frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 21 February 2025
DOI 10.3389/fpubh.2025.1472050

OPEN ACCESS

EDITED BY
Lené Levy-Storms,
University of California, Los Angeles,
United States

REVIEWED BY
Lang Chen,
Santa Clara University, United States
Sung- Cho,
Seoul National University, Republic of Korea

*CORRESPONDENCE
Zhou Xiang
✉ xiangzhou@scu.edu.cn
Xin Duan
✉ dxbaal@hotmail.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 28 July 2024
ACCEPTED 10 February 2025
PUBLISHED 21 February 2025

CITATION
Chen Z, Liu H, Zhang Y, Xing F, Jiang J, Xiang Z and Duan X (2025) Identifying major depressive disorder among US adults living alone using stacked ensemble machine learning algorithms.
Front. Public Health 13:1472050.
doi: 10.3389/fpubh.2025.1472050

Identifying major depressive disorder among US adults living alone using stacked ensemble machine learning algorithms

Zhao Chen[1†], Hao Liu[1†], Yao Zhang[2†], Fei Xing[3], Jiabao Jiang[1], Zhou Xiang[1,4*] and Xin Duan[1,5*]

[1]Department of Orthopedic Surgery, Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, China, [2]Department of Emergency Management, Sichuan Center for Diseases Control and Prevention, Chengdu, China, [3]Department of Pediatric Surgery, Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, China, [4]Department of Orthopedics Surgery, West China Sanya Hospital, Sichuan University, Sanya, Hainan, China, [5]Department of Orthopedic Surgery, The Fifth People's Hospital of Sichuan Province, Chengdu, Sichuan, China

Background: It has been increasingly recognized that adults living alone have a higher likelihood of developing Major Depressive Disorder (MDD) than those living with others. However, there is still no prediction model for MDD specifically designed for adults who live alone.

Objective: This study aims to investigate the effectiveness of utilizing personal health data in combination with a stacked ensemble machine learning (SEML) technique to detect MDD among adults living alone, seeking to gain insights into the interaction between personal health data and MDD.

Methods: Our data originated from the US National Health and Nutrition Examination Survey (NHANES) spanning 2007 to 2018. We finally selected a set of 30 easily accessible variables encompassing demographic profiles, lifestyle factors, and baseline health conditions. We constructed a SEML model for MDD detection, incorporating three conventional machine learning algorithms as base models and a Neural Network (NN) as the meta-model. Furthermore, SHapley Additive exPlanations (SHAP) analysis was used to explain the impact of each predictor on MDD.

Results: The study included 2,642 adult participants who lived alone, of whom 10.6% (279 out of 2,642) had a PHQ-9 score of 10 or above, indicating the presence of MDD. The performance of our SEML model was robust, with an area under the curve (AUC) of 0.85. Further analysis using SHAP revealed positive correlations between the occurrence of MDD and factors such as sleep disorders, number of prescription medications, need for specific walking aids, leak urine during nonphysical activities, chronic bronchitis, and Healthy Eating Index (HEI) scores for sodium. Conversely, age, the Family Monthly Poverty Level Index (FMMPI), and HEI scores for added sugar showed negative correlations with MDD occurrence. Additionally, a U-shaped relationship was noted between the occurrence of MDD and both sleep duration and Body Mass Index (BMI), as well as HEI scores for dairy.

Conclusion: The study has successfully developed a predictive model for MDD, specifically tailored for adults living alone using a stacked ensemble technique, enhancing the identification of MDD and its risk factors among adults living alone.

KEYWORDS

major depressive disorder, adults living alone, stacked ensemble technique, machine learning, US NHANES

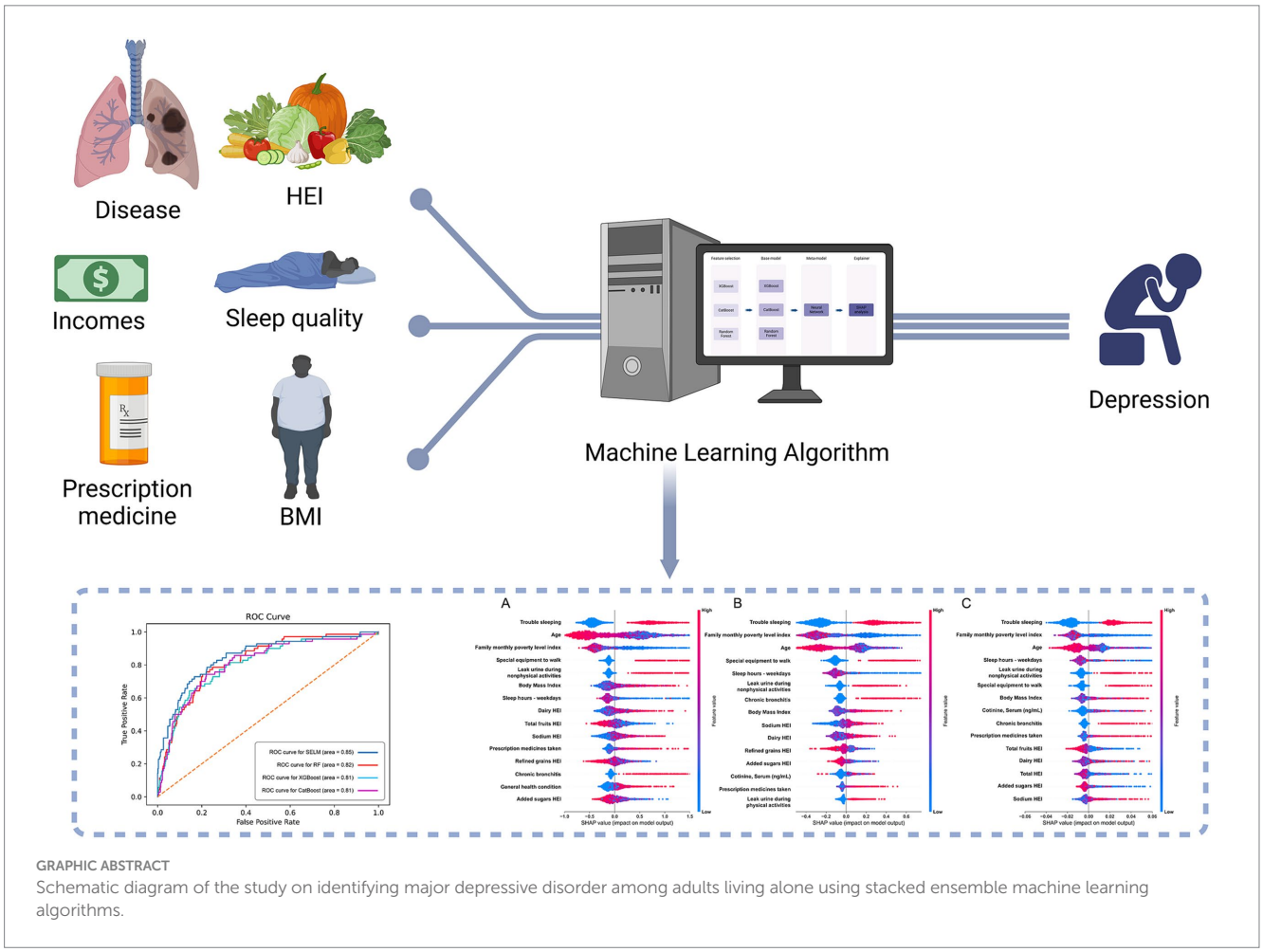
1 Introduction

The past decade has witnessed a surge in the number of individuals living alone in the United States (U.S.), with a significant increase of 4.7 million to a total of 37.9 million from 2012 to 2022. This shift has garnered considerable attention due to the well-established link between living arrangements and mental health outcomes, particularly concerning Major Depressive Disorder (MDD). Recent research findings indicate a higher prevalence of MDD among adults living alone in the U.S. (6.4%) compared to those residing with others (4.1%) (1).

