Machine Learning-Based Prediction Model for Depression Tendency in Chinese Elderly

Hongxiang Xu, Xiangxiu Yao
Xi'an Eurasia University, Xi'an, Shaanxi, China

Abstract: As global aging accelerates, the incidence of depression among the elderly has risen significantly, posing a serious threat not only to their quality of life but also to their overall physical health and well-being. While timely identification and effective intervention are crucial for addressing this issue, existing research often falls short in providing a systematic analysis of the various influencing factors. This study utilizes the China Health and Retirement Longitudinal Study (CHARLS) dataset to thoroughly explore the multiple factors that contribute to depression in older adults. During the data preprocessing phase, we meticulously ensured data quality by analyzing the relationships between variables through correlation coefficient heatmaps and conducting chi-square tests to assess independence. Following this, we constructed and optimized several predictive models, including random forests, decision trees, and naive Bayes, fine-tuning their parameters for enhanced performance. Ultimately, the experimental results demonstrated that the optimized Support Vector Machine (SVM) model excelled in predicting depression in the elderly, providing robust theoretical support and empirical evidence that can inform targeted strategies for improving the mental health and overall well-being of older adults.

Keywords: Decision Tree Model; Random Forest Model; Naive Bayesian Model; SVM Model

1. Introduction

In today's society, the acceleration of global aging has made mental health issues among the elderly increasingly prominent, with elderly depression becoming a phenomenon that cannot be ignored. Elderly depression not only severely affects quality of life but also poses potential threats to physical health. Therefore, it is particularly important to deeply study the causes, manifestations, and intervention measures related to elderly depression. This research aims to explore the multiple factors influencing elderly depression, using scientific methods to provide theoretical support and practical guidance for developing effective prevention and intervention strategies, thereby enhancing the mental health of the elderly and promoting their happiness and well-being in later life.

2. A Review of Studies

In recent years, the incidence of depression among the elderly has gradually increased, becoming a widely recognized issue. Many scholars have explored its influencing factors through related research. Merril et al [1]. found through data analysis from 2011 and 2013 that depression is significantly related to caregiving, family structure, and socioeconomic status. Research by Njai et al. found a significant association between physiological changes in older adults and depressive symptoms, particularly concerning changes in physical function and residual force, which have a substantial impact on mental health [2]. Zhu et al. conducted a systematic evaluation and discovered that non-pharmacological interventions, such as Tai Chi, show positive effects in alleviating depressive symptoms in the elderly. This not only improves their mental health but also promotes the establishment of social connections [3]. Cordero's research further emphasizes the importance of non-pharmacological interventions (such as psychological counseling and social activities) in the treatment of late-life depression, providing a theoretical basis for developing effective treatment plans [4].

Catalano et al. conducted a systematic review that revealed the psychophysiological indicators of late-life depression, indicating that depressive symptoms in older adults are

closely related not only to psychological states but also to physical health. This finding deepens our understanding of the mechanisms behind depressive symptom onset [5]. Nishida et al. focused on the severe depressive characteristics in the prodromal phase of neurodegenerative dementia, highlighting the importance of identifying and intervening in this population [6]. Research by Elsheikh et al. indicates that genetic factors play a crucial role in the occurrence of late-life depression and the response to antidepressant medications, offering new insights for personalized treatment [7].

Liu et al. investigated the relationship between high suicide risk and late-life depression, finding that various factors, including a lack of social support, pose threats to the mental health of older adults [8]. Guo et al. explored the historical link between the Great Famine and late-life depression, revealing the profound impact of social and environmental factors on mental health [9]. Giehl et al. focused on the accessibility and user experience of digital media among elderly patients with depression, emphasizing the need for digital interventions tailored to older users [10]. Additionally, Moyano et al. studied the effects of electroconvulsive therapy in treating late-life depression, suggesting that this method may provide an effective alternative treatment in certain cases [11]. Finally, Bandyopadhyay highlighted the close relationship between loneliness and late-life depression, calling for attention to the social networks and mental health of older adults [12]. In summary, these studies provide important evidence for understanding the multifaceted influencing factors of depression in the elderly, suggesting that comprehensive intervention measures should be adopted to promote the mental health of older adults.

