The Identification research of bipolar disorder based on CNN

Qiu Sun¹, Qixuan Yue¹, Feng Zhu² and Kunxian Shu¹, a

¹Chongqing Key Laboratory on Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China
²Innovative Drug Research and Bioinformatics Group, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China.

Abstract. Bipolar disorder (BD), as a type of mood disorder, refers to the disease that has both manic episodes and depressive episodes. At present, the clinical diagnosis and treatment of bipolar disorder is often unsatisfactory, the disease always has a high rate of misdiagnosis and recurrence. The specific pathogenesis of bipolar disorder is still unclear, but related studies have shown that bipolar disorder is highly heritable, so a large number of studies have been carried out around the genomics of the disease, including some research combined with machine learning. Motivated by these research, we use the Single Nucleotide Polymorphism (SNP) which through the Genome-wide association analysis (GWAS) obtained as molecular genetic markers, combined with Convolutional Neural Network (CNN) constructing a recognition model for bipolar disorder. The experimental results show that the overall recognition rate can up to 79%, compared to similar studies, our model has achieved excellent performance. In addition, unlike traditional GWAS for finding pathogenic sites, our study first attempted to use it as a feature selection method of data preprocessing in machine learning, the experimental results prove the feasibility of this method. We believe that our research will provide some new ideas for the precision diagnosis and treatment of bipolar disorder in the future.

1. INTRODUCTION

Bipolar disorder, which also named bipolar affective disorder, refers to a situational repetitive pathological disorder that often occurs in emotions from extreme excitement or mania to severe depression, usually accompanied by disturbances in thinking and behavior frequent psychiatric features (delusions and hallucinations) [1]. Relevant epidemiological survey data show that the prevalence of BD is increasing year by year[2-4], but there is still a low level of the overall recognition rate and treatment rate of the disease[5-6], patients spend an average of ten years between the initial symptoms and the formal diagnosis[7]. There may be several reasons. Firstly, the specific pathogenesis of BD is still unclear, it’s only known to be a complex disease affected by environmental and genetic factors. Secondly, the current diagnostic and therapeutic mechanisms of BD still have certain limitations, the proportion of misdiagnosis is large. The diagnosis category of BD has been a hot spot in the research of this disease, and there are also many disputes. The phenotypic definition of bipolar disorder at this stage is based entirely on clinical features, the rate of misdiagnosis is high, this method lacks a validated diagnostic test[8], for example, which can be used for diagnostic tests of many physical diseases, so it is necessary to explore the relationship between the pathogenic mechanism of BD and external phenotypic symptoms from a new perspective.

Although the specific pathogenesis of bipolar disorder remains unclear, there is sufficient evidence
to show the contribution of genetics to the risk of disease[9-10]. A twin study of BD revealed up to 90% of genetic factors[11], which is the highest among other types of mental disorders[12]. In fact, one of the main goals of the molecular genetic approach to mental illness is the improvement in diagnostic classification. As the most common type of human heritable variation, single nucleotide polymorphism (SNP) is widespread in the human genome and affects the genetic traits of organisms, the nature of the SNP itself determines its suitability for the genetic anatomy to complex traits and gene recognition based on population. Genome-wide association analysis (GWAS) as one of the most popular genomics research methods, is to use the millions of SNPs as molecular genetic markers for control or association analysis, and finds genetic variants that affect complex traits by comparison[13].

In recent years, machine learning has been widely used in fields such as image recognition, speech recognition and natural language processing, has achieved remarkable results. Due to its powerful feature extraction, complex model construction and image processing capabilities, it has also been gradually applied to the analysis of massive biomedical data[14]. In the case of BD, researchers used support vector machines (SVM) to construct a classification model based on neuroimaging data[15] to distinguish between bipolar disorder and major depression; a research team in Taiwan used random forests to construct a genetic risk prediction model for bipolar disorder[16] to predict the risk of illness. Unlike traditional machine learning algorithms that require well-defined features, deep learning can use the deep network structure to extract feature and learn automatically. However, in the case of literature, the outcome of using neural networks combined with genetic data to study bipolar disorder is still to be enriched, the only one is about the BD classification model based on exome SNPs[17].

