EAI Endorsed Transactions

on Pervasive Health and Technology Research Article EAI.EU

An Interpretable Stacking Ensemble Model with SHAP for Geriatric Depression Prediction: Analysis of the NHANES 2005-2023 Database

An interpretable stacking ensemble model for geriatric depression prediction

Manman Cui [1,2,a※], Xian Li [1,b※], Wei Gong [1,2,c*]

[1] School of Medical Information Engineering, Guangzhou University of Chinese Medicine,
[2] Intelligent Chinese Medicine Research Institute, Guangzhou University of Chinese Medicine

Abstract

OBJECTIVE: Leveraging multimodal data from the 2005-2023 National Health and Nutrition Examination Survey (NHANES) database, this study aims to develop a predictive method for the geriatric depression that combines high predictive accuracy with good interpretability, thereby providing support for in-depth exploration of the pathogenesis and risk factors of geriatric depression.
METHODS: Data from 8760 participants aged 65 and older in the NHANES database from 2005-2023 are utilized to develop and validate the stacking ensemble predictive model. Depression is assessed using the Patient Health Questionnaire-9 (PHQ-9) total score meeting or exceeding 10. Before the model construction, this work employs the normalization of training data and test data, Synthetic Minority Over-sampling Technique - Random Under-Sampling (SMOTE-RUS) hybrid sampling strategy to address the class imbalance, and the recursive feature elimination method based on the random forest (RFE-RF) for feature selection. A stacking ensemble predictive framework for depression is constructed based on the primary learners (Random Forest, SVM, XGBoost, and Logistic Regression) and meta-learners (SVM and Logistic Regression). Finally, the interpretable machine learning technique SHapley Additive exPlanations (SHAP) is used to visualize the model predictive outputs.
RESULTS: The XGBoost model demonstrated outstanding performance on the test set in terms of AUC (83.92%), while the Random Forest (RF) model excelled in sensitivity (71.05%). Subsequently, a specifically designed RFE-Stacking ensemble model, using RF and XGBoost as the primary learner and the SVM as the meta-learner, is developed. In comparison, this stacking ensemble model exhibits the best predictive performance with the biggest AUC (85.14%) and the highest sensitivity (78.71%). The SHAP interpretation reveals that general health condition, frequency of oral pain in the past year, marital status, history of mental health consultations in the past year, and frequency of urine leakage are the top five most influential factors in predicting the depression risk.
CONCLUSION: This stacking ensemble model enhances the performance of both the primary learners and the meta-learners. This verifies the feasibility and effectiveness of the proposed model in predicting the geriatric depression. This work integrating the stacking ensemble model with SHAP offers valuable clinical references for assessing the risk of depressive symptoms, which is beneficial to develop the personalized depression interventions and preventions in the elderly.

Keywords: Geriatric Depression, Predictive Model, Stacking Ensemble Learning, SHAP

Received on 14 May 2025, accepted on 10 December 2025, published on 28 January 2025

doi:10.4108/eetpht.11.11671

※These authors contributed equally to this work and should be considered co-first authors.
* Corresponding author: Gong Wei: gongwei@gzucm.edu.cn
[a]cmm@gzucm.edu.cn, [b]llixianlily@163.com

1. Introduction

As a prevalent mental disorder, depression exerts severe negative impacts on the health and life quality of the older adults, while significantly elevating their risks of suicide and the burden of comorbid physical illnesses, such as cardiovascular diseases [1]. Statistics indicate that the global prevalence of depression among the elderly reached a concerning 31.74% in 2021 [2]. In recent years, alongside the accelerating trend of global population aging, this rate continues to exhibit a strong upward trajectory [3]. The high incidence of geriatric depression and its serious consequences have garnered widespread attention from both the medical communities and the academic communities [4]-[6]. Consequently, developing efficient and accurate risk prediction models for geriatric depression is crucial for enabling early intervention and alleviating the pressure on healthcare systems.

Traditional researches on predicting the geriatric depression risk has predominantly relied on questionnaire-based socio-structural analyses or conventional statistical methods. These approaches often focus on identifying individual risk factors and employ multivariate linear regression to explore their associations with the depression [7]-[8], followed by the theoretical analysis at the sociological level. However, these methods, largely based on linear assumptions, fail to adequately capture the complex, non-linear relationships among variables. Consequently, their predictive accuracy and generalization capability are greatly limited when handling the high-dimensional, non-linear medical data.

With advancements in artificial intelligence, machine learning techniques, leveraging their powerful pattern recognition and inherent ability to model non-linear relationships, have demonstrated significant potential in predicting the geriatric depression. Many current machine learning-based predictive models frequently utilize multiple machine learning algorithms and compare their performance. Literature reviews suggest that Logistic Regression, Support Vector Machines, Random Forest, and the eXtreme Gradient Boosting (XGBoost) model are among the most frequently applied algorithms [9]. Typically, ensemble models based on decision trees (such as Random Forest and the XGBoost model) exhibit superior performance [9], specially demonstrating particular advantages in handling high-dimensional data and complex feature interactions [10]. Nonetheless, single models often face performance bottlenecks, including insufficient accuracy and limited generalization ability. Furthermore, the opaque "black-box" nature of their decision-making processes substantially hinders clinical trust and adoption of their outputs by healthcare practitioners [11].

To address these challenges, this study introduces an ensemble learning strategy and aims to develop an interpretable Stacking ensemble model that combines high sensitivity, strong generalization capability, and good interpretability, thereby providing a novel data-driven solution for the early prevention and control of the geriatric depression. Specifically, this work first trains multiple individual learners. The models demonstrating optimal performance in sensitivity (to minimize missed diagnoses) and AUC (reflecting overall discriminative power) are selected as the primary learner [4]. Their predictions are then aggregated to form a new feature space and input into the meta-learners composed of linear models for the final learning stage. At last, we obtain the optimal stacking ensemble model by comparing the performance metrics of meta-learners. This design not only harnesses the strengths of diverse non-linear models but also effectively controls the overall model complexity and overfitting risk through the linear meta-learner, thereby enhancing performance while ensuring robustness. To mitigate the impact of the "black-box" decision process on the clinical translation and application of the predictive model, this paper employs the explainable AI tool SHapley Additive exPlanations (SHAP) [12] to provide both global explanations and local explanations for the aforementioned optimal Stacking ensemble model, enhancing transparency and trust in its outputs. Experimental results verify the feasibility and the effectiveness of the proposed model, indicating its great latent potential as a valuable tool for clinical practice in the geriatric depression.