To effectively prevent and control elderly depression, it is crucial to conduct in-depth research on its pathogenesis and influencing factors. This study aims to utilize the China Health and Retirement Longitudinal Study (CHARLS) dataset to explore the multiple factors affecting the incidence of depression in older adults. By comprehensively applying machine learning algorithms, including logistic regression, random forests, decision trees, and support vector machines (SVM), this research seeks to develop efficient and accurate predictive models, providing theoretical support and practical guidance for the prevention and intervention of elderly depression. This study not only holds significant academic value but also has far-reaching practical implications for enhancing the mental health of the elderly and promoting their happiness and well-being in later life.

3. Empirical Research

3.1 Indicator Design

The data comes from the China Health and Elderly Care Panel Survey (CHARLS), which was conducted by the National School of Development of Peking University, which collects data from middle-aged and older adults through questionnaires and interviews, and regularly tracks them, covering health, economic and social interactions, etc., to provide data support for aging research and policy. By using Python to carry out data cleaning, data transformation, data standardization and other operations on the data used, due to the large amount of data, the columns with missing values, outliers and vacant values are deleted and finally obtained a dataset composed of 5405 rows and 16 columns, as shown in the Table 1:

Table 1. Describes the Data Variables

Variable type	Column	Remark
Dependent variable	Mental health conditions	A CES-D10 score of 0-9 indicates no depression and is assigned a value of 0; A score of ≥ 10 indicates a depressive condition and is assigned a value of 1
Argument	gender	0 for female; 1 indicates male
	marry	0 means unmarried; 1 indicates married
	drinkl	0 means no drinking habits; 1 On the contrary
	smoken	0 means no smoking; 1 On the contrary
	retire	0 means not retired; 1 On the contrary
	satlife	The value ranges from 1 to 5
	edu	The value is 1-4 according to the uneducated, unfinished primary school to primary school graduation, junior high school graduation to technical secondary school graduation, junior college graduation to doctoral graduation
	sleep_category	0 to 5 hours for sleep deprivation, 5 to 9 hours of sleep normal, 9 to 12 sleep excess, and extra 12

		hours of sleep exceeding the standard are assigned as 0-3
	exercise	0 indicates lack of basic mobility; 1 On the contrary
	ins	0 means none; 1 On the contrary
	pension	0 means none; 1 On the contrary
	disease_history	If there are diseases such as hypertension and diabetes, it is expressed as a hidden danger in the body, and the value is assigned as 1; Otherwise, it is 0
	social_history	If there are behaviors such as visiting doors, playing mahjong, etc., it means that there are social activities, and the value is assigned as 1; Otherwise, it is 0
	age	Unit years
	hchild_group	If the number of surviving children is greater than 1, the value is 1; Otherwise, it is 0

3.2 Feature Selection

In feature selection methods, correlation coefficient feature selection and chi-square test feature selection are two commonly used techniques. This paper will utilize both correlation coefficients and chi-square tests for feature selection:

(1) Correlation Coefficient Selection

To eliminate the mutual influence among various quantitative features, this paper explores the correlations among quantitative features and presents a correlation heatmap of the quantitative features, as shown in the Figure 1:

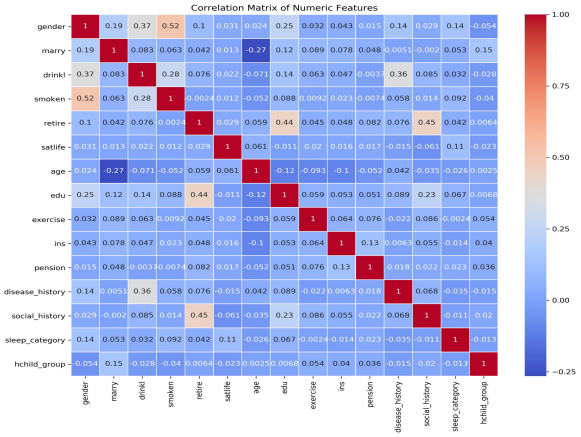

Figure 1. Diagram of the Thermodynamic Correlation Coefficient of a Numerical Variable

As can be seen from Figure 1, there is an obvious linear relationship between different numerical variables in the dataset. Among them, there was a significant positive correlation between alcohol consumption status and health status, and smoking status, education status, and social status also showed a strong positive correlation, suggesting that there may be a link between these indicators.

(2) Chi-Square Test Feature Selection

The chi-square test is used to evaluate the independence of two categorical variables. In feature selection, it can assess the correlation between features and the target variable by calculating the chi-square values and p-values for the categorical variables. The results are shown in the Table 2.

