

RESEARCH

Open Access



Prediction of depressive disorder using machine learning approaches: findings from the NHANES

Thien Vu^{1,2*}, Research Dawadi¹, Masaki Yamamoto¹, Jie Ting Tay¹, Naoki Watanabe¹, Yuki Kuriya¹, Ai Oya¹, Phap Ngoc Hoang Tran¹ and Michihiro Araki^{1,3,4,5*}

Abstract

Background Depressive disorder, particularly major depressive disorder (MDD), significantly impact individuals and society. Traditional analysis methods often suffer from subjectivity and may not capture complex, non-linear relationships between risk factors. Machine learning (ML) offers a data-driven approach to predict and diagnose depression more accurately by analyzing large and complex datasets.

Methods This study utilized data from the National Health and Nutrition Examination Survey (NHANES) 2013–2014 to predict depression using six supervised ML models: Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), and Light Gradient Boosting Machine (LightGBM). Depression was assessed using the Patient Health Questionnaire (PHQ-9), with a score of 10 or higher indicating moderate to severe depression. The dataset was split into training and testing sets (80% and 20%, respectively), and model performance was evaluated using accuracy, sensitivity, specificity, precision, AUC, and F1 score. SHAP (SHapley Additive exPlanations) values were used to identify the critical risk factors and interpret the contributions of each feature to the prediction.

Results XGBoost was identified as the best-performing model, achieving the highest accuracy, sensitivity, specificity, precision, AUC, and F1 score. SHAP analysis highlighted the most significant predictors of depression: the ratio family income to poverty (PIR), sex, hypertension, serum cotinine and hydroxycotinine, BMI, education level, glucose levels, age, marital status, and renal function (eGFR).

Conclusion We developed ML models to predict depression and utilized SHAP for interpretation. This approach identifies key factors associated with depression, encompassing socioeconomic, demographic, and health-related aspects.

Keywords Depression, Depressive disorder, Supervised machine learning, Logistic regression, Random forest, Naïve bayes, Support Vector Machine (SVM), eXtreme Gradient Boost (XGBoost), Light Gradient Boosted Machine (LightGBM), Shapley Additive exPlanations (SHAP)

*Correspondence:

Thien Vu
thienvuyd01@gmail.com; thien-vu@nibiohn.go.jp
Michihiro Araki
araki@nibiohn.go.jp; michihiroaraki@gmail.com

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Introduction

Depressive disorder, specifically major depressive disorder (MDD), place a significant burden on individuals and society. These psychiatric disorders are distinguishable by pervasive feelings of sadness, emptiness, or hopelessness that significantly interfere with a person's daily activities. These disorders are highly prevalent worldwide and have a significant impact on individuals' ability to do daily tasks, their overall well-being, their cognitive function, and employment status [1]. Depression can cause significant limitations in personal, social, and occupational domains, making it an important public health concern that requires effective methods for prediction and intervention [1, 2]. Depression is the primary cause of impairment, as determined by Years Lived with Disability (YLDs), and the fourth most significant contributor in the worldwide burden of disease, according to the World Health Organization (WHO) [3]. The Global Burden of Disease (GBD) study shows that depressive disorders account for a substantial portion of total Disability Adjusted Life Years (DALYs) and YLDs, with a trend of increasing burden over time [4]. An increase in the global economic burden of mental health diseases is anticipated, with a particular emphasis on the burden of depressive disorders. This highlights the urgent need for effective strategies to address the rising burden of depressive disorders and improve access to mental health care services.

Traditional statistical methods have long been employed in the prediction and analysis of depression. These methods typically involve hypothesis-driven approaches that use predefined models to understand the relationships between various risk factors and depression outcomes. Despite being helpful, they may not capture the complex, non-linear relationships between different risk factors and depression, limiting their effectiveness in fully understanding and predicting the disorder.

Recent studies have shown that machine learning (ML) has become a revolutionary tool for predicting and diagnosing diseases [5–7]. It offers several advantages over traditional statistical methods [6]. Unlike conventional approaches that test hypotheses derived from theories, ML focuses on discovering hidden patterns and interactions within large datasets [6]. This capability enables ML to analyze complex, non-linear relationships among variables, leading to more accurate and nuanced predictions of depression risk [8].

One of the major challenges in adequately addressing MDD is identifying affected individuals and ensuring appropriate and timely treatment. MDD symptoms are internally experienced, and often go undetected. The application of ML in epidemiological studies and public health has revolutionized the approach to depression prediction and early intervention [9, 10]. ML algorithms can process vast amounts of data from electronic health

records (EHRs), biometric markers, and patient characteristics to identify individuals at risk of developing depression [9, 11]. This ability to analyze complex, multidimensional data sets with greater precision than traditional methods make ML particularly valuable in the context of population-level health studies.

When considering the broader context of an individual or population, depression is influenced by a multitude of factors. The critical question arises: which of these identified risk factors are the most significant, and how do they contribute to the formation of the predicted outcome?

Through this study, we aim to assess the role of machine learning (ML) in epidemiological research. Additionally, we strive to explore all potential risk factors for depressive disorders (DD) based on large-scale National Health and Nutrition Examination Survey (NHANES) data from the year 2013–2014 using supervised ML. Furthermore, we aim to evaluate the contribution of each risk factor to the development of DD.

Methods

Study participants

The study employed data from NHANES 2013–2014, which is a dependable and extensive random sample designed to evaluate the health and nutritional condition of the US population (www.cdc.gov/nchs/nhanes). Participants were questioned in their residences, and had physical and laboratory examinations at a mobile examination center (MEC). The National Centre for Health Statistics Research Ethics Review Board granted ethics approval (Protocol # 2013-14) for all the study procedures, and all subjects provided signed informed consent. The studies conducted adhered to NHANES guidelines and regulations. The study specifically eliminated those who were under the age of 18 and those who provided incomplete or insufficient responses.

Data collection

Data for the study were collected from various sources within NHANES 2013–2014, including: Demographic Data, Examination Data, Laboratory Data, and Questionnaire Data. The following variables were employed:

- Age was calculated from the date of the interview to the date of birth.
- Gender was coded as male or female by NHANES personnel.
- Race was assessed by two questions: “Do you identify as Hispanic, Latino, or Hispanic?” and “What race do you consider yourself?” Based on the responses, race was categorized into six groups: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, and Other Race.