2. Methods

2.1 Data and Variables

The study focuses on data from participants aged 65 and older in the NHANES database from 2005-2023. Depression is screened for using the Patient Health Questionnaire-9 (PHQ-9) in NHANES. This questionnaire is widely used for the rapid screening and preliminary identification of depression, systematically assessing the participants' mental and psychological state over the past two weeks. Compared to similar screening tools, it demonstrates higher sensitivity and specificity in diagnostic efficacy. In this work, if any of the 9 questions in the Patient Health Questionnaire-9 (PHQ-9) [13] had a missing response, the individual's data would be excluded. After applying these exclusions, a total of 9,060 samples from individuals aged 65 and above are ultimately included. Specifically, depression would be defined based on whether the total score of the 9 PHQ-9 questions is greater than 10. Participants with a PHQ-9 total score meeting or exceeding 10 are classified into the depression group (coded as 1), while those below this threshold are classified into the non-depression group (coded as 0).

The variables incorporated into the study consisted of four parts: (1) Socio-demographic characteristics: such as age, gender, educational level, and marital status; (2) Dietary nutrition data: including energy, protein, vitamins, and carbohydrate intake; (3) Laboratory examination data: such as blood pressure, cholesterol levels, insulin levels, and Body Mass Index (BMI); (4) Questionnaire data: including hearing status, history of chronic diseases, alcohol consumption, and history of prescription drug use. It is noteworthy that the NHANES (National Health and Nutrition Examination Survey) database exhibits the following typical

characteristics: (1) High-dimensional features: It contains multi-dimensional data including demographics, diet, physical examinations, and laboratory tests. (2) Data complexity: It has diverse variable types (such as positively skewed distribution laboratory data and continuous variables) and contains missing values. (3) Class imbalance: The number of patients is usually much smaller than that of healthy individuals. Directly using such data for model training would cause the model to be severely biased towards the majority class, resulting in very poor predictive performance for the minority class.

Therefore, a series of data preprocessing operations are required on the original dataset before model training. Specifically, this paper proposes a multi-stage data preprocessing process in Subsection 2.2 to reduce data complexity, applies a hybrid sampling method in Subsection 2.3 to address the class imbalance problem, and introduces feature engineering in Subsection 2.4 to mitigate the overfitting risk caused by high-dimensional features, all while ensuring model predictive performance and reducing model complexity.

2.2 Data Preprocessing

To ensure data quality and model reliability, this study adopts a multi-stage data pretreatment process, primarily optimizing data quality through Removal of Duplicate Variables and Missing Values, Multiple Imputation by Chained Equations (MICE) [14], and Data Normalization. The specific process is illustrated in Figure 1.

(1) Removal of Duplicate Variables and Missing Values: In order to mitigate multicollinearity effects, duplicate variables representing the same information but with different units are deleted. Variables with missing values exceeding 50% across the four main database modules are also removed.

(2) Multiple Imputation by Chained Equations (MICE) for missing values in retained variables: In complex missing data scenarios, particularly with large-scale datasets, the data often exhibit multidimensional and missing-at-random characteristics. To fully leverage the available information and enhance data quality of the NHANES database, this study employs the Multiple Imputation by Chained Equations (MICE) [14] for handling missing values in preserved variables. The MICE operate by constructing conditional distribution models between variables to generate multiple imputed datasets for joint inference. Specifically, an iterative algorithm establishes regression models for each target variable based on other completely observed variables. During each iteration, the parameters of predictive models are updated using currently imputed variable values, thereby generating new imputations. This process iteratively refines the conditional distribution models across variables, producing multiple complete datasets through collaborative imputation.

(3) Data Standardization: As abovementioned, the NHANES database has diverse variable types. In order to improve the training efficiency and accuracy of machine learning models, this work performs a series of data standardization operations. On one hand, as to the positively skewed distribution laboratory data (such as cholesterol and blood components), we apply the logarithmic transformation to make the data distribution more uniform, presenting characteristics closer to normal distribution, thereby improving the performance of the model. On the other hand, as to the continuous variables (like energy, protein, and dietary fiber intake) are subjected to Min-Max normalization, scaling the data to the [0, 1] interval while preserving the original data distribution characteristics, which aims at eliminating dimensional differences between data features and making the numerical ranges of different features consistent.

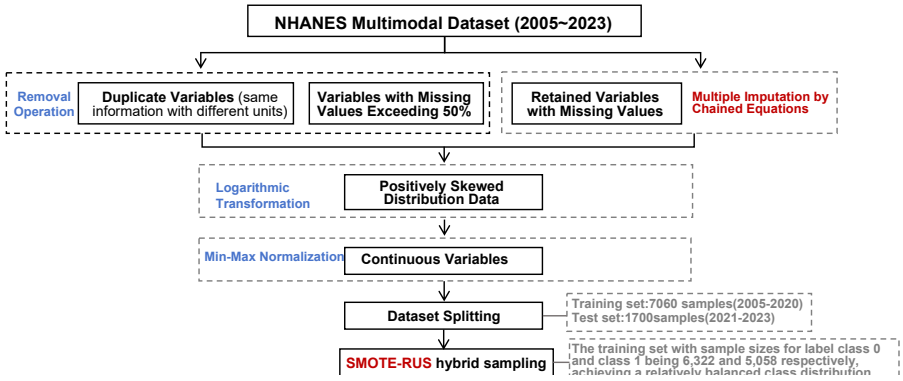


Figure 1. Data Preprocessing and Feature Selection

2.3 SMOTE-RUS Hybrid Sampling

Class imbalance is a prevalent issue in healthcare datasets such as NHANES. Resampling techniques are pivotal in mitigating this challenge. Commonly employed methods include the Synthetic Minority Over-sampling Technique (SMOTE) [15], Random Under-Sampling (RUS) [16], and more advanced hybrid sampling methods like Synthetic Minority Over-sampling Technique-Edited Nearest Neighbours (SMOTE-ENN) and Adaptive Synthetic Sampling (ADASYN). However, it is important to note that SMOTE-ENN suffers from high computational costs in high-dimensional spaces, where the "curse of dimensionality" destabilizes the very notion of nearest neighbors. Meanwhile, ADASYN is highly sensitive to noise due to its data-generating mechanism. Given the high dimensionality, complexity, and inherent noise of the NHANES dataset, SMOTE-ENN and ADASYN methods are suboptimal for addressing imbalance in the NHANES dataset.

To tackle the class imbalance in the NHANES training data, this study adopts a hybrid sampling strategy that combines SMOTE with RUS (SMOTE-RUS). In this framework, SMOTE enhances the representation of the minority class, while RUS counteracts the dominance of the majority class. The application of RUS not only rectifies the class distribution skew, thereby significantly improving subsequent model training speed, but also helps reduce the influence of potential noise and redundant samples within the majority class. Although RUS carries a risk of discarding some information, the introduced randomness can serve as a form of regularization. This helps prevent the model from overfitting to specific nuances of the majority class, ultimately enhancing the model's generalization capability.

It should be noted that a preprocessing pipeline should be constructed to strictly confine the aforementioned sampling operations to the training set. Therefore, it could ensure the model to learn from a balanced data distribution, while simultaneously ensuring its final evaluation is performed on a test set that represents the real-world, unaltered distribution. This approach fundamentally prevents information leakage and evaluation bias attributable to improper data preprocessing.