Based on the high genetics of BD and the application of machine learning and deep learning in the study of BD, we have carried out our research. We screened the original samples by GWAS, and encoded the genotype samples into image form by using numbers instead of bases, then combined with convolutional neural network to construct models to identify bipolar disorder cases and healthy controls. This is also from the genomic level to explore the factors associated with the phenotypic symptoms of bipolar disorder. We hope that our research can provide some new ideas for the existing treatment and diagnosis methods for BD and realize the precision medicine for it furtherly.

2. Materials and Methods

2.1. Research Overview
The overall research process is shown in Figure 1. We selected qualified samples and SNPs that may have pathogenic effects for BD from original cases and control sample by GWAS, and then used digital instead of bases to encode samples which contained several SNP genotypes into pictures, that is one sample corresponds to one picture, divided these pictures into training set, verification set and test set, trained and tested through a CNN model to achieve the final classification purpose.
2.2. Genome-wide Association Analysis

2.2.1 Quality Control
The original data was applied from WTCCC (The Wellcome Trust Case Control Consortium) which recorded the genotype of 1998 cases of bipolar disorder and 3004 controls at 500568 SNPs. The raw data stores the genotypes of all samples in units of chromosomes, we used linux-shell script to reassemble original data into genotype data with per-sample individuals. We excluded samples with a missing data more than 3% and the similarity up to 99%, got 1868 cases and 2938 healthy controls; variants with studywise missing data proportion more than 5%, or MAF less than 5% but studywise missing data proportion more than 1%, or a Hardy-Weinberg disequilibrium value of p less than 5.7x10^-7 were excluded, finally got 46912 qualified SNPs[18].

2.2.2 Feature Selection
More than hundreds of thousands SNPs obtained through quality control, but only a small number of them are associated with the risk of BD. From the perspective of machine learning, there are only a small number of features associated with sample classification label, and there are a large number of irrelevant features and redundant features. In order to avoid dimensionality disasters and reduce learning difficulty, feature selection is needed. That is, in this study, through the GWAS selected some of the variants associated with the risk of BD as feature sites.

We use PLINK to analyze chromosomes one by one, which needs to be explained about the treatment of sex chromosomes: the X chromosome is divided into two segments: a segment different from the Y chromosome named chromosome 23, and a segment homologous to the Y chromosome is treated as a pseudoautosomal chromosome[19], named chromosome 24. Table 1 shows the partial results of the association analysis (taking chromosome 1 as an example) , similar chromosomes 2 to 24 have corresponding correlation results. The results of the partial multi-hypothesis test p value adjustment are not exhaustive due to space reasons.

| Snp name    | UNADJ   | BONF   | FDR_BH   |
|-------------|---------|--------|----------|
| rs1685085   | 6.91E-10| 2.53E-05| 2.53E-05|
| rs2335466   | 5.70E-8 | 0.002082| 0.001041|
| rs1682948   | 1.62E-06| 0.0593 | 0.01977 |
The column of UNADJ represents the unadjusted asymptotic p value of the original analysis. Here we use the adjusted p value of FDR_BH[20] to screen the SNPs associated with the risk of bipolar disorder. The Manhattan map drawn according to the FDR_BH value is shown in Figure 2. Different from the extremely strict correction of BONF, FDR_BH is a method proposed by Benjamini and Hochberg to adjust the p-value by controlling the false discovery rate. It is the most lenient correction method, which allows more false positives to exist, but reduces false negatives at the same time, thus avoiding the defect of BONF's overcorrecting[21]. The specific correction method is as follows: the unadjusted p values are sorted from small to large, the maximum p value remains unchanged, and the remaining p values are multiplied by a coefficient (the total number of sites / the order of the p value), if the corrected p value <0.05, the loci is considered to have a significant association with the disease. In order to avoid the omission of small risk sites that may have a combined pathogenic effect for BD, we further expanded the threshold of p by a factor of 10 and screened a total of 4376 sites with p value< 0.5.