- Education level was grouped by the question: “What is the highest grade or level of school you have completed or the highest degree you have received?” Responses were divided into five groups: less than grade 9th, 9–11th grade, high school graduate, some college or associate degree, and college graduate or above.
- Marital status was categorized as married, widowed, divorced, separated, never married, and living with a partner.
- PIR refers to the ratio of family income to the family’s appropriate poverty threshold. Annual family income was categorized into four levels: low (income < \$20,000), low-medium ($\$20,000 \leq \text{income} < \$75,000$), medium-high ($\$75,000 \leq \text{income} < \$99,000$), and high income (income > \$100,000).
- Body mass index (BMI) was calculated as weight in kilograms divided by height in meters squared (kg/m^2).
- Blood pressure was measured using a mercury sphygmomanometer, with two consecutive readings of systolic blood pressure (SBP) and diastolic blood pressure (DBP) taken at 5-minute intervals. The mean of the two readings was calculated for analysis.
- The estimated glomerular filtration rate (eGFR) was calculated using the Modification of Diet in Renal Disease formula: $\text{eGFR} = 186 * \text{creatinine}^{(-1.154)} * \text{age}^{(-0.203)} * (0.742 \text{ if female})$.
- Both smoking and drinking habits were separated into three categories: current, past, and never.
- Physical activity was categorized as mild, moderate, or vigorous.
- Hypertension was defined as a resting SBP ≥ 140 mmHg, a DBP ≥ 90 mmHg, or the use of hypertensive medication.
- Diabetes was defined as a two-hour glucose tolerance test result ≥ 200 mg/dL, plasma fasting glucose ≥ 126 mg/dL, glycohemoglobin $\geq 6.5\%$, or the use of diabetic medication.
- Dyslipidemia was defined as total cholesterol ≥ 200 mg/dL, triglycerides ≥ 150 mg/dL, low-density lipoprotein cholesterol ≥ 140 mg/dL, high-density lipoprotein cholesterol < 40 mg/dL, or the use of lipid-lowering medication.

Depressive disorder assessment

Depression was assessed using the Patient Health Questionnaire (PHQ-9), a nine-item depression scale that is widely used to screen, diagnose, monitor, and measure the severity of depression. The PHQ-9 screening instrument consists of nine questions about depression symptoms experienced during the past 2 weeks followed by a single question that assesses associated impairment. Each question is scored on a scale from 0 to 3, with 0 indicating

the absence of the symptom, 1 indicating its presence for several days, 2 indicating its presence for more than half of the days and 3 indicating its presence nearly every day. The result of the PHQ-9 questionnaire consists of scores ranging from 0 to 27. The PHQ-9 is widely used in various areas, such as psychiatric hospitals, primary care, and the general population, and has demonstrated good reliability and validity in assessing depression severity [12, 13]. A score equal to or exceeding 10 is frequently employed as a benchmark for indicating moderate to severe depression. Studies have discovered that a PHQ-9 score of 10 or higher is both extremely sensitive and specific in identifying severe depression [14]. In this study, we employed a PHQ-9 score of 10 as the threshold for diagnosing individuals with depressive disorder.

Data pre-processing

We conducted a comprehensive analysis of the NHANES dataset to predict depression status using various statistical and machine learning techniques. Initially, we filtered and cleaned the data, creating a cumulative depression score and categorizing individuals based on their depression levels. We also derived variables for smoking, drinking, physical activity, family income, blood pressure, hypertension, dyslipidemia, diabetes, and renal function.

The data was described using the mean and standard deviation for symmetric numerical variables, and the median and interquartile range for asymmetric numerical variables. Categorical variables were described using frequency and percentage. To analyse differences in participants characteristics between those with and without depression, we utilized a range of statistical tests, including Student’s t-tests, Mann-Whitney U tests, and Chi-squared tests, depending on the suitability for each variable.

Features selection

The selection of features (variables) was conducted in a systematic manner to retain only the most relevant predictors for the model development. Initially, variables with over 50% missing data were excluded to minimize potential bias caused by imputation. We used Random Forest-based imputation for handling missing data [15]. Following this, a correlation matrix was employed to identify and remove variables exhibiting high multicollinearity, defined as having a correlation coefficient exceeding 0.8. Subsequently, Least Absolute Shrinkage and Selection Operator (LASSO) regression was used to shrink the coefficients of less influential predictors towards zero, effectively eliminating them from the model. The final set of features used for model development is presented in Supplementary Table S2.

Model development

The significant imbalance could lead to the model being biased toward the majority class (without depression), which would reduce its effectiveness in predicting depression cases. To mitigate this issue, we reduced the number of samples from the majority class (without depression) to match the number of samples in the minority class (depression). By using Random Under-sampling, we ensured that both classes were equally represented during model training. This approach was chosen for its simplicity and effectiveness in improving model performance when dealing with imbalanced data. Next, we applied one-hot encoding to convert categorical variables into a binary format and normalization to scale the numerical features to prepare the data for the machine learning models. In this study, the Z-score normalization method was used for data normalization. Specifically, this technique centers each numerical feature by subtracting the mean and then scales it by dividing by the standard deviation [5, 6]. The dataset was split into training and testing sets with an 80% and 20% ratio, respectively. We trained multiple models, including LR, RF, NB, SVM, XGBoost, and LightGBM. The selection of LR, RF, NB, SVM, XGBoost, and LightGBM is based on several key considerations, including diversity, proven effectiveness, handling of different data types, and computational efficiency. We choose a diverse set of algorithms to capture both linear and non-linear relationships in the data. Each model brings unique strengths: LR is a staple for risk prediction due to its simplicity and interpretability [16], Naive Bayes efficiently handles categorical features [17], SVM is suitable for high-dimensional data [18, 19], RF and gradient boosting models capture non-linear relationships and complex feature interactions [20, 21]. XGBoost, a gradient boosting method, is known for its high efficiency and accuracy, making it a powerful tool for both classification and regression tasks in the context of medical research [22]. Additionally, LightGBM's computational efficiency makes it ideal for large-scale datasets [23]. This diverse set of models was chosen to leverage their complementary strengths, providing a comprehensive evaluation across different algorithm types. During training, we used 5-fold cross-validation to ensure robustness. After training, we tested the models on the test set and calculated performance metrics such as accuracy, sensitivity, specificity, precision, area under the receiver operating characteristic curve (AUC) and F1-score.

The hyperparameter optimization process involved the systematic tuning of key parameters for each machine learning model. The study utilized grid search-based hyperparameter tuning in combination with 5-fold cross-validation to evaluate different combinations of parameters and ensure stability and minimizing the impact of

random variations. The objective was to identify the optimal configuration that maximized model performance, specifically focusing on metrics such as accuracy and AUC. The best model for predicting depression was identified by comparing performance metrics on the test set.

Model explanation

SHAP scores are used to explain the output of our depression prediction model. SHAP values provide a common measure for interpreting how each feature in a model contributes to prediction.

We analysed feature importance using SHAP values. This thorough approach allowed us to pinpoint key factors influencing depression and evaluate how these factors contribute for predicted outcome, ultimately enhancing our understanding of the dataset and the underlying relationships within the data. The analysis was conducted using the statistical software R (version 4.4.0).

Results

Participants' characteristics

The study evaluated the characteristics of 5,372 participants, divided into those without depression (4,861, 90.5%) and those with depression (511, 9.5%). The results show significant differences between the two groups on a range of demographic, socio-economic and health-related factors.

Participants with depression were older with high proportion of females and low level of education. However, non-Hispanic Asians were the more prevalent among those without depression.