Major Depressive Disorder (MDD) is a prevalent mental health conditions marked by persistent mood lows, diminished interest, and a variety of affective, cognitive, somatic, and behavioral symptoms. These symptoms can profoundly impair psychosocial functioning and greatly diminish quality of life (2). The global prevalence of MDD has steadily risen over the past three decades, affecting approximately 5% of the adult population (3). It ranks as one of the leading causes of disability worldwide, contributing significantly to the overall burden of disease (4).

The therapy of MDD poses a challenge due to its heterogeneous nature (5). Key to improving patient prognosis is early detection and intervention. Unfortunately, stigma surrounding mental health assessments, inadequate mental health resources, and the common practice of concealing symptoms complicate the timely recognition and prediction of MDD (6). Although numerous studies have developed predictive models for MDD (6–10) to enhancing the likelihood of detection, there is currently a dearth of MDD predictive models specifically tailored for adults living alone.

Machine learning has demonstrated remarkable effectiveness in medical prediction (11), and is increasingly used in medical diagnostics (12). Machine learning models can adaptively learn from data to identify complex, nonlinear patterns (13). Furthermore, these models offer high interpretability, allowing researchers to understand model behavior and how decisions are made through various visualization and explanatory techniques (14, 15). Ensemble learning is a machine learning paradigm that enhances prediction accuracy by aggregating the predictions from several base models. It reduces the risk associated with individual models by combining their opinions, typically resulting in more accurate and robust predictions (16). The most common ensemble strategies include voting, averaging, stacking, and boosting. Among them, stacking has demonstrated strong predictive ability in tackling complex problems



GRAPHIC ABSTRACT
Schematic diagram of the study on identifying major depressive disorder among adults living alone using stacked ensemble machine learning algorithms.

(17, 18). Nevertheless, to date, there is a dearth of research on the application of stacking in predicting the presence of MDD among adults living alone.

Therefore, we propose a stacked ensemble machine learning (SEML) model, utilizing data from the US National Health and Nutrition Examination Survey (NHANES), to predict MDD in adults living alone. We compared the performance of our SEML model with single machine learning models. Furthermore, we explored the interaction between predictors and the presence of MDD via the application of SHapley Additive exPlanations (SHAP) analysis. We aim to make a contribution to the field of MDD prediction in adults living alone and provide valuable insights for potential interventions and treatments.

2 Methods

2.1 Study design and participants

This cross-sectional study utilizes publicly available data from the US NHANES, an ongoing health-related initiative conducted periodically by the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). The NCHS Ethics Review Board (ERB) approved the protocol for US NHANES, and all participants provided written informed consent (19).

Our study included data from respondents surveyed between 2007 and 2018. Living arrangements were evaluated by the number of people in the household, with only one person defined as living alone. Respondents under 18 years old, living with others, or with incomplete data were excluded.

2.2 Definition of MDD

MDD was evaluated using the Patient Health Questionnaire-9 (PHQ-9), which is based on the diagnostic criteria for MDD illustrated in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The PHQ-9 has been widely recognized for its accuracy in screening for MDD (20), with a score of 10 or higher indicating clinically significant symptoms (21). Therefore, in this study, we utilized a PHQ-9 score of 10 as the threshold for defining MDD.

2.3 Selection of predictors

The initial phase of the study included 72 feature variables related to demographic profiles, lifestyle factors, and baseline health conditions. Subsequently, a rigorous feature selection process was conducted, resulting in the identification of a subset of 30 variables for constructing the final predictive model. Within the initial pool of 72 variables, demographic characteristics encompassed participants' age, gender, race, marital status, citizenship status, educational level, employment status, and income level. Lifestyle factors encompassed dietary quality, physical activity level, sleep patterns, and smoking habits. Dietary quality was assessed using the Healthy Eating Index-2020 (HEI-2020) (22), following the methodology proposed by Zhan et al. (23). Physical activity level was quantified by calculating participants' weekly metabolic equivalent task (MET) minutes, derived from multiplying the MET values of specific physical activities by their respective weekly frequency and duration. The determination of the remaining variables relied on participants' detailed responses to interview questions. Baseline health status was assessed based on participants' medication usage and the presence of various diseases.

2.4 Statistical analysis

In our study, the raw data from the US NHANES database was used to construct machine learning models. Continuous variables were presented as means with standard deviations (SD), while categorical variables were displayed as counts and percentages. The statistical significance of differences in continuous and categorical variables was evaluated using independent t-tests and Chi-square tests, respectively, with a significance threshold set at a two-sided p-value of less than 0.05.

2.5 Model construction

In our study, a stratified sampling method was adopted to ensure an even distribution of MDD cases across different groups. The dataset was split into a training set, comprising 80% of the participants, and a test set, comprising the remaining 20%. To create the SEML model, we incorporated three algorithms as base models, including eXtreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and Random Forest (RF). Additionally, a Neural Network (NN) was designated as the meta-model (Figure 1). Both XGBoost and CatBoost belong to the Gradient Boosting Decision Trees (GBDT) algorithm, which employs ensemble learning to combine multiple decision trees using an additive model for predictions. RF represents another ensemble learning approach that enhances predictive power through voting or averaging the results of several decision trees. NN, a deep learning technique, process data by simulating the connections between biological neurons, making them adept at capturing complex nonlinear relationships. In this study, we integrated the aforementioned methods through SEML technique to fully leverage the strengths of each model and improve overall predictive performance.

Firstly, we employed the base models to select the top 30 variables in terms of importance out of the initial set of 72 variables. The feature importance was calculated by normalizing and summing the importance values from three different models. Both 10-fold cross-validation and Bayesian optimization were utilized by us for hyperparameter tuning and model evaluation (Details of the model hyperparameter settings are provided in the Supplementary materials). To enhance the predictive accuracy of the models, we performed normalization on the dataset. Additionally, the technique of Synthetic Minority Over-sampling Technique (SMOTE) was employed to mitigate class imbalance between MDD and non-MDD instances in the training data.

During the training phase, we utilized the predictions generated by each base model as input features for training the meta-model. When it came to the testing phase, the base models that had been trained were used to predict the test set. These predictions were then incorporated as input features to make the final prediction using the meta-model.

The assessment of model performance in this study employed the area under the Receiver Operating Characteristic (ROC) curve as the

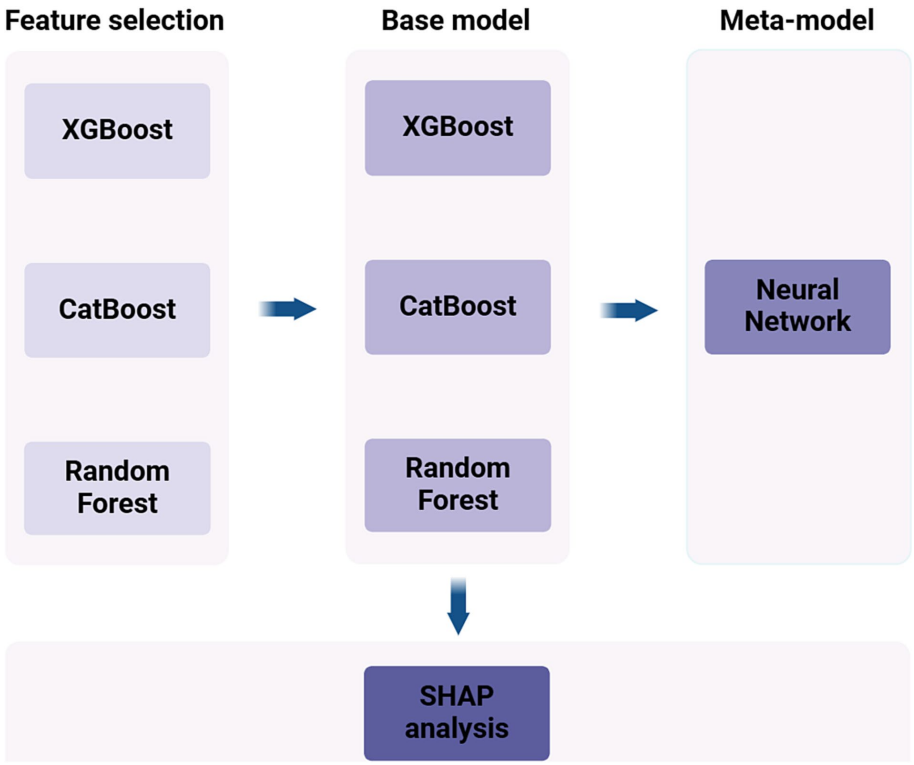


FIGURE 1
Flowchart of machine learning model construction.

primary metric. Sensitivity, specificity, Youden index, and F1-score were also calculated at the optimal threshold to provide a comprehensive assessment. Furthermore, to elucidate the individual variables' contributions and their importance in the model's predictions, SHapley Additive exPlanations (SHAP) analysis was conducted.