![Figure 2. Manhattan Map of BD](image)

### 2.3. Base Coding

Through the above treatment, the sample we obtained is a genotype form of 4376 SNPs, the order of the SNP in each sample is same. We know that CNN has powerful function for feature extraction of pictures, and it act on the spatial invariance of pictures. We encoded each sample as a picture, because of the limited genotypes at the same SNP and the similar genotypes at the same SNP for different sample, the similarity in space of the sample picture can be constructed, It is possible to reveal the combined pathogenic risk of loci on different chromosomes for bipolar disorder furtherly. According to the combination of bases, the alleles of a certain site may have the following 10 cases: AA, TT, CC, GG, AT, AC, AG, TC, TG, CG, we use the number 0, 1 instead of base to code. According to $2^3 <10 <2^4$, each base-pairs requires 4 digits for encoding, the corresponding form is shown in Table 2:

| Base pair | Digital coding |
|-----------|----------------|
| AA        | 0101           |
| TT        | 1010           |
| CC        | 0000           |
| GG        | 1111           |
| AT        | 0110           |
| AC        | 0100           |
| AG        | 0111           |
| TC        | 1000           |
| TG        | 1011           |
| CG        | 0011           |

0, 1 represents the pixel value of the picture, then the genotype of each SNP can be represented by
4 pixel values. Furtherly, we encoded the sample into a 134x134 single-channel bmp format picture, not enough pixels are filled with 0. Which is different from the jpg format picture, the image stored by bmp format without any compression[22], completely retaining the information of the original data in this study, and compared to the genotype sample, the image sample occupies less storage space and is easier to read. Each two rows of pixel values represent the genotype information of each 134 snps, the specific coding format is shown in Figure3:

To verify the validity of GWAS for feature selection, we randomly selected the same number of 4376 sites to encode 134x134 images in the same way.

2.4. CNN model

2.4.1. Composition of the data set

Since the convolutional neural network is sensitive to the balance of sample class in dataset, and a large number of similar samples may cause overfitting[23], in order to avoid these problems, we use a series of data enhancement techniques such as flipping, rotation to expand the number of sample and ensure balanced sample class, got 22416 samples finally. The data set is finally divided into a training set, a validate set and a test set by a ratio of 6:2:2, and the number of BD samples and control samples is equal in each set. Note that here we will construct two types of data sets, hereinafter referred to as DB1 and DB2. The SNPs of the samples in DB1 were screened by GWAS, the SNPs of samples in DB2 were randomly selected. The two types of datasets were identical except for the way of screening the featured loci.

2.4.2. Neural network structure

Since 4 pixels stand one SNP, in order to learn the useful features of the sample, the size of the convolution kernel in model should not be too large; the pool-layer uses max-pool techniques. The structure of each layer of the convolutional neural network in this study is shown in Table 3:

| Layer  | Filter size | Number of filter | Strides | Padding |
|--------|-------------|------------------|---------|---------|
| input  | /           | /                | /       | /       |
| conv1  | 3x3         | 8                | 1       | valid   |
| pool1  | 2x2         | /                | 2       | valid   |
| conv2  | 3x3         | 16               | 1       | valid   |
| pool2  | 2x2         | /                | 2       | valid   |
| fc3    | /           | 128              | /       | /       |
| fc4    | /           | 2                | /       | /       |

In addition to setting a reasonable model structure, the selection of the corresponding hyperparameters will also affect the performance of CNN model, we adjust the value of the hyperparameter by the performance of the model on the validate set. To improve the overfitting, we added L2 regularization to the loss function and introduced dropout[24] in the fully connected layer. The settings of final parameters in this study are shown in Table 4:
Table 4. Parameter setting

| Parameter     | Value          |
|---------------|----------------|
| Learning rate | (0.0001,500,0.96) |
| Min_Batch_size| 64             |
| Drop_out      | 0.4(fc3)       |
| w1            | 0.004(fc3)     |

3. Experiment and results

We trained and tested our model under DB1 and DB2 respectively. The loss function and accuracy curve of the model on the training set are shown in Figure 4 and Figure 5. The performance on DB1 is robust than DB2 obviously.