There were fewer married individuals and more divorced individuals in the depression group. Similarly, current smokers were more prevalent among those with depression. Physical activity levels were lower, and very-low family income was more common in the depression group.

Health measures showed that the depression group had higher systolic blood pressure and BMI. Biochemical differences included lower eGFR, higher apolipoprotein B, lower HDL-cholesterol, and higher triglyceride, fasting glucose, and two-hour glucose levels.

Moreover, serum and urine cotinine and hydroxycotinine levels were significantly higher among participants with depression. Furthermore, participants with depression tended to have higher prevalence of hypertension, diabetes and dyslipidemia.

These findings indicate significant differences in demographics, socioeconomic status, lifestyle factors, biochemical parameters, and health conditions between participants with and without depression, highlighting the multifaceted nature of depression and its association with various risk factors, illustrated in Table 1.

Table 1 Characteristics of study participants with and without depression

	Overall (N= 5372)	Without depression (N= 4861)	Depression (N= 511)	p-value
Age, years	47 [32, 63]	46 [31, 63]	52 [36, 63]	< 0.001
Sex, n (%)				< 0.001
Male	2585 (48.1%)	2416 (49.7%)	169 (33.1%)	
Female	2787 (51.9%)	2445 (50.3%)	342 (66.9%)	
Race, n (%)				< 0.001
Mexican American	753 (14.0%)	676 (13.9%)	77 (15.1%)	
Other Hispanic	481 (9.0%)	423 (8.70%)	58 (11.4%)	
Non-Hispanic White	2315 (43.1%)	2091 (43.0%)	224 (43.8%)	
Non-Hispanic Black	1087 (20.2%)	979 (20.1%)	108 (21.1%)	
Non-Hispanic Asian	562 (10.5%)	544 (11.2%)	18 (3.52%)	
Other Race - Multi-Racial	174 (3.2%)	148 (3.04%)	26 (5.09%)	
Education level, n (%)				< 0.001
Less than 9th grade	357 (6.6%)	293 (6.0%)	64 (12.5%)	
9-11th grade	683 (12.7%)	578 (11.9%)	105 (20.5%)	
High school graduate	1130 (21.0%)	1012 (20.8%)	118 (23.1%)	
Some college or AA degree	1591 (29.6%)	1443 (29.7%)	148 (29.0%)	
College graduate or above	1292 (24.1%)	1237 (25.4%)	55 (10.8%)	
Marital status, n (%)				< 0.001
Married	2637 (49.1%)	2449 (50.4%)	188 (36.8%)	
Widowed	369 (6.9%)	312 (6.4%)	57 (11.2%)	
Divorced	582 (10.8%)	484 (10.0%)	98 (19.2%)	
Separated	151 (2.8%)	124 (2.6%)	27 (5.3%)	
Never married	955 (17.8%)	872 (17.9%)	83 (16.2%)	
Living with partner	358 (6.7%)	322 (6.6%)	36 (7.0%)	
Smoking status, n (%)				< 0.001
Current	1056 (19.7%)	891 (18.3%)	165 (32.3%)	
Former	1199 (22.3%)	1077 (22.2%)	122 (23.9%)	
Never	3117 (58.0%)	2893 (59.5%)	224 (43.8%)	
Drinking status, n (%)				0.411

Table 1 (continued)

	Overall (N= 5372)	Without depression (N= 4861)	Depression (N= 511)	p-value
Current	3771 (70.2%)	3410 (70.2%)	361 (70.6%)	
Former	687 (12.8%)	615 (12.7%)	72 (14.1%)	
Never	914 (17.0%)	836 (17.2%)	78 (15.3%)	
Physical activity, n (%)				< 0.001
Mild	1320 (24.6%)	1107 (22.8%)	213 (41.7%)	
Moderate	3713 (69.1%)	3441 (70.8%)	272 (53.2%)	
Vigorous	339 (6.3%)	313 (6.4%)	26 (5.1%)	
Family income, n (%)				< 0.001
Very-Low	1210 (22.5%)	1014 (20.9%)	196 (38.4%)	
Low-Medium	2586 (48.1%)	2341 (48.2%)	245 (47.9%)	
Medium-High	442 (8.2%)	427 (8.8%)	15 (2.9%)	
High	919 (17.1%)	890 (18.3%)	29 (5.7%)	
Ratio of family income to poverty	2.10 [1.04, 4.05]	2.19 [1.07, 4.19]	1.27 [0.78, 2.40]	< 0.001
SBP, mm Hg	122 (17.3)	122 (17.1)	125 (19.0)	0.001
DBP, mm Hg	68.9 (12.6)	68.9 (12.6)	69.3 (12.4)	0.538
BMI, kg/m²	29.0 (7.23)	28.7 (6.91)	31.6 (9.38)	< 0.001
BUN, mg/dL	13.3 (6.08)	13.2 (5.68)	13.5 (8.99)	0.543
Creatinine (mg/dL)	0.914 (0.525)	0.86 [0.72;1.01]	0.82 [0.71;1.00]	0.038
eGFR (mL/min/1.73m²)	91.8 (26.3)	92.2 (26.0)	88.5 (29.0)	0.008
AST (U/L)	22 [19, 27]	22 [19, 27]	22 [19, 28]	0.583
ALT (U/L)	20 [16, 28]	20 [16, 28]	20 [15, 28]	0.501
Apolipoprotein (B) (mg/dL)	89.2 (25.0)	88.8 (24.7)	93.9 (26.8)	0.006
HDL-Cholesterol (mg/dL)	52.7 (15.9)	52.9 (16.0)	51.0 (15.3)	0.010
LDL-cholesterol (mg/dL)	110 (34.8)	110 (34.5)	112 (37.2)	0.397
Triglyceride (mg/dL)	94 [64, 141.2]	92 [63, 136]	115 [78, 178]	< 0.001
Total Cholesterol(mg/dL)	188 (41.8)	187 (41.8)	191 (41.5)	0.107
Fasting Glucose (mg/dL)	99 [92, 108]	99 [92, 108]	102 [94, 116]	< 0.001
Two Hour Glucose (mg/dL)	117 (49.4)	117 (49.2)	126 (51.2)	0.026
Cotinine, Serum (ng/mL)	59.3 (133)	55.0 (129)	100 (161)	< 0.001

Table 1 (continued)

	Overall	Without depression	Depression	p-value
	(N=5372)	(N=4861)	(N=511)	
Hydroxycotinine, Serum (ng/mL)	23.9 (61.7)	21.8 (59.6)	43.9 (76.7)	<0.001
Total Cotinine, urine (ng/mL)	730 (1960)	660 (1790)	1510 (3170)	0.002
Hydroxycotinine, urine (ng/mL)	1440 (4320)	1300 (4110)	2980 (6040)	0.001
Diabetes, n (%)	830 (15.5%)	706 (14.5%)	124 (24.3%)	<0.001
Dyslipidemia, n (%)	3251 (60.5%)	2897 (59.6%)	354 (69.3%)	<0.001
Hypertension, n (%)	1914 (35.6%)	1651 (34.0%)	263 (51.5%)	<0.001

Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; BMI, body mass index; BUN, Blood urea nitrogen; eGFR, estimated glomerular filtration rate; AST, Aspartate aminotransferase; ALT, alanine aminotransferase; HDL, high-density lipoprotein; LDL, low-density lipoprotein

The performance metrics of the six machine learning models applied in prediction of depression give the overview of their effectiveness, as shown in Table 2. XGBoost generally had the highest metrics across all categories, indicating strong overall performance. It showed the best accuracy (0.69), good sensitivity (0.68), and high specificity (0.71), meaning it balanced well between true positive rate and true negative rate. Additionally, it had the highest AUC (0.69), indicating good discriminatory power, and its F1-score was also the highest (0.69), suggesting it maintained a good balance between precision and recall. Naive Bayes also performed well, with similar accuracy (0.68), sensitivity (0.70), and AUC (0.68). However, it slightly lagged behind XGBoost in specificity (0.67) and F1-score (0.69). SVM and Logistic Regression were close contenders but generally fell short in some areas compared to XGBoost and Naive Bayes. Random Forest and LightGBM did not perform as well as the other models in most metrics. XGBoost appeared to be the best-performing model based on the provided metrics. It offered the highest accuracy, strong sensitivity, and specificity, along with the highest AUC and F1-score. Therefore, this indicated that XGBoost was likely the most reliable model for this classification task, offering a good balance between all evaluated metrics. See Supplementary Table S1 for more information about model performance metrics and Supplementary Figure S1 for ROC curves of all machine learning models.

Figures 1 and 2 both highlight the importance and impact of different features on the model's predictions using SHAP values. These visualizations, including a SHAP feature importance plot and a SHAP summary heat plot, offer critical insights into the model's behavior

and the factors influencing its prediction. See Supplementary Table S2 for more information about variables' names.

The SHAP feature importance plot ranks the features based on their average absolute SHAP values, highlighting those with the most significant impact on the model's prediction. The top features identified are PIR, which is the most important, indicating its highest impact on the model's predictions. This is followed by sex.2 (female) and hypertension, highlighting their significant roles. sHCOT (Serum Hydroxycotinine), sCOT (Serum Cotinine) and BMI also show substantial importance, indicating their strong influence on the model. Additionally, educ.5 (Education Level: College graduate or above) significantly impacts the model's prediction. Features such as glucose levels, age, marital status (divorced status) have moderate importance, meaning they still contribute meaningfully but to a lesser extent than the top features. Other features have the smaller impact on the model's predictions.

The SHAP summary plot provides insights into how various features influence the model's predictions. It uses colors to indicate feature values, with yellow representing low values and blue representing high values. The SHAP values on the horizontal axis indicate the impact of each feature on the model's output. Negative SHAP values suggest a protective effect against depression, while positive SHAP values indicate a higher likelihood of depression.

One of the most important features is the PIR. Lower PIR values (yellow dots) generally push the model's predictions higher, indicating that lower family income increases the likelihood of depression. Conversely, higher PIR values (blue dots) decrease this likelihood, acting as a protective factor, with the wide spread of SHAP values showing PIR's strong and consistent impact on the model. Gender is another significant categorical variable. Females have a higher chance of depression, as the positive SHAP values suggest an increased likelihood of depression compared to the male group.

Similarly, hypertension push the model's predictions higher, indicating a higher likelihood of depression, while not having hypertension is more protective.

Marital status (masts.3), particularly being divorced, shows that divorced individuals have a higher likelihood of depression, as indicated by the positive SHAP values, while not being in this group shows a lower likelihood.

Education level (educ.5), categorized as college graduate or above, acts as a protective factor. Higher education levels have negative SHAP values, indicating a lower likelihood of depression, while lower education levels increase the risk.

However, BMI shows varied impacts. For serum cotinine levels (sHCOT), serum cotinine (sCOT), and glucose levels (Glu), higher values increase the likelihood

Table 2 Performance metrics of different machine learning approaches

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	F1_Score
LR	0.66 (0.59–0.73)	0.64 (0.55–0.73)	0.68 (0.59–0.77)	0.66 (0.57–0.75)	0.66 (0.59–0.72)	0.65 (0.57–0.72)
RF	0.65 (0.59–0.72)	0.60 (0.50–0.69)	0.71 (0.61–0.79)	0.67 (0.57–0.77)	0.65 (0.59–0.72)	0.63 (0.55–0.71)
NB	0.68 (0.62–0.75)	0.70 (0.61–0.78)	0.67 (0.57–0.76)	0.68 (0.59–0.77)	0.68 (0.62–0.75)	0.69 (0.62–0.76)
SVM	0.68 (0.61–0.75)	0.65 (0.55–0.74)	0.72 (0.63–0.80)	0.69 (0.60–0.78)	0.68 (0.62–0.75)	0.67 (0.59–0.74)
XGBoost	0.69 (0.63–0.75)	0.68 (0.59–0.77)	0.71 (0.62–0.79)	0.70 (0.60–0.79)	0.69 (0.63–0.75)	0.69 (0.61–0.76)
LightGBM	0.62 (0.55–0.69)	0.64 (0.54–0.73)	0.61 (0.51–0.70)	0.62 (0.52–0.71)	0.62 (0.55–0.69)	0.63 (0.55–0.70)

Abbreviation: AUC, Area Under the Curve; LR, Logistic Regression; RF, Random Forest; NB, Naïve Bayes; SVM, Support Vector Machine; XG Boost, eXtreme Gradient Boost; Light-GBM, Light Gradient Boosted Machine

