, and NMF all...
mixed BPD and SCZ samples though they loosely identified some local control group regions.

Similarly, MDS mixed all samples in a ball shape though it had somewhat an identified control group. Also consistent with their clustering results, both kernel PCA and random projection both had the worst visualization performance: they completely mixed the three groups and made them indistinguishable.

All methods in the study indicated that BPD and SCZ samples were not distinguishable from SNP data analysis and knowledge-based visualization. Such a result pushed us to reconsider the traditional classification of BPD and SCZ, which was based on psychiatric or psychological behaviors. They both might have shared the same pathogenesis on behalf of genetics. That was probably why the misdiagnosis between BPD and SCZ widely existed [30]. If so, the misdiagnosis would be unavoidable in the clinic because of their pathogenesis. We employed pathway analysis to address this concern to seek if they shared the same or similar genetic variations by the end of this study.

### C. BPD AND SCZ PHENOTYPE DISCRIMINANT ANALYSIS

We further employed support vector machines (SVM) to validate the non-separability of BPD and SCZ by dropping all control samples and queried how well we could discriminate the two phenotypes under a binary classification environment. SVM has a good advantage in investigating data separability issue for its built-in geometric interpretation in classification. If input data is linearly separable, SVM can achieve almost perfect classification performance (e.g. ∼100% accuracy) under linear kernels [45]. If data is non-linearly non-separable, SVM can only achieve a mediocre or even poor performance (e.g., <70% accuracy) under a nonlinear kernel such as a Gaussian kernel [45], [46].

Figure 4 showed the SVM classification performance under ‘RBF’ kernels on the datasets with different percentages of top-ranked SNPs selected by NSVA in terms of accuracy, recall, precision, and F1-scores. We partitioned the original BPD and SCZ samples as 80% training and 20% test in classification. We found that there was almost no difference between all kinds of nonlinear and linear kernels: all kernels achieved less than 65% accuracy. Almost all F1 scores were less than 60%. Our F1 scores were unbiased because of the balanced distribution of BPD and SCZ [47].

Furthermore, all low recalls were <55% indicated that the ratios of correctly predicted positive samples were approximate to a random diagnostic case with 50% recall. All the machine learning results from SNP data suggested that BPD and SCZ could not be distinguishable well for their same pathogenesis on behalf of genetics. It echoed the previous manifold learning analysis that showed that BPD and SCZ could be non-separable in phenotype detection.

### D. PHENOTYPE DISCRIMINANT ANALYSIS FOR CONTROL AND SCZ/BPD

Figure 5 illustrated that the binary classification performance of the same rbf-SVM model to distinguish control samples with respect to SCZ and BPD respectively under the same
classification setting as Figure 4. It indicated that the SVM model can easily distinguish SCZ from normal samples with a high diagnostic accuracy (e.g. $\geq 98\%$) by only using top 10% NSVA-selected SNPs. Similarly, rbf-SVM can diagnose BPD samples from controls with $>96\%$ accuracy under the same top 10% selected SNPs. Both had a correspondingly strong recall, precision and F1-score support.

The results suggest that separating controls with SCZ/BPD can be a linearly non-separable problem that tolerated few misclassifications. It indicated that a BPD/SCZ patient had quite different genetic expressions than a normal sample. The slightly better diagnosis accuracy of classifying SCZ samples with controls than that of classifying BPD samples with controls seemed to suggest that that BPD can be a subtype of SCZ but with some ‘blurring signals’. However, separating SCZ with BPD can be a nonlinear non-separable problem for its low machine learning performance because they shared the same or similar genetic roots. It was also worthwhile to point out that better or at least equivalent diagnostic accuracy can be achieved under the top-selected SNPs under NSVA. It again provided a strong support for the effectiveness of the data-driven feature selection.

E. THE MISDIAGNOSIS BETWEEN SCZ AND BPD

Clinically, SCZ can have a very large spectrum of symptoms that cover almost all symptoms of BPD [29], [30]. Recent psychiatry research reported that there existed a large percentage of misdiagnosis between SCZ and BPD [29]. Singh and Rajout reported 61% misdiagnosis rates between SCZ and BPD by looking for psychotic biological markers [30]. Hirschfeld et al. reported that the misdiagnosis rate for BPD could reach as high as 69% [48]. Shen et al. investigated the misdiagnosis of BPD and found that the misdiagnosis rate can reach 76.8% [48]. Witt et al. conducted a genome-wide study and revealed the genetic overlap between SCZ, BPD, and MDD [49]. Chen et al. conducted another Genome-wide association study (GWAS) for SCZ, BPD and SCZ and conclude that BPD and MDD (major depression) both could be the subtypes of SCZ [50]. The results strongly supported our results that the misdiagnosis would be unavoidable in the clinic because of their same or similar genetic root. Correspondingly SCZ and BPD were non-separable on behalf of machine learning using SNP data.

