BMC Psychiatry

RESEARCH

Open Access

Predicting depression risk with machine learning models: identifying familial, personal, and dietary determinants

Yankai Dong[1], Huiping Wen[1], Chengpu Lu[1], Jinyang Li[1] and Qiang Zheng[1*]

Abstract

Background The pathogenesis of depression is highly complex, therefore, the development of predictive models using readily available clinical parameters to identify individuals at risk of adverse depressive outcomes holds significant clinical value.

Method 7108 participants from the United States National Health and Nutrition Examination Survey were collected. A total of 11 machine learning models were employed, including CatBoost, Decision Tree, Gradient Boosting Tree, LightGBM (LGB), Logistic Regression (LR), Lasso, Naive Bayes, Neural Network, Random Forest (RF), Support Vector Machine, and XGBoost, with comparisons made against the generalized linear regression model. Model performance was rigorously assessed using receiver operating characteristic (ROCs), calibration curves, and decision curves analysis. Feature importance was interpreted through Shapley Additive exPlanations to identify key influencing factors at the whole level and interpret individual heterogeneity through instance-level analysis.

Results Significant differences in overall characteristics were observed between depressed patients and healthy controls. The RF model demonstrated superior performance, followed by Lasso, XGBoost, and LGB models, which also showed relatively high predictive accuracy. The training set AUC values for the RF, Lasso, XGBoost, and LGB models were 0.998, 0.713, 0.723, and 0.804, respectively, while their corresponding test set AUC values were 0.705, 0.719, 0.714, and 0.687. Based on variable importance ranking from RF, Lasso, XGBoost, and LGB models, we identified eight key predictors: body mass index, education level, marital status, annual family income, family income to poverty ratio, trouble sleeping, composite dietary antioxidant index, and dietary inflammatory index. These variables were integrated to develop a comprehensive statistical model for predicting depression risk.

Conclusion We developed a robust predictive model for assessing depression risk, incorporating eight clinically accessible predictors. This model demonstrates reliable predictive performance for depression onset and provides valuable reference for clinical decision-making. Clinical trial number is not applicable.

Keywords Depression, Machine learning, Risk factors, Prediction model, NHANES

*Correspondence:
Qiang Zheng
smiling-qiang@163.com
[1]Modern Industrial College of Traditional Chinese Medicine and Health, Lishui University, Lishui, Zhejiang, China

Introduction

Depression, a global health crisis affecting over 300 million individuals worldwide, represents one of the most significant public health challenges [1]. Depression is clinically characterized by persistent low mood and marked reduction in interest or pleasure in daily activities. Epidemiological projections indicate that by 2030, depression is anticipated to surpass the cumulative burden of all cardiovascular diseases, emerging as the leading cause of global disability [2]. Identifying modifiable risk factors can help implement targeted interventions to alleviate depression. Nevertheless, the inherent heterogeneity in depression's pathophysiology and etiology poses substantial challenges for effective prevention and treatment approaches [3].

Previous studies have typically employed traditional statistical methods, such as regression analysis, to explore the relationship between depression and specific variables [4]. Several studies have identified factors associated with depression, including diabetes [5], volatile organic compounds [6], socioeconomic indicators [7], long work hours [8], body mass index (BMI) [9], composite dietary antioxidant index (CDAI) [10], dietary inflammatory index (DII) [11, 12], sleep disturbances [13], ratio of family income to poverty(FIRP) [14] and others.

Machine learning (ML) has developed into an effective computational approach for extracting insights from data, serving as a valuable prediction tool that widespread used in multiple engineering and medical fields. ML demonstrate higher performance in prediction tasks compared to traditional statistical methods [15]. ML algorithms have shown promise in early disease detection by predicting risks [16–18]. Researchers have applied ML algorithms to predict depression in young children [19], postpartum depression in new mothers [20], depression in stroke patients [21], and college students [22]. However, there remains a scarcity of studies using ML algorithms to predict depression risk in adults, considering factors such as diet, family, physical activity, and physiology.

Thus, this study aims to evaluate the performance of various ML algorithms and identify significant factors influencing depression risk among Americans, and compare ML algorithms with traditional statistical methods.

Methods

Study population

As a large-scale cross-sectional study, National Health and Nutrition Examination Survey (NHANES) has received ethical approval and employs scientific sampling methods to select representative samples of the U.S. population for health examinations and questionnaires. In this study, data from three consecutive cycles (2011–2012, 2013–2014, and 2015–2016) were utilized, along with available depression information. Initially, 29,902 participants were included, but the analysis was restricted to 7,108 participants based on the following exclusion criteria: (1) age < 20 years or pregnant women; (2) missing PHQ-9 questionnaire data; (3) incomplete demographic information, such as annual family incomes, education, marital status, or age; (4) unavailable or unreliable dietary information; and (5) missing height or BMI data, among others.

Dietary inflammatory index (DII)

NHANES collects dietary data through 24-hour recall interviews conducted at the Mobile Examination Center (MEC). Participants' nutritional intake was estimated using the average of two reliable 24-hour recalls or a single recall when only one was deemed reliable, excluding supplements and medications. This study calculated the DII using 23 components, including alcohol, vitamin E, beta-carotene, carbohydrates, folic acid, dietary fiber, energy, magnesium, total monounsaturated fatty acids, niacin, protein, iron, selenium, cholesterol, total saturated fatty acids, total fat, vitamin A, vitamin B12, riboflavin, vitamin B6, caffeine, vitamin C, and zinc. The DII was computed using the following equation [23]: DII=(Z score' × the inflammatory effect score of each dietary component).

$$\text{Z score} = (\text{daily mean intake} - \text{global daily mean intake}) / \text{standard deviation}$$

$$Z\ \text{score}' = \text{Z score} \rightarrow (\text{converted to a percentile score}) \times 2 - 1$$

Composite dietary antioxidant index (CDAI)

The development of the CDAI follows the methodology described previously [24, 25]. The CDAI was computed by aggregating the intake levels of six antioxidants - vitamin A, zinc, selenium, vitamin C, vitamin E, and magnesium - exclusively from dietary sources, as follows:

$$\text{CDAI} = \sum_{i=1}^{6} \frac{\text{Xi-Ui}}{\text{Si}}$$

Here, *Xi* is the antioxidant i consumed per day, and *Ui* is the average value of *Xi* in the entire cohort; *Si* was the standard deviation for *Ui*.

Metabolic dysfunction indicators

Blood pressure was measured in the MEC. Fasting blood glucose, triglyceride, total cholesterol and uric acid levels were determined by enzymatic method using an automatic analyzer. The detailed testing protocol is provided on the NHANES website (https://wwwn.cdc.gov/nchs/nhanes/index.htm). Mean arterial pressure (MAP)

is diastolic pressure + 1/3 × (systolic pressure - diastolic pressure) [26]. The triglyceride-glucose (TyG) index is calculated as ln (fasting triglyceride [mg/dl] × fasting blood glucose [mg/dl]/2). In addition, we constructed a metabolic score (MS) as the sum of the z-transformed values of the four factors of total cholesterol, uric acid, MAP and TyG [27].

Machine learning

Selection of candidate variables and predictors

Participant demographics were summarized as mean ± SD (continuous variables) or counts/percentages (categorical variables). Principal component analysis (PCA) and orthogonal partial least squares-discriminant analysis (OPLS-DA) differentiated depressed and non-depressed individuals. Feature selection was performed using the Kruskal-Wallis test (non-normal data) and chi-square test (categorical data), retaining variables with $P < 0.05$ for ML. Model performance was evaluated via receiver operating characteristic (ROC), decision curve and calibration curve, with the best-performing model selected for depression risk prediction. Shapley Additive exPlanations (SHAP) analysis identified key predictors by quantifying each feature's contribution to predictions based on game-theoretic Shapley values. Here, the model acts as a "game", and features as "players". A positive SHAP value indicates a feature increased predicted risk, while a negative value decreased it; larger absolute values denote greater importance. This approach provides both global (dataset-wide) and local (individual prediction) interpretability.