![Figure 4. The acc and loss on DB1](image)

![Figure 5. The acc and loss on DB2](image)

The confusion matrix corresponding to the test results is shown in Table 5, corresponding value of P, R and F1-score shown in Table 6. At the same time it can be calculated that the overall recognition rate of the model under DB1 and DB2 is 79% and 41.6% respectively. Obviously, the samples through feature selection by GWAS has a better performance than the samples selected features randomly. It can also be seen from Table 6 that the F1 score about BD and control are basically consistent under DB1, which indicates that our model is robust and the prediction results are generally not biased a certain category.

Table 5. Confusion matrix of test results

| Category | BD(predicted) | Control(predicted) |
|----------|---------------|---------------------|
| BD       | 1658/1054     | 581/1188            |
| control  | 362/1428      | 1879/810            |

Table 6. The classification score of model

| Category | Precision(P) | Recall (R) | F1-score |
|----------|--------------|------------|----------|
| BD       | 0.82/0.43    | 0.74/0.47  | 0.78/0.45|
| control  | 0.76/0.41    | 0.84/0.37  | 0.80/0.39|
4. Conclusion and Dissscuss
We designed a convolutional neural network model and first used a molecular genetic marker (snp) derived from genome-wide association analysis to distinguish between BD patients and healthy populations, and converted genotype samples into image formats through base-coding, making it more suitable for analysis with CNN, which reduces the difficulty of data processing. Our model can obtain the highest accuracy of 79% on the test set. The experimental results also show that the effectiveness of GWAS as a feature selection method.

To summarize our research, the following points are for further discussion: firstly, the accuracy of our model on the test set is generally low, which indicates that the generalization ability of the model needs to be further improved. As for the model itself, we have adopted a series of measures such as regularization, dropout, and data enhancement to reduce the risk of overfitting. In addition, the physical difficulty of bipolar disorder recognition may explain the accuracy which need to be improved, samples with complex diseases are diverse, it is possible that the labels we assign to the model are not sufficient to reflect the true complexity of the disease, so the ability of the model which learn the features from the training set to determine unknown samples is limited. As a complex disease, bipolar disorder its occurrence and development are related to genetic factors and environmental factors, but our research only used the locus information[25], this maybe an important factor affecting the classification accuracy. A study similar to ours is CAGI's BD exome challenge[26], which identifies bipolar disorder patients and healthy controls using unscreened exome snp, the recognition rate is concentrated in the accuracy range of 0.4 to 0.65. Secondly, although there is no evidence to suggest this, our model may have learned characteristics that are not related to the biological factors of disease classification. Unfortunately, there is no corresponding developed tool in the field of genome prediction to understand such a black box model[26]. Thirdly, the samples used in this study are all from the same study, lacking the verification of more external data, and the sample of the study is only a small sample of the disease, the distribution of the data does not represent a distribution in the real world. Therefore, there is still a certain distance to apply the results obtained in this study and similar studies to the clinic, but this possibility is not excluded. Lastly, we believe that the greatest significance of this study is to predict the risk of those who progress to bipolar disorder, allowing early intervention rather than challenging existing clinical diagnostic systems, as the labels given in the data sets used in this study are still relies on existing clinical systems rather than data driving. For this research, the possible improvement direction in the future is to consider the establishment of quantitative system standards for environmental factors, to quantify the amount of environmental factors and genetic information into a series of characteristics, integrate multi-party data sources and combine clinical research, through deep learning or a series of machine learning method trains, tests, and builds more robust model to provide a more comprehensive and in-depth analysis of the underlying mechanisms of the disease.

References
[1] R.H. Yolken, E.F. Torrey. Viruses, schizophrenia, and bipolar disorder. CMR, 8, 1: 131-45 (1995)
[2] F. Navarromateu, J. Alonso, L. Cew, et al. The association between psychotic experiences and disability: results from the WHO World Mental Health Surveys. Acta Psychiat Scand, 136, 1: 74-84 (2017)
[3] R.A. Marrie, R. Walld, J.M. Bolton, et al. Increased incidence of psychiatric disorders in immune-mediated inflammatory disease. J Psychosom Res, 101: 17 (2017)
[4] A. Fagiolini, A. Coluccia, G. Maina, et al. Diagnosis, Epidemiology and Management of Mixed States in Bipolar Disorder. Cns Drugs, 29, 9: 1-16 (2015)
[5] P. Lichtenstein, B.H. Yip, C. Bjökk, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. The Lancet, 373, 9659: 234-239 (2016)
[6] S. Paterniti, J.C. Bisserbe. Pharmacotherapy for bipolar disorder and concordance with treatment guidelines: survey of a general population sample referred to a tertiary care service. BMC
Psychiatry, 13: 211 (2013)

[7] J.D. Lish, S. Dimemenean, P.C. Whybrow, et al. The National Depressive and Manic-Depressive Association (DMDA) survey of bipolar members. J Affect Disord, 31, 4: 281-294 (1994)

[8] C. Miller, M.S. Bauer. Excess mortality in bipolar disorders. Curr Psychiat Rep, 16, 11: 1-7 (2014)

[9] N. Carmiol, J.M. Peralta, L. Almasy, et al. Shared genetic factors influence risk for bipolar disorder and alcohol use disorders. Eur Psychiat, 29, 5: 282-287 (2014)

[10] S.H Chang, L. Gao, Z. Li, et al. BD gene: a genetic database for bipolar disorder and its overlap with schizophrenia and major depressive disorder. Biol Psychiat, 74, 10: 727-33 (2013)

[11] N. Pato Carlos, L. Kennedy James, A. Bauer, et al. Genetics of Bipolar Disorder. CNS Spectrums, 4, 6: 22-29 (2013)

[12] P.F. Sullivan, M.J. Daly, and M. O'donovan. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat ReV Genet, 13, 8: 537-551 (2012)

[13] P. Sklar, S. Ripke, L.J. Scott, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet, 43: 977-83 (2011)

[14] R. Miotto, F. Wang, S. Wang, et al. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform (2017)

[15] N.F. Jie, M.H. Zhu, X.Y Ma, et al. Discriminating Bipolar Disorder From Major Depression Based on SVM-FoBa: Efficient Feature Selection With Multimodal Brain Imaging Data. IEEE T AUTON MENT DE, 7, 4: 320-331 (2017)

[16] L.C. Chuang, P.H. Kuo. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. Sci Rep, 7, 39943 (2017)

[17] S. Lakshman, R.R. Bhat, V. Viswanath, et al. Deep bipolar: Identifying genomic mutations for bipolar disorder via deep learning. Hum Mutat, 38, 9: 1217 (2017)

[18] P.R. Burton, D.G Clayton, L.R. Cardon, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447, 7145: 661-78 (2007)

[19] Wellcome Trust Case Control Consortium. Data formats and interpretation[EB/OL]. https://www.wtccc.org.uk/info/data_formats.html (2018)

[20] Y. Benjamini, Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, 57, 1: 289-300 (1995)

[21] Q. Pan. Multiple Hypotheses Testing Procedures in Clinical Trials and Genomic Studies. Frontiers in Public Health, 1, 1: 63 (2013)

[22] Y.W. Song, Y.E. Jian-Fang. Analysis of BMP Format File and Display Algorithm. Modern Electronics Technique (2011)

[23] D.M. Hawkins. The problem of overfitting. Cheminform, 35, 19: 1 (2004)

[24] N. Srivastava, G. Hinton, A. Krizhevsky, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res, 15, 1: 1929-1958 (2014)

[25] D. Librenzagarcia, B.J. Kotzian, J. Yang, et al. The impact of machine learning techniques in the study of bipolar disorder: a systematic review. Neurosci Biobehav R, 80: 538 (2017)

[26] R. Daneshjou, Y. Wang, Y. Bromberg, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum Mutat, 38 (2017)