

The Research on the Key Pathogenic Factors of Depression

Shaozong Liang ^a

*Oriental College of International Trade and Foreign Languages, Haikou University of Economics,
Haikou, Hainan, 571127, China*

Keywords: Depression, Depression Factors, Depression Causes, Random Forest.


Abstract: Depression is a significant public health concern, posing critical risks such as suicide. This study investigates the key pathogenic factors of depression using clinical data sourced from the Kaggle website, comprising 413768 instances with 15 variables, including socioeconomic, lifestyle, and medical factors. The Random Forest algorithm is employed to analyze the correlation between these variables and chronic medical conditions (depression). The model identified income as the most critical predictor of depression, with an importance score of 0.668, followed by the number of children and history of substance abuse. Descriptive analysis further revealed that low-income groups, individuals with lower education levels, and non-smokers exhibited higher rates of chronic medical conditions. These findings underscore the pivotal role of socioeconomic and lifestyle factors in the prevalence of depression, offering valuable insights for targeted healthcare interventions and preventive strategies. While the current model demonstrates moderate predictive accuracy, future research should focus on expanding the dataset, incorporating additional clinical variables, and comparing results across multiple machine learning models to enhance predictive performance and generalizability. The approach holds promise for advancing the understanding of depression and improving mental health outcomes globally.

1 INTRODUCTION

Depression has emerged as a public health concern in contemporary society, affecting millions of individuals worldwide. As a chronic and recurrent disease, depression is one of the most critical risk factors for suicide. Nearly one billion people worldwide are affected by mental health issues, and the diagnosis rate of depression among teenagers surges (Nguyen, T.-L. & Lee, 2025). The suicide rate among elderly patients with depression accounts for 27.2% of global suicide deaths (Szűcs, Cohen & Reynolds, 2025). The incidence of depression is increasing in recent years. Epidemiological studies show that among the 25 major causes of the total burden of disease, depression ranks 13th in the disability adjusted life year and 2nd in the disability adjusted life year (Ao et al., 2024). A number of depression patient can be treated with drugs, but there are still 15%~33% of patients who are ineffective after drug treatments (Fei, Cheng & Shen, 2024). Depression is a common mental disorder in clinic at

present. The main clinical manifestations of depression are depression, slow thinking, weakened will activity and decreased interest. The latest research shows that the prevalence of depression is about 27.6%, and the number of patients with depression in the world has reached 350 million, and is increasing year by year. The recurrence rate of depression is high, and people are prone to self injury or suicidal behavior, which has brought serious damage to individuals, families and society (Ma, 2024). In the United States, the lifetime prevalence of major depression is 21% in women and 11% -13% in men. It is the main cause of suicide. Among the top 10 deaths in the United States, it is reported that nearly 50000 people commit suicide every year. The morbidity and mortality associated with depression make it the number one cause of disability in the world and cause a huge economic burden on society in terms of productivity loss (Kuerban, 2024).

Depression is a genetic disease, the probability of depression in relatives of patients with depression, First-Degree Relatives, is higher than that of the

^a <https://orcid.org/0009-0004-4901-1584>

general population. Wang's et al. (2024) study also shows that the incidence of depression in families with depression, alcohol dependence or antisocial personality is significantly higher than that in the control population, indicating that the interaction between gene and environment play a significant role in the pathogenesis of depression. People with high frequency of interaction with patients with depression have higher recognition of biogenetic interpretation and childhood adversity interpretation (Zhang, 2024). The onset of depression lasts for more than 2 weeks each time, even for 1 year or several years, and depression is easy to relapse (Chen et al., 2024). The survey shows that the incidence of sleep disorders in patients with depression is 55.00%, indicating that more than half of patients with depression are accompanied by sleep disorders, and once accompanied by insomnia, the incidence and severity of the remaining physical symptoms of patients would also increase (Li, Shi & Zhang, 2023). Depression patients may have suicidal ideation of different intensities, and even suicidal behavior, the suicide rate of depression patients is about 10%~15%. Projections indicate that by 2030, depression will account for the highest proportion of disability-adjusted life years (DALYs) lost globally (Zhong, 2024). Therefore, it is of significance to study the key pathogenic factors of depression for the prevention of depression.

The study aims to leverage clinical data to investigate depression, with all patient data sourced from the Kaggle website. Using the Random Forest algorithm to explore the correlation between various risk factors and chronic medical conditions, focusing on depression. By analyzing key determinants of depression, the research seeks to provide insights for targeted interventions and preventive strategies, ultimately contributing to improved mental health outcomes.