2.4 RFE-RF Feature Selection

Despite of the data preprocessing, the dataset still contains 147 feature variables. Such a high-dimensional feature set would not only significantly increase the computational burden and reduce training efficiency but also, due to the curse of dimensionality, introduce substantial noise and redundancy. This greatly heightens the risk of model overfitting and compromises its generalizability. Therefore, implementing effective feature dimensionality reduction becomes a critical step in building a robust predictive model.

Compared with the Lasso regression, the recursive feature elimination (RFE) method [17] based on tree models is insensitive to feature collinearity, typically yielding more stable and reproducible screening results. Therefore, this work adopts a RFE method based on random forest (i.e., RFE-RF) for feature selection. Specifically, this method employs a systematic and iterative process, guided directly by model performance, to identify the most predictive feature subset. As depicted in Figure 2, the specific procedure is given as follows:

(1) Establishing a Performance Benchmark: An initial Random Forest model is trained using all 147 features, and its predictive accuracy is established as the performance benchmark. Then, we calculate the importance of all features according to their information gain and rank them based on the feature importance.

(2) Iterative Feature Evaluation: The least important features are temporarily removed at a time. A new Random Forest model is then trained using the remaining feature subset, and its predictive accuracy is recorded.

(3) Performance-Based Elimination Decision: The accuracy of the new model is compared against the current benchmark. If the removal of least important features resulted in an accuracy was higher than the current benchmark, that features would be permanently removed. Conversely, if the accuracy decreased, the features would be retained in the set.

(4) Cycling and Convergence: Steps 2 and 3 are repeated, and the highest achieved accuracy at each iteration is updated as the new performance benchmark. This process continues until the model's predictive accuracy could no longer be improved and stabilized, at which point the optimal feature subset is determined.

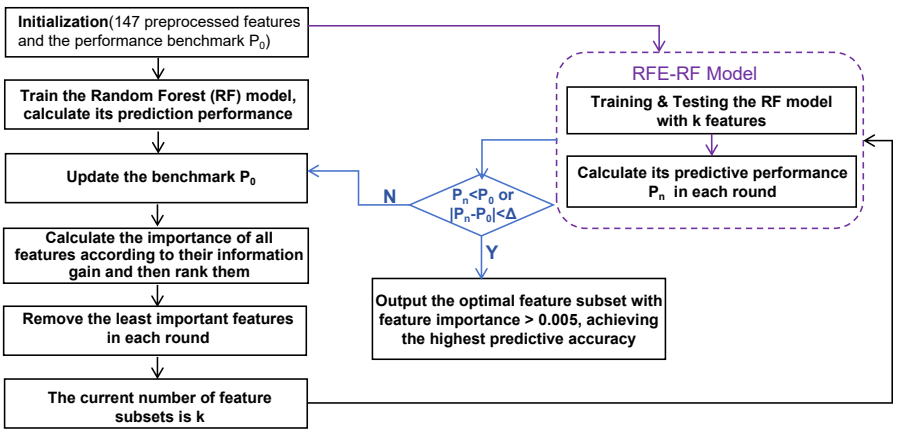


Figure 2. RFE-RF feature selection process

Through the process, this refined feature set is retained and used for all subsequent model construction, ensuring the efficiency and robustness of the final model.

2.5 Model Construction

As illustrated in Figure 3, four individual learners are constructed based on Random Forest, Support Vector Machines (SVM), XGBoost, and Logistic Regression. The hyperparameters of each individual learner would be tuned through grid search and manual fine-tuning. In scenarios such as depression prediction, maximizing the clinical value of identifying potential patients (high sensitivity, i.e., high recall rate) is usually much higher than avoiding misjudgment of healthy individuals (high specificity). Given that, we select individual learners with optimal performance in the sensitivity and AUC to form the primary learner. The predictions from these primary learners are combined to create a new feature space. Notably, using linear models as meta-learners not only mitigates overfitting risks but also creates complementary synergy with the complex non-linear primary learners, thereby reducing overall system complexity [18]. Considering that, we then build stacking ensemble models by employing the Logistic Regression (LR) model and the SVM model as meta-learners, respectively, to identify the optimal predictive model. Finally, we apply SHAP, an explainable machine learning framework, to provide both global and local interpretations of the optimal model's predictions.

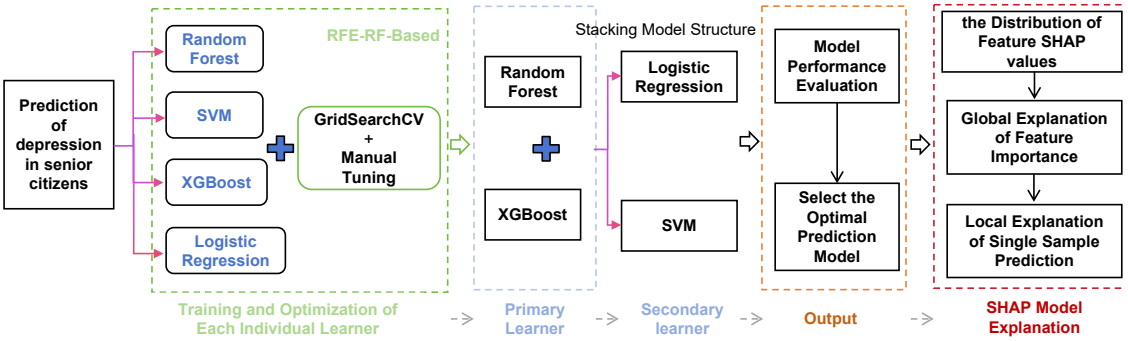


Figure 3. The model construction process

3. Experimental Results and Analysis

In this section, four RFE-based individual learners (i.e., RFE-based Logistic Regression, RFE-based Random Forest, RFE-based SVM, and RFE-based XGBoost) are trained. Then two RFE-based stacking ensemble models are built by employing Logistic Regression and SVM as meta-learners, respectively. At last, the visualization analysis is given by the explainable machine learning framework SHAP, with both the global explanations and the local explanations of the optimal model's predictions.

The model evaluation and output interpretation in this study are all implemented in PyCharm 2024.