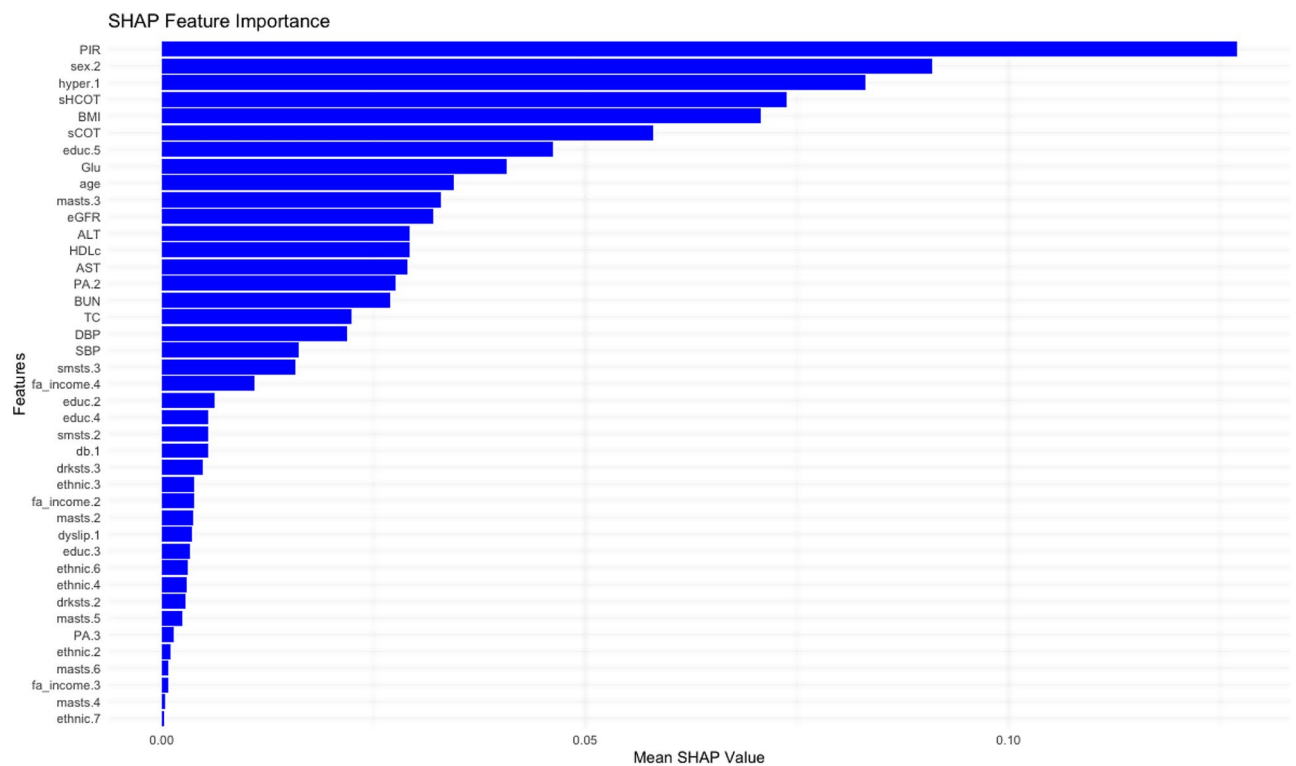


Fig. 1 The contribution levels of all variables to depression based on SHAP values. The global significance of each feature in the model is illustrated in the SHAP (blue) bar plot. It provides an overview of the features' impact on the model's output by displaying the mean absolute SHAP value for each feature. A feature (variable) is represented by each bar in the plot, and the length of the bar indicates the extent of the feature's contribution to Depression.

of depression, as shown by positive SHAP values, while lower values are protective.

Age is another crucial numerical variable that had a varied impact. Lower eGFR (Estimated Glomerular Filtration Rate) values were harmful, increasing the likelihood of depression, while higher eGFR values were protective. Figure 2 showed heat plot of all variables to depression, however, it doesn't clearly focus on each variable. Figures 3, and 4 provide more information of important variables for both categorical and numerical data.

These insights help in understanding the model's behavior and identifying the most important factors influencing the prediction.

Discussion

In this study, we applied machine learning (ML) approaches to predict depression using big data from the NHANES cycle 2013–2014. We used the PHQ-9 score with a cut-off point of 10 to dichotomize depressive disorder. The input factors were derived from demographic data, examination data, laboratory data,

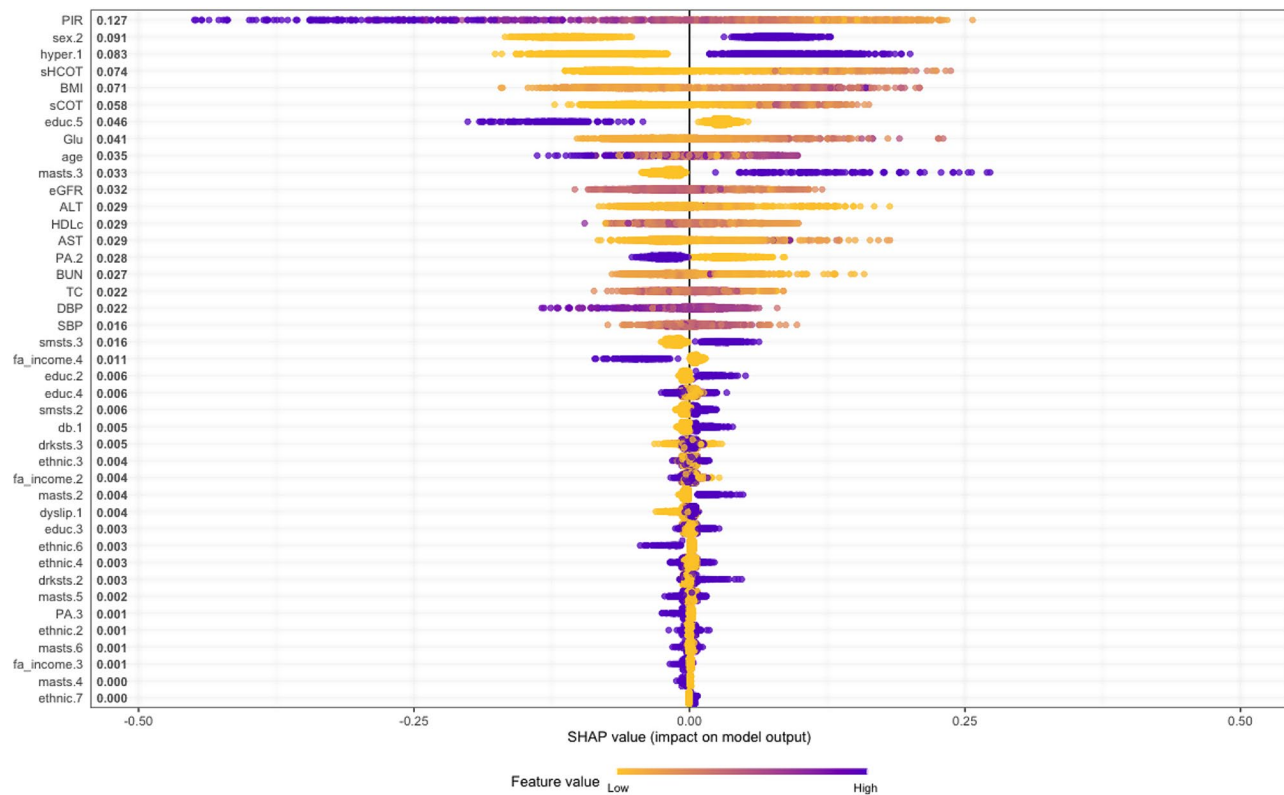


Fig. 2 The heat plot on SHAP values The relationships between the feature (variable) and Depression are revealed by the heat plot of SHAP values. The relationship between the value of a specific feature and its impact on prediction can be fundamentally understood through this. Each data point is associated with a specific participant and their corresponding Shapley value for a specific feature. The Shapley value, which is represented on the x-axis, and the feature's prominence, which is represented on the y-axis, determine the position of a data point on this plot

and questionnaire data of the study participants. We employed six supervised models: Logistic Regression, Random Forest, Naïve Bayes, XGBoost, and LightGBM to predict depression. AUC and F1-score were used as critical indicators for evaluating model performance, with XGBoost emerging as the best-performing model. It offered the highest accuracy, sensitivity, specificity, along with the highest AUC and F1- score, indicating its reliability for this classification task.