On the other hand, it was possible that BPD and SCZ had the same genetic root, but demonstrated some differences in clinical symptoms from a precision medicine viewpoint. This was because a same psychiatric disease even from a same genetic root can have different subtypes on different people due to different environments and personal differences [30].

F. EMBEDDING SPACE PHENOTYPE DIAGNOSIS

To further validate the effectiveness of t-SNE, we compared it with other peers in phenotype diagnosis by integrating it with state-of-the-art machine learning. We first employed random forests (RF) to evaluate the two-dimensional embeddings of the manifold learning methods and their peers [51].
It answered the query: ‘which embeddings will be more informative in determining a given sample is a psychiatric disorder one or a normal one? under the RF model?’

We employed the five-fold cross-validation and calculated the average AUC (area under the receiver operating characteristic curve) value to measure diagnosis output quality. The higher the AUC value, the better quality of the embedding on behalf of phenotype diagnosis.

Figure 6 showed the average AUC values of the RF model by using the two-dimensional embeddings of the datasets with different percentages of top-ranked SNPs selected by NSVA. The reason we employed the RF model lied in its stability and overfitting robustness in phenotype detection than its peer models [51]–[59]. t-SNE achieved a leading performance compared to its peers. Consistent with the previous results in clustering, t-SNE only needed about 20% top-ranked SNPs to attain the best phenotype diagnosis.

The pathways implied that BPD and SCZ would share the same genetic roots rather than two different types of genetic diseases. Their associated pathways were mainly involved in central nervous system activities. It was almost impossible to classify the pathways according to two separate BPD and SCZ related groups. On the other hand, it suggested the genes associated with the selected ASM-SNPs were most likely those related to the nervous system regulation or connection. The possible mutations of the genes would be the shared pathogenesis for BPD and SCZ.

In other words, BPD could be a subtype of SCZ but have a different or sometimes similar psychiatric behaviors. The psychiatric or psychological behavior-based BPD and SCZ classifications seemed to be inconsistent with our pathway and machine learning analysis results. We provided detailed pathway analysis information as follows.

### G. SNP Driven Pathway Enrichment Analysis

To query whether BPD and SCZ shared the same pathogenesis, we employed pathway analysis to investigate the biological functions of genes corresponding to the top-ranked 30% ASM-SNPs in KEGG pathways. The reason for us to select the top 30% ASM-SNPs was the phenotype clustering and diagnosis analysis confirmed about top 20%-40% ASM-SNPs could achieve the best clustering and diagnosis performance.

Figure 7 illustrated the pathway analysis results. It indicated that the genes associated with the SNPs were significantly enriched in eight pathways. The eight pathways included GABAergic synapse, Glutamatergic synapse, Circadian entrainment, Cell adhesion molecules, ECM-receptor interaction, Axon guidance, Nicotine addiction, and AMPK signaling pathways. All the pathways were directly involved in important biological functions of the nervous system, such as neurotransmitters promoting and suppression, synaptic formation, and differentiation. They also interfered with the production of high-energy compounds and related functionalities.

The pathways implied that BPD and SCZ would share the same genetic roots rather than two different types of genetic diseases. Their associated pathways were mainly involved in central nervous system activities. It was almost impossible to classify the pathways according to two separate BPD and SCZ related groups. On the other hand, it suggested the genes associated with the selected ASM-SNPs were most likely those related to the nervous system regulation or connection. The possible mutations of the genes would be the shared pathogenesis for BPD and SCZ.

In other words, BPD could be a subtype of SCZ but have a different or sometimes similar psychiatric behaviors. The psychiatric or psychological behavior-based BPD and SCZ classifications seemed to be inconsistent with our pathway and machine learning analysis results. We provided detailed pathway analysis information as follows.

Gamma-aminobutyric acid (GABA) in the GABAergic synapse pathway was the most abundant inhibitory neurotransmitter in the mammalian central nervous system [60]. Craddock et al. found substantial evidence between variation...
in GABAA receptor genes (GABRB3) and high risk for the schizoaffective disorder, bipolar type \( (p = 3.8 \times 10^{-6}) \) [61].