Machine learning models

The dataset was randomly partitioned into training and test sets (70%–30% split). The training set, comprising 70% of the data, was utilized for model selection and hyperparameter tuning. During the training phase, 10-fold cross-validation was employed, wherein the training set was divided into 10 subsets. In each iteration, 9 subsets were used for training while the remaining subset served as the validation set. The final model accuracy was determined by averaging the performance metrics obtained from these 10 iterations. The remaining 30% of the original dataset was allocated as the test set for model evaluation. In this study, the optimal model for each algorithm was selected based on the highest area under the ROC. 11 machine learning models were employed for training and testing, including CatBoost, Lasso, Gradient Boosting Tree (GBM), Decision Tree (DT), Naive Bayes (NB), Support Vector Machines (SVM), LightGBM (LGB), Neural Network (NN), XGBoost, Logistic Regression (LR), and Random Forest (RF). The Friedman test and paired pairwise Wilcoxon test were used to compare the performance of different machine learning models under cross-validation. Additionally, regression analysis was performed on the key features identified by machine learning to examine their associations with depression and other phenotypes.

Statistical analysis

Statistical analysis in this study was performed using EmpowerStats (http://www.EmpowerStats.com) and R (http://www.r-project.org), with a significance threshold of $P < 0.05$. In alignment with NCHS analysis guidelines, sample weights were applied to all estimates to ensure better representation of the non-institutionalized civilian population in the United States. A generalized linear regression model was utilized to calculate the odds ratio (OR) or β and 95% confidence intervals for the relationships between key features and depression or related phenotypes, with subgroup analyses also conducted.

Result

Clinical characteristics

Among 29,902 adult participants, 22,794 were excluded due to age under 20, missing PHQ-9 scores, pregnancy, incomplete demographic information, or other reasons. After further excluding missing covariate data, 7,108 participants remained for analysis (Fig. 1). The average age of the 7,108 participants was 48.125 ± 17.575 years, with 3,280 being Male. A total of 2,574 participants (36.21%) were classified as having depression. Compared to healthy participants (BMI: 29.020 ± 6.852), those with depression had a higher BMI (30.629 ± 8.016), a greater proportion of widowed (8.24%), divorced (14.69%), and separated individuals (5.28%), and a higher percentage of females (1,417, 55.05%). Depressed participants were also older (48.355 ± 17.670 years). Additionally, their energy intake (2,081.985 ± 1,044.290 kcal) was significantly lower than that of healthy individuals (2,181.748 ± 986.629 kcal). The FIRP was also significantly lower (1.989 ± 1.501), with a higher prevalence of failing kidneys (5.71%) and kidney stones (13.02%), and a lower proportion of high-income individuals (Supplementary Table 1).

Feature profiling

The PCA score plot (2 components) and wavelet power spectra for depression patients and normal participants were displayed in Fig. 2A and B, respectively. Similarly, the OPLS-DA score plot (2 Components, R2 = 0.00108, Q2=−0.00172) and wavelet power spectra were presented in Fig. 2C and D, respectively. Both PCA and OPLS-DA reveal significant differences in overall features between depression patients and normal individuals, indicating that the onset of depression was associated with multiple factors. The variable importance in projection for each feature were shown in Fig. 2E, with money spent on

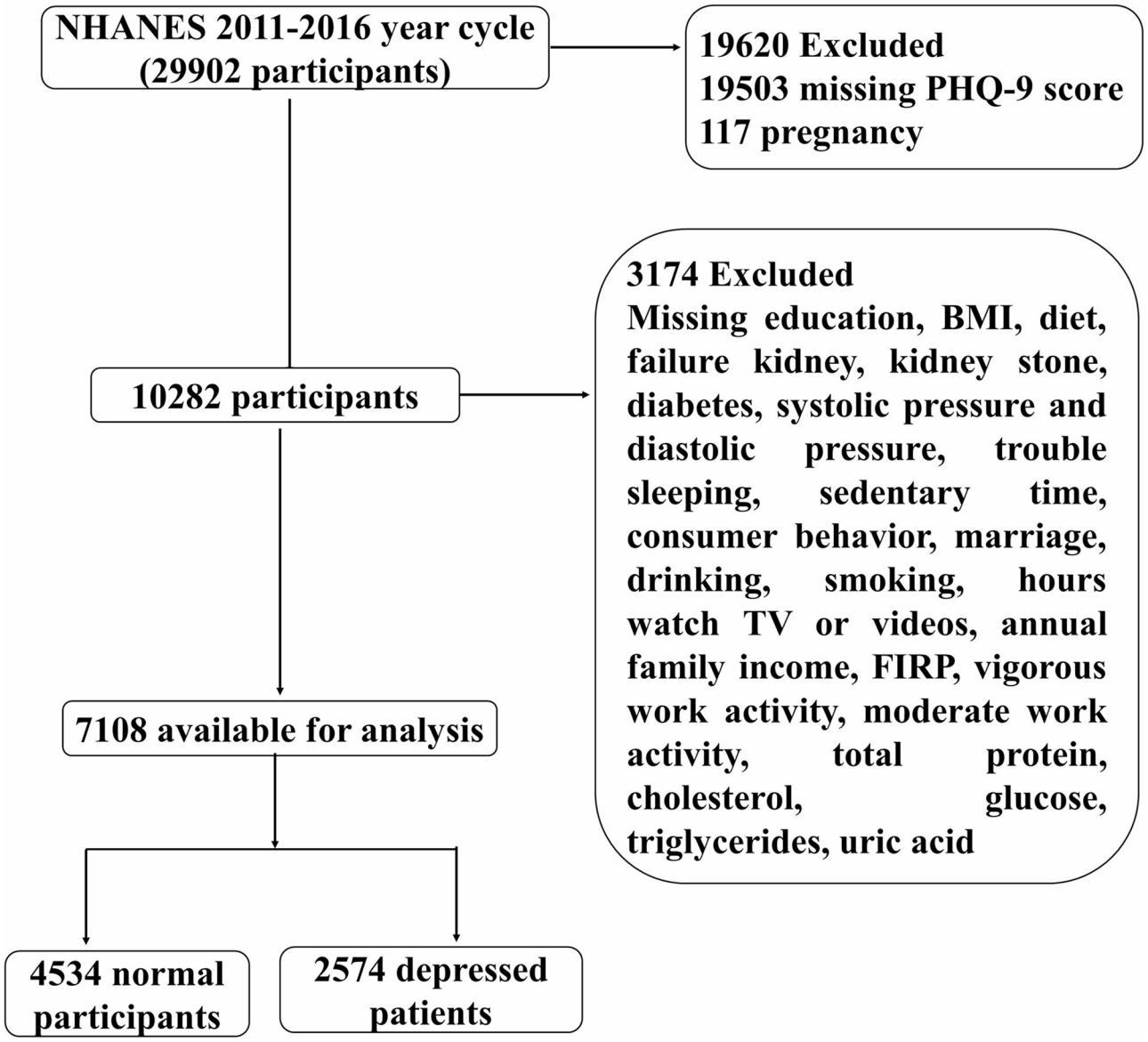

Fig. 1 NHANES participant selection flowchart

eating out, energy, money spent at supermarket/grocery store ranking relatively high.