All procedures were implemented in Python version 3.12.4.[1]

3 Study results and findings

3.1 Participant inclusion

38,788 respondents were initially selected from the 2007–2018 US NHANES data. 9,429 participants under the age of 18 were excluded from the study. A further 26,717 participants were excluded due to incomplete data or residing with others. Finally, a total of 2,642 participants were included in the study (Figure 2).

3.2 General characteristics

Among these participants, 10.6% (279/2,642) of them had a PHQ-9 score of 10 or higher. No significant differences were observed in age ($p = 0.2$), marital status ($p = 0.287$), race ($p = 0.644$) and citizenship status ($p = 0.315$) between participants with and without MDD. Notable differences were found in gender ($p < 0.001$), education level ($p < 0.001$), employment status ($p < 0.001$) and family monthly poverty level index ($p < 0.001$; Table 1).

3.3 Included predictors

The final set of 30 feature variables integrated into the predictive model could be categorized as follows: demographic characteristics, including DMDMARTL, RIDAGEYR, IND235, and INDFMMPI; lifestyle factors, including HEI2020_ALL, HEI2020_DAIRY, HEI2020_ADDEDSUGAR, HEI2020_FATTYACID, HEI2020_FRT, HEI2020_GREENNBEAN, HEI2020_SODIUM, HEI2020_REFINEDGRAIN, HEI2020_SATFAT, HEI2020_SEAPLANTPRO, HEI2020_TOTALFRT, HEI2020_TOTALPRO, HEI2020_VEG and HEI2020_WHOLEGRAIN, LBXCOT, PAD680, TPA, SLD012, and SLQ050; Baseline health status, which encompassed BMXBMI, RXDDAYS, RXDCOUNT, PFQ054, MCQ160k, KIQ046, and HSD010. For more detailed information on each feature, please refer to Table 2.

3.4 Performance of machine learning models

The ROC curve analysis revealed that SEML model performed the best among the evaluated models, as evidenced by an AUC value of 0.85 (95% CI 0.84–0.88), Followed by the RF model, which achieved an AUC of 0.82 (95% CI 0.80–0.83), while the AUC values for the CatBoost and XGBoost models were 0.81 (95% CI 0.78–0.84) and 0.81 (95% CI 0.77–0.85), respectively (Figure 3). Moreover, when considering the optimal

1 https://www.python.org/downloads/release/python-3124/

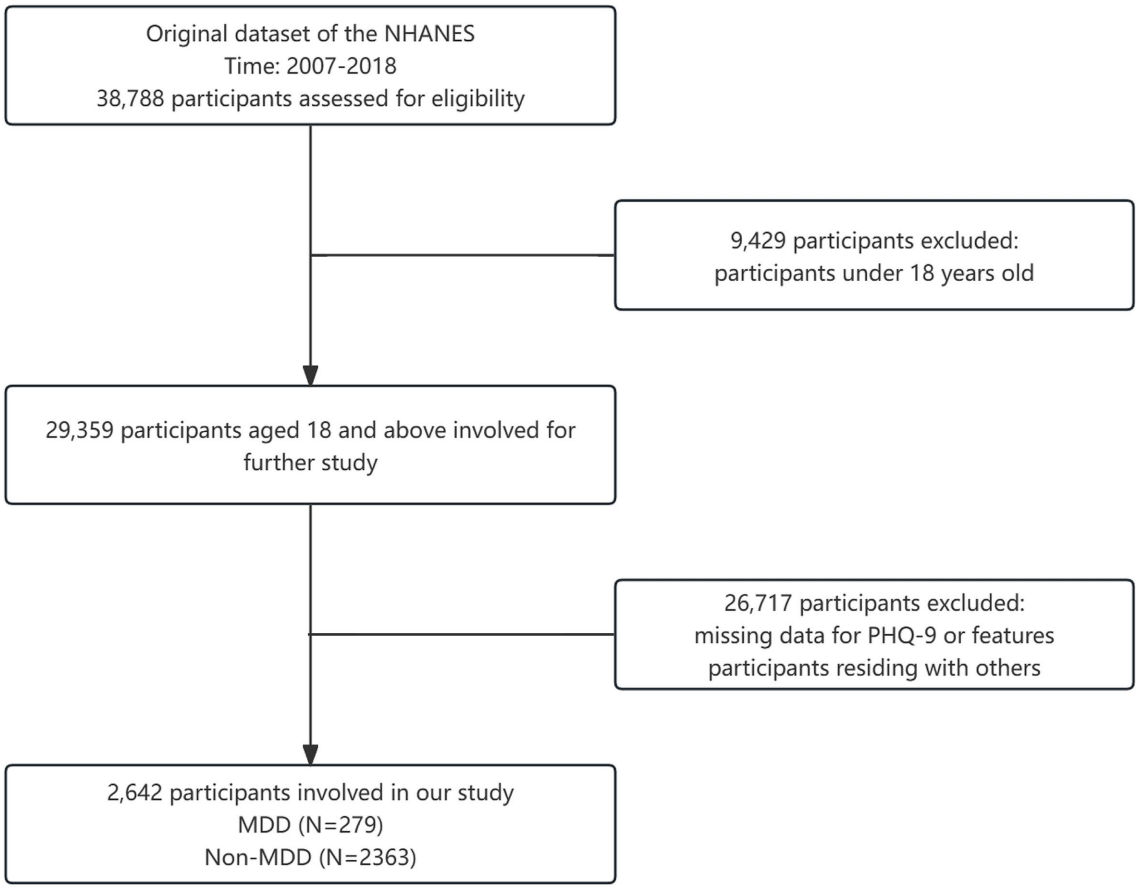


FIGURE 2
Flowchart of participants inclusion.

threshold, the SEML model exhibited superior performance in terms of sensitivity, Youden index, and F1 score, with values of 0.79, 0.57, and 0.50, respectively. Furthermore, the SEML model demonstrated a specificity of 0.78, which was marginally lower than the specificity of the XGBoost model, recorded at 0.86. Detailed results were recorded in Table 3.

3.5 Importance of predictive features

In this study, we utilized SHAP analysis to determine the importance ranking of the features included in the base models. We summarized the top 15 features in terms of importance in the base models, and the detailed analysis results are presented in Figure 4. Across all three base models, the top 15 features consistently included SLQ050, SLD012, RIDAGEYR, INDFMMPI, BMXBMI, PFQ054, KIQ046, MCQ160K, HEI2020_SODIUM, HEI2020_DAIRY, and HEI2020_ADDSUGAR, amounting to a total of 12 features. Please refer to Table 3 for detailed interpretation regarding the code.

3.6 Features impact on MDD

By conducting SHAP analysis, we had effectively delineated the positive and negative impacts exerted by each feature on the occurrence of MDD. As illustrated in Figure 4, a set of features, namely SLQ050, PFQ054, KIQ046, BMXBMI, RXDCOUNT, HEI2020_DAIRY, HEI2020_SODIUM, MCQ160K, and HSD010 exhibited a positive correlation with the occurrence of MDD, whereas features such as RIDAGEYR, INDFMMPI, and HEI2020_ADDEDSUGAR manifested a negative correlation. Subsequent SHAP dependency analysis further corroborated these findings, as depicted in Figures 5, 6. Specifically, an augmentation in the values of features such as RXDCOUNT and HEI2020_SODIUM was concomitant with an escalated risk of developing MDD. Conversely, an escalation in the values of RIDAGEYR, INDFMMPI, and HEI2020_ADDEDSUGAR was linked to a diminished risk of MDD. Additionally, the dependence plot of SLQ012, BMXBMI, and HEI2020_DAIRY revealed a U-shaped relationship between their feature values and the occurrence of MDD. These findings aligned with the data presented in the SHAP summary plots, which depicted the influence of these features on MDD occurrence as a relatively complex interplay of both positive and negative impacts.