Glutamate in the Glutamatergic synapse pathway was the major excitatory neurotransmitter in the mammalian central nervous system [62]. Kohlrausch et al. found that high expression of GRIN2B \((\text{rs}7311519)\) was an implication of the pathogenesis of obsessive-compulsive disorder (OCD) [63]. Wu et al. also found that GRIN2B could be related to the orbitofrontal cortex, anterior cingulate cortex, and thalamus with OCD [64].

The Cell adhesion pathway played an essential role in the early development of the nervous system, not only making brain functions highly coordinated, but also maintaining synaptic formation through the connection between cell-cell adhesions. Reissner et al. discovered that NLGN1 \((\text{rs}16833234)\) could encode proteins on the nerve cells and participate in the synapse formation and remodeling [65].

The Extracellular matrix (ECM) pathway was related to cell functions such as adhesion, migration, and differentiation, which had specific effects on the formation of synapses. Suidan et al. found that the VWF \((\text{rs}1800385)\) in the endothelial cells interfered with the nervous system through the permeability of the blood-brain barrier (BBB) [66].

In addition, the scavenger receptor CD36 \((\text{rs}7999490)\) could activate brain microglia related to brain pathologies such as Alzheimer’s and malaria. CD36 may influence the supplement and metabolism of FA polyunsaturated fatty acids (PUFAs), which were important for brain function and mostly derived from the plasma [67].

The Axon guidance represented a critical stage in the formation of the neuronal network. In a two-phase case-control study in northeast China, Wang et al. verified the correlation between genes in the axon guidance pathway and schizophrenia [68].

Besides, the nicotine addiction and AMPK signaling pathways were found to be related to the activity of neurotransmitters in the brain [15], [69]. People with a symptom of severe depression, schizophrenia, or bipolar disorder commonly had signs of a circadian entrainment disorder [70].

In sum, our pathway analysis indicated BPD and SCZ shared the same genetic root. All the enrichment pathway analysis showed that they were from the same set of pathways and genes. It was likely that BPD could be a subtype of SCZ rather than an independent disease type from our bioinformatics analysis. The result was consistent with those from previous machine learning, i.e., BPD and SCZ were not separable under low-dimensional embedding and SVM diagnosis. Thus, their misdiagnosis in the clinic would be unavoidable. On the other hand, the meaningful pathways derived from selected ASM-SNP also supported the effectiveness of the proposed feature selection algorithm.

IV. DISCUSSION AND CONCLUSION
In this study, we presented a comprehensive novel investigation for BPD and SCZ by using ASM-SNP data. We proposed a data-driven feature selection for SNP data: non-negative singular value approximation (NSVA) [14]. We then examined major manifold learning algorithms along with its
peers (e.g., sparse PCA) in ASM-SNP phenotype clustering, visualization, and diagnosis. We employed pathway enrichment analysis to conduct pathogenesis analysis for BPD and SCZ besides demonstrating the effectiveness of the proposed feature selection algorithm. The related data and codes are available at https://github.com/condor2020/neuro_ASM.

Our studies demonstrated the effectiveness of the proposed SNP feature selection. Along with pathway analysis, it provided new insights into genetic etiology analysis for BPD and SCZ. We further showed that almost all manifold learning and its peers could achieve their best performance in SNP phenotype clustering and diagnosis by only using the top-ranked SNPs under the proposed feature selection method.

Our studies also showed that the superiority of t-SNE in phenotype clustering, visualization, and diagnosis for SNP data, besides the possible advantage of manifold learning over decomposition-based dimension reduction methods (e.g., NMF). The good performance of t-SNE seemed to lie in its built-in advantages in capturing intrinsic global data structure besides maintaining good locality in embedding [22], [41]. t-SNE’s normalized Gaussian and t-distribution-based similarity calculations and K-L distance-based embedding computing made it capture the global and local structure of data well in embedding and robust to data with noise [20].

Compared to t-SNE, LLE suffered from losing global data geometric structure tracking in embedding though keeping a good local data structure in embedding [23]. It also demonstrated instability in phenotype clustering for high-dimensional SNP data with some level of noise for its special embedding calculations that maintained the local linear relations in each neighbor. The decomposed based decomposition dimension reduction methods (e.g., NMF) generally would show a lower level performance in SNP phenotype clustering, visualization, and diagnosis, compared to their manifold learning peers (e.g., t-SNE). For example, NMF’s parts-based learning mechanism might easily let NMF lose global data structure tracking in learning and result in poor phenotype visualization and clustering.