Performance of multiple machine learning models

Utilizing features with $P < 0.05$, we performed extensive model training and evaluation. In the training set, the AUC for all models ranged from 0.574 to 0.998, and with the RF model achieving the highest performance (AUC 0.998) (Fig. 3A), and a certain degree of overfitting may be present. Calibration curves for multiple machine learning models in the training set (Fig. 3B) revealed that the XGBoost model displayed excellent calibration properties. Decision curve analysis (DCA) of the training set (Fig. 3C) indicated that both RF and SVM models yielded larger net benefit areas. For the test set, the area under the ROC ranged from 0.576 to 0.719, with LR and Lasso regression both achieving the highest performance (AUC: 0.719) (Fig. 3D). The test set calibration curves (Fig. 3E) showed optimal performance for the Lasso model, while DCA (Fig. 3F) demonstrated comparable net benefit areas across most models. The AUC for the 11 machine learning algorithms on the training set were as follows: CatBoost (0.857), DT (0.637), GBM (0.724), LGB (0.804), LR (0.713), Lasso (0.713), NB (0.716), NN (0.574), RF (0.998), SVM (0.994), and XGBoost (0.723). On the test set, the corresponding AUC values were: CatBoost (0.716), DT (0.641), GBM (0.713), LGB (0.687), LR (0.719), Lasso (0.719), NB (0.693), NN (0.578), RF (0.705), SVMe (0.673), and XGBoost (0.714). Integrated the results of ROC, DCA, and calibration curves, we selected

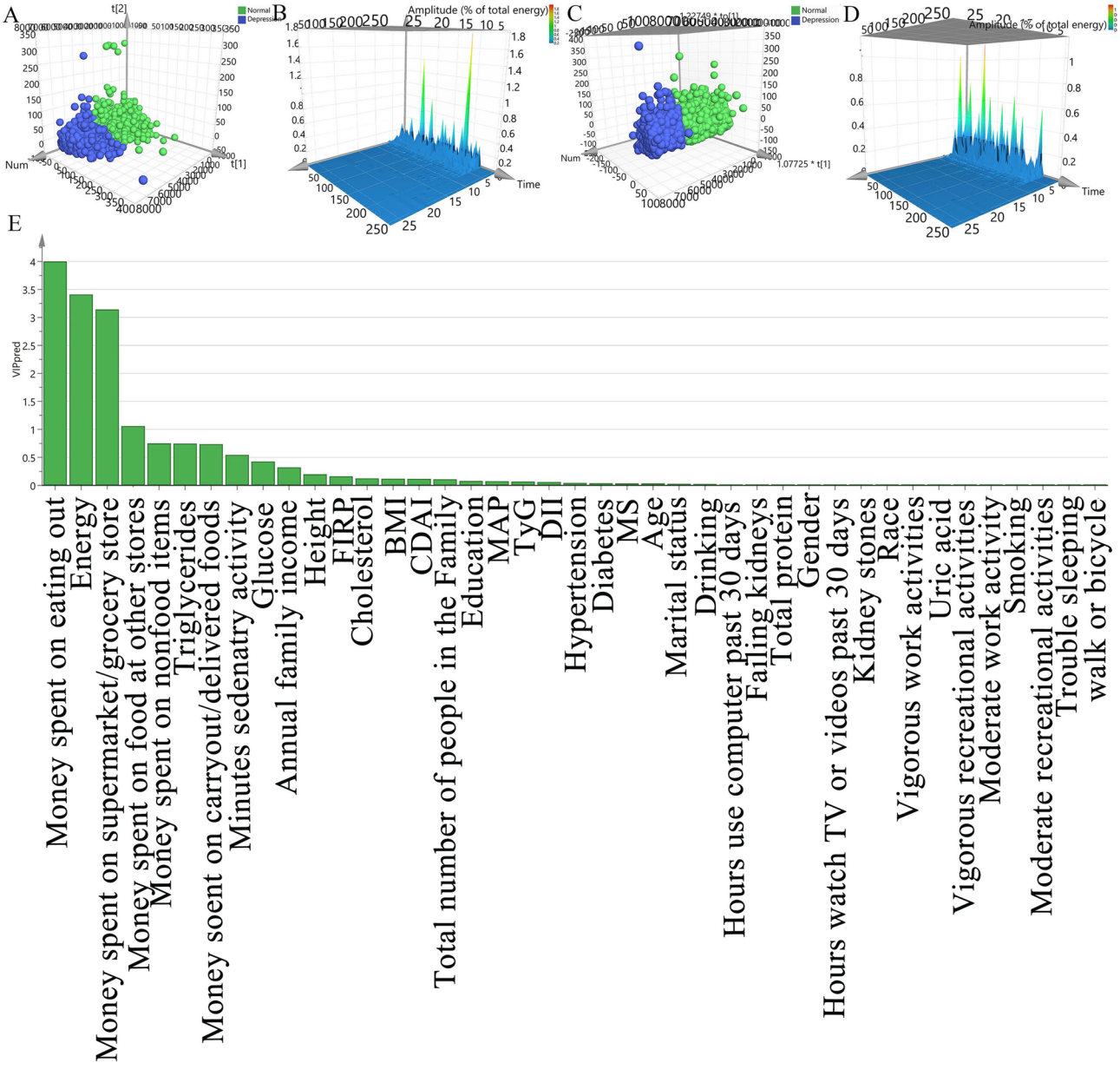

Fig. 2 Score charts of the PCA and OPLS-DA models. **A** Differences between the depression and normal groups based on the PCA model. **B** Wavelet power spectrum of the PCA model. **C** Differences between the depression and normal groups based on the OPLS-DA model. **D** Wavelet power spectrum of the OPLS-DA model. (**E**) VIP values of different features

LGB, Lasso, RF, and XGBoost as the top-performing algorithms, which were subsequently used for feature importance analysis and selection. The statistical tests were shown in Supplementary Fig. 1.

Variable importance and variable interpretation

The feature importance based on mean decrease accuracy and mean decrease gini were shown in Fig. 4A. The out-of-bag error curve indicated that the RF model is stable and yields accurate results (Fig. 4B), with the top 15 features ranked by importance scores displayed in Fig. 4C. Comprehensive analysis revealed that trouble sleeping, FIRP, annual family income, CDAI, and DII were among the most significant features.

Using the Lasso algorithm (Figs. 4D, E, F), a total of 28 features emerged as important features. The important features identified by LGB were shown in Fig. 5A, and the top 5 features were trouble sleeping, FIRP, BMI, height, marital status. As illustrated in Fig. 5B, for a specific sample, the expected value f(x) is −0.524, higher than E(f(x)) (−0.669), suggesting that this individual's depression level exceeds the average. SHAP interpretation attributes this

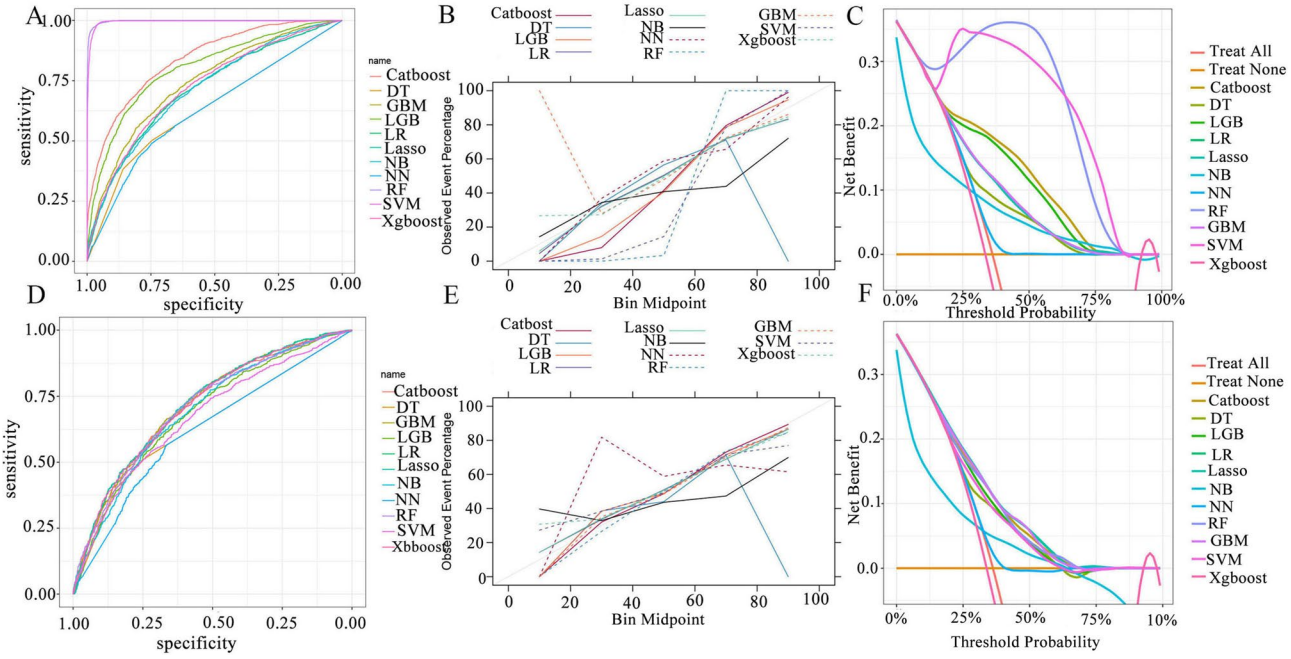

Fig. 3 Performance evaluation curves for 11 machine learning models in identifying depression. **A** ROC curve analysis in the training data. **B** Calibration curve analysis in the training data. **C** Decision curve analysis in the training data. **D** ROC curve analysis in the test data. **E** Calibration curve analysis in the test data. **F** Decision curve analysis in the test data

higher prediction primarily to variables such as annual family income, FIRP, and education. Among the 15 key features, trouble sleeping had the most significant impact, indicating that more severe sleep issues correlated with a higher risk of depression (Fig. 5C). The important features identified by XGBoost were shown in Fig. 5D, and the top 5 features were trouble sleeping, FIRP, BMI, education, smoking. In Fig. 5E, the expected value f(x) for a specific sample is 0.397, higher than E(f(x)) (0.362), indicating an above-average depression level for this individual. This is primarily influenced by annual family income, FIRP, and triglycerides. In Fig. 5F, trouble sleeping again emerged as the most significant factor affecting depression onset.