4 Discussion

Despite the existence of numerous studies utilizing machine learning methods to predict the occurrence of MDD (24–26), there

10.3389/fpubh.2025.1472050

TABLE 1 Baseline characteristics according to depression or not.

	Non-depression (*N* = 2,363)	Depression (*N* = 279)	Total (*N* = 2,642)	*p*-value
Gender				
Male	1,057 (44.7%)	113 (40.5%)	1,170 (44.3%)	<0.001
Female	1,306 (55.3%)	166 (59.5%)	1,472 (55.7%)	
Age (year)				
Mean (SD)	60.2 (16.5)	56.4 (14.9)	59.8 (16.4)	0.2
Median [Min, Max]	63.0 [20.0, 80.0]	59.0 [20.0, 80.0]	63.0 [20.0, 80.0]	
Education level				
Less than 9th grade	133 (5.6%)	26 (9.3%)	159 (6.0%)	<0.001
9-11th grade	268 (11.3%)	50 (17.9%)	318 (12.0%)	
High school graduate/GED or equivalent	572 (24.2%)	59 (21.1%)	631 (23.9%)	
Some college or AA degree	735 (31.1%)	95 (34.1%)	830 (31.4%)	
College graduate or above	655 (27.7%)	49 (17.6%)	704 (26.6%)	
Marital status				
Married or Living with partner	75 (3.2%)	8 (2.9%)	83 (3.1%)	0.287
Widowed, divorced or separated	1,619 (68.5%)	204 (73.1%)	1823 (69.0%)	
Never married	669 (28.3%)	67 (24.0%)	736 (27.9%)	
Race				
Mexican American	209 (8.8%)	28 (10.0%)	237 (9.0%)	0.644
Other Hispanic	189 (8.0%)	27 (9.7%)	216 (8.2%)	
Non-Hispanic White	1,192 (50.4%)	135 (48.4%)	1,327 (50.2%)	
Non-Hispanic Black	598 (25.3%)	73 (26.2%)	671 (25.4%)	
Other Race - Including Multi-Racial	175 (7.4%)	16 (5.7%)	191 (7.2%)	
Citizenship status				
Citizen by birth or naturalization	145 (6.1%)	22 (7.9%)	167 (6.3%)	0.315
Not a citizen of the US	2,218 (93.9%)	257 (92.1%)	2,475 (93.7%)	
Employment status				
Employed	1,027 (43.5%)	81 (29.0%)	1,108 (41.9%)	<0.001
Unemployed	1,336 (56.5%)	198 (71.0%)	1,534 (58.1%)	
Family monthly poverty level index				
Mean (SD)	2.27 (1.51)	1.58 (1.34)	2.20 (1.51)	<0.001
Median [Min, Max]	1.81 [0, 5.00]	1.11 [0, 5.00]	1.69 [0, 5.00]	

remains a lack of research specifically focused on predicting MDD among adults living alone. Furthermore, many prediction models rely heavily on comprehensive clinical evaluations and laboratory data (9, 27), which limits their applicability. To address these limitations, this study is the first to focus on the population of adults living alone, utilizing easily accessible predictive variables from the US NHANES database to construct an MDD prediction model. This model incorporates 30 features related to demographic characteristics, lifestyle factors, and baseline health conditions. By employing a stacked ensemble technique, the model achieved an AUC value of 0.85, providing a new tool for predicting MDD in adults living alone.

To elucidate the contribution of each variable in the model's predictions, we conducted SHAP analysis, a method based on game theory. The core idea is to calculate Shapley values to quantify each feature's contribution to the prediction outcome, thereby enhancing the model's transparency and interpretability. This helps researchers and decision-makers better understand the impact of

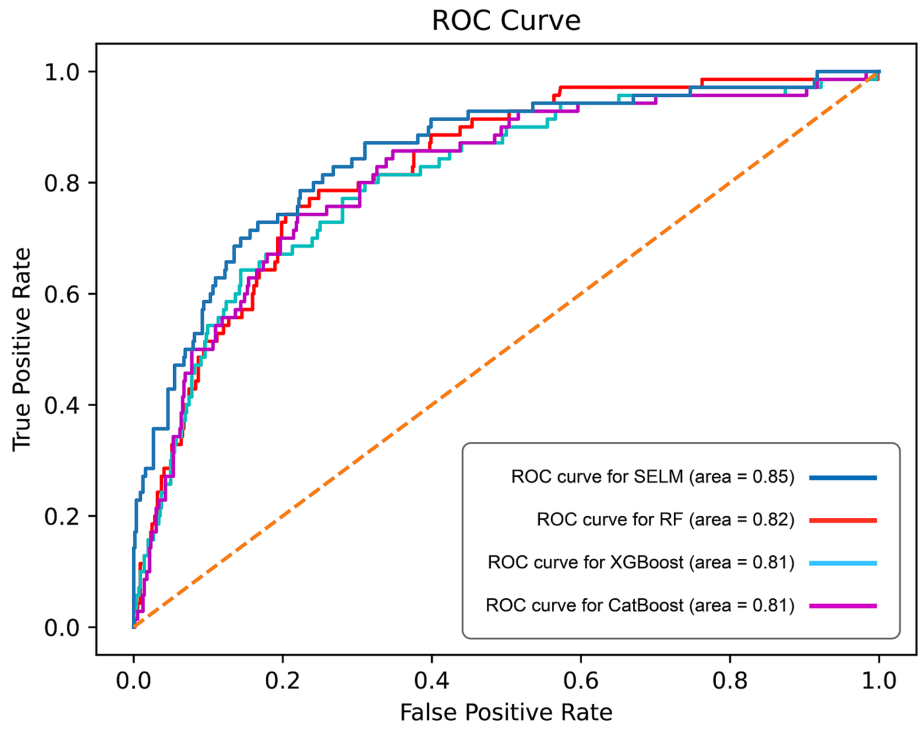


FIGURE 3
Curves of receiver operating characteristic (ROC) of different algorithms in test set. ROC, receiver operating characteristic; AUC, Area Under the Curve.

different factors on the prediction results. Provides valuable insights for potential intervention and treatment strategies for MDD in adults living alone (28).

Compared to previous studies focused on the general adult population, our research identifies that, in addition to common influencing factors such as age, sleep, and income, daily sodium intake, added sugars, and dairy consumption are also significant factors affecting MDD in adults living alone. Furthermore, physical health factors such as mobility issues, urinary incontinence during non-physical activities, and chronic bronchitis also have important impacts on MDD (2, 27).

Firstly, within the demographic variables incorporated into our final predictive model, a significant correlation was observed between age and MDD occurrence. The SHAP dependence curve demonstrated a progressive decline in MDD risk with advancing age as participants aged 50 and above. Subsequently, as participants reached the age range of 60 to 65, we observed a shift in SHAP values from positive to negative, suggesting a potential protective effect of advanced age on the occurrence of MDD. This observation aligns with previous research by Villamil et al., who reported a significant reduction in MDD prevalence among women over 60 and men over 65 (29). Based on these findings, we can infer that individuals under 60 living alone face a higher risk of MDD compared to their older counterparts. Moreover, the Family Monthly Mean Poverty Index (FMMPI) of the participants emerged as a significant predictive feature. The SHAP dependence curve revealed a progressive decline in MDD risk as FMMPI increased within the range of 0 to 2. Previous studies have consistently demonstrated that lower income levels are typically associated with a heightened risk of MDD (30–32).Our findings further support this perspective, underscoring that adults living alone with a FMMPI below 2 are more vulnerable to MDD.