It was noted that we used the Euclidean distance to model sample similarity in t-SNE in this study. We also tried all kinds of different distances in our t-SNE based phenotype visualization and found that correlation and Manhattan distance can achieve comparable clustering performance as the Euclidean distance. However, it remained unknown which distance metric will be optimal. Furthermore, it was not clear how to choose an optimal perplexity that controls neighbor size in t-SNE [20], [22], [27]. We are interested in investigating optimal distance metric inference via kernel-based learning in similarity modeling of t-SNE besides seeking an optimal perplexity according to data entropy [71], [72].

More importantly, our machine learning and pathway analysis results both showed that the misdiagnosis between SCZ and BPD could be unavoidable. Our pathway analysis showed that SCZ and BPD could share the same pathological root in genetics. The BPD could be a subtype of SCZ rather than an independent psychiatric disease [73]. That may be why they share many similar or the same symptoms in the clinic. Our visualization studies from manifold learning along with its peers also strongly suggested that the SNP data of SCZ and BPD could not be separable under low-dimensional embeddings. It was further verified by BPD and SCZ phenotype support vector machine discriminant analysis. The result should shed light on the existing psychiatry research to reexamine the existing behavior-based classification for BPD and SCZ.

On the other hand, we will continue to investigate the clinic ‘misdiagnosis’ between SCZ and BPD from bioinformatics and epigenetics standing point to look for possible further division of SCZ [46]. For example, investigating the possible impacts of individual SNPs on SCZ subtype generation and classification can be a more practical approach to unveil SCZ pathological subdivisions along with new learning techniques such as LIME and ResNet for high-dimensional SNP data [74], [75]. In addition, we will consider the use of EEG and fMRI data to comprehensively explore the pathogenesis of mental illness along with bioinformatics data [76], [77].

COMPETING INTERESTS
The authors declare that no conflict of interests exist regarding the publication of this article.

AUTHORS’ CONTRIBUTIONS
Han designed the methods. Li, Liu, and Han analyzed the data. Liu and Han wrote and finalized this article.

REFERENCES
[1] M. C. O’Donovan, “Identification of loci associated with schizophrenia by genome-wide association and follow-up,” Nature Genet, vol. 40, no. 9, pp. 1053–1055, Sep. 2008, doi: 10.1038/ng.201.
[2] M. A. R. Ferreira, “Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder,” Nature Genet, vol. 40, no. 9, pp. 1056–1058, Sep. 2008, doi: 10.1038/ng.209.
[3] N. Naidoo, Y. Pawitan, R. Soong, D. N. Cooper, and C.-S. Ku, “Human genetics and genomics a decade after the release of the draft sequence of the human genome,” Hum Genomics, vol. 5, no. 6, pp. 577–622, Dec. 2011, doi: 10.1186/1479-7364-5-6-577.
[4] C. C. Chung and S. J. Chanock, “Current status of genome-wide association studies in cancer,” Hum. Genet, vol. 130, no. 1, pp. 59–78, Jul. 2011, doi: 10.1007/s00439-011-1030-9.
[5] L. O. Bryzgalov, E. E. Korbolina, I. I. Brusentsso, E. Y. Leberfarb, N. P. Bondar, and T. I. Merkulova, “Novel functional variants at the GWAS-implicated loci might confer risk to major depressive disorder, bipolar affective disorder and schizophrenia,” BMC Neurosci., vol. 19, no. S1, p. 22, Apr. 2018, doi: 10.1186/s12868-018-0414-3.
[6] D. M. Ruderfer, “Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes,” Cell, vol. 173, no. 7, pp. 1705–1715, Jun. 2018, doi: 10.1016/j.cell.2018.05.046.
[7] X. Xiao, H. Chang, and M. Li, “Molecular mechanisms underlying non-coding risk variations in psychiatric genetic studies,” Mol Psychiatry, vol. 22, no. 4, pp. 497–511, Apr. 2017, doi: 10.1038/mp.2016.241.
[8] O. B. Smeland, “Genome-wide analysis reveals extensive genetic overlap between schizophrenia, bipolar disorder, and intelligence,” Mol Psychiatry, vol. 2019, pp. 1–18, Jul. 2019, doi: 10.1038/s41380-018-0332-x.
[9] T. 1000 Genomes Project Consortium, “A global reference for human genetic variation,” Nature, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.
[10] T. International SNP Map Working Group, “A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms,” Nature, vol. 409, no. 6822, pp. 928–933, Feb. 2001, doi: 10.1038/35057149.
[57] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.

[58] An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cristianini Nello, 2000.