3.1 Dataset Splitting

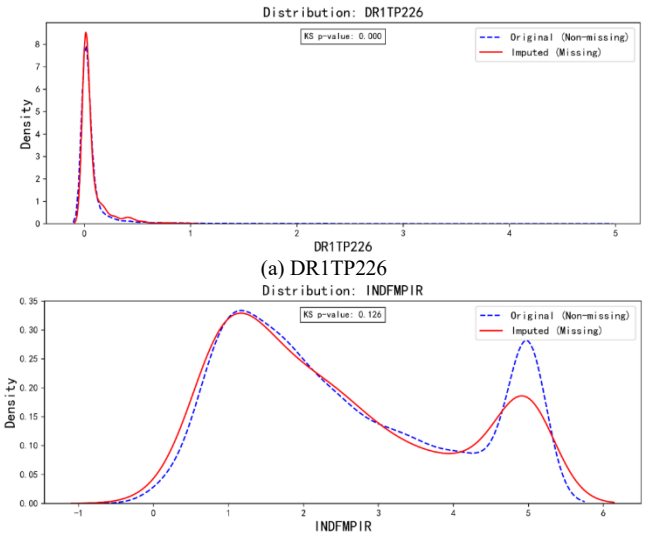
To evaluate the external validity and temporal generalizability of the proposed stacking ensemble model, we split the dataset with a rigorous temporal validation strategy. Specifically, data from the 2005-2020 NHANES cycle (n=7,060) are used as the training set for model development and hyperparameter tuning, while the temporally independent data from the 2021-2023 NHANES cycle (n=1,700) are designated as the test set. This approach simulates a realistic application scenario where the model is applied to a future population. All model performance metrics, including AUC, accuracy, and F1-score, are reported exclusively on this external test set. The performance on this temporally separate external data serves as the primary evidence for the model's external consistency.

3.2 MICE, Hybrid Sampling and Feature Selection

3.2.1 MICE

To scientifically evaluate the appropriateness of the imputation method MICE, this study employs a combined approach of data visualization and statistical testing. At first, an intuitive assessment is conducted by plotting density comparison charts of the data before and after imputation. For cases where the density plots indicate noticeable discrepancies, the Kolmogorov-Smirnov (KS) non-parametric test is further applied for quantitative validation.

Taking the variables DR1TP226 and INDFMPIR as examples, their density comparison charts reveal distinct patterns, as shown in Figure 4. For DR1TP226, the post-imputation data curve (represented by a solid red line) shows a high degree of overlap with the original data curve (represented by a blue dashed line), indicating that the distribution of the imputed missing values is largely consistent with the non-missing portion of the original data. In contrast, the density comparison charts for INDFMPIR exhibits some observable differences. Consequently, the KS test is performed, calculating its D statistic (the maximum distance between the two sample cumulative distribution functions) and the corresponding p-value. The results show a very low D statistic of 0.001, accompanied by a p-value of 0.126, which exceeds the common significance level of 0.05. This strongly suggests that the distributions before and after imputation are highly similar. It can therefore be concluded that the imputation method adopted in this study is effective and appropriate, as it successfully preserves the original statistical characteristics of the variables without introducing significant distributional bias due to the handling of missing values.



(a) DR1TP226

(b) INDFMPIR

Figure 4. Density comparison charts and KS non-parametric test

3.2.2 Hybrid Sampling

Following the application of the SMOTE-RUS hybrid sampling, the total sample size of the training set is adjusted to 11,380. The sample counts for class 0 and class 1 become 6,322 and 5,058, respectively, achieving a state of relative class balance at the training set level.

3.2.3 Feature Selection

After the RFE-RF feature selection, the study ultimately identifies an optimal feature subset composed of 24 key features (with the feature importance greater than 0.005). This specific feature combination yields the highest predictive accuracy of 83.05%. The names and importance rankings of these features are detailed in Table 1.

Table 1. Feature importance ranking

Variable Name	Variable Description	Ranking	Importance
HUQ010	General health condition	1	0.189
OHQ620	Frequency of oral pain in the past year	2	0.123
KIQ005	Frequency of Urine leakage	3	0.077
MCQ160A	Diagnosis of Arthritis	4	0.052
KIQ044	Urinary Leakage or Loss of Control due to urgency or pressure during urination	5	0.049
KIQ042	Urinary Leakage due to coughing, or other physical activities	6	0.048
DMDMARTZ	Marital Status	7	0.046
DMDEDUC2	Education Level	8	0.044
HUQ00	History of Mental Health Consultations in the past year	9	0.042
LBXHA	Hepatitis A Antibody	10	0.037
INDFMMPC	Monthly poverty line category for households	11	0.034
LBDBANO	Number of basophil granulocyte	12	0.019
INDFMPIR	Household income and poverty ratio	13	0.018
DBQ095Z	Types of salt used	14	0.017
BPQ020	Does the patient have hypertension before	15	0.017
DR1TTHEO	Theobromine	16	0.016
LBDEONO	Number of eosinophils	17	0.015
BPQ080	Blood cholesterol levels	18	0.013
DIQ010	Whether the patient has diabetes	19	0.008
MCQ010	Does the patient have asthma before	20	0.008
INDFMMPI	Monthly poverty level index of households	21	0.007
HSQ590	Does the patient have the AIDS virus infection in the Blood	22	0.006
MCQ160F	Has the patient had a stroke before	23	0.006
DR1TNIAC	Niacin	24	0.006

In order to ensure the model robustness, this paper further calculates the Variance Inflation Factors (VIFs) of the 24 selected key features to test the inter-feature multicollinearity. As shown in Figure 5, the vast majority of features exhibit VIF values close to 1, indicating negligible linear dependence among them. A small subset of features—namely INDFMMPC, INDFMHR, and INDFMMPI, which are related to household income and poverty levels—show VIF values ranging from 2.14 to 3.47. This suggests mild multicollinearity among them. Nevertheless, this is an expected outcome given the inherent correlations between such socioeconomic indicators in reality. it should be noted that, since the degree of correlation is sufficiently low as to not pose a threat to model integrity, retaining these features helps capture a more comprehensive socioeconomic context.

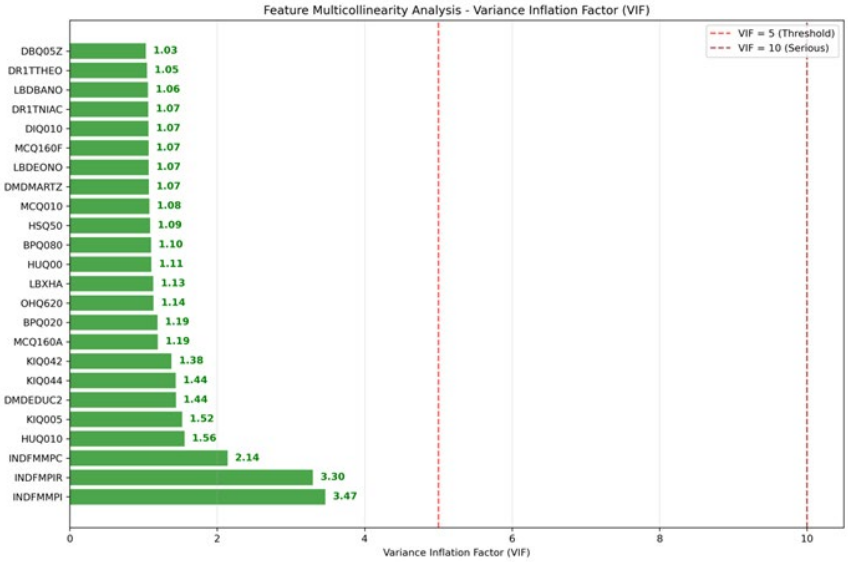


Figure 5. Variance Inflation Factor analysis

3.3 Performance Indicators

3.3.1 Accuracy

Accuracy reflects the proportion of samples correctly classified by the model to the total number of samples. High accuracy means that the model has good overall classification ability. Its expression is as follows,

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where *TP* is the number of True Positive, *TN* is the number of True Negative, *FP* is the number of False Positive, *FN* is the number of False Negative.