Regarding lifestyle characteristics, our findings underscore that both sleep quality and dietary quality are pivotal predictors of the occurrence of MDD. Prior study by Zhang et al. has identified sleep disorders as risk factors for secondary depression (33). Additionally, research by Baglioni et al. has shown that individuals with insomnia face a doubled risk of developing depression compared to those without sleep problems (34). Our current analysis employing SHAP reveals a positive correlation between sleep disorders and MDD occurrence, highlighting sleep disorders as significant risk factors for MDD among adults living alone. Another feature reflecting participants' sleep status is sleep duration. In this study, the SHAP dependence curves displayed a U-shaped relationship between sleep duration and MDD occurrence. Notably, the lowest SHAP values are observed within a sleep duration of 7–8 h. This finding aligns with the conclusions drawn by Zhang et al., that individuals with an 8-h sleep duration exhibit the lowest risk of depression (35). Consequently, modifying sleep duration among adults living alone may potentially yield a substantial reduction in the risk of MDD occurrence. Regarding dietary quality, we employed the HEI scores to quantify the intake of various components. Through SHAP analysis, significant correlations were identified between the HEI scores for added sugar, dairy, and sodium components and the occurrence of MDD among adults living alone. Firstly, an inverse association was observed with the occurrence of MDD when HEI scores for added sugar exceeded 5. This finding aligns

TABLE 2 Explanation of the predictors used in this study.

Code	Variable type	Label	Involvement/Exclusion
BMXBMI	Continuous variables	Body Mass Index (kg/m**2)	Involvement
BPQ020	Categorical variables	Ever told you had high blood pressure	Exclusion
DIQ010	Categorical variables	Doctor told you have diabetes	Exclusion
DMDCITZN	Categorical variables	Citizenship status	Exclusion
DMDEDUC2	Categorical variables	Education level - Adults 20+	Exclusion
DMDMARTL	Categorical variables	Marital status	Involvement
HEI2020_ADDEDSUGAR	Continuous variables	Added sugars HEI	Involvement
HEI2020_ALL	Continuous variables	Total Healthy Eating Index (HEI)	Involvement
HEI2020_DAIRY	Continuous variables	Dairy HEI	Involvement
HEI2020_FATTYACID	Continuous variables	Fatty acids HEI	Involvement
HEI2020_FRT	Continuous variables	Whole fruits HEI	Involvement
HEI2020_GREENNBEAN	Continuous variables	Greens and beans HEI	Involvement
HEI2020_REFINEDGRAIN	Continuous variables	Refined grains HEI	Involvement
HEI2020_SATFAT	Continuous variables	Saturated fats HEI	Involvement
HEI2020_SEAPLANTPRO	Continuous variables	Seafood and plant proteins HEI	Involvement
HEI2020_SODIUM	Continuous variables	Sodium HEI	Involvement
HEI2020_TOTALFRT	Continuous variables	Total fruits HEI	Involvement
HEI2020_TOTALPRO	Continuous variables	Total protein foods HEI	Involvement
HEI2020_VEG	Continuous variables	Total vegetables HEI	Involvement
HEI2020_WHOLEGRAIN	Continuous variables	Whole grains HEI	Involvement
HSD010	Categorical variables	General health condition	Involvement
HSQ500	Categorical variables	SP have head cold or chest cold	Exclusion
HSQ510	Categorical variables	SP have stomach or intestinal illness?	Exclusion
HSQ520	Categorical variables	SP have flu, pneumonia, ear infection?	Exclusion
HSQ571	Categorical variables	SP donated blood in past 12 months?	Exclusion
IND235	Categorical variables	Monthly family income	Involvement
INDFMMPC	Categorical variables	Family monthly poverty level category	Exclusion
INDFMMPI	Continuous variables	Family monthly poverty level index	Involvement
INQ012	Categorical variables	Income from self-employment	Exclusion
INQ020	Categorical variables	Income from wages/salaries	Exclusion
INQ030	Categorical variables	Income from Social Security or RR	Exclusion
INQ060	Categorical variables	Income from other disability pension	Exclusion
INQ080	Categorical variables	Income from retirement/survivor pension	Exclusion
INQ090	Categorical variables	Income from Supplemental Security Income	Exclusion
INQ132	Categorical variables	Income from state/county cash assistance	Exclusion
INQ140	Categorical variables	Income from interest/dividends or rental	Exclusion

(Continued)

TABLE 2 (Continued)

Code	Variable type	Label	Involvement/Exclusion
INQ150	Categorical variables	Income from other sources	Exclusion
KIQ022	Categorical variables	Ever told you had weak/failing kidneys	Exclusion
KIQ042	Categorical variables	Leak urine during physical activities	Exclusion
KIQ046	Categorical variables	Leak urine during nonphysical activities	Involvement
LBXCOT	Continuous variables	Cotinine, Serum (ng/mL)	Involvement
MCQ010	Categorical variables	Ever been told you have asthma	Exclusion
MCQ160a	Categorical variables	Doctor ever said you had arthritis	Exclusion
MCQ160b	Categorical variables	Ever told had congestive heart failure	Exclusion
MCQ160c	Categorical variables	Ever told you had coronary heart disease	Exclusion
MCQ160d	Categorical variables	Ever told you had angina/angina pectoris	Exclusion
MCQ160e	Categorical variables	Ever told you had heart attack	Exclusion
MCQ160f	Categorical variables	Ever told you had a stroke	Exclusion
MCQ160g	Categorical variables	Ever told you had emphysema	Exclusion
MCQ160k	Categorical variables	Ever told you had chronic bronchitis	Involvement
MCQ160l	Categorical variables	Ever told you had any liver condition	Exclusion
MCQ160m	Categorical variables	Ever told you had thyroid problem	Exclusion
MCQ220	Categorical variables	Ever told you had cancer or malignancy	Exclusion
MCQ300A	Categorical variables	Close relative had heart attack?	Exclusion
MCQ300B	Categorical variables	Close relative had asthma?	Exclusion
MCQ300C	Categorical variables	Close relative had diabetes?	Exclusion
OCQ150	Categorical variables	Type of work done last week	Exclusion
PAD680	Continuous variables	Minutes sedentary activity	Involvement
PFQ049	Categorical variables	Limitations keeping you from working	Exclusion
PFQ054	Categorical variables	Need special equipment to walk	Involvement
RIAGENDR	Categorical variables	Gender	Exclusion
RIDAGEYR	Continuous variables	Age in years at screening	Involvement
RIDRETH1	Categorical variables	Race/Hispanic origin	Exclusion
RPA	Continuous variables	Recreational physical activities	Exclusion
RXDCOUNT	Categorical variables	Number of prescription medicines taken	Involvement
RXDDAYS	Categorical variables	Number of days taken medicine	Involvement
RXDUSE	Categorical variables	Taken prescription medicine, past month	Exclusion
SLD012	Continuous variables	Sleep hours - weekdays or workdays	Involvement
SLQ050	Categorical variables	Ever told doctor had trouble sleeping?	Involvement

(Continued)

TABLE 2 (Continued)

Code	Variable type	Label	Involvement/Exclusion
SMQ680	Categorical variables	Used tobacco/nicotine last 5 days?	Exclusion
SMS	Categorical variables	Smoking status	Exclusion
TPA	Continuous variables	Total physical activity	Involvement
WPA	Continuous variables	Work-related physical activity	Exclusion

HEI, Health Eating Index.

TABLE 3 Performance evaluation of machine learning algorithms.

	AUC	Sensitivity	Specificity	Youden index	F1-score
SEML	0.85	0.79	0.78	0.57	0.50
CatBoost	0.81	0.74	0.78	0.52	0.47
XGBoost	0.81	0.64	0.86	0.50	0.46
RF	0.82	0.74	0.80	0.54	0.45

AUC, Area Under the Curve; SEML, Stacking Ensemble Machine Learning; CatBoost, Categorical Boosting; XGBoost, eXtreme Gradient Boosting; RF, Random Forest.