Hub features detection based on machine learning

After overlap analysis, the hub features, including BMI, education, marital status, annual family income, FIRP, trouble sleeping, CDAI, and DII were identified (Fig. 6A). To further evaluate the potential diagnostic efficacy of these features for depression, ROC curves were plotted based on the levels of each feature in the two groups. Trouble sleeping, FIRP, and annual family income exhibited relatively high AUC values of 0.6255, 0.6191, and 0.6195, respectively. These eight features were combined as a biomarker group to predict depression, and the resulting ROC curve yielded an AUC of 0.7026 (Fig. 6B), which suggests that individuals who are unmarried, experience trouble sleeping, have a high BMI, high DII, low CDAI, low annual family income, and low FIRP may face a slightly higher risk of depression.

The relationship between hub feature levels and SHAP values was illustrated in Fig. 7A. The SHAP values of BMI, CDAI, DII, and annual family income fluctuate with feature levels, while those of FIRP and education decrease as feature levels increase. Trouble sleeping and being never married were associated with relatively high SHAP values. The interaction relationships between the SHAP values of different features were shown in Fig. 7B. Interactions were observed between DII and education, CDAI and money spent on food at other stores, trouble sleeping and vigorous recreational activities, FIRP and hours use computer past 30 days, annual family income and CDAI, marital status and CDAI, education and CDAI, and BMI and money spent on food at other stores.

Association between hub features and depression

In univariate generalized linear model analysis, DII, CDAI, BMI, FIRP, marital status, education, annual family income, and trouble sleeping were significantly associated with depression (Fig. 8). Furthermore, smooth curve fitting confirmed a negative association between CDAI, FIRP, and the prevalence of depression, as well as a positive association between DII, BMI, and the prevalence of depression. Correlations between CDAI, FIRP, DII, BMI, and depression across different populations were also illustrated in Fig. 9.

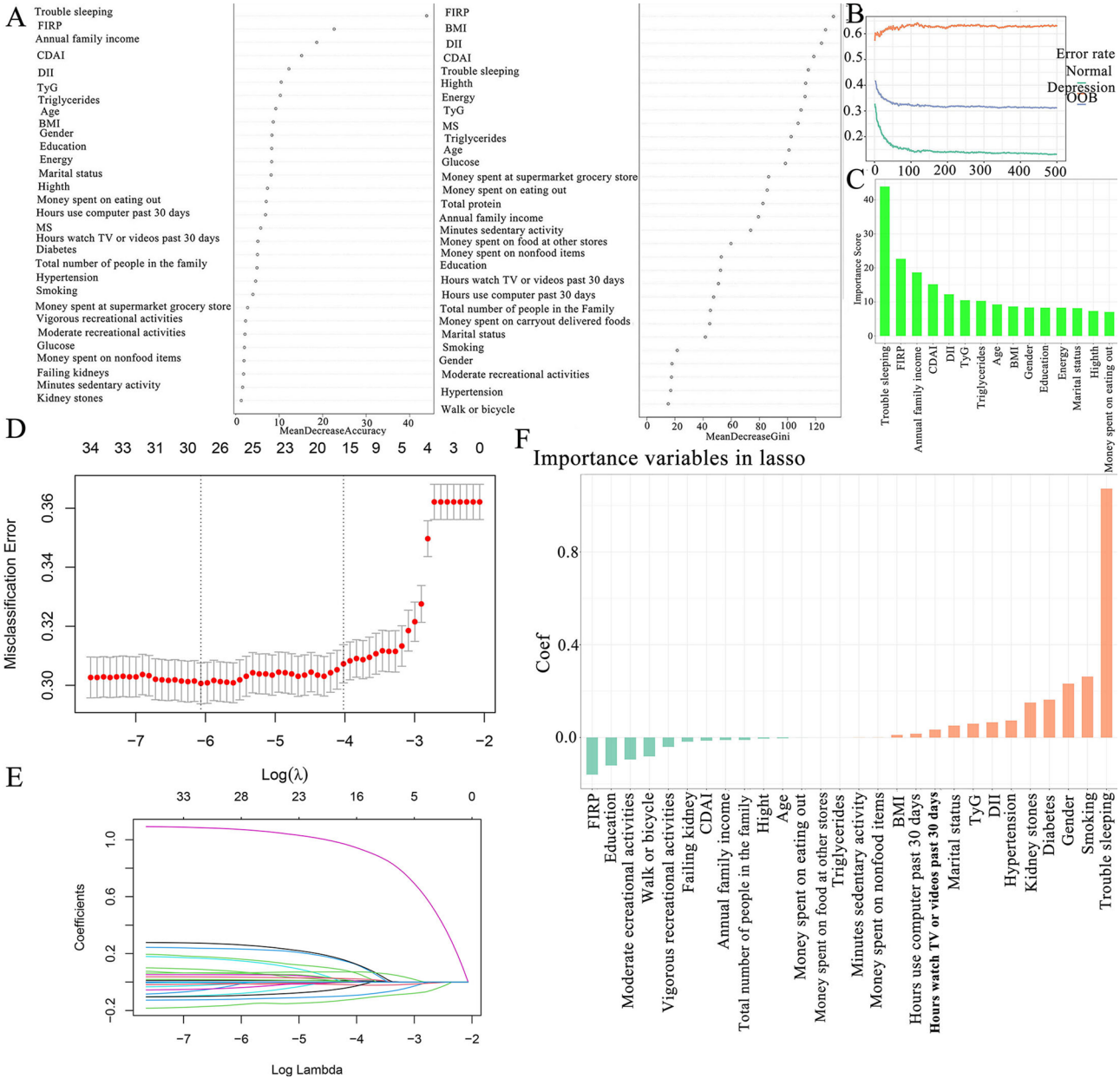

Fig. 4 RF algorithm for identifying important features. **A** Feature importance ranking based on mean decrease accuracy and mean decrease Gini. **B** Out-of-bag error curve. **C** The top 15 important features. Lasso algorithm for identifying important features. **D**-**E** Variable selection using Lasso regression. **F** Regression coefficients of important features

Previous studies had demonstrated that depressed patients often exhibited impaired energy metabolism, characterized by relatively low energy levels [28, 29], decreased MS [30], and increased TyG index [31]. Additionally, the prevalence of hypertension was relatively high among depression patients, and vice versa [32], with subsequent increases in MAP [33]. In our study, no significant difference in MAP was observed between depressed patients and normal individuals, potentially due to variations in the study population. Consequently, MAP was selected as the reference. We identified exposure variables (DII, CDAI, BMI, FIRP) that significantly impact physiological functions and conducted a generalized linear regression analysis with TyG, energy, MS, and MAP, as depicted in Fig. 10. The results reveal a highly significant negative correlation between DII and energy, and a highly significant positive correlation between CDAI and energy.