2 METHODS

2.1 Data Source

The data utilized in this paper for predicting depression were obtained from clinical cases on the Kaggle website (Sasaki, 2024). The dataset provides the basis for correlations between factors and depression. It contains 413768 sample cases and 15 variables, including age, marital status, number of

children, smoking status, physical activity level, employment status, alcohol consumption, dietary habits, sleep patterns, education level, income, history of mental illness, history of substance abuse, family history of depression, and chronic medical conditions, determined through medical diagnostic records.

92% of the 413768 instances were randomly selected as samples with no missing values. Using the Random Forest algorithm, which employs all available clinical features, to judge the samples as having chronic medical conditions or not having chronic medical conditions.

2.2 Method Introduction

The method used in this study is the Random Forest algorithm. The Random Forest model operates on ensemble learning principles, constructing multiple decision trees through dual randomization: bootstrap sampling and random feature subsets during node splitting. Predictions are aggregated via majority voting, reducing overfitting risks. Its advantages are tolerance for missing values and interpretable feature importance rankings. The model excels in handling heterogeneous variables without manual preprocessing, while maintaining computational efficiency through parallelized tree training.

Chronic medical conditions are classified as target variables and the other 14 factors are divided into feature columns, which explores the correlation between various variables in the table and chronic medical conditions (depression), calculates the importance of the characteristics of each variable, and predicts whether a person has depression.

3 RESULTS

3.1 Descriptive Analysis

Extract 3 variables: education level, income, and smoking status. Figures 1, 2 and 3 ordinates represent the number of abscissas.

Figure 1, whose abscissas represent 5 education levels from high school to PhD, shows whether the variable Education level has the number of Chronic Medical Conditions. The sample with Bachelor's degree and high school education had the highest number of chronic medical conditions.

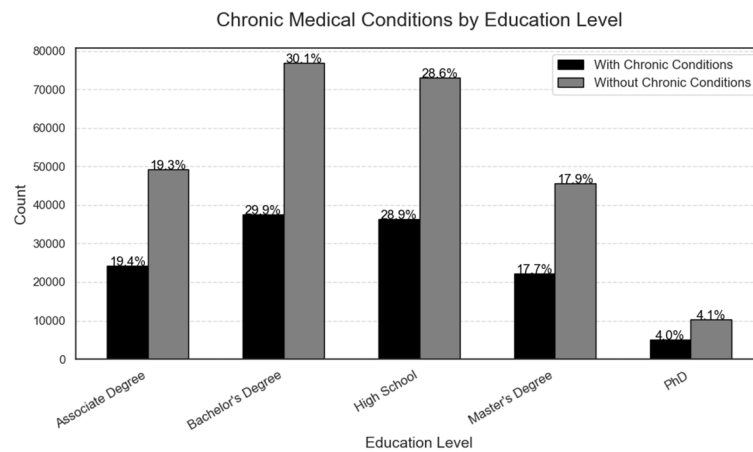


Figure 1: Education levels has the number of Chronic Medical Conditions (Photo/Picture credit: Original).

The Figure 2, whose abscissas represent 3 income levels from low to high, shows whether the variable Income has the number of chronic medical conditions, with the low-income group having the highest number of chronic medical conditions.

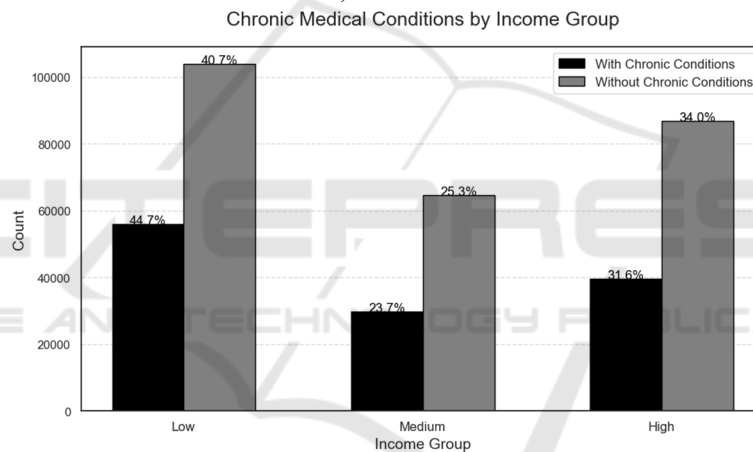


Figure 2: Income groups has the number of chronic medical conditions (Photo/Picture credit: Original).

Figure 3, whose abscissas represent 3 smoking conditions, shows whether the variable Smoking Status has the number of chronic medical conditions, with the non-smoking group having the highest number of chronic medical conditions.