To explain the model, we applied SHAP values to identify the most important variables contributing to the risk of depression. The top variables included socioeconomic factors like PIR, education level, marital status, demographic factors such as age, sex, and health-related factors such as hypertension, BMI, blood glucose, eGFR and consumption of nicotine products (serum cotinine & serum hydroxycotinine).

Our findings highlight the superiority of ML models in leveraging all input data to build predictive models. Traditional analysis typically confirms one or a few predictors with the predicted outcome, but our approach demonstrates the advantage of using data-driven ML techniques to gain a comprehensive view of risk factors and their contributions to depression.

The most significant risk factor identified was PIR. Lower PIRs, indicating higher poverty, were associated with significantly higher rates of depressive symptoms. This is consistent with previous studies examining the nonlinear associations between PIR and various health outcomes, which found that lower PIRs were linked to higher vulnerability to adverse health outcomes, including mental health issues like depression [24, 25]. These findings emphasize the importance of considering income levels in public health strategies [24, 25].

We found that women were more likely to have depression compare to men. This aligns with numerous studies showing that women are more likely to suffer from depression than men [26, 27]. Hormonal changes related to the menstrual cycle, pregnancy, and menopause, along with chronic stressors and social discrimination, contribute to this higher prevalence [28].

Both hypertension and depression are linked to increased sympathetic nervous system activity and decreased parasympathetic activity, leading to elevated blood pressure and increased risk of cardiac arrhythmias [29]. Depression is associated with unhealthy behaviors such as smoking, physical inactivity, increased alcohol consumption, poor nutrition, and poor sleep, which are

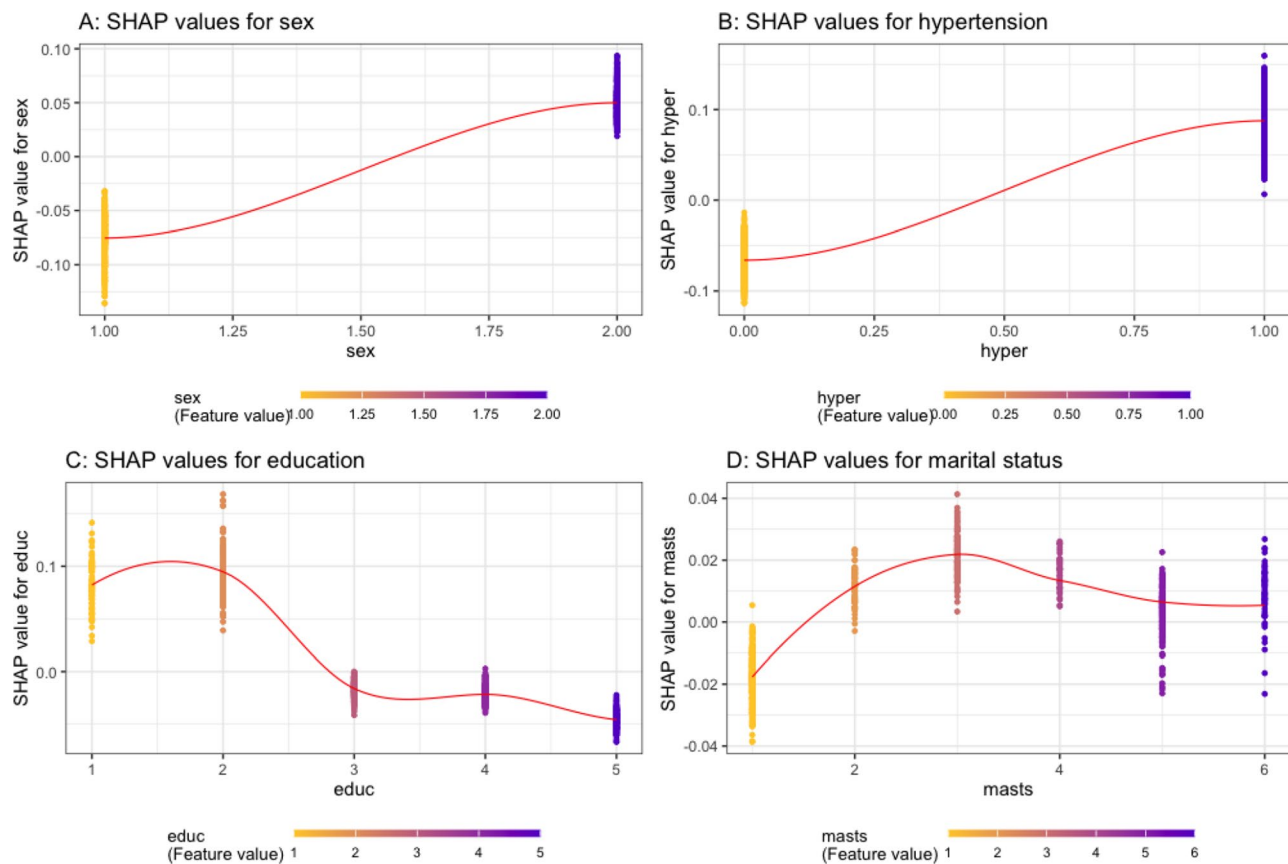


Fig. 3 The impact of categorical variables on depression

known risk factors for hypertension. The heightened sympathetic tone in depressed individuals may contribute to poor blood pressure control and exacerbate hypertension [30, 31]. Elevated cortisol levels due to depression can also promote vascular changes leading to sustained high blood pressure [32].

BMI and depression exhibit intricate relationships, as indicated by our findings in Fig. 4B. Individuals with a BMI of less than 20 exhibit a high SHAP value. Nevertheless, participants with a BMI of 20–30 have negative SHAP values, which indicates that they are less likely to experience depression. Conversely, individuals with a BMI greater than 30 exhibit a significantly elevated SHAP value, which is associated with a favorable correlation with depression. Malnutrition and health issues associated with low body weight can impact mood and mental health [33]. Obesity is often associated with lifestyle factors such as poor diet, physical inactivity, and sleep disturbances, all of which are risk factors for depression [33, 34]. Conversely, depression can lead to changes in appetite and physical activity, contributing to weight gain and obesity. Therefore, personalized intervention and treatment strategies tailored to specific BMI levels are necessary for optimal outcomes [35–37].

Individuals with lower education levels tend to have higher rates of depression compared to those with higher educational attainment. Education plays a protective role against depression through various socioeconomic pathways. Higher education levels are associated with better economic and social resources, which help individuals manage and mitigate depressive episodes. Education also influences socioeconomic position, leading to better employment opportunities, higher income, and greater social status, all contributing to lower depression rates [38, 39].

The relationship between glucose levels and depression involves multiple biological, psychological, and lifestyle factors. Insulin resistance and high blood sugar levels can stimulate the release of stress hormones like cortisol, linked to depression. Both high blood sugar levels and depression are associated with chronic inflammation, with inflammatory markers such as C-reactive protein (CRP) and interleukin-6 (IL-6) often elevated in individuals with both conditions [40, 41].