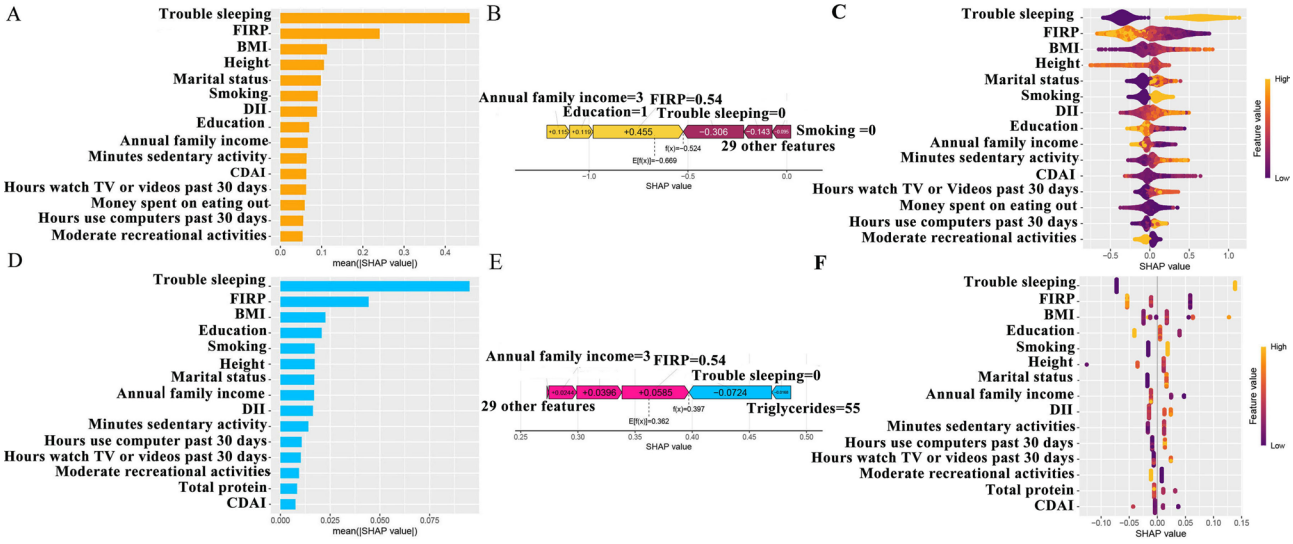

Fig. 5 LGB and XGBoost algorithms for identifying important features. **A** Bar plot of feature importance based on LGB. **B** Evaluation of SHAP values across all features for an individual instance based on LGB. **C** Beeswarm plot of feature effects based on LGB -SHAP. **D** Bar plot of feature importance based on XGBoost. **E** Evaluation of SHAP values across all features for an individual instance based on XGBoost. **F** Beeswarm plot of feature effects based on XGBoost-SHAP

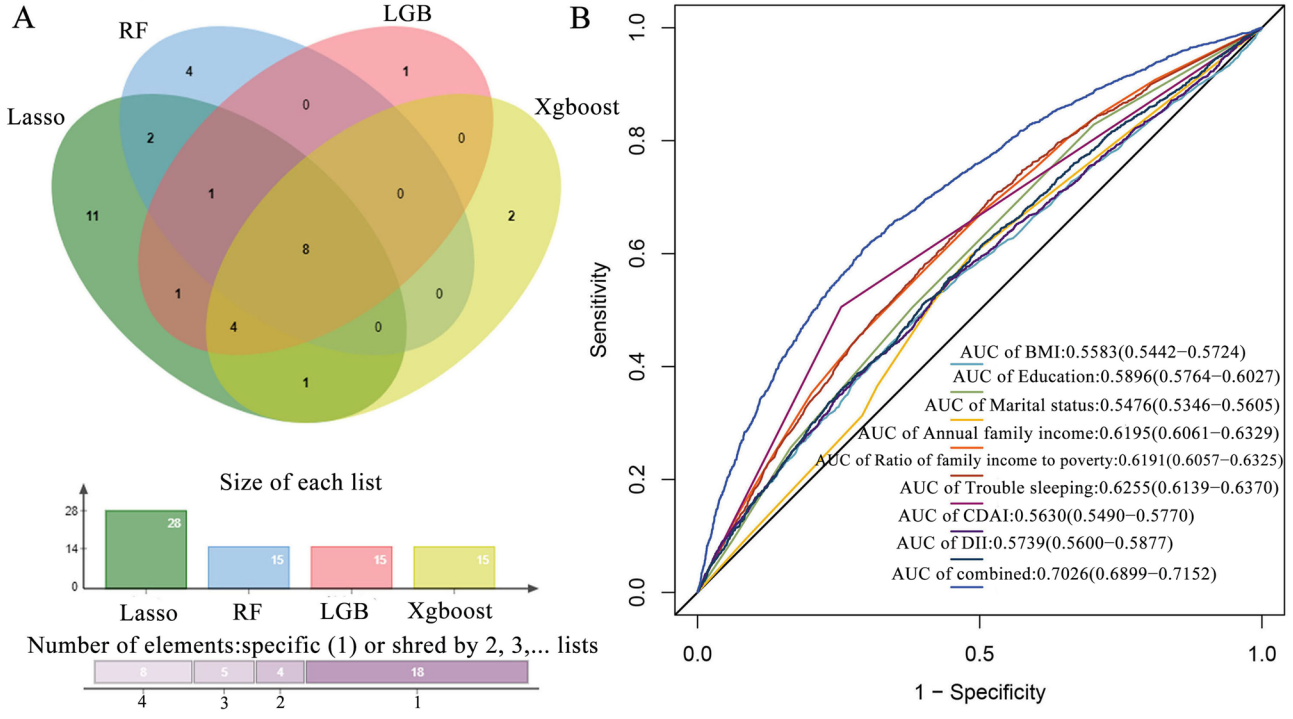

Fig. 6 Venn diagram illustrating the overlapping hub features among the top features identified by LGB, Lasso, RF, and XGBoost models (**A**), and ROC curve analysis for the hub features (**B**)

Discussion

We employed ML algorithms to effectively identify and describe key features for depression. These techniques were utilized to detect the presence of depression within the NHANES database spanning 2011 to 2016. In our comprehensive analysis, the XGBoost, LGB, Lasso, and RF algorithms demonstrated exceptional performance in handling the dataset's complexity, with RF achieving an AUC of 0.998, LGB an AUC of 0.804. The RF model may exhibit a certain degree of overfitting. A potential reason is that although RF's bagging mechanism reduces overfitting compared to individual DT, it remains highly susceptible when handling high-cardinality categorical variables. Similar observations have been reported in