Table 2. Chi-Square Test Results

Column	Chi-squared	P-value
gender	115.122	0.000
marry	32.181	0.000
drinkl	31.059	0.000
smoken	24.561	0.000
retire	177.864	0.000
satlife	447.373	0.000
edu	138.203	0.060
sleep_category	222.755	0.000
exercise	6.051	0.003
ins	3.216	0.042
pension	3.793	0.034
disease_history	2.772	0.010
social_history	66.046	0.000
age	43.262	0.007
hchild_group	356.988	0.821

Overall, the independent variables gender, marry, drinkl, retire and satlife had a significant impact on the model, and their p values were all less than 0.05. However, the influence of independent variables on the model hchild_group was not significant, and the P value was larger, indicating that the correlation between the independent variables and y was weak.

4. Model Selection and Training

4.1 Model Prediction Studies

In life, the first thing to pay attention to is whether the elderly suffer from depression,

and if so, to study the degree of depression. Therefore, this paper classifies the depression scale into a dichotomous with or without depression, and makes model predictions based on the above definitions. In this chapter, the data are divided into training set and test set in a 7:3 ratio, and the decision tree, naive Bayesian, random forest, and support vector machine models are used to analyze and predict the data in a 7:3 ratio.

4.2 Predictions Based on Naive Bayes

The Naive Bayes-based prediction results are shown in Table 3, with a recall rate of 84.3% and an F1 value of 81.7%, which is a good performance with a prediction accuracy of 69.1%. However, as can be seen from Figure 2, the number of correct classifications for category 0 is small. In the test set, only 458 samples with a true 0 were correctly predicted and 183 were incorrectly predicted, and the overall prediction effect was good.

As shown in Figure 3, the area under the ROC curve for the optimized Naive Bayes model is 0.7231, indicating that the model has a certain ability to distinguish between positive and negative samples; the closer the value is to 1, the better the performance. Meanwhile, the overall accuracy is 0.691, reflecting the model's overall performance, with higher values indicating better predictive effectiveness. The blue dashed line represents the baseline. Overall, the ROC curve and P-R curve jointly assess the effectiveness of the model in classifying elderly depression patients, with the ROC curve focusing on classification ability and the P-R curve emphasizing the balance between precision and recall.

Table 3. Naive Bayes Prediction Data Results

Dependent variable	Accuracy	Precision: 95% CI	Recall	F1-Score
CES－D2	69.1%	[0.689,0.712]	84.3%	81.7%

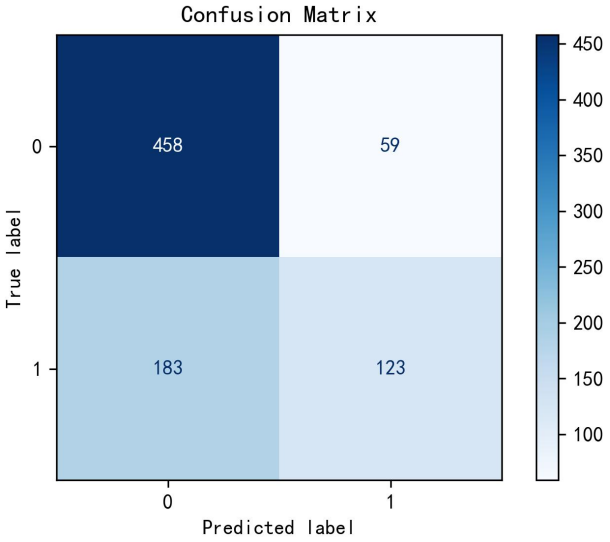

Figure 2. Confusion Matrix of the Naive Bayes Prediction

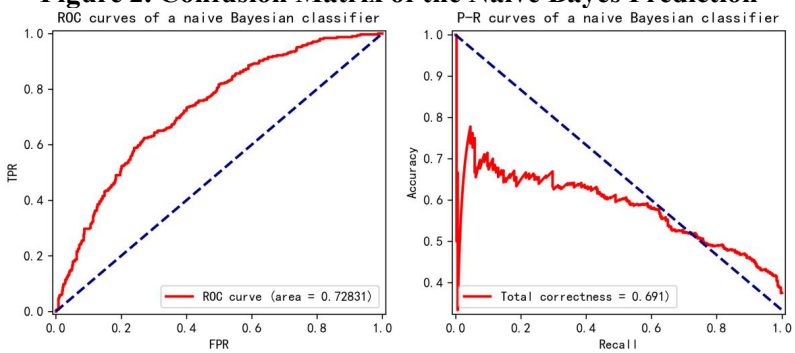

Figure 3. ROC Curve and P-R Curve of Naive Bayes Prediction

4.3 Prediction Based on Decision Trees

The prediction results based on the decision tree are shown in Table 4, with an accuracy of 65.9% for the test set, which is a decline compared to the Naive Bayes method. As

illustrated in Figure 4, the number of correctly classified instances for category 0 is 383, while there are 170 mispredictions, leading to a decrease in both recall rate and F1 value. Overall, the predictive performance of the decision tree model is unsatisfactory.