The complex relationship between depression and age was discovered, as illustrated in Fig. 4E. Depression was low among participants under the age of 50, as evidenced by their negative SHAP value. The SHAP value was positive for participants aged 50–75, with the highest value

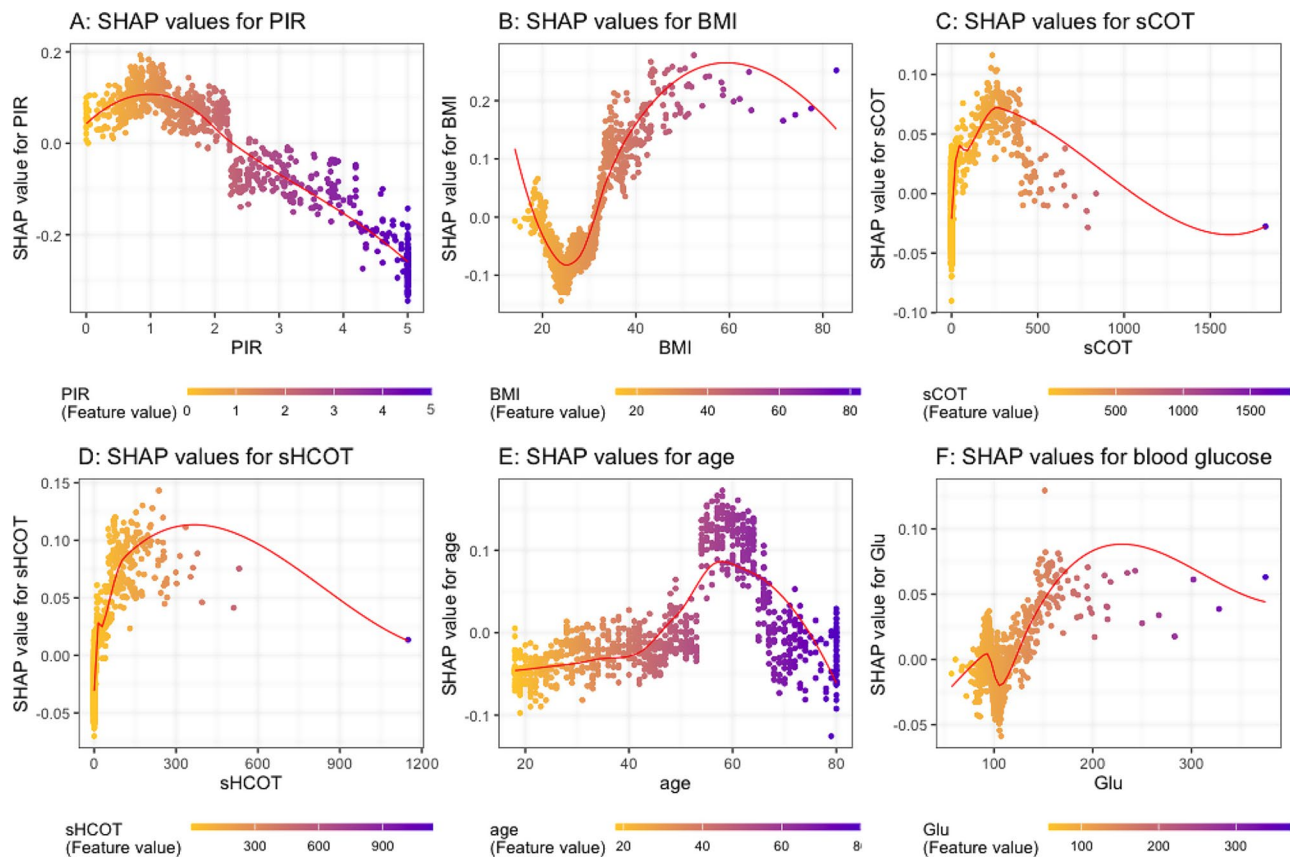


Fig. 4 The impact of numerical variables on depression

observed in the group of participants aged 60 (55–65 years) who were experiencing elevated levels of depression. The SHAP value becomes negative when the age exceeds 75 years.

Marital status is significantly associated with the prevalence of major depression. Married individuals generally report lower rates of depression compared to those who are single, divorced, or widowed. Married individuals often report higher levels of subjective well-being, which is associated with lower depression rates [42–44].

Poor kidney function, indicated by lower eGFR, is associated with higher rates of depression. Individuals with chronic kidney disease (CKD) are more likely to experience depressive symptoms compared to those with normal kidney function. The prevalence of depression increases as kidney function declines, with the highest rates observed in patients with end-stage renal disease (ESRD) undergoing dialysis [45, 46].

Cotinine and hydroxycotinine, biomarkers for nicotine exposure, are linked to depression through neuroinflammation, nicotine dependence, and oxidative stress. Understanding these relationships can help in developing targeted interventions for individuals with depression who are also exposed to nicotine [47–50].

Strength and limitation

In summary, the study utilized data from NHANES 2013–2014, which is a robust and comprehensive dataset including a wide range of demographic, health, and laboratory information. This broad dataset enabled a detailed and multifaceted analysis of depression predictors. Additionally, the use of SHAP values for model interpretation was a significant strength, providing clear and understandable explanations of how each feature contributes to the model's predictions. This helped identify the most important variables influencing depression risk. Furthermore, the study successfully identified key risk factors for depression, such as PIR, education level, marital status, age, sex, hypertension, BMI, blood glucose, eGFR and nicotine products. This comprehensive identification can provide the basis for targeted public health interventions.

However, this study has several limitations that should be acknowledged. First, as it is based on cross-sectional data, the findings cannot establish causal relationships between the identified risk factors and depression, which limits the ability to draw conclusions about the temporal dynamics of these associations. Longitudinal studies are necessary to confirm these relationships and understand their progression over time. Second, some variables, particularly those related to socioeconomic status,

physical activity, smoking, and drinking habits, were self-reported, introducing the potential for recall bias and inaccuracies. Third, the use of imputation to handle missing data, while effective, may not fully capture the true values, potentially leading to residual bias. Additionally, the exclusion of variables with over 50% missing data might have resulted in the loss of important information. Finally, addressing class imbalance through random undersampling, while effective in balancing the dataset, may limit the generalizability of the findings to broader populations where the prevalence of depression is lower. Furthermore, undersampling may introduce bias by underrepresenting specific subgroups or feature combinations within the control group, which could affect the robustness and interpretability of the model.

Conclusion

Our study demonstrates the advantages of leveraging ML models to predict depression by using comprehensive datasets. By identifying key risk factors, these models provide valuable insights into the multifaceted nature of depression and emphasize the significance of considering socioeconomic, demographic, and health-related factors in understanding and addressing this complex condition.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-025-02903-1>.