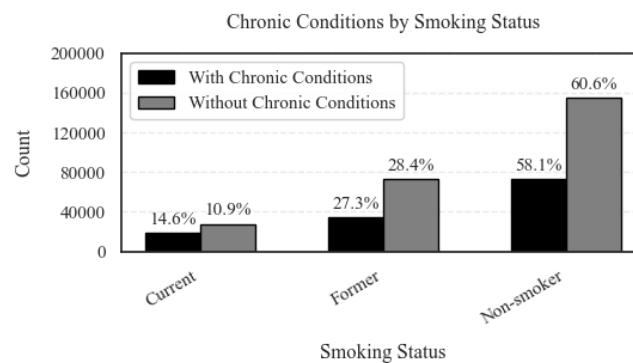


Figure 3: Smoking Status has the number of chronic medical conditions (Photo/Picture credit: Original).

3.2 Random Forest Results

Table 1: Model Evaluation

	Precision	Recall	F1-score	Support
0	67.36%	74.35%	70.68%	70664
1	33.04%	25.99%	29.10%	34401
accuracy			58.52%	105065
macro avg	50.20%	50.17%	49.89%	105065
weighted avg	56.12%	58.52%	57.07%	105065

0 = No chronic medical conditions, 1 = Having chronic medical conditions. For 0, precision is 67.36%, recall is 74.35%, and the F1-score is 70.68%. For 1, precision is 33.04%, recall is 25.99%, and the F1-score is 29.10%. The macro-average precision is 50.20%, recall is 50.17%, and F1-score is 49.89%. The weighted-average precision is 56.12%, recall is 58.52%, and F1-score is 57.07%. The accuracy of the model is 58.52%, indicating that the model is reliable.

Train the model using a Random Forest Classifier and calculate the importance score for each feature. Table 2 shows all feature importance scores in descending order.

Table 2: Importance Score

Feature	Importance
Income	0.668
Number of Children	0.071
History of Substance Abuse (Yes)	0.027
Family History of Depression (Yes)	0.025
History of Mental Illness (Yes)	0.025
Alcohol Consumption (Moderate)	0.016
Sleep Patterns (Poor)	0.016
Alcohol Consumption (Low)	0.015
Sleep Patterns (Good)	0.013
Dietary Habits (Moderate)	0.013
Education Level (High School)	0.013
Marital Status (Married)	0.012
Physical Activity Level (Moderate)	0.012
Education Level (Bachelor's Degree)	0.011
Education Level (Master's Degree)	0.009
Marital Status (Widowed)	0.009
Physical Activity Level (Sedentary)	0.008
Dietary Habits (Unhealthy)	0.008
Smoking Status (Non-smoker)	0.007
Smoking Status (Former)	0.007
Employment Status (Unemployed)	0.006
Marital Status (Single)	0.005
Education Level (PhD)	0.005

The importance score of Income is 0.668, the highest. The most important feature for predicting depression is: Income.

4 DISCUSSION

The Random Forest model demonstrated strengths in this study, including its capability to handle high-dimensional data with mixed feature types while maintaining robustness against missing values and outliers (Breiman, 2001). The disadvantage is limited model interpretability, struggling to quantify the specific impact direction of individual features on prediction outcomes. It is sensitive to class imbalance, it may affect the prediction accuracy of minority classes if the target variable "chronic medical conditions" is severely imbalanced (Zhang, Liu, & Wang, 2020).

Introducing a weighted iterative mechanism to further enhance classification performance in XGBoost, or incorporating external datasets (such as public health databases) to expand sample diversity and avoid sampling bias, can optimize the model (Zhang et al., 2020; Zhang, Chen, & Li, 2021). The limitations of this study's design are reliance on self-reported questionnaires and samples drawn from a single database, which affect the model's generalizability (Wong et al., 2022).

In the future, electronic health records, wearable device data, and genomic information can be integrated to develop a multidimensional predictive model (Topol, 2019).

5 CONCLUSION

This study demonstrates the reliability of predicting whether an individual has depression based on a comprehensive set of factors, including marital status, number of children, smoking status, physical activity level, employment status, alcohol consumption, dietary habits, sleep patterns, education level, income, history of mental illness, history of substance abuse, family history of depression, and chronic medical conditions. Using the Random Forest algorithm, income is identified as the most important predictor of depression, with the importance score of 0.668, followed by the number of children and history of substance abuse. Descriptive analysis further revealed that low-income groups, individuals with lower education levels, and non-smokers exhibited higher rates of chronic medical conditions. These findings emphasize the critical role of socioeconomic

and lifestyle factors in the development of depression, providing valuable insights for targeted healthcare interventions and preventive strategies. Significant differences in the distribution between the new data and the training data (such as the lifestyle habits of different populations in different regions) will decrease the performance of the model. Not considering key variables in medical diagnosis, relying solely on questionnaire data limits clinical practicality. While the current model achieves moderate predictive accuracy, future research should focus on expanding the dataset, incorporating additional clinical variables, and comparing results across multiple machine learning models to improve predictive performance and generalizability. The approach holds promise for advancing our understanding of depression.