3.3.2 Recall rate (Sensitivity)

Recall rate, also known as sensitivity, refers to the probability of correctly predicting depression in all samples with actual depression. High recall rate contributes to reducing false negatives and lowering the risk of missed diagnosis. Its expression is as follows,

$$R = \frac{TP}{TP+FN} \quad (2)$$

3.3.3 Specificity

Specificity refers to the probability of correctly predicting healthy persons in all samples that does not actually have depressive symptoms. High specificity contributes to reducing false positives and avoiding unnecessary medical interventions. Its expression is as follows,

$$Sp = \frac{TN}{TN+FP} \quad (3)$$

3.3.4 ROC curve

ROC curve is a curve graph with false positive rate (FPR) as the horizontal axis and true positive rate (TPR) as the vertical axis. Its derivative indicator AUC evaluates model performance by quantifying the area under the curve. High AUC value means that the model has good comprehensive performance. The expressions of FPR and TPR are

$$FPR = \frac{FP}{TN+FP}, TPR = \frac{TP}{FN+TP} \quad (4)$$

3.4 Model Evaluation

The performance metrics of four RFE-based individual learners and two stacking ensemble models are compared in

Table 2. We first conduct a horizontal performance comparison of four individual learners to establish performance benchmarks. As shown, the RFE-based Random Forest learner attains a highest sensitivity of 71.05% and a higher AUC of 83.42% among these four individual learners; the RFE-based XGBoost learner has a superiority of AUC (83.92%), accuracy (85.22%), and specificity (88.41%), but with a lowest sensitivity (62.28%). Thus, the RFE-based Random Forest learner and the RFE-based XGBoost learner are naturally selected to form the primary learners. As aforementioned, the LR model and the SVM model are employed as meta-learners, respectively, to build two stacking ensemble models, i.e., the RFE-based Stacking LR learner and the RFE-based Stacking SVM learner. From Table 2, we can clearly see that, compared to the RFE-based Stacking LR learner, the RFE-based Stacking SVM learner has better performance in terms of sensitivity (78.71%) and AUC (85.14%), both of which are significant to the depression prediction. Therefore, we would employ the SVM model as the optimal meta-learner.

Table 2. Comparison of performance metrics across models on the testing dataset

Models	Sensitivity (Recall Rate)	AUC	Accuracy	Specificity
RFE-based Logistic Regression (LR)	63.45	83.03	80.88	82.30
RFE-based Random Forest	**71.05**	83.42	81.65	81.59
RFE-based SVM	69.74	82.23	79.20	79.18
RFE-based XGBoost	62.28	**83.92**	**85.22**	**88.41**
RFE-based Stacking LR	73.54	84.87	80.27	78.62
RFE-based Stacking SVM	**78.71**	**85.14**	82.12	80.46

Then we conduct the vertical performance comparison between the primary learners and the stacking ensemble models. As shown, compared to the RFE-based Random Forest learner and the RFE-based XGBoost learner, the RFE-based Stacking SVM learner can achieve a 7.66% and 16.43% improvement in sensitivity and a 1.71% and 1.22% improvement in AUC, respectively. In order to capture the true positives missed by XGBoost and improve the sensitivity, the meta-learner inevitably misjudges some of the true negatives (i.e., healthy individuals) correctly classified by XGBoost as the positives. Therefore, although the sensitivity of the RFE-based Stacking SVM learner has significantly increased, its specificity has plummeted from 88.41% of XGBoost to 80.46%. The significant decrease in the specificity of the RFE-based Stacking SVM learner means that more errors are made in the majority of healthy individuals, which directly lowers the overall accuracy. Figure 6 shows the comparison of ROC curves for the six RFE-based learners.

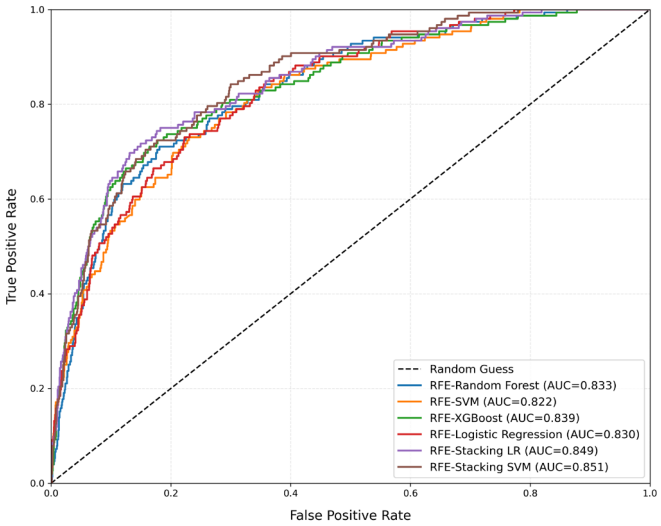
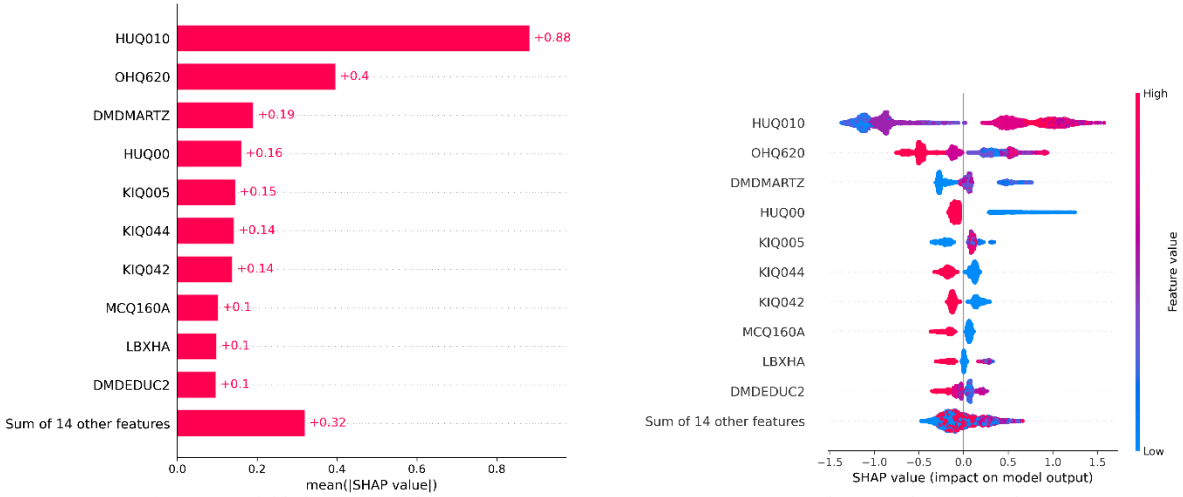