Table 4. Prediction Results of the Decision Tree

Dependent variable	Accuracy	Precision: 95% CI	Recall	F1-Score
CES－D2	65.9%	[0.648,0.673]	82.7%	80.3%

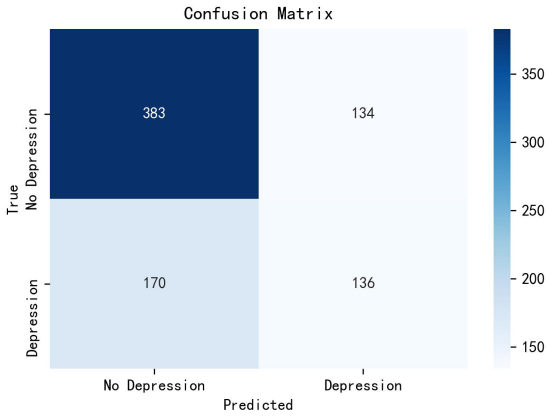

Figure 4. Confusion Matrix of the Decision Tree Prediction

As shown in Figure 5, the area under the ROC curve for the optimized decision tree model is 0.61, indicating that the model has moderate ability in distinguishing between positive and negative samples. This reflects the effectiveness of the classification task but also suggests that there is room for improvement.

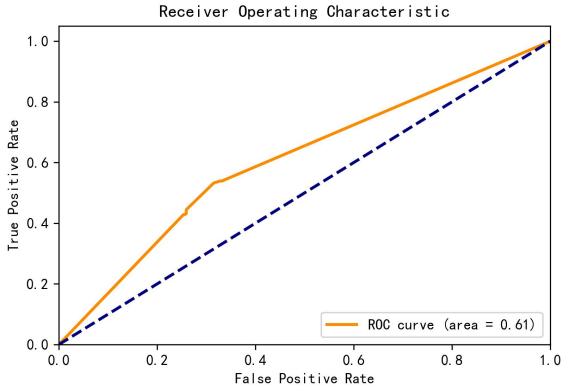

Figure 5. ROC Curve of the Decision Tree Prediction.

Table 5. Prediction Results of the Random Forest

Dependent variable	Accuracy	Precision: 95% CI	Recall	F1-Score
CES－D2	75.0%	[0.737,0.762]	88.4%	84.7%

4.4 Prediction Based on Random Forest

The prediction results using the random forest model are presented in Table 5. With the number of trees set to 600, the model achieves an accuracy of 75.0%, which is an improvement over the previous two models, and both recall rate and F1 value also perform well. As depicted in Figure 6, for category 0, there are 402 correctly classified instances and 162 mispredictions, indicating good predictive performance. However, the difference between correct and incorrect classifications for category 1 is relatively small

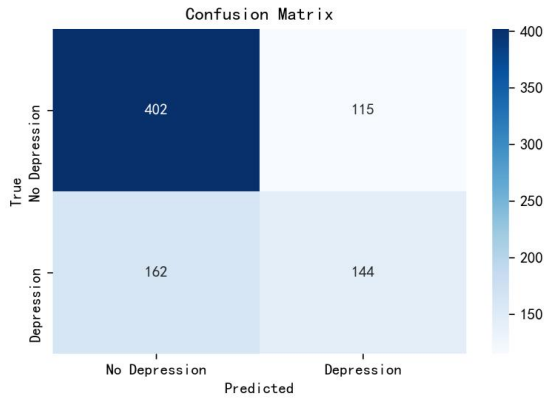

Figure 6. Confusion Matrix of the Random Forest Prediction

Figure 7 illustrates the bar chart of feature importance assessment, showing the relative importance of different features in the model's predictions. Age emerges as the most significant feature, with an importance value exceeding 0.35, indicating its substantial impact on the prediction outcomes. Following age, sleep quality and education level are also important, though their significance is lower than that of age. Other features, such as gender, alcohol history, and retirement status, have relatively low importance in the model.