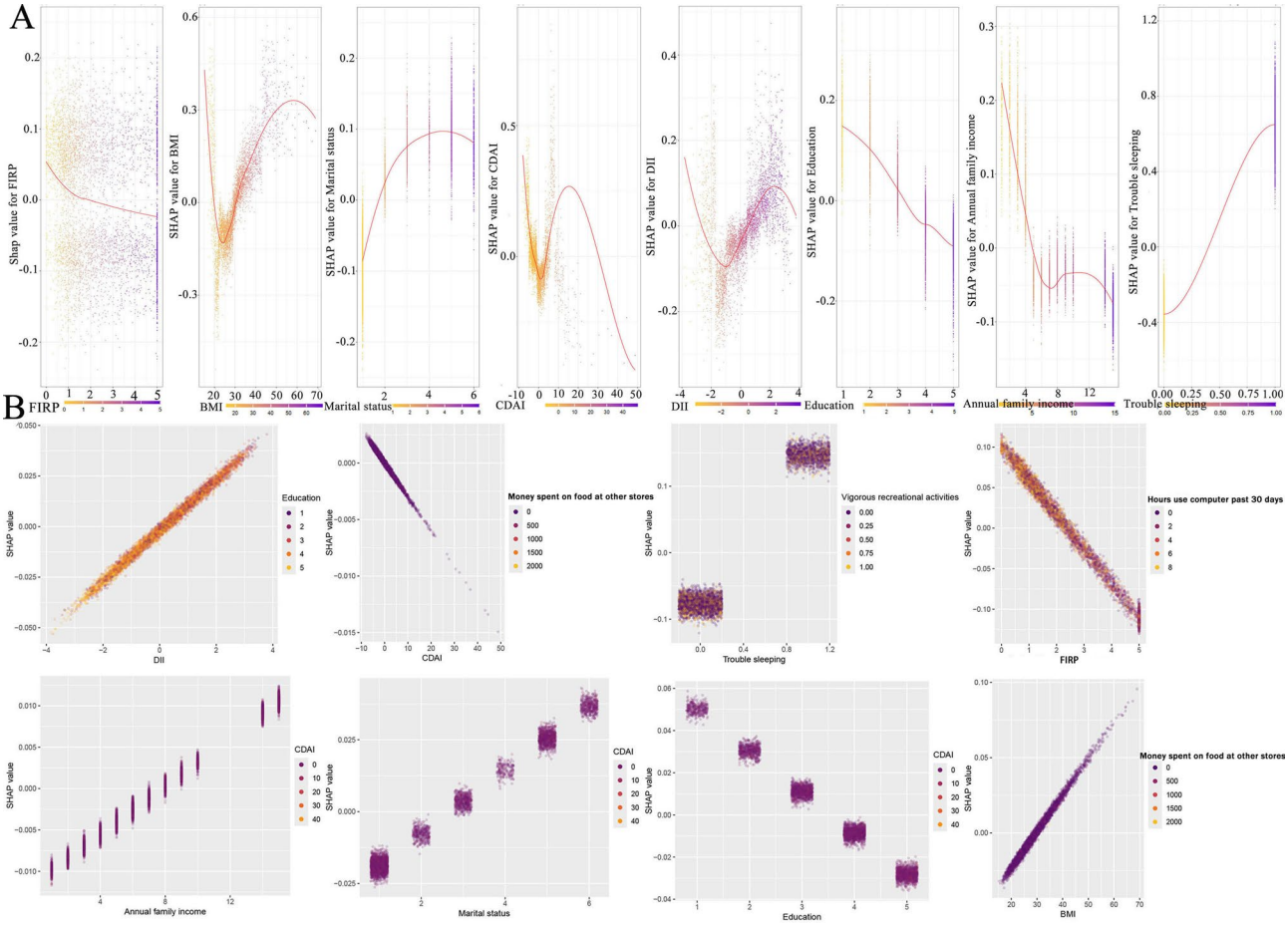

Fig. 7 The relationship between hub feature levels and SHAP values (**A**), and the interaction effect diagram of hub features (**B**)

previous studies [34, 35]. However, XGBoost achieved an AUC of 0.723, while Lasso yielded an AUC of 0.713, both below the 0.8 threshold. Potential contributing factors may include the linear constraints inherent in Lasso and possible over-regularization during XGBoost parameter tuning, both of which could have limited their expressive power. Similar limitations have been documented in previous study [36].

As a sophisticated branch of artificial intelligence, ML employs advanced mathematical algorithms to analyze and classify patterns within diverse datasets, thereby facilitating decision-making processes. This study highlights several unique advantages of ML algorithms. Firstly, it leverages questionnaire data, population data, and laboratory indicators from NHANES, integrating multi-source data into ML model, thereby eliminating the need for new data collection. Additionally, we conducted model training and evaluation on an extensive dataset, emphasizing the specificity of individual participants.

Based on the four best-performing algorithms (XGBoost, Lasso, LGB, and RF), the eight most significant features (BMI, education, marital status, annual family income, FIRP, trouble sleeping, CDAI, and DII) were identified. These findings provide profound insights into how these features influence the risk trajectory of depression, thereby improving the interpretability of our model and guiding future research directions. Among these features, DII, BMI, trouble sleeping, and living alone or being unmarried are identified as potential risk factors, while higher education levels, CDAI, higher family income, and FIRP are recognized as potential protective factors.

Previous studies have demonstrated that factors such as DII [37], CDAI [38], BMI [39], education [40], marital status [41] exhibit significant correlations with depression, although most of these studies relied on traditional statistical methods. Our study suggests that the association between DII and depression may be more pronounced in populations with high BMI, diabetes, failing kidney, hypertension, kidney stone, or marital separation, indicating that patients with diabetes, hypertension, and failing kidney should reduce consumption of pro-inflammatory foods to lower their risk of developing depression.

Pro-inflammatory diets demonstrate significant associations with adverse mental health outcomes and psychiatric disorders. Among older U.S. adults, dietary

Variables	OR (95% CI)	p−Value
DII	1.222 (1.176−1.271)	<0.001
CDAI	0.947 (0.933−0.961)	<0.001
BMI	1.030 (1.023−1.037)	<0.001
FIRP	0.769 (0.744−0.793)	<0.001
Trouble sleeping		
No	1 (reference)	
Yes	2.907 (2.444−3.458)	<0.001
Marital status		
Married	1 (reference)	
Widowed	1.600 (1.320−1.930)	<0.001
Divorced	1.920 (1.650−2.250)	<0.001
Separated	2.510 (1.950−3.240)	<0.001
Never married	1.360 (1.190−1.540)	<0.001
Living with partner	1.410 (1.180−1.690)	<0.001
Education		
Less than 9th grade	1 (reference)	
9−11th grade	1.146 (0.861−1.525)	0.356
High school graduate/GED or equivalent	0.783 (0.599−1.024)	0.082
Some college or AA degree	0.700 (0.532−0.922)	0.015
College graduate or above	0.362 (0.281−0.465)	<0.001
Annual family income		
$ 0 to $ 4,999	1 (reference)	
$ 5,000 to $ 9,999	0.939 (0.691−1.275)	0.685
10,000 to $14,999	0.831 (0.626−1.104)	0.201
$15,000 to $19,999	0.798 (0.599−1.064)	0.124
$20,000 to $24,999	0.595 (0.448−0.790)	<0.001
$25,000 to $34,999	0.504 (0.385−0.660)	<0.001
$35,000 to $44,999	0.494 (0.374−0.652)	<0.001
$45,000 to $54,999	0.409 (0.306−0.548)	<0.001
$55,000 to $64,999	0.455 (0.334−0.619)	<0.001
$65,000 to $74,999	0.367 (0.264−0.511)	<0.001
$75,000 to $99,999	0.308 (0.229−0.414)	<0.001
$100,000 and Over	0.242 (0.184−0.318)	<0.001