Figure 6. ROC curves of six RFE-Based models on the test dataset

3.5 SHAP Visualization Analysis

In this subsection, we would provide both global explanations and local explanations of the optimal model's predictive results. Based on the above analysis, the RFE-based Stacking SVM learner as the optimal depression predictive model is selected for feature analysis.



a) SHAP variable importance plot

b) SHAP beeswarm plot

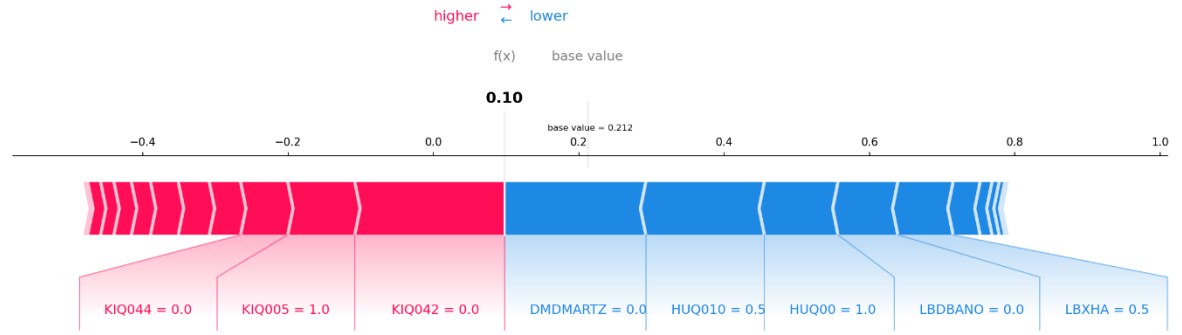
Figure 7. Global SHAP explanation. a) SHAP variable importance plot, b) SHAP beeswarm plot. (Note: In the beeswarm plot, blue-colored points represent low assigned feature values, while red-colored points represent high assigned feature values. The following parentheses provide the corresponding encoding explanations for each feature: (i) HUQ010 denotes self-rated general health status (Excellent: 0 / Good: 0.25 / Fair: 0.5 / Poor: 0.75 / Very Poor: 1). (ii) OHQ620 indicates the frequency of oral pain in the past year (Very often: 0 / Fairly often: 0.25 / Occasionally: 0.5 / Hardly ever: 0.75 / Never: 1). (iii) DMDMARTZ describes marital status (Married/Living with partner: 0 / Widowed, divorced, or separated: 0.2 / Never married: 0.4). (iv) KIQ005 represents the frequency of urine leakage (Never: 0 / Less than once a month: 0.25 / A few times a month: 0.5 / A few times a week: 0.75 / Every day and/or night: 1). (v) HUQ00 indicates whether the individual saw a mental health professional in the past year (Yes: 0 / No: 1). (vi) KIQ044 specifies urine leakage due to urgency or pressure (Yes: 0 / No: 1). (vii) KIQ042 indicates urine leakage associated with activities like coughing or exercise (Yes: 0 / No: 1). (viii) MCQ160A shows the diagnosis of arthritis (Yes: 0 / No: 1). (ix) LBXHA denotes Hepatitis A antibody test result (Positive: 0 / Negative: 1). (x) DMDEDUC2 indicates the education level (No diploma: 0 / Junior high school: 0.5 / High school: 0.5 / College: 0.75 / Bachelor's degree or above: 1).)

Figure 7 presents the global explanations based on SHAP, visualizing the impact of features on the overall model. Figure 7a) shows the SHAP variable importance plot, depicting the average impact of each feature on the model's prediction results from a global perspective. From Figure 6a), we can see that the top 10 variables influencing the geriatric depression status were: General Health Condition (HUQ010, +0.88); Frequency of Oral Pain in the past year (OHQ620, +0.4); Marital Status (DMDMARTZ, +0.19); History of Mental Health Consultations in the past year (HUQ00, +0.16); Frequency of Urine Leakage (KIQ005, +0.15); Urinary leakage or loss of control due to urgency or pressure during urination (KIQ044, +0.14); Urinary leakage due to coughing, exercise, or other activities (KIQ042, +0.14); Diagnosis of Arthritis (MCQ160A, +0.1); Hepatitis A Antibody (LBXHA, +0.1); and Education Level (DMDEDUC2, +0.1). Among them, the SHAP importance value of general health status (+0.88) far exceeds other features and is the most important predictive basis for the model. This indicates that the subjective evaluation of one's own general health by the elderly is the strongest indicator of the risk of depression. In addition, the importance of "Sum of 14 other features" reaches 0.32. This indicates that, although individual long tail features have a weak impact, their contributions to the model cannot be ignored, providing rich supplementary information.

Figure 7b) gives the SHAP beeswarm plot, a deepening of the variable importance plot, depicting how each feature of each sample affects the model prediction and revealing the relationship between feature values and the direction of influence (positive/negative). From Figure 6b), it can be seen that (1) General Health Condition (HUQ010): The higher the assigned feature values (red representing poorer self-rated health), the more biased the SHAP values are towards the right, which indicates the positive impact of General Health Condition on pushing up the risk of depression. (2)

Frequency of Oral Pain in the past year (OHQ620): Similar to General Health Condition, high assigned feature values (red representing high pain frequency) are densely distributed in the positive SHAP value area. (3) Marital Status (DMDMARTZ): The distribution of these feature values (blue for unmarried/ divorced /widowed, red for married) is relatively scattered, but the blue dots appear more on the left side (protective effect), while red ones have a higher proportion on the right side (risk effect). (4) Education Level (DMDEDUC2): Low assigned feature values (blue representing low education level) are more distributed in the positive SHAP value area, while high eigenvalues (red representing high education level) are more distributed near the central axis or negative SHAP value area. (5) Other features: The different values of the 14 other features (blue and red) are widely distributed in the positive and negative SHAP value regions, and do not show a simple linear relationship.