REFERENCES

- Ao, Y., Guo, Y. Q., Guo, R. J., et al., 2024. Analysis of traditional Chinese medicine syndrome elements, accompanying symptoms, and their correlations in 1063 patients with depression. *Global Traditional Chinese Medicine*, 17(11), 2250 – 2256.
- Breiman, L., 2001. Random forests. In *Machine Learning*, 45, 5-32.
- Chen, J., Li, X., Zhang, T., et al., 2024. The impact of early trauma experiences in patients with depression. In *Psychology Monthly*, 19(06), 19-21. DOI: 10.19738/j.cnki.psy.2024.06.006
- Fei, Y. X., Cheng, S. F., & Shen, Y. P., 2024. Changes and influencing factors of cognitive function in patients with depression during modified electroconvulsive therapy. *Zhejiang Clinical Medicine*, 26(6), 837 – 839.
- Kuerban, K., 2024. *Differences in clinical characteristics and influencing factors between melancholic and non-melancholic depression patients* [Master ' s thesis, Xinjiang Medical University]. CNKI. https://kns.cnki.net/kcms2/article/abstract?v=5ykJdPmCibI82sUyBlyincZu8P6NPZVvct7NlmKGXWGBJICeKs7s4IPBKHloWPyhGCj5caFHR-b-alfMbqFbAVgImDtI719jIC8FVJGWIXVBs3bXdHcxQRhtBHx0ITB1uGG8xKoQsPxSM_8s9Cles3DNwvwsEUutFDxCQAN2BrYgn_H9xZPBw=&uniplatform=NZKPT&language=CHS
- Li, S., Shi, L., Zhang, W., 2023. Sleep disorder incidence and its influencing factors in patients with depression. In *World Journal of Sleep Medicine*, 10(06), 1302-1305.
- Ma, B. X., 2024. Analysis of related factors of anhedonia in patients with depression and research on traditional Chinese medicine prescriptions and syndromes [Doctoral dissertation, Tianjin University of Traditional Chinese Medicine].
- Nguyen, T.-L., & Lee, J.-Y., 2025. Kindness as a public health action. *Communications Medicine*, 5(4), 112 – 125.
- Sasaki, T. 2024. What causes depression? Causal inference. Kaggle. <https://www.kaggle.com/code/sasakitetsuya/what-causes-depression-causal-inference>
- Szücs, A., Cohen, R., & Reynolds, C. F., 2025. Investigating direct and moderating effects of social connectedness and perceived social support on suicidal ideation in depressed aging adults: A prospective study. In *Biological Psychiatry Global Open Science*, 4(2), 102–115.
- Topol, E. J., 2019. High-performance medicine: the convergence of human and artificial intelligence. In *Nature Medicine*, 25(1), 44-56.
- Wang, X., Xiao, C., Li, L. L., et al., 2024. Prevalence and influencing factors of depression in Heilongjiang Province. *Journal of Qiqihar Medical College*, 45(11), 1060 – 1064.
- Wong, K. C., Luo, X., & Zhang, Q., 2022. Cross-cultural validation of machine learning models in healthcare: A systematic review. In *The Lancet Digital Health*, 4, e158–e167.
- Zhang, J., 2024. The mechanism and intervention of causal explanations affecting the stigma of depression [Doctoral dissertation, Southwest University]. <https://doi.org/10.27684/d.cnki.gxndx.2024.002253>
- Zhang, L., Chen, H., & Li, M., 2021. Enhancing model generalizability through multi-source health data integration. In *Journal of Biomedical Informatics*, 123, 104567.
- Zhang, Y., Liu, Y., & Wang, J., 2020. Boosting ensemble learning for imbalanced data classification: A hybrid approach. In *IEEE Transactions on Knowledge and Data Engineering*, 33, 4567–4581.
- Zhong, W. S., 2024. Analysis of living conditions of college students with depression. *Health Education and Health Promotion*, 19(4), 430 – 433. <https://doi.org/10.16117/j.cnki.31-1974/r.202404430>