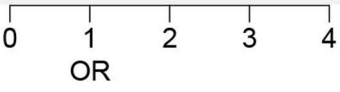

Fig. 8 Association between hub features and depression

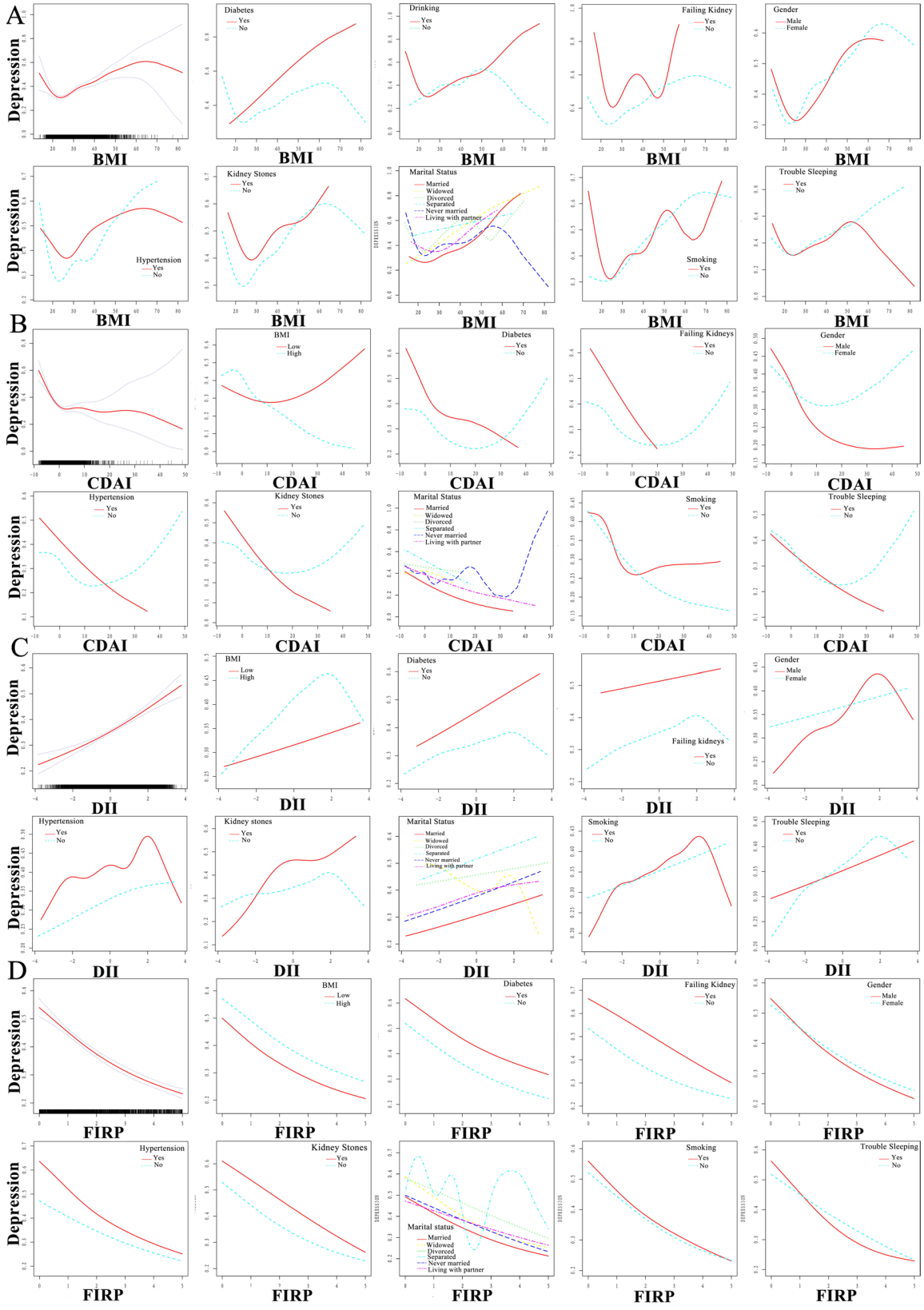

Fig. 9 Smooth curve fitting analysis. **A** Nonlinear relationship between BMI and depression. **B** Nonlinear relationship between CDAI and depression. **C** Nonlinear relationship between DII and depression. **D** Nonlinear relationship between FIRP and depression

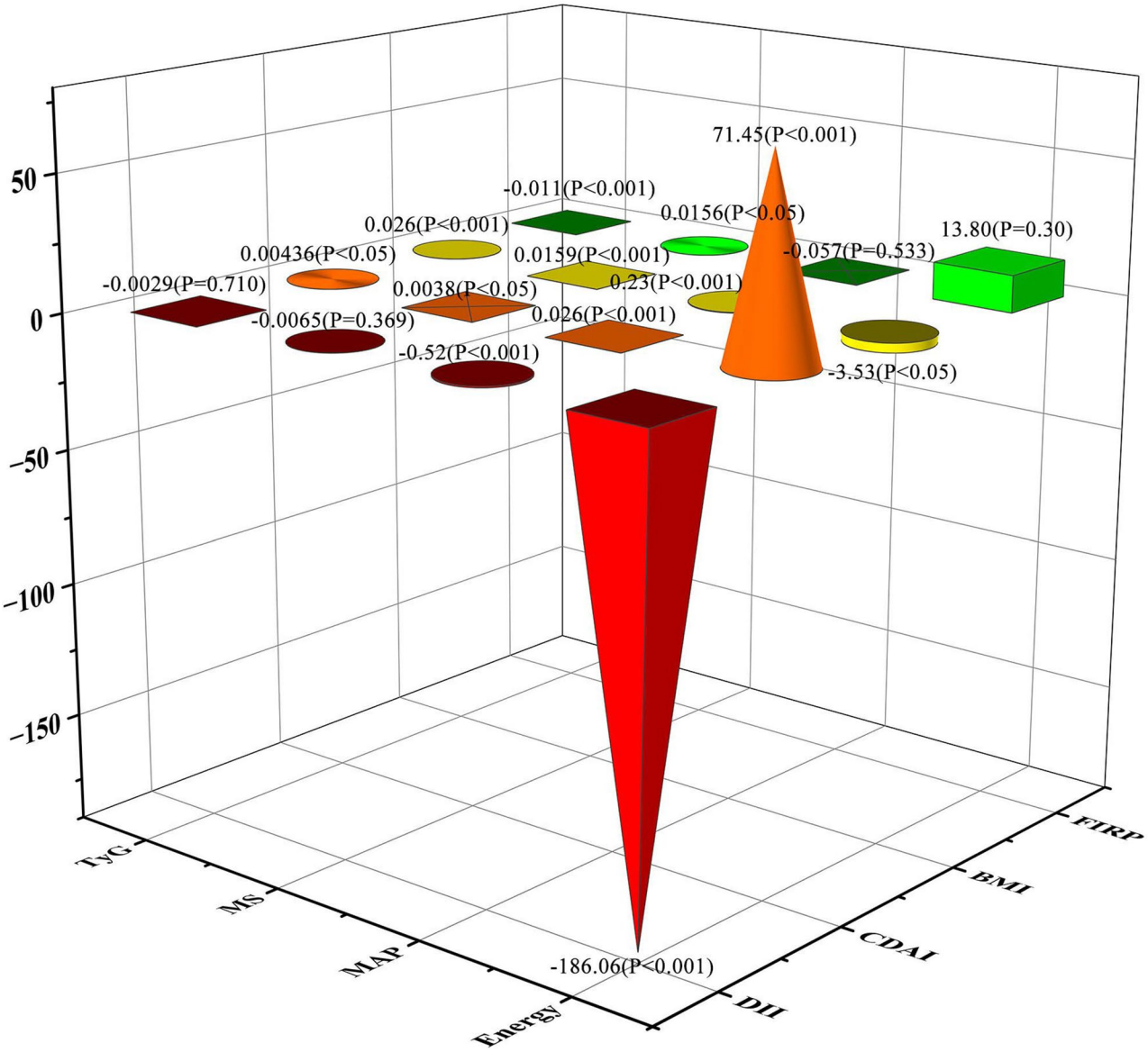

Fig. 10 The influence of hub features (DII, CDAI, BMI, FIRP) on different phenotypes (energy, MAP, MS, TyG)

inflammatory potentially shows positive correlations with increased incidence of depressive symptoms. Mechanistically, highly pro-inflammatory diets may elevate depression risk through both direct promotion of systemic inflammatory processes and indirect BMI-mediated pathways, with inflammation serving as the pivotal mediator in the obesity-depression relationship. Obesity significantly elevates circulating pro-inflammatory cytokines, and under chronic stress conditions, these elevated cytokines promote depression through multiple neurobiological pathways including disruption of neurotransmitter synthesis and impairment of neural signal transduction [42].

Based on four ML algorithms, we identified trouble sleeping as the most significant factor for depression among all evaluated variables. Sleep is an indispensable component of human life. With societal development and increasing psychological stress, sleep disturbances have become a prevalent issue in modern populations. Previous study have demonstrated a significant association between sleep problems and depression [43], while vitamin C supplementation has been shown to effectively improve sleep quality [44], suggesting that antioxidant nutrients may mitigate depression risk. This study observed that increased CDAI was associated with reduced depression likelihood in individuals with trouble sleeping. Therefore, populations experiencing trouble sleeping may benefit from dietary interventions to elevate CDAI, potentially lowering depression susceptibility.

Looking ahead, continuous monitoring and clarification of the selected features will provide valuable insights for experts, enabling them to draw well-founded conclusions rather than merely accepting algorithmic outputs. Additionally, we aim to validate the model's performance by expanding the database and improving the interpretability of the interface between clinical study and ML models. This study also has certain limitations. Firstly, there is a lack of longitudinal follow-up for the same cohort. Secondly, there is no external dataset of comparable scale for validation. We plan to address these issues in future research. Moreover, the inherent biases of cross-sectional studies, potential information biases in diagnosing depression using the PHQ-9 scale are additional limitations.

Conclusion

This study effectively employed a ML strategy to investigate the key features to predict prevalence of depression among NHANES participants from 2011 to 2016. The use of SHAP enhanced the interpretability of the model's prediction results, providing deeper insights into the factors predicting depression. By integrating advanced analytical techniques with improved interpretability, this approach addresses the typical "black box" problem in ML, enabling a more detailed exploration of the relationships between diverse features and depression.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12888-025-07182-8.

Supplementary Material 1.

Acknowledgements
We are grateful to all the staff who participated in the research.