The SHAP force plot provides local explanations for individual predictions. The visualization results for two randomly selected samples are shown below in Figure 8. From Figure 8a), the 130389th individual has a predicted depression risk probability of 0.1, which is obviously lower than the model's base value (average predicted probability) of 0.21. This indicates that the model predicts this sample as not having depressive symptoms, which is consistent with the sample's actual label. The SHAP force plot reveals that key features contributing to this prediction are stable Marital Status (DMDMARTZ=0.0), good General Health Condition (HUQ010=0.5), and no History of Mental Health Consultations (HUQ00=1), which increased the probability of being classified as non-depressed. While high frequency of Urine Leakage (KIQ005=0.0) and Urinary Leakage or Loss of Control upon Urgency (KIQ044=0.0) or Physical Activities (KIQ042=0.0) increase the probability of depression.



a) SHAP force plot of the 130389th individual

b) SHAP force plot of the 130413th individual

Figure 8. Local SHAP Explanation

From Figure 8b), the 130413th individual has a predicted depression risk probability of 0.41, which is significantly higher than the model's base value of 0.21. Thus this sample would be predicted as having a high risk of depressive symptoms. The SHAP force plot reveals that key features that increase the risk probability of depression are Frequency of oral pain in the past year (OHQ620 = 0.5), Urine Leakage upon Physical Activities (KIQ042=0.0), the lack of a stable Marital Status (DMDMARTZ = 0.2), while the key features that decrease the risk probability of depression are good General Health Condition (HUQ010=0.25), No Urine Leakage due to Urgency (KIQ044 = 1.0), No history of Arthritis (MCQ160A = 1.0), and No history of Mental Health Consultations (HUQ00 = 1.0).

4. Discussion

This study focuses on the study of an interpretable predictive model for the geriatric depression and the explanation of its predictive mechanisms, systematically integrating a relatively comprehensive technical framework encompassing

data preprocessing, feature engineering, machine learning modeling, and interpretability analysis.

4.1 Discussion of Data Processing and Feature Engineering

To develop a predictive model with enhanced accuracy and robustness, we integrate the multi-source, heterogeneous data from the NHANES database (2005-2023). Beyond basic scale data, we incorporate enhanced modalities including proteomics, metabolomics, genomics, and environmental exposomics. This approach is designed to transcend the limitations of unimodal data, enabling the model to capture the complex feature space of depression through the integrated "Biopsychosocial" model [19], thereby improving the accuracy of assessing multi-risk factor contributions. For data preprocessing, we employ the MICE method to handle missing values and a hybrid sampling technique (SMOTE-RUS) to mitigate class imbalance, ensuring a high-quality dataset for model training. Given the complex etiology of depression, which involves numerous non-linear relationships and intricate interaction effects between risk factors and clinical outcomes, we introduce the Random Forest algorithm to capture such complex patterns in the feature selection stage. In specific, we utilize a Random Forest-driven RFE process, not the Lasso regression which is constrained by its inherent linearity assumption and sensitivity to multicollinearity, to identify features involved in these non-linear and interactive associations with depression. Using the RFE-RF feature selection, this study predicts 24 key features.

4.2 Discussion of Model Performance

During the primary learner construction phase, the predictive performance metrics of four individual learners are compared in Table 2. As shown, the RFE-based Logistic Regression learner and the RFE-based XGBoost learner exhibit similar "high specificity, low sensitivity" patterns, indicating that they are more inclined to reduce misdiagnosis (False Positive) during the learning process, but at the cost of increasing misdiagnosis (False Negative); While the RFE-based Random Forest learner exhibits the opposite tendency, with its higher sensitivity indicating that it is more committed to identifying all potential patients, even if it may misjudge some healthy individuals. In summary, no individual learners can simultaneously achieve the optimal values for both sensitivity and specificity.

During the secondary learner construction phase, the RFE-based stacking ensemble model combines multiple individual learners with different inductive biases, and the meta learner is able to identify samples that are incorrectly judged by one model but correctly judged by another model through learning from the outputs of the primary learners. Therefore, the stacking ensemble models could integrate the advantages of both "high specificity" model (i.e., the RFE-based XGBoost learner) and "high sensitivity" model (i.e., the RFE-based Random Forest learner) for decision-making, thus forming a more robust and generalizable decision boundary and achieving optimal performance on AUC. Nevertheless, in order to achieve high sensitivity (reduce missed diagnoses), the stacking ensemble model has to relax the criteria for determining positive cases, which inevitably leads to more false positives and lowers the specificity. Given that healthy individuals make up the majority of the dataset, the significant decrease in the specificity directly lowers the overall accuracy.

In the depression screening scenarios, high sensitivity is crucial. The RFE-based stacking SVM model has increased sensitivity from the optimal 71.05% of the individual learners to 78.71%, which is a significant clinical value improvement. In other words, the RFE-based stacking SVM model can identify approximately 7 more depression patients out of every 100 screened individuals. These patients who have been 'rescued' can receive timely intervention and treatment opportunities, avoiding the deterioration of their condition. It has immeasurable value for individuals, families, and even society. Although the accuracy and the specificity of the RFE-based stacking SVM model are not the highest, its clinical value cannot be simply denied. Its task is not to make a final diagnosis, but to efficiently screen as many high-risk individuals as possible from the population, and then hand them over to professional doctors for secondary diagnosis through interviews, scales, etc. Therefore, the RFE-based stacking SVM model could be regarded as an optimal tool for achieving the clinical goal of "maximizing the discovery of potential depression patients".

4.3 Discussion of SHAP Visualization Analysis

These SHAP-related figures together form a complete model interpretability system from global to local, from feature importance to individual predictive interpretation.

As presented in Figure 7, the global feature importance analysis illustrates the core predictors identified by the model for depression risk. Consistent with the findings in Table 1, the model assigns the most importance to general health condition (HUQ010) (mean $|SHAP| \approx 0.88$). This computationally validates the significant comorbidity between physical health and mental health in the elderly, establishing it as the primary basis for the model's risk assessment. Furthermore, the prominence of oral pain frequency (OHQ620) and marital status (DMDMARTZ) as key predictors underscores the critical roles of chronic pain as a physiological stressor and the social support system, respectively. It is noteworthy that several urinary incontinence-related features (KIQ005, KIQ044, KIQ042) rank within the top ten, revealing that functional impairment and diminished quality of life stemming from urological health issues constitute a strong, yet often overlooked, predictive signal for the geriatric depression. In summary, the feature importance extracted by the proposed data-driven model takes full considerations of biomedical, psychological,

and socio-environmental risk dimensions to form a more completed assessment framework.

The SHAP beeswarm plot in Figure 7b) offers a global interpretation of the model's decision-making mechanism, revealing complex non-linear patterns between feature values and prediction output. Specifically, (1) The worse the self-assessment results of General Health Condition (HUQ010), the higher the risk of depression. It shows a very clear and clinically intuitive monotonic positive correlation. This strong alignment with clinical consensus significantly improves the credibility of the proposed predictive model. (2) Oral Pain Frequency (OHQ620) also exhibits a distinct risk gradient: the frequency of oral pain is positively correlated with SHAP values, confirming the role of chronic physical discomfort as a physiological driver of depression risk. (3) The distribution pattern of Education Level (DMDEDUC2) reveals its protective role: low education levels (low feature values) are strongly associated with the high-risk region (positive SHAP values), while higher education levels (high feature values) are more concentrated in risk-neutral or protective regions. This suggests its positive influence potentially mediated through pathways like socioeconomic status and health literacy. (4) While some heterogeneity exists, the impact of Marital Status (DMDMARTZ) shows a clear overall trend: being married/in a partnership (blue) is significantly associated with negative SHAP values (protective effect), whereas non-married statuses (red/purple) clearly increase risk. This aligns with the theoretical model that views spousal support as a psychological resilience resource [20]. (5) For binary features (e.g., KIQ044, MCQ160A), the distinct separation of their feature values in the SHAP value space provides visual evidence that the corresponding disease states are clear sensitizing factors for increasing the depression risk.