Supplementary Material 1

Acknowledgements

We would like to express our gratitude to all the participants who generously volunteered for the National Health and Nutrition Examination Survey. This study was supported by Japan Science and Technology Agency (JST) COI-NEXT 315 Grant number JPMJPF2018 to M.A.

Author contributions

Study concept and design: T.V., M.A.; data analysis and interpretation: T.V., M.Y.; drafting of the manuscript: T.V.; supervision: M.A.; reviewing and editing: T.V., R.D., T.J.T., N.W., Y.K., A.O., N.H.P.T., M.A. All authors critically revised and approved the final version of the manuscript.

Funding

This study was supported by Japan Science and Technology Agency (JST) COI-NEXT 315 Grant number JPMJPF2018 to M.A.

Data availability

The datasets used in this study are publicly available on the official CDC's website: https://www.cdc.gov/nchs/nhanes/about/erb.html?CDC_AAref_Val=https://www.cdc.gov/nchs/nhanes/irba98.htm.

Declarations

Ethics approval and consent to participate

Ethics approval for this study was granted by the National Centre for Health Statistics Research Ethics Review Board (Protocol # 2013-14). Since this study involves secondary data analysis, the original informed consent provided during primary data collection included permission for secondary use, eliminating the need for additional participant consent. Participants' privacy was protected by anonymizing or de-identifying the data to prevent

identification. Further details on NHANES ethics approval are available on the CDC's official website: https://www.cdc.gov/nchs/nhanes/about/erb.html?CDC_AAref_Val=https://www.cdc.gov/nchs/nhanes/irba98.htm.

Consent for publication

Not applicable.

Relevant guidelines and regulations

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, 3-17 Senrioka-shinmachi, Osaka, Settsu 566-0002, Japan

²Department of Cardiac Surgery, Cardiovascular Center, Cho Ray Hospital, Ho Chi Minh, Vietnam

³Graduate School of Medicine, Kobe University, 54 Shogoin-Kawaharacho, Sakyo-ku, Kyoto 606-8507, Japan

⁴Graduate School of Science, Technology and Innovation, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe, Hyogo 657-8501, Japan

⁵Department of Preventive Cardiology, National Cerebral and Cardiovascular Center, Suita, Osaka 564-8565, Japan

Received: 9 August 2024 / Accepted: 29 January 2025

Published online: 17 February 2025

References

1. Steger MF, Kashdan TB. Depression and everyday social activity, belonging, and well-being. *J Couns Psychol.* 2009;56(2):289–300. <https://doi.org/10.1037/a0015416>.
2. Santomauro DF et al. Nov., Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021, [https://doi.org/10.1016/S0140-6736\(21\)02143-7](https://doi.org/10.1016/S0140-6736(21)02143-7)
3. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/depression>.
4. Reddy MS. Depression: the disorder and the Burden. *Indian J Psychol Med.* Jan. 2010;32(1):1–2. <https://doi.org/10.4103/0253-7176.70510>.
5. Vu T, et al. Machine learning approaches for stroke risk prediction: findings from the Suita Study. *J Cardiovasc Dev Dis.* Jul. 2024;11:207. <https://doi.org/10.3390/jcdd11070207>.
6. Martin-Morales A, Yamamoto M, Inoue M, Vu T, Dawadi R, Araki M. Predicting Cardiovascular Disease Mortality: Leveraging Machine Learning for Comprehensive Assessment of Health and Nutrition Variables. *Nutrients.* Sep. 2023;15(18):3937. <https://doi.org/10.3390/nu15183937>.
7. Thanh NT, Luan VT, Viet DC, Tung TH, Thien V. A machine learning-based risk score for prediction of mechanical ventilation in children with dengue shock syndrome: a retrospective cohort study. *PLoS ONE.* Dec. 2024;19(12):e0315281. <https://doi.org/10.1371/journal.pone.0315281>.
8. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci Rep.* Jan. 2021;11(1):1980. <https://doi.org/10.1038/s41598-021-81368-4>
9. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. in *Artificial Intelligence in Healthcare*. Elsevier; 2020:25–60. <https://doi.org/10.1016/B978-0-12-818438-7.00002-2>.
10. Nickson D, Meyer C, Walasek L, Toro C. Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. *BMC Med Inf Decis Mak.* Nov. 2023;23(1):271. <https://doi.org/10.1186/s12911-023-02341-x>.
11. Squires M, et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inf.* Dec. 2023;10(1). <https://doi.org/10.1186/s40708-023-00188-6>.
12. Tomitaka S, et al. Distributional patterns of item responses and total scores on the PHQ-9 in the general population: data from the National Health and Nutrition Examination Survey. *BMC Psychiatry.* Dec. 2018;18(1):108. <https://doi.org/10.1186/s12888-018-1696-9>.

13. Sun Y, Fu Z, Bo Q, Mao Z, Ma X, Wang C. The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital. *BMC Psychiatry*. Dec. 2020;20(1):474. <https://doi.org/10.1186/s12888-020-02885-6>.
14. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. *J Gen Intern Med*. Sep. 2001;16(9):606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
15. Wright MN, Ziegler A. Ranger: a fast implementation of Random forests for high Dimensional Data in C++ and R. *J Stat Softw*. 2017;77(1). <https://doi.org/10.18637/jss.v077.i01>.
16. Boateng EY, Abayie DA. A review of the logistic regression model with emphasis on Medical Research. *J Data Anal Inform Process*. 2019;07(04):190–207. <https://doi.org/10.4236/jdaip.2019.74012>.
17. Langarizadeh M, Moghbeli F. Applying naive bayesian networks to Disease Prediction: a systematic review. *Acta Informatica Med*. 2016;24(5):364. <https://doi.org/10.5455/aim.2016.24.364-369>.
18. Son Y-J, Kim H-G, Kim E-H, Choi S, Lee S-K. Application of support Vector Machine for Prediction of Medication Adherence in Heart failure patients. *Healthc Inf Res*. 2010;16(4):253. <https://doi.org/10.4258/hir.2010.16.4.253>.
19. Unnikrishnan P, Kumar DK, Poosapadi Arjunan S, Kumar H, Mitchell P, Kawasaki R. Development of Health parameter model for risk prediction of CVD using SVM. *Comput Math Methods Med*. 2016;2016:1–7. <https://doi.org/10.1155/2016/3016245>.
20. Bader M, Abdelwanis M, Maalouf M, Jelinek HF. Detecting depression severity using weighted random forest and oxidative stress biomarkers. *Sci Rep*. Jul. 2024;14(1):16328. <https://doi.org/10.1038/s41598-024-67251-y>.
21. Xin Y, Ren X. Predicting depression among rural and urban disabled elderly in China using a random forest classifier. *BMC Psychiatry*. Feb. 2022;22(1):118. <https://doi.org/10.1186/s12888-022-03742-4>.
22. Dong T, et al. Cardiac surgery risk prediction using ensemble machine learning to incorporate legacy risk scores: a benchmarking study. *Digit Health*. Jan. 2023;9. <https://doi.org/10.1177/20552076231187605>.
23. Yang H, Chen Z, Yