An Influence Factors Analysis and Prediction of Mental Disorder

Rui Rui *

School of Public Health, Anhui Medical University, Anhui, 230022, China

* Corresponding Author Email: hiyan@ldy.edu.rs

Abstract. The prevalence of mental disorders continues to rise, necessitating increasingly stringent diagnostic requirements. This study conducted a comprehensive visualization analysis of raw data, performing correlation analyses between dependent variables and all independent variables, as well as pairwise relationships among all independent variables. The primary objective was to compare the performance of predictive models, such as logistic regression and random forest, in mental health diagnostics and to identify the most significant influencing factors for different disorders. The results indicate that factors associated with depression, bipolar I disorder, and bipolar II disorder are all correlated with emotional states. Aggressiveness emerged as the most prominent factor in bipolar I disorder, while suicidal ideation was identified as a significant factor in bipolar II disorder. Depression demonstrated strong correlations with self-loathing and self-harm tendencies. Symptomatology reflects alterations in emotional states. Among the models evaluated, the random forest algorithm demonstrated superior accuracy, stability, and sensitivity. Consequently, it demonstrates superior performance.

Keywords: Mental disorders; logistic regression; random forest.

1. Introduction

According to the research of the World Health Organization, the attack rate of mental disorders is as high as 12.5%, which is a significant disorder of thinking and emotional regulation and behavioral disorder, including Schizophrenia, eating disorders, disruptive behavior disorders, antisocial personality disorders, and neurodevelopmental disorders, anxiety disorder, depression, bidirectional affective disorder, and post-traumatic stress disorder [1]. The disease prevention and treatment programs are relatively mature. At the same time, most people fail to get effective medical treatment or foxes, and among the most common are depression and bipolar disorder (BD) [2]. This study investigates the predictive modeling of mental disorders through multimodal data analysis. The paper explores potential correlations among these features by examining visual presentation patterns' characteristics. It assesses how different data distribution modalities influence predictive outcomes—utilizing clinical behavioral indicators - including mood swings, sleep disturbances, and suicidal ideation - as key predictive variables.

As a common mental disorder, depression is at risk of being underappreciated. Studies have shown that more women are affected by depression than men. Still, often, long periods of depression or loss of happiness or interest in activities are seen as normal mood swings and more pronounced emotional perceptions in women [3]. The patient and their families had no intention to actively participate in the treatment. BD is even worse: treatment coverage is low, patients are often misdiagnosed, stigma and discrimination are widespread, and social conflicts intensify [4]. BD is characterized by mood swings from extreme depression to extreme mania. Its patients often have symptoms such as drinking and smoking, sleep disorders, and inattention. This leads to BD being diagnosed as an emotional disorder or concentration disorder [5]. There are two main types of bipolar disorder [6]. One experienced one and multiple manic episodes accompanied by depressive episodes, depression became more and more common over time and was called bipolar disorder type I, and the other with one or more hypomanic episodes and at least one depressive episode was named bipolar disorder type II. Compared with the global average age, people with bipolar disorder live more than 15 years less than their life expectancy [7]. Therefore, these diseases deserve the attention of the public.

This study examines depression, bipolar I disorder, bipolar II disorder, and healthy controls as dependent variables. Using logistic regression, least absolute shrinkage and selection operator (LASSO), and random forest models, it analyzes those independent variables. The analysis identifies key influencing factors, correlations, and predictive effects of these variables. The paper employs machine learning classification models (decision trees, random forests, neural networks) to establish an analytical framework.

This study investigates sleep disorders and emotion dysregulation mechanisms through neurobiological marker analysis. That develops an early-warning model using logistic regression and random forest algorithms to detect preclinical behavioral signals. This framework assists clinicians in identifying bipolar depression prodromes, enabling timely interventions to enhance treatment outcomes and quality of life.

2. Methods

2.1. Data Source

The data used in this study came from Kaggle, and the dataset is owned by Chirag Desai. Mental Disorders Classification has a usability score of 8.24 and a total of 32,000 views. The dataset offers observations of 120 in the CVS format and has no missing values.

2.2. Variable and Data Preprocessing

In this study, variables are mapped to symbols such as x_1 to be used in statistical formulas like regression models. Ranges and categories are standardized: Likert scales (1-4) are used to quantify subjective frequencies, self - assessment scales (1-10) evaluate mental states, and binary variables (0/1) denote presence. Diagnoses adhere to DSM - 5 criteria for clinical comparability, and objective metrics like BMI for anorexia screening enhance data reliability by flagging outliers (BMI < 18.5).

Table 1. Mental Health Assessment Variables with Mathematical Symbol Mapping

Symbol	Variable Name	Explanation
x_1	Patient Number	Participant ID, representing 120 unique samples.
x_2	Sadness	Frequency of recent sadness: 1=usually, 2=most often, 3=sometimes, 4=seldom.
x_3	Euphoric	Frequency of euphoric feelings, scored similarly to Sadness.
x_4	Exhausted	Fatigue level, with options identical to Sadness.
x_5	Sleep disorder	Frequency of insomnia: 1=frequent, 4=rare.
x_6	Authority Respect	Respect for authority: 0=No (e.g., frequent breakdowns), 1=Yes.
x_7	Try Explanation	Participants' self-explanation of complex scenarios (non-quantitative).
x_8	Aggressive Response	Aggressive behavior: 0=absent, 1=present (e.g., verbal/physical conflict).
x_9	Ignore Move On	Avoidance behavior: 0=confront, 1=ignore, and proceed.
x_{10}	Nervous/Breakdown	Anxiety/breakdown tendency: 0=no breakdowns, 1=frequent breakdowns.
x_{11}	Admit Mistakes	Self-reflection: 0=deny mistakes, 1=admit mistakes.
x_{12}	Suicidal thought	Suicidal ideation: 0=absent, 1=present (based on statements/behaviors).
x_{13}	Mood Swing	Intensity of mood swings: 1=stable, 10=extreme fluctuation.
x_{14}	Overthinking	Overthinking tendency: 0=absent, 1=present (e.g., persistent rumination).
x_{15}	Anorexia	Anorexia risk: flagged if BMI<18.5.
x_{16}	Sexual Activity	Sexual activity frequency: 1=no activity, 10=high frequency.
x_{17}	Concentration	Focus ability: 1=unable to concentrate, 10=highly focused.
x_{18}	Optimism	Optimism level: 1=extremely pessimistic, 10=extremely optimistic.
y	Expert Diagnose	Diagnosis by psychiatrists based on DSM-5 criteria.

Table 1 adopts a three-line academic format to prioritize readability and facilitate analysis. There are 19 variables in total: patient number, multiple symptom-related variables, and expert diagnosis, where expert diagnosis serves as the dependent variable, patient number is numeric, and the rest function as independent variables. Table 1 provides details on different quantification methods.

2.3. Machine Learning Models

This study employs the following core models and algorithms: Correlation analysis utilizes the Pearson correlation coefficient to quantify linear relationships between variables, with results visualized via heatmaps (color gradients indicate correlation strength) [8]. The logistic regression model treats disease classification (e.g., binary or multiclass) as the dependent variable and emotional indicators (e.g., depression, optimism) as independent variables. Its core formula is:

$$P(Y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}} \tag{1}$$

With parameters estimated through maximum likelihood [9]. The random forest algorithm, based on a Bagging framework, generates multiple decision trees via bootstrap sampling (node splitting criteria include the Gini index:

$$Gini(D) = \sum_{K=1}^{K} P_K^2 \tag{2}$$

Final predictions are aggregated through majority voting to enhance classification performance [10, 11]. The algorithm structure is simplified as parallel tree training and result integration.

3. Results and Discussion

3.1. Data Distribution

After processing the original data, the most significant mean value was selected and presented in Table 2. This table visually illustrates the differences in the data before processing, as depicted in the figure 1 below.

Table 2. Average value of different thoughts

Symptom	BD I	BDII	Depression	Normal
Sadness	1.9	3	3.3	1.9
Euphoric	2.4	1.5	1.4	2.5
Exhausted	2.3	2.8	3.2	2.2
Sleep disorder	2.6	2.6	2.6	1.9
Concentration	3.8	3.4	4.6	5.7
Optimism	6.2	3.2	3.2	5.1
Sexual Activity	6.5	3.4	4.1	5.1
Mood Swing	0.9	1	0	0.03
Suicidal thoughts	0.4	0.7	0.7	0.07
Anorxia	0.5	0.5	0.3	0.2
Authority Respect	0.2	0.5	0.2	0.7
Try Explanation	0.5	0.5	0.4	0.5
Aggressive Response	0.9	0.5	0.3	0.4
Ignore Move On	0.2	0.5	0.5	0.5
Nervous Breakdown	0.6	0.7	0.6	0.2
Admit Mistakes	0.3	0.5	0.5	0.6
Overthinking	0.5	0.5	0.7	0.3

The section classifies variables according to their properties and survey methods: Figure 1 focuses on dependent variables and other variables, which were classified by Likert scales (1-4 points), self-assessment scales (1-10 points), and binary variables (0/1).

Figure 1 illustrates the distribution of diagnostic categories, which are predominantly clustered within the range of 28 to 31. It reduces data noise and makes small associations between independent and dependent variables easier to detect. It simultaneously reduces the interference of multicollinearity by stabilizing the distribution of dependent variables, which can lower the risk of collinearity among independent variables. For instance, in multiple linear regression, if the volatility of Y is relatively low, it may constrain the correlations among independent variables, thereby minimizing excessive Variance Inflation Factor (VIF). This point will become more apparent in the subsequent section.

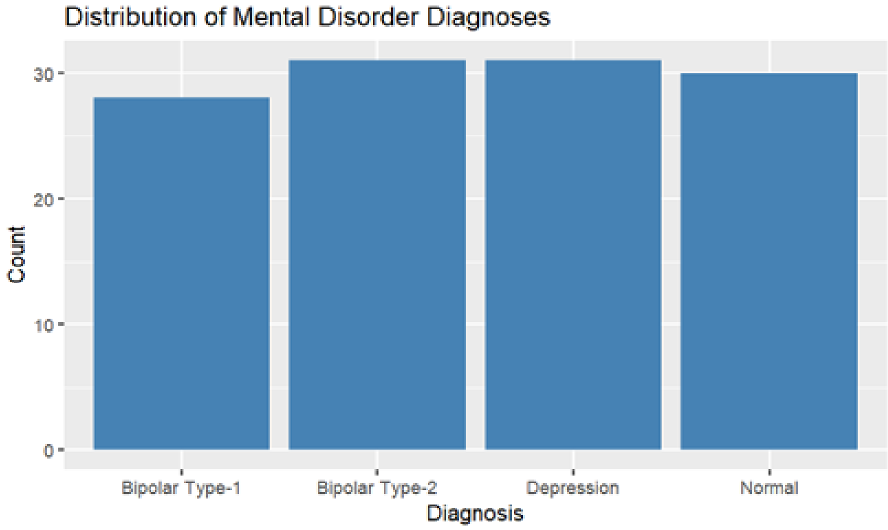

Fig. 1 The basic distribution of mental illness

Figure 2 is researched by Likert scales. According to Figure 2 and Table 2, BD I, BD II, Major Depressive Disorder, and Normal present significant heterogeneity in emotions or symptoms (exhaustion, sadness, sleep disorders, and euphoria). Bipolar disorder type 1 is characterized by a high ecstatic mood (possibly associated with the manic phase) and significant sleep disturbances, accompanied by low sadness and low fatigue, suggesting that its pathological mechanism may involve overactivation of the dopaminergic system and circadian dysregulation. In contrast, bipolar disorder type 2 showed the coexistence of low ecstatic emotions and high sadness emotions, and the degree of sleep disturbance was comparable to that of type 1, but fatigue was lower than that of depression, suggesting that intermittent energy recovery during hypomania may partially buffer the somatization of depressive symptoms. The core features of depression are persistent high sadness, severe fatigue, and generalized sleep disorders (such as early waking or sleepiness), which conform to the typical symptom-logical framework of unipolar depression, and the neurobiological basis may involve abnormal functional connectivity of the prefrontal limbic system and imbalance of serotonin transmitters.

Some of the effects are shown in Figure 3. The quantitative analysis presented in Figure 3 and Table 2 indicates that the average level of optimism among patients with bipolar II disorder and depression is significantly lower compared to that of the normal group. Moreover, both groups demonstrate a clear tendency toward negative future expectations. Further analysis revealed that the optimistic value in bipolar type I was significantly higher than in other groups, which aligns with symptoms associated with prolonged manic episodes. A significant difference in concentration was observed between the two groups: concentration scores exhibited substantial fluctuations among patients with depression, indicating that their cognitive function might be influenced by mood states. Conversely, the experimental group demonstrated lower overall attention scores, consistent with the core symptom of attention deficit. Additionally, sexual activity scores showed the widest distribution in bipolar type II patients, with a considerable relative gap, potentially linked to the alternating hypersexuality during manic phases and diminished sexual interest during depressive phases.

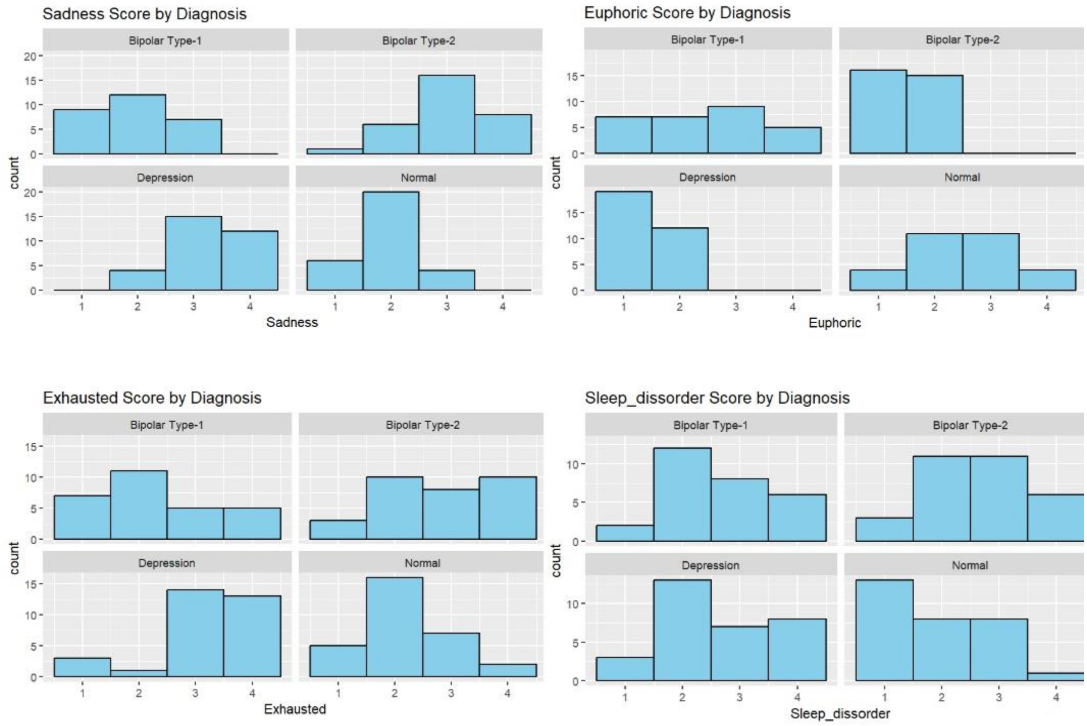

Fig. 2 Significant heterogeneity in emotions

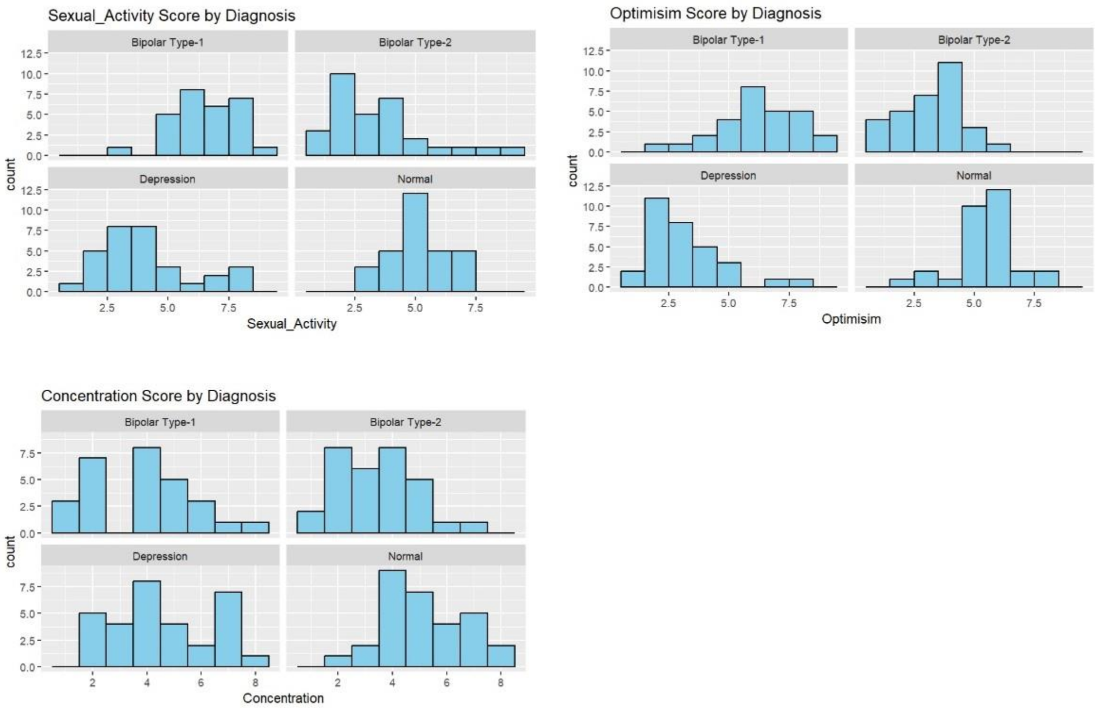

Fig. 3 Significant heterogeneity in symptoms

Finally, binary variables were subjected to binary analysis, focusing primarily on behavioral social functioning and risk-related indicators. As illustrated in Table 2, BDI aggression represents the most prominent response to emotional fluctuations and impaired social functioning (e.g., respect for authority), which aligns with the irritable characteristics observed during the manic phase. Regarding self-regulation abilities, the general population tends to admit mistakes more readily and exhibits less overthinking, whereas overthinking is most severe among individuals with depression, potentially due to rumination. Additionally, both BDII and depression patients exhibit higher levels of suicidal ideation, with depression-related fatigue and appetite issues being more pronounced, warranting particular attention. Although BDI presents a relatively lower risk of suicide, vigilance is still required during the manic-depressive transition period.

3.2. Correlation Matrix

The research results are presented in Figure 4, which elucidates critical interaction mechanisms among emotional, physiological, and cognitive functions. First, emotions can serve as drivers of risky behaviors. The strong positive correlation between sadness and suicidal ideation (r = 0.8) demonstrates that negative emotions are central driving factors of suicidal behaviors, suggesting the necessity of blocking pathways leading to emotional deterioration. Additionally, the high positive correlation between sleep disturbance and fatigue (r = 0.7) implies that sleep interventions, such as Cognitive Behavioral Therapy for Insomnia (CBT-I), could simultaneously alleviate chronic fatigue. Furthermore, the study identified psychological protective mechanisms in patients; the strong negative correlation between optimism tendency and sadness (r = -0.8) confirms the buffering effect of positive cognition on emotional collapse, indicating that enhancing positive thinking may contribute to treating mental disorders. Conversely, the negative correlation between attention and overthinking (r = -0.7) highlights the resource competition caused by rumination, which leads to cognitive imbalance. Therefore, integrating rumination-blocking techniques (e.g., attention diversion) with goal-directed training (e.g., working memory reinforcement) may optimize cognitive resource allocation efficiency.

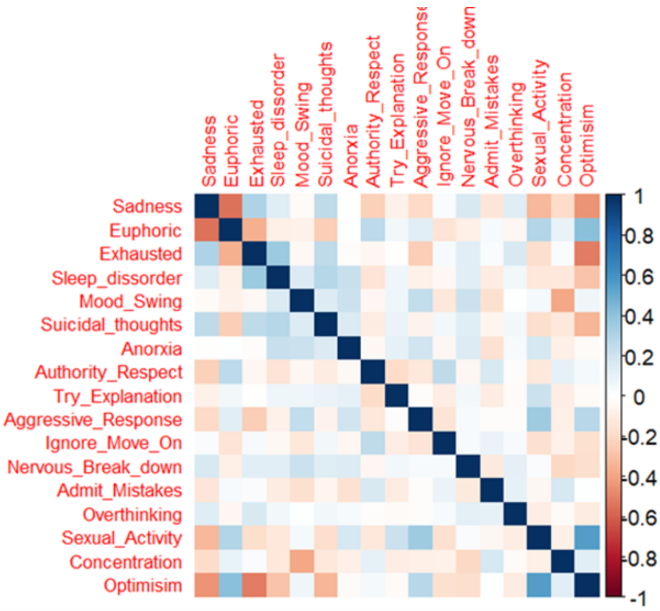

Fig. 4 Correlation matrix

3.3. Logistic Regression Models

According to Table 3, the model's overall accuracy is 74.5%; however, significant variations exist across different categories. Notably, the missed diagnosis rate for BD I reached 63.64%, indicating that the model struggled to effectively identify patients with this condition. Specifically, six cases were misclassified as BD II, and seven cases were incorrectly labeled as Normal. The high false-positive distribution rate may be attributed to overlapping features between BD I, BD II, and Normal categories (e.g., mood swings and sleep disturbances). BD II exhibited a high risk of false negatives, with two cases being misdiagnosed as Depression, potentially leading to delayed treatment. In contrast, Depression demonstrated the best performance in terms of specificity, with only two cases misclassified as Normal, suggesting that the model could effectively capture key symptoms. However, six cases of Depression are also misclassified as Normal, which results in missed diagnoses and real patients being erroneously categorized as "normal." After excluding collinearity as a factor, the model's accuracy significantly decreases to 0.564, reflecting its mediocre overall performance.

Table 3. Logical Regression Model Results

Predictions	BipolarType-1	BipolarType-2	Depression	Normal
Bipolartype-1	8	6	1	7
BipolarType-2	0	22	2	0
Depression	0	0	22	2
Normal	0	0	6	18
Model Accuracy: 0.745 Residual Deviance: 10.19775 AIC: 118.1978				

3.4. Random Forests Models

Therefore, this paper selects an alternative model. As illustrated in Table 4, the overall accuracy of the random forest model (Accuracy = 82.98%) demonstrates strong predictive capability (an accuracy higher than 80% is generally regarded as satisfactory). Despite the presence of class imbalance in the dataset (No Information Rate = 25.53%), where the most frequent category constitutes approximately 25.5% of the data, the model's accuracy significantly surpasses this baseline ($p < 2.2e-16$), indicating that the model indeed possesses predictive power. The Kappa value, which ranges from 0.61 to 0.80 for "Substantial agreement" and from 0.81 to 1.0 for "Almost perfect," currently stands at 0.7725, nearing the upper limit of "Substantial." This suggests a high level of consistency within the model.

Table 4. Random Forests Models Results

Reference/Prediction	BipolarType-1	BipolarType-2	Depression	Normal
BipolarType-1	14	2	0	0
BipolarType-2	4	22	0	1
Depression	0	0	21	2
Normal	4	0	3	21

As illustrated in Figure 5 and Table 5, the various disease models within the presented framework exhibit distinct characteristics and challenges. The BD I model demonstrates a risk of insufficient sensitivity, leading to potential missed diagnoses. Specifically, its sensitivity is 63.64%, with a missed diagnosis rate of 36.36%, indicating that approximately one-third of actual patients may remain unidentified. However, this model exhibits an exceptionally high specificity of 97.22%, effectively excluding non-BD I case. Conversely, the BD II model displays high sensitivity (91.67%) but is accompanied by a moderate false-positive rate. While its missed diagnosis rate is relatively low at 8.33%, the positive predictive value (PPV) of 81.48% suggests a limited number of misjudgments. The Depression model achieves the best overall performance among the three, with a sensitivity of 87.50%, specificity of 97.14%, and a PPV of 91.30%. These metrics indicate a low missed diagnosis rate (12.50%) and an even lower false-positive rate (2.86%), making it highly reliable for clinical prioritization. This superior performance may stem from the model's capacity to accurately capture critical symptoms such as suicidal tendencies and persistent low mood. Lastly, the Normal model has a PPV of 75.00%, implying that 25% of cases predicted as normal are misclassified. For instance, depression may be erroneously diagnosed as normal, potentially delaying treatment for genuine patients. Therefore, it is essential to review these predictions in conjunction with additional diagnostic indicators.

Table 5. Key Features of the Random Forest Model

Class	BipolarType-1	BipolarType-2	Depression	Normal
Sensitivity	0.6364	0.9167	0.8750	0.8750
Specificity	0.9722	0.9286	0.9714	0.9000
Pos Pred Value	0.8750	0.8148	0.9130	0.7500
NegPred Value	0.8974	0.9701	0.9577	0.9545
Prevalence	0.2340	0.2553	0.2553	0.2553
Detection Rate	0.1489	0.2340	0.2234	0.2234
Detection Prevalence	0.1702	0.2872	0.2447	0.2979
Balanced Accuracy	0.8043	0.9226	0.9232	0.8875

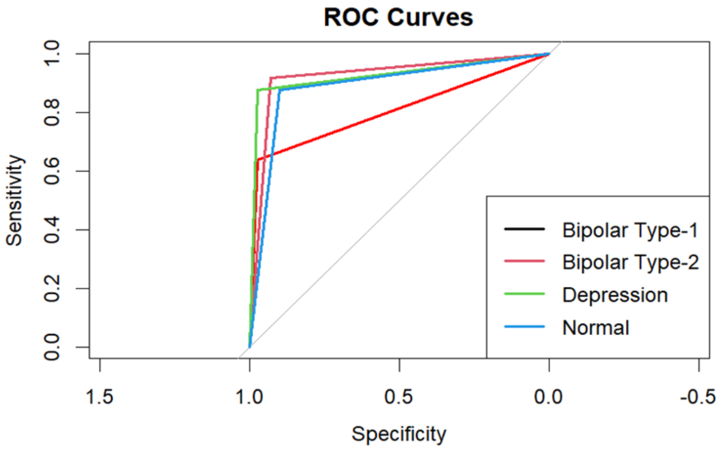

Fig. 5 ROC Curves

All in all, the model exhibits strong atopic properties and appropriately calibrated sensitivity. The following are the advantages of the models summarized in this section: The depression model demonstrates high sensitivity and specificity, providing reliable prediction outcomes and precise identification of key symptoms. The BD II model features high sensitivity and a low rate of missed diagnoses, effectively minimizing the omission of genuine cases during preliminary screening. The BD I model showcases strong specificity, enabling accurate exclusion of non-cases and reducing the likelihood of misdiagnosis.

4. Conclusion

The objective of the study is to compare the performance of predictive models, including logistic regression and random forest, concerning mental disorders and to identify the most significant influencing factors for different diseases. Through a comprehensive and detailed analysis of various models, it was found that the random forest model demonstrates significantly favorable performance in disease diagnosis under specific conditions. At the same time, this study identified that the most significant influencing factor for Bipolar Type 1 was the pronounced aggression resulting from mood swings. For Bipolar Type 2, the most significant factor was the suicidal tendency triggered by mood fluctuations. Depression exhibited a strong correlation with persistent low mood and a marked tendency toward self-hatred and self-injury. However, Bipolar Type 1 and Bipolar Type 2 demonstrated high similarity in terms of mood swings (e.g., alternating between manic and depressive states). In contrast, it is challenging to differentiate these bipolar types from normal conditions based on baseline characteristics such as sleep disorders, which warrants further investigation in future research. It is anticipated that the findings of this study will influence the development of relevant mental health policies and encourage both governmental bodies and stakeholders from various sectors to focus on and address pertinent issues.

References

[1] Prigerson H G, Boelen P A, Xu J. Validation of the new DSM-5-TR criteria for prolonged grief disorder and the PG-13-Revised (PG-13-R) scale. World Psychiatry, 2021, 20(1): 96-106.

[2] WANG Hanjia, WEI Yanyan, Qian Zhenying, LI Qingwei, Wang Jijun. Research progress of retinal structural changes in patients with mental disorders. Journal of Shanghai Jiao Tong University (Med Edition), 2019, 45(02): 247-252.

[3] Woody C A, Ferrari A J, Siskind D J, Whiteford H A, Harris M G. A systematic review and meta-regression of the prevalence and incidence of perinatal depression. J Affect Disord, 2017, 219: 86-92.

[4] Global Burden of Disease 2019 Mental Disorders collaborator. The global, regional, and national burden of 12 mental disorders in 204 countries and territories during 1990-2019: A systematic analysis of the 2019 Global Burden of Disease Study. Lancet Psychiatry, 2022, 9: 137-150.

[5] Plans L, Barrot C, Nieto E, et al. Association between completed suicide and bipolar disorder: A systematic review of the literature. J Affect Disord, 2019, 242: 111-122.

[6] Kwun Nam Chan J, Ho Yi Yong C, Sau Man Wong C, Yu Hai Chen E, Chung Chang W. Life expectancy and potential years of life lost in bipolar disorder: a systematic review and meta-analysis. British Journal of Psychiatry, 2022, 221(3): 567-576.

[7] Diagnostic and Statistical Manual of Mental Disorders, 5th ed, Text Revision (DSM-5-TR). American Psychiatric Association Publishing, Washington, DC, pp 140-141.

[8] Li Feng, ZHU Xunlin, FENG Sanying. Cooling "heat map": Careful use of Pearson correlation coefficient. Statistical Theory & Practice, 2024, 548(12): 68-72.

[9] Ji Huimin, BAI Ya-Mei, SONG Yu-lei, ZHANG Xue-Qing, XU Gui-hua, WANG Xiao-feng. Application of machine learning algorithm in risk warning of stress injury in ICU patients. Journal of Nursing, 2025, 40(05): 126-129.

[10] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[11] Zhang Xiao, Li Chao, Dou Yanguang, et al. Prediction of organic carbon content distribution in surface sediments of East China Sea using stochastic forest algorithm. Marine Geology and Quaternary Geology, 2025, 1-17.