4.5 Prediction Based on Support Vector Machine

The Support Vector Machine (SVM) demonstrates the best performance among the predictive models. As shown in Table 6, its accuracy, recall rate, and F1 score are all higher than those of the previous three methods, indicating that this model is more suitable for the selected data. From Figure 8, we can see that for category 0, there are 746 correctly classified instances and 367 mispredictions. Despite the overall performance, the prediction results for category 0 remain less than ideal, which may be due to the uneven distribution of the data,

as the range for category 0 is relatively large. Other metrics perform well, suggesting that the SVM model is overall applicable.

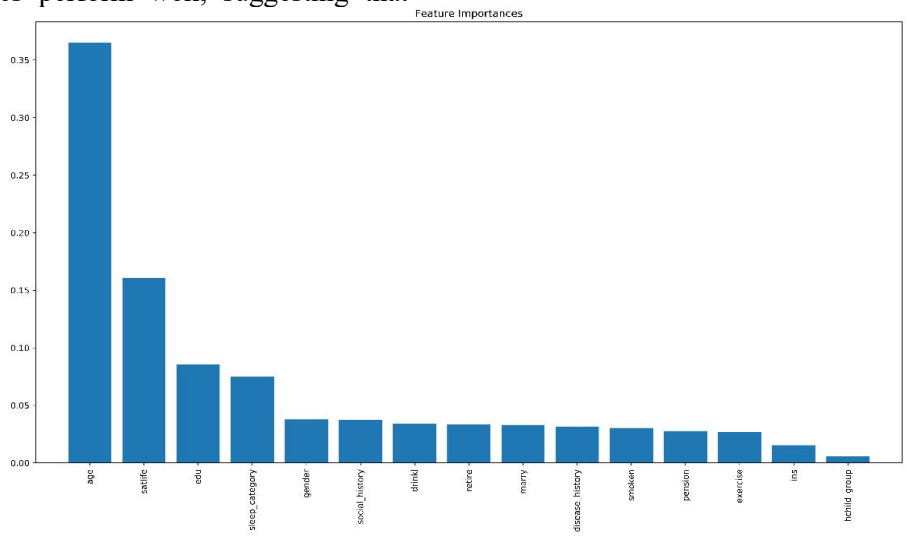

Figure 7. Bar Chart of Feature Importance Evaluation

Table 6. Prediction Results of the Support Vector Machine

Dependent variable	Accuracy	Precision: 95% CI	Recall	F1-Score
CES－D2	78.1%	[0.762,0.795]	91.8%	86.5%

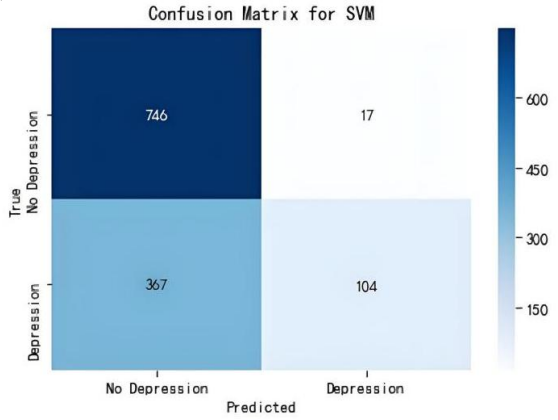

Figure 8. Confusion Matrix of the Support Vector Machine Prediction

As shown in Figure 9, the area under the ROC curve for the optimized SVM model is 0.77, indicating a strong ability to distinguish between positive and negative samples. The blue dashed line represents the baseline for random guessing, and the curve is significantly above this line, further emphasizing the model's effectiveness. These results demonstrate that the SVM model exhibits good classification performance in predictions, providing confidence for subsequent analyses.

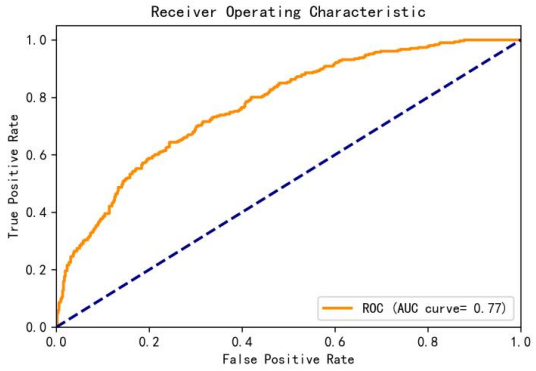

Figure 9. ROC Curve of the Support Vector Machine Prediction

4.6 Performance Comparison

Comparing the performance of different models is crucial in machine learning, as it evaluates how well multiple models perform on the same task to identify the most suitable one. In this study, the comparison is made across five aspects, including accuracy, precision, recall, F1 score, and ROC AUC. The specific results can be found in the Table 7.