[59] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[60] J. R. Chalfiaux and A. G. Carter, “GABAB receptor modulation of synaptic function,” Current Opinion Neuropsychol., vol. 21, no. 2, pp. 339–344, Apr. 2011, doi: 10.1016/j.conb.2011.02.004.

[61] N. Craddock et al., “Erratum: Strong genetic evidence for a selective influence of GABAA receptors on a component of the bipolar disorder phenotype,” Mol. Psychiatry, vol. 15, no. 11, p. 1211, Nov. 2010, doi: 10.1038/mp.2010.62.

[62] C. H. Kim, J. Lee, J.-Y. Lee, and K. W. Roche, “Metabotropic glutamate receptors: Phosphorylation and receptor signaling,” J. Neurosci. Res., vol. 86, no. 1, pp. 1–10, Jan. 2008, doi: 10.1002/jnr.21437.

[63] F. B. Kohlrausch, I. G. Giori, F. B. Melo-Felippe, T. Vieira-Fonseca, L. G. C. Velarde, J. B. de Salles Andrade, and L. F. Fontenelle, “Association of GRIN2B gene polymorphism and obsessive compulsive disorder and symptom dimensions: A pilot study,” Psychiatry Res., vol. 243, pp. 152–155, Sep. 2016, doi: 10.1016/j.psychres.2016.06.077.

[64] K. Wu, G. L. Hanna, P. Easter, J. L. Kennedy, D. R. Rosenberg, and P. D. Arnold, “Glutamate system genes and brain volume alterations in pediatric obsessive-compulsive disorder: A preliminary study,” Psychiatry Res., Neuroimag., vol. 211, no. 3, pp. 214–220, Mar. 2013, doi: 10.1016/j.pscychresns.2012.07.003.

[65] C. Reissner, M. Klose, R. Fairless, and M. Missler, “Maturational analysis of the neurexin/neuroligin complex reveals essential and regulatory components,” Proc. Nat. Acad. Sci. USA, vol. 105, no. 39, pp. 15124–15129, Sep. 2008, doi: 10.1073/pnas.0801639105.

[66] G. L. Suidan, A. Brill, S. F. De Meyer, J. R. Voorhees, S. M. Cifuni, J. E. Cabral, and D. D. Wagner, “Endothelial von willebrand factor promotes blood–brain barrier flexibility and provides protection from hypoxia and seizures in mice,” Arterioscler Thromb Vasc Biol., vol. 33, no. 9, pp. 2112–2120, Sep. 2013, doi: 10.1161/ATVBAHA.113.301362.

[67] N. A. Abumrad, M. Ajmal, K. Pothakos, and J. K. Robinson, “CD36 expression and brain function: Does CD36 deficiency impact learning ability?” Prostaglandins Other Lipid Mediators, vol. 77, nos. 1–4, pp. 77–83, Sep. 2005, doi: 10.1016/j.prostaglandins.2004.09.012.

[68] Z. Wang, P. Li, T. Wu, S. Zhu, L. Deng, and G. Cui, “Axon guidance pathway genes are associated with schizophrenia risk,” Expierim. Therapeutic Med., vol. 16, no. 6, pp. 4519–4526, 2018, doi: 10.1016/j.exer.2017.11.087.

[69] M. Mozrzymeswic and R. F. Tyndale, “Pharmacogenetics: A tool for identifying genetic factors in drug dependence and response to treatment,” Addiction Sci. Clin. Pract., vol. 5, no. 2, pp. 17–29, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3120126/

[70] A. Wirz-Justice, V. Bromundt, and C. Cajochen, “Circadian disruption and psychiatric disorders: The importance of entrainment,” Sleep Med. Clin., vol. 4, no. 2, pp. 273–284, Jun. 2009, doi: 10.1016/j.smcl.2009.01.008.

[71] H. Han and K. Men, “How does normalization impact RNA-seq disease diagnosis?” J. Biomed. Informat., vol. 85, pp. 80–92, Sep. 2018, doi: 10.1016/j.jbi.2018.07.016.

[72] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, “Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning,” Nature Methods, vol. 14, no. 4, pp. 414–416, Apr. 2017, doi: 10.1038/nmeth.4207.

[73] X. Chen, F. Long, B. Cai, and G. Chen, “A novel relationship for schizophrenia, bipolar and major depressive disorder Part 5: A hint from chromosome 5 high density association screen,” Amer. J. Transl. Res., vol. 9, no. 5, pp. 2473–2491, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5446530/

[74] M. T. D. F. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in Proc. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144.