Authors' contributions
YKD carried out the research and wrote the main manuscript. HPW, CPL, JYL, QZ edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
This work received no funding.

Data availability
The data used in this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 25 March 2025 / Accepted: 4 July 2025
Published online: 30 September 2025

References

1. van Mierlo LA, Scheffers M, Koning I. The relative relation between body satisfaction, body investment, and depression among Dutch emerging adults. J Affect Disord. 2021;278:252–8.
2. Wang JJ, et al. The influence of body investment on depression in Chinese college students: A moderated mediating effect. Int J Mental Health Promotion. 2022;24(1):39–50.
3. Nam SM, et al. Discovery of Depression-Associated Factors From a Nationwide Population-Based Survey: Epidemiological Study Using Machine Learning and Network Analysis. J Med Internet Res. 2021;23(6):e27344.
4. Kim SS, Gil M, Min EJ. Machine learning models for predicting depression in Korean young employees. Front Public Health. 2023;11:1201054.
5. Chieh A, et al. Depression prevalence, screening, and treatment in adult outpatients with type 1 and type 2 diabetes: A nationally representative cross-sectional study (National ambulatory medical care survey 2014–2019)-2019). J Affect Disord. 2025;368:471–6.
6. Ma T, et al. Expose to volatile organic compounds is associated with increased risk of depression: A cross-sectional study. J Affect Disord. 2024;363:239–48.
7. Badini I, et al. Associations between socioeconomic factors and depression in Sri lanka: the role of gene-environment interplay. J Affect Disord. 2023;340:1–9.
8. Li Z, et al. Effect of Long Working Hours on Depression and Mental Well-Being among Employees in Shanghai: The Role of Having Leisure Hobbies. Int J Environ Res Public Health. 2019;16(24):498.
9. Chen YJ, et al. Depression among Tibetan residents in the Southeastern region of Qinghai-Tibet plateau: a cross-sectional study. Sci Rep. 2025;15(1):313.
10. Zhang T, et al. The impact of Composite Dietary Antioxidant Index on the relationship between eczema and depression symptoms in US adults. Front Nutr. 2024;11:1470833.
11. Luo L, et al. The association between dietary inflammation index and depression. Front Psychiatry. 2023;14:1131802.
12. Zhang PP, et al. Dietary inflammatory index is associated with severe depression in older adults with stroke: a cross-sectional study. Br J Nutr. 2024;132(2):162–8.
13. Huang J, et al. Associations between smoking, sex steroid hormones, trouble sleeping, and depression among US adults: a cross-sectional study from NHANES (2013–2016). BMC Public Health. 2024;24(1):1541.
14. Li EG, Ai FZ, Liang CG. A machine learning model to predict the risk of depression in US adults with obstructive sleep apnea hypopnea syndrome: a cross-sectional study. Front Public Health. 2024;11:1348803.
15. Geng Z, et al. Development and validation of a machine learning-based predictive model for assessing the 90-day prognostic outcome of patients with spontaneous intracerebral hemorrhage. J Translational Med. 2024;22(1):236.
16. Dhasaradhan K, Jaichandran R. Performance analysis of machine learning algorithms in heart disease prediction. Concur Engineering-Research Appl. 2022;30(4):335–43.
17. Cha Y, Kagalwala MN, Ross J. Navigating the frontiers of machine learning in neurodegenerative disease therapeutics. Pharmaceuticals. 2024;17(2):1–14.
18. Hügle M, et al. Applied machine learning and artificial intelligence in rheumatology. Rheumatol Adv Pract. 2020;4(1):rkaa005.
19. Qasrawi R, et al. Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis. Jmir Formative Res. 2022;6(8):e32736.
20. Cellini P, et al. Machine learning in the prediction of postpartum depression: a review. J Affect Disord. 2022;309:350–7.
21. Zuo WW, Yang XL. Network-based predictive models for artificial intelligence: an interpretable application of machine learning techniques in the assessment of depression in stroke patients. BMC Geriatr. 2025;25(1):193.
22. Gil M, Kim SS, Min EJ. Machine learning models for predicting risk of depression in Korean college students: Identifying family and individual factors. Front Public Health. 2022;10:1023010.
23. Shivappa N, et al. Designing and developing a literature-derived, population-based dietary inflammatory index. Public Health Nutr. 2014;17(8):1689–96.

24. Wright ME, et al. Development of a comprehensive dietary antioxidant index and application to lung cancer risk in a cohort of male smokers. Am J Epidemiol. 2004;160(1):68–76.
25. Yu YC, et al. Composite dietary antioxidant index and the risk of colorectal cancer: findings from the Singapore Chinese health study. Int J Cancer. 2022;150(10):1599–608.
26. Zheng YY, et al. Associations of dietary inflammation index and composite dietary antioxidant index with preserved ratio impaired spirometry in US adults and the mediating roles of triglyceride-glucose index: NHANES 2007–2012. Redox Biol. 2024;76:103334.
27. Fritz J, et al. The association of excess body weight with risk of ESKD is mediated through insulin resistance, hypertension, and hyperuricemia. J Am Soc Nephrol. 2022;33(7):1377–89.
28. Jiang MR, Wang LY, Sheng H. Mitochondria in depression: The dysfunction of mitochondrial energy metabolism and quality control systems, vol. 30. Cns Neuroscience & Therapeutics; 2024.
29. Yoon SI, et al. Nutrient Inadequacy in Korean Young Adults with Depression: A Case Control Study. Nutrients. 2023;15(9):2195.
30. de la Torre-Luque A, et al. Metabolic dysregulation in older adults with depression and loneliness: The ATHLOS study. Psychoneuroendocrinology. 2021;123:104918.
31. Behnoush AH, et al. The importance of assessing the triglyceride-glucose index (TyG) in patients with depression: A systematic review. Neurosci Biobehav Rev. 2024;159:105582.
32. Alhawari H, et al. Hypertension and depression among medical students: is there an association? Heliyon. 2022;8(12):e12319.
33. Olive LS, et al. Depression, stress and vascular function from childhood to adolescence: A longitudinal investigation. Gen Hosp Psychiatry. 2020;62:6–12.
34. Hu H, et al. Identification of Potential Biomarkers for Group I Pulmonary Hypertension Based on Machine Learning and Bioinformatics Analysis. Int J Mol Sci. 2023;24(9):8050.
35. Liu FZ, et al. Identification of immune-related genes in diagnosing atherosclerosis with rheumatoid arthritis through bioinformatics analysis and machine learning. Front Immunol. 2023;14:1126647.
36. Xu Q, et al. Prediction of Atrial Fibrillation in Hospitalized Elderly Patients With Coronary Heart Disease and Type 2 Diabetes Mellitus Using Machine Learning: A Multicenter Retrospective Study. Front Public Health. 2022;10:842104.
37. Ren R, et al. Depressive symptoms mediate the association between dietary inflammatory index and sleep: A cross-sectional study of NHANES 2005–2014. J Affect Disord. 2025;372:117–25.
38. Luo JJ, et al. Association of composite dietary antioxidant index with depression and all-cause mortality in middle-aged and elderly population. Sci Rep. 2024;14(1):9809.
39. Mulugeta A, et al. Depression increases the genetic susceptibility to high body mass index: evidence from UK biobank. Depress Anxiety. 2019;36(12):1154–62.
40. Dang TNH, Sukontamarn P. Education and geriatric depression in vietnam: investigating gender differences using path analysis. Ageing Int. 2023;48(4):1204–20.
41. Sutapa P, et al. The relationship between the level of physical fitness and the level of depression in elderly age based on gender and marital status. Retos-Nuevas Tendencias En Educ Fisica Deporte Y Recreacion. 2024;53:36–44.
42. Ma YX, et al. Role of BMI in the Relationship Between Dietary Inflammatory Index and Depression: An Intermediary Analysis. Front Med. 2021;8:748788.
43. Barsha RAA, Hossain MB. Trouble Sleeping and Depression Among US Women Aged 20 to 30 Years. Preventing Chronic Disease; 2020. p. 17.
44. Wang SH, et al. Association between vitamin C in serum and trouble sleeping based on NHANES 2017–2018. Sci Rep. 2024;14(1):9727.

Publisher's Note