Table 7. Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
NB Model	69.1%	[0.689,0.712]	84.3%	81.7%	69.1%
DT Model	65.9%	[0.648,0.673]	82.7%	80.3%	61.0%
RF Model	75.0%	[0.737,0.762]	88.4%	84.7%	75.9%
SVM Model	78.1%	[0.762,0.795]	91.8%	86.5%	77.5%

According to Table 7, considering all indicators, the SVM model is the best choice as it performs best in terms of accuracy, recall, F1-score, and AUC.

5. Conclusion

Based on the CHARLS dataset of the China Health and Elderly Care Longitudinal Survey, this study deeply analyzed the impact of depression in the elderly population. Through data preprocessing, descriptive analysis, feature engineering and model selection, the SVM model was successfully constructed, and its performance was better than that of Naive Bayes, decision trees and random forests. Studies have found that the incidence of depression in the elderly group is more than one-third, and their mental health needs to be paid attention to. Feature engineering screened out significant factors such as gender, marital status, and drinking habits. It is recommended that relevant institutions strengthen mental health publicity and intervention, and provide timely psychological counseling and treatment for the elderly in need. At the same time, the quality of life and social activities of the elderly should be improved to help them develop a positive attitude towards life. Medical and elderly care service institutions need to enhance the ability to identify and diagnose depression and intervene in a timely manner. In addition, future research should explore more influencing factors and interventions to improve prediction accuracy and support effective prevention and intervention strategies.

Acknowledgement

This work was supported by the 2023 Shaanxi Province College Students' Innovation and Entrepreneurship Training Program 'Smart Comfort, Caring for the Elderly—Smart Nursing Home' (Project Number: S202312712035)

References

[1] Merril S, Honge C G, Hal K. Perceived availability of future care and depressive symptoms among older adults in China: evidence from CHARLS. BMC Geriatrics,2020,20(1):31.

[2] Njai S B, Hinks A, Patterson A M, et al. Residual force enhancement is not altered while force depression is amplified at the cellular level in old age. The Journal of Experimental Biology, 2024,

[3] Zhu F, Wang Y, Yin S, et al. The effect of Tai Chi on elderly depression: A systematic review and meta-analysis of randomized controlled trials. Frontiers in Psychology, 2024, 15 1489384-1489384.

[4] Cordero A D. Exploring non-pharmacological interventions for older adults with depression. Asian Journal of Psychiatry, 2025, 103 104313-104313.

[5] Catalano L, Panico F, Trojano L, et al. Psychophysiological indices of late-life depression: A systematic review. Brain Research, 2024, 1849 149361.

[6] Nishida H, Takamiya A, Kudo S, et al. Characteristics of severe late-life depression in the prodromal phase of neurodegenerative dementia. The American Journal of Geriatric Psychiatry: Open Science, Education, and Practice, 2025, 5 10-20.

[7] Elsheikh M S S, Marshe S V, Men X, et al. Polygenic score analyses on antidepressant response in late-life depression, results from the IRL-GRey study. The Pharmacogenomics Journal, 2024, 24 (6): 38-38.

[8] Liu F, Ye J, Wei Y, et al. Factors associated with a high level of suicide risk among patients with late-life depression: A cross-sectional study from a tertiary psychiatric hospital in Guangzhou China. BMC Geriatrics, 2024, 24 (1): 933-933.

[9] Guo L, Sang B, Li S, et al. From starvation to depression: unveiling the link between the great famine and late-life depression. BMC Public Health, 2024, 24 (1): 3096-3096.

[10] Giehl C, Chatsatrian M, Vollmar C H, et al. Exploring accessibility, user experience and engagement of digital media among older patients with depression: A pilot and observational screening study protocol of the DiGA4Aged study. BMJ Open, 2024, 14 (11): e086779.

[11] Moyano P B, Lenart S K, Amoussou R J, et al. Prediction of electroconvulsive therapy response and remission in late-life depression: A review. Swiss Medical Weekly, 2024, 154 3684.

[12] Bandyopadhyay S. Loneliness and depression in older adults: Living well in older age. Medicine, 2024, 52 (11): 719-724.