5. Conclusion

This work proposed an interpretable stacking ensemble model with SHAP for depression prediction in the participants aged 65 and older of the NHANES database. In specific, the stacking ensemble model, with the RF model and the XGBoost model as the primary learners and SVM as the meta-learners, exhibited the best sensitivity and robustness performance. SHAP analysis revealed the five key factors for affecting the depression are general health condition, frequency of oral pain in the past year, marital status, history of mental health consultations in the past year, and frequency of urine leakage. This proposed model not only serves as a powerful tool to screen the geriatric depression, but also breaks through the "black-box" nature of the traditional model's decision-making processes, providing a reliable and transparent clinical decision-making basis. Future studies could adopt larger database to futher verify the effectiveness of the proposed model in the clinical applications.

Acknowledgments

This work was supported in part by Guangdong Medical Research Fund Project under Grant A2024156, in part by General Cultivation Project of Humanities and Social Sciences Project of Guangzhou University of Chinese Medicine in 2023 under Grant 2023YBPY10, and in part by 2024 Teaching Quality and Education Reform Project Program at Guangzhou University of Chinese Medicine (University-level).

References

[1] Yangfan Wang, Yinhuan Hu, Shaoyu Lu, Sha Liu, Xiaodong Feng, Hui Wang, Analysis and prediction on disease burden of depressive disorders among elderly population in China from 1990 to 2021, Chinese Journal of Prevention and Control of Chronic Diseases, 2025, 33(3): 178-184.

[2] Yosef Zenebe, Bay Akele, Mulugeta W/Selassie, Mogesie Necho, Prevalence and determinants of depression among old age: a systematic review and meta-analysis, Annals of General Psychiatry, 2021, 20(1):55.

[3] Ying Cheng, Yu Fang, Jinxin Zheng, Shiyang Guan, Meiti Wang, Wu Hong, The burden of depression, anxiety and schizophrenia among the older population in ageing and aged countries: an analysis of the Global Burden of Disease Study 2019, General Psychiatry, 2024, 37(1):e101078.

[4] Pengcheng Miao, Beier Lu, Rongji Ma, Yongkang Qian, Chenhua Hu, Hualing Chen, Ru Fan, Biyun Xu, Bingwei Chen, Identification of patients with senile depression by interpretable machine learning model-based on the US National Health and Nutrition Examination Survey, Modern Preventive Medicine, 2024,51(05):781-787.

[5] Hongbo Sun, Youcai Zhou, Xianqiang Zhang, Zhongxin Liang, Jinlan Chen, Ping Zhou, Xinjie Xue, Uncovering unseen ties: a network analysis explores activities of daily living limitations and depression among Chinese older adults, Frontiers in Aging Neuroscience, 2025, 17:1527774.

[6] Qian Wu, Jian Feng, Chenwei Pan, Risk factors for depression in the elderly: An umbrella review of published meta-analyses and systematic reviews, Journal of Affective Disorders, 2022. 307: 37-45.

[7] Zhuang Zhuang, Yunjing Zhang, Canyang Li, Ziwen Wang, Qiyuan Lv, The impact of self-rated health and self-care ability and their interaction on depression symptoms in the elderly, Modern Preventive Medicine, 2025, 52(03): 497-502.

[8] Heyan Xu, Dandan Geng, Yuna Wang, Yujia Chen, Lei Shi, Ning Du, Ziqiang He, Li Kuang, Association between living habits and depressive symptoms among the elderly in China, Journal of Chongqing Medical University, 2025,50(01):114-122.

[9] Minwei Gong, Jiaqi Shi, Jian Wu, Review on machine learning methods in predicting the risk of depression, Chinese Journal of Medical Physics, 2024,41(06):776-781.

[10] Yuqi Zhang, Analysis and prediction model development of factors affecting depressive symptoms in elderly individuals aged 65 and above in China, Master Dissertation, China Medical University, 2024.

[11] Hui Nie, Xiaoyan Wu, Detecting Depression Factors with Gradient Boosting Tree and Explainable Machine Learning Model SHAP. Data Analysis and Knowledge Discovery, 2024, 8(3): 41-52.

[12] Ahmed M. Salih, Zahra Raisi-Estabragh, Ilaria Boscolo

Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, Gloria Menegaz, A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Advanced Intelligent Systems, 2024, 7: 2400304.

[13] Zhen Zeng, Qiao Li, Eric D Caine, Yemisi Takwoingi, Baoliang Zhong, Yongsheng Tong, K K Cheng, Wenjie Gong, Prevalence of and optimal screening tool for postpartum depression in a community-based population in China, Journal of Affective Disorders, 2024, 348:191-199.

[14] Bruno Legendre, Damiano Cerasuolo, Olivier Dejardin, Annabel Boyer, How to deal with missing data? Multiple imputation by chained equations: recommendations and explanations for clinical practice, Nephrologie and Therapeutique, 2023, 19(3):171-179.

[15] Amadi G. Udu, Marwah T. Salman, Maryam K. Ghalati, Andrea Lecchini-Visintini, David R. Siddle, Hongbiao Dong, Emerging SMOTE and GAN variants for data augmentation in imbalance machine learning tasks: a review, IEEE ACCESS, 2025, 13: 113838-113853.

[16] Anbazhagan Mahadevan, Michael Arock, A class imbalance-aware review rating prediction using hybrid sampling and ensemble learning, Multimedia Tools and Applications, 2021, 80(5): 6911-6938.

[17] Weikuan Jia, Meili Sun, Jian Lian, Sujuan Hou, Feature dimensionality reduction: a review, Complex and Intelligent Systems, 2022, 8(3): 2663-2693.

[18] Annapoorani Selvaraj, Lakshmi Mohandoss, Enhancing Depression Detection: A Stacked Ensemble Model with Feature Selection and RF Feature Importance Analysis Using NHANES Data, Applied Sciences-Basel, 2024, 14(16):7366.

[19] Earvin S. Tio, Melissa C. Misztal, Daniel Felsky, Evidence for the biopsychosocial model of suicide: a review of whole person modeling studies using machine learning, Frontiers in Psychiatry, 2024, 14: 1294666.

[20] Yuxin Wang, Yuan Qiu, Liya Ren, Hao Jiang, Meijia Chen, Chaoqun Dong, Social support, family resilience and psychological resilience among maintenance hemodialysis patients: a longitudinal study, BMC Psychiatry, 2024, 24(1):76.