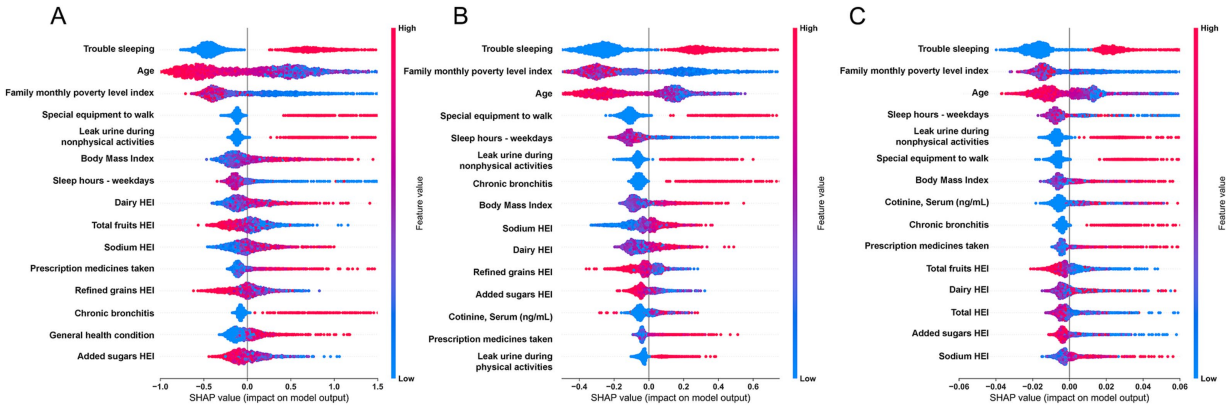


FIGURE 4
SHapley Additive exPlanations (SHAP) summary plot of features in the base models. **(A)** CatBoost; **(B)** XGBoost; **(C)** Random Forest.

with prior research, which indicates that excessive intake of added sugars can have adverse effects on mental health and increase the risk of developing MDD (36, 37). Secondly, for sodium intake, a positive correlation was observed with the occurrence of MDD when HEI scores exceeded 2, indicating a potential link between low sodium diets and increased risk of MDD in adults living alone. This finding is consistent with findings from animal studies, which suggest that insufficient sodium intake may induce depressive symptoms, alleviated by sodium supplementation (38, 39). However, it is noteworthy that while these results from animal experiments exist, current limited human studies only support this inverse relationship between sodium intake and depression among females (40). Thus, the present findings still require further validation through larger, more rigorously designed studies. Lastly, regarding dairy intake, a U-shaped relationship was found between dairy intake and MDD occurrence. Specifically, the lowest contribution to MDD risk was observed when HEI scores reached 5, emphasizing that both excessive and insufficient dairy intake are closely associated with the occurrence of MDD in adults living alone. Given the current insufficiency of research on the relationship between dairy products and depression, particularly lacking longitudinal studies targeting the US population, we cannot yet determine a causal relationship between the two. These findings underscore the necessity for future research.

Moreover, this study has confirmed a series of baseline health conditions as key predictors for MDD. we systematically identified multiple health indicators closely associated with MDD occurrence. Firstly, body mass index (BMI), a common measure assessing whether an individual's weight falls within a healthy range, shows a significant correlation with MDD risk. SHAP dependence curves indicate that elevated BMI (BMI > 25 kg/m^2) is positively correlated with an increased risk of MDD. aligning with findings from previous epidemiological and clinical studies, which have shown that individuals who are overweight or obese have a higher risk of MDD (41–43). Conversely, a moderate BMI (18–25 kg/m^2) appears to have a protective effect. This highlights the critical role of weight management in preventing MDD among adults living alone. Secondly, the number of prescription medications has also been identified as an important predictive

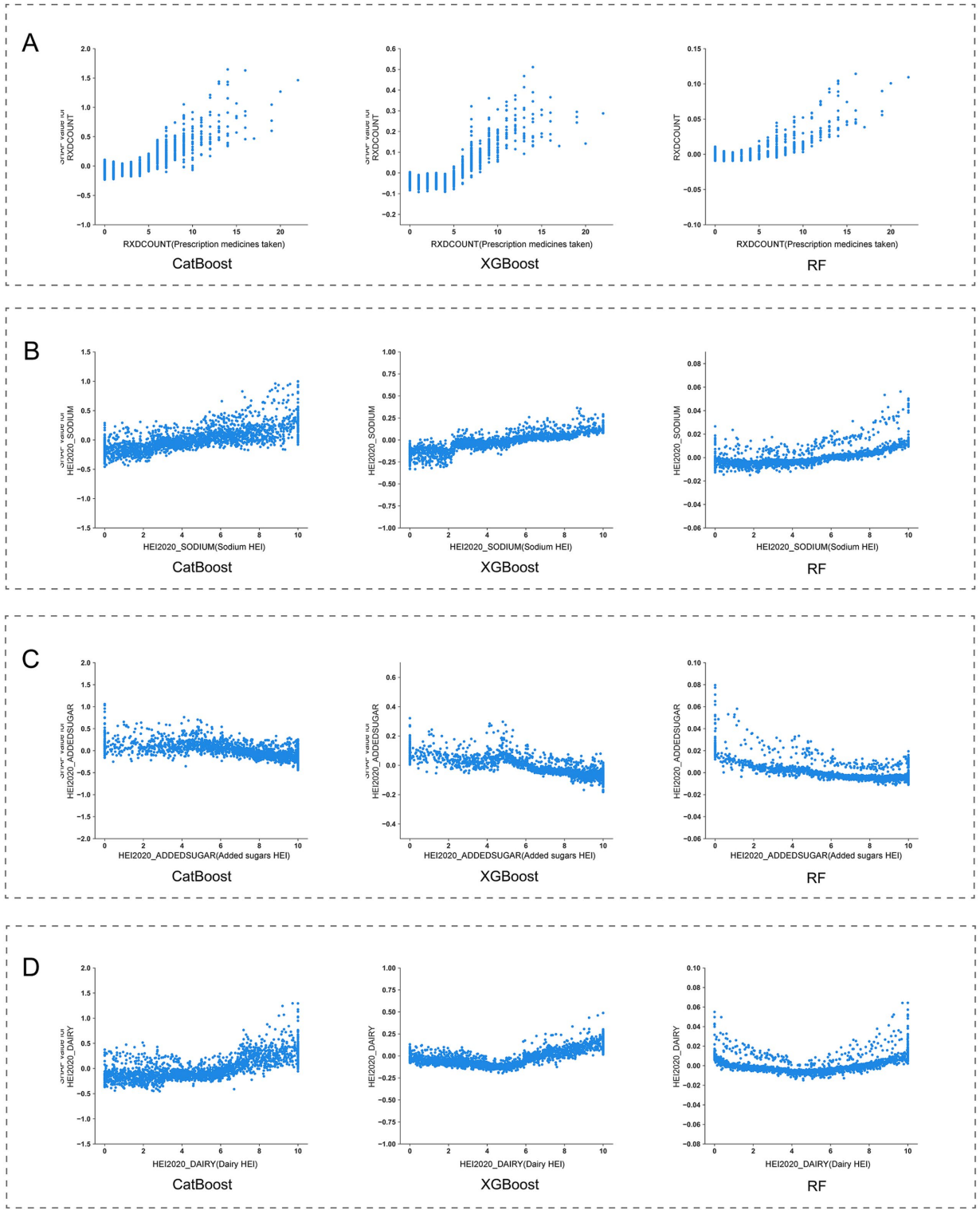


FIGURE 5
SHapley Additive exPlanations (SHAP) dependence plot of features in the base models. **(A)** RIDAGEYR; **(B)** BMXBMI; **(C)** SLD012; **(D)** INDFMMPI.

feature. Generally, an increase in the number of prescription medications indicates potentially more complex health conditions among participants. In this study, we observed that the risk of MDD significantly rises when the number of prescription medications reaches five or more, indicating the need for special attention to mental health issues among adults living alone who require multiple medications. Additionally, the study found significant associations between the risk of MDD and the need for specific walking aids, symptoms of leak urine during nonphysical activities, and chronic bronchitis conditions.

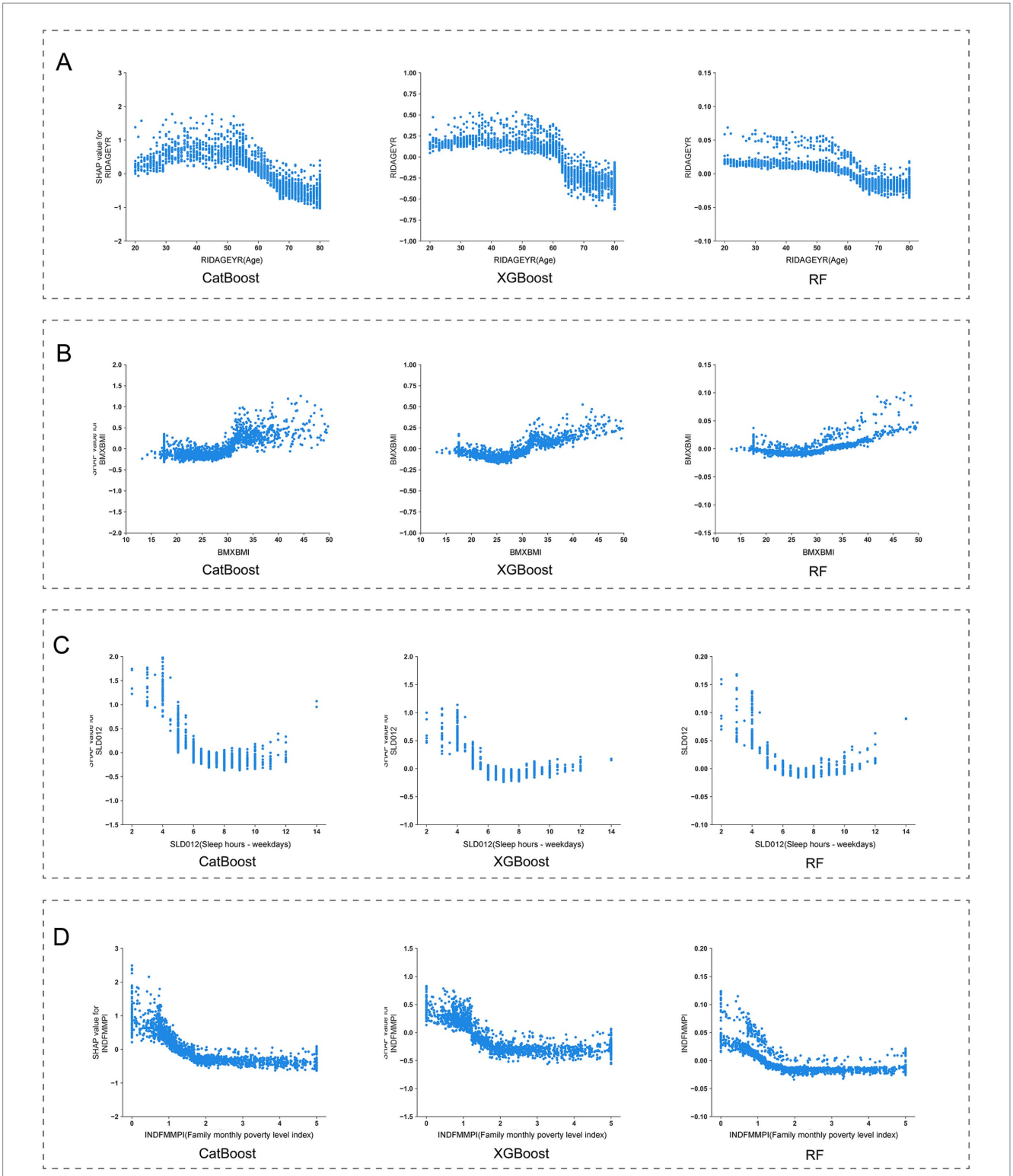


FIGURE 6
SHapley Additive exPlanations (SHAP) dependence plot of features in the base models. **(A)** RXDCOUNT; **(B)** HEI2020_SODIUM; **(C)** HEI2020_ADDEDSUGAR; **(D)** HEI2020_DAIRY.

Changes in these participants' physiological functions may indirectly increase the risk of MDD by affecting their quality of life. However, the causal relationships between these factors and MDD require further longitudinal research to be confirmed.

In summary, the findings presented above demonstrate that our model's results are interpretable and meaningful. These outcomes not only corroborate the algorithm's efficacy in predicting MDD but also offer a reference for the development of public health policies.

There are several limitations found in the current study. Firstly, the data utilized in this research was sourced exclusively from the US NHANES database, which contains information on the US population only. Therefore, caution ought to be applied when generalizing the

predictive model to other countries. Secondly, the cross-sectional design of the study necessitates a high-quality prospective research to delve into the causal relationship between the identified predictors and the occurrence of MDD.

In conclusion, this study has successfully constructed a predictive model for MDD specifically tailored for adults living alone by applying stacked ensemble technique. Through SHAP analysis, the research thoroughly dissected the complex interplay among various predictors and MDD, providing a scientific reference for the development of personalized and effective intervention strategies.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.cdc.gov/nchs/nhanes.

Ethics statement

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the [patients/ participants OR patients/participants legal guardian/next of kin] was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

ZC: Conceptualization, Writing – original draft, Writing – review & editing. HL: Writing – original draft, Writing – review & editing. YZ: Methodology, Software, Writing – review & editing. FX: Formal analysis, Methodology, Writing – review & editing. JJ: Formal analysis, Methodology, Writing – review & editing. ZX: Funding acquisition, Supervision, Writing – review & editing. XD: Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (32371420 and 82202705), the Project of the Science and Technology Department of Sichuan Province (grant nos. 2023YFS0162 and 2023NSFSC1738), Sichuan University-Luzhou Municipal People's Government Strategic Cooperation Project (2022CDLZ-19), Sichuan Provincial Cadre Health Research Project (2023-401), and Sanya Science and Technology Innovation Project (2022KJCX09).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2025.1472050/full#supplementary-material

References

1. Mykyta L. Living alone and feelings of depression among adults age 18 and older. *Natl Health Stat Report*. (2024):1–11. doi: 10.15620/cdc:136451

2. Nam SM, Peterson TA, Seo KY, Han HW, Kang JI. 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:e27344. doi: 10.2196/27344

3. World Health Organization. Depressive disorder (depression). (2023). Available at: https://www.who.int/news-room/fact-sheets/detail/depression (accessed [June 1, 2024]).

4. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the global burden of disease study 2010. *Lancet*. (2012) 380:2197–223. doi: 10.1016/S0140-6736(12)61689-4

5. Hasler G. Pathophysiology of depression: do we have any solid evidence of interest to clinicians? *World Psychiatry*. (2010) 9:155–61. doi: 10.1002/j.2051-5545.2010.tb00298.x

6. Qu Z, Wang Y, Guo D, He G, Sui C, Duan Y, et al. Identifying depression in the United States veterans using deep learning algorithms, nhanes 2005-2018. *BMC Psychiatry*. (2023) 23:620. doi: 10.1186/s12888-023-05109-9

7. Oh J, Yun K, Maoz U, Kim TS, Chae JH. Identifying depression in the national health and nutrition examination survey data using a deep learning algorithm. *J Affect Disord*. (2019) 257:623–31. doi: 10.1016/j.jad.2019.06.034

8. Lee JY, Won D, Lee K. Machine learning-based identification and related features of depression in patients with diabetes mellitus based on the Korea national health and nutrition examination survey: a cross-sectional study. *PLoS One*. (2023) 18:e288648. doi: 10.1371/journal.pone.0288648

9. Zhang C, Chen X, Wang S, Hu J, Wang C, Liu X. Using catboost algorithm to identify middle-aged and elderly depression, national health and nutrition examination survey 2011-2018. *Psychiatry Res*. (2021) 306:114261. doi: 10.1016/j.psychres.2021.114261

10. Gomes S, von Schantz M, Leocadio-Miguel M. Predicting depressive symptoms in middle-aged and elderly adults using sleep data and clinical health markers: a machine learning approach. *Sleep Med*. (2023) 102:123–31. doi: 10.1016/j.sleep.2023.01.002

11. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. (2023) 388:1201–8. doi: 10.1056/NEJMra2302038

12. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. (2016) 375:1216–9. doi: 10.1056/NEJMp1606181

13. Velagapudi L, Saiegh FA, Swaminathan S, Mouchtouris N, Khanna O, Sabourin V, et al. Machine learning for outcome prediction of neurosurgical aneurysm treatment: current methods and future directions. *Clin Neurol Neurosurg*. (2023) 224:107547. doi: 10.1016/j.clineuro.2022.107547

14. Tsai SF, Yang CT, Liu WJ, Lee CL. Development and validation of an insulin resistance model for a population without diabetes mellitus and its clinical implication: a prospective cohort study. *Eclinicalmedicine*. (2023) 58:101934. doi: 10.1016/j.eclinm.2023.101934

15. Goodwin NL, Nilsson S, Choong JJ, Golden SA. Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. *Curr Opin Neurobiol*. (2022) 73:102544. doi: 10.1016/j.conb.2022.102544

16. Ensemble based systems in decision making. *Ieee Circuits and Systems Magazine*. (2006) 6:21–45. doi: 10.1109/MCAS.2006.1688199

17. Zhou T, Jiao H. Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educ Psychol Meas*. (2023) 83:831–54. doi: 10.1177/00131644221117193

18. Gupta A, Jain V, Singh A. Stacking ensemble-based intelligent machine learning model for predicting post-covid-19 complications. *N Gener Comput*. (2022) 40:987–1007. doi: 10.1007/s00354-021-00144-0

19. About the national health and nutrition examination survey. (2024). Available at: https://www.cdc.gov/nchs/nhanes/about/erb.html (accessed [June 20, 2024]).

20. Levis B, Benedetti A, Thombs BD. Accuracy of patient health questionnaire-9 (phq-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ*. (2019) 365:l1476. doi: 10.1136/bmj.l1476

21. Kroenke K, Spitzer RL, Williams JB. The phq-9: validity of a brief depression severity measure. *J Gen Intern Med*. (2001) 16:606–13. doi: 10.1046/j.1525-1497.2001.016009606.x

22. Shams-White MM, Pannucci TE, Lerman JL, Herrick KA, Zimmer M, Meyers MK, et al. Healthy eating index-2020: review and update process to reflect the dietary guidelines for americans,2020-2025. *J Acad Nutr Diet*. (2023) 123:1280–8. doi: 10.1016/j.jand.2023.05.015

23. Zhan JJ, Hodge RA, Dunlop AL, Lee MM, Bui L, Liang D, et al. Dietaryindex: a user-friendly and versatile r package for standardizing dietary pattern analysis in epidemiological and clinical studies. *Biorxiv*. (2024), 120:165–1174. doi: 10.1016/j.ajcnut.2024.08.021

24. Liang J, Huang S, Jiang N, Kakaer A, Chen Y, Liu M, et al. Association between joint physical activity and dietary quality and lower risk of depression symptoms in us adults: cross-sectional nhanes study. *JMIR Public Health Surveill*. (2023) 9:e45776. doi: 10.2196/45776

25. Nawrin SS, Inada H, Momma H, Nagatomi R. Twenty-four-hour physical activity patterns associated with depressive symptoms: a cross-sectional study using big data-machine learning approach. *BMC Public Health*. (2024) 24:1254. doi: 10.1186/s12889-024-18759-5

26. Li E, Ai F, Liang C. 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*. (2023) 11:1348803. doi: 10.3389/fpubh.2023.1348803

27. Cho SE, Geem ZW, Na KS. Predicting depression in community dwellers using a machine learning algorithm. *Diagnostics (Basel)*. (2021) 11:11. doi: 10.3390/diagnostics11081429

28. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to shap analysis: explaining supervised machine learning model predictions in drug development. *Clin Transl Sci*. (2024) 17:e70056. doi: 10.1111/cts.70056

29. Villamil E, Huppert F, Melzer D. Low prevalence of depression and anxiety is linked to statutory retirement ages rather than personal work exit: a national survey. *Psychol Med*. (2006) 36:999–1009. doi: 10.1017/S0033291706007719

30. Schaakxs R, Comijs HC, van der Mast RC, Schoevers RA, Beekman A, Penninx B. Risk factors for depression: differential across age? *Am J Geriatr Psychiatry*. (2017) 25:966–77. doi: 10.1016/j.jagp.2017.04.004

31. Liu J, Yan F, Ma X, Guo HL, Tang YL, Rakofsky JJ, et al. Prevalence of major depressive disorder and socio-demographic correlates: results of a representative household epidemiological survey in Beijing, China. *J Affect Disord*. (2015) 179:74–81. doi: 10.1016/j.jad.2015.03.009

32. Hinata A, Kabasawa K, Watanabe Y, Kitamura K, Ito Y, Takachi R, et al. Education, household income, and depressive symptoms in middle-aged and older japanese adults. *BMC Public Health*. (2021) 21:2120. doi: 10.1186/s12889-021-12168-8

33. Zhang M, Ma Y, Du L, Wang K, Li Z, Zhu W, et al. Sleep disorders and non-sleep circadian disorders predict depression: a systematic review and meta-analysis of longitudinal studies. *Neurosci Biobehav Rev*. (2022) 134:104532. doi: 10.1016/j.neubiorev.2022.104532

34. Baglioni C, Battagliese G, Feige B, Spiegelhalder K, Nissen C, Voderholzer U, et al. Insomnia as a predictor of depression: a meta-analytic evaluation of longitudinal epidemiological studies. *J Affect Disord*. (2011) 135:10–9. doi: 10.1016/j.jad.2011.01.011

35. Bulloch AGM, Williams JVA, Lavorato DH, Patten SB. The depression and marital status relationship is modified by both age and gender. *J Affect Disord*. (2017) 223:65–8. doi: 10.1016/j.jad.2017.06.007

36. Knüppel A, Shipley MJ, Llewellyn CH, Brunner EJ. Sugar intake from sweet food and beverages, common mental disorder and depression: prospective findings from the Whitehall ii study. *Sci Rep*. (2017) 7:6287. doi: 10.1038/s41598-017-05649-7

37. Hu D, Cheng L, Jiang W. Sugar-sweetened beverages consumption and the risk of depression: a meta-analysis of observational studies. *J Affect Disord*. (2019) 245:348–55. doi: 10.1016/j.jad.2018.11.015

38. Grippo AJ, Moffitt JA, Beltz TG, Johnson AK. Reduced hedonic behavior and altered cardiovascular function induced by mild sodium depletion in rats. *Behav Neurosci*. (2006) 120:1133–43. doi: 10.1037/0735-7044.120.5.1133

39. Morris MJ, Na ES, Johnson AK. Mineralocorticoid receptor antagonism prevents hedonic deficits induced by a chronic sodium appetite. *Behav Neurosci*. (2010) 124:211–24. doi: 10.1037/a0018910

40. Goldstein P, Leshem M. Dietary sodium, added salt, and serum sodium associations with growth and depression in the u.s. general population. *Appetite*. (2014) 79:83–90. doi: 10.1016/j.appet.2014.04.008

41. Wang R, He Y, Deng Y, Wang H, Zhang Y, Feng J, et al. Body weight in neurological and psychiatric disorders: a large prospective cohort study. *Nature Mental Health*. (2024) 2:41–51. doi: 10.1038/s44220-023-00158-1

42. He K, Pang T, Huang H. The relationship between depressive symptoms and bmi: 2005–2018 nhanes data. *J Affect Disord*. (2022) 313:151–7. doi: 10.1016/j.jad.2022.06.046

43. Luppino FS, de Wit LM, Bouvy PF, Stijnen T, Cuijpers P, Penninx BWJH, et al. Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. *Arch Gen Psychiatry*. (2010) 67:220–9. doi: 10.1001/archgenpsychiatry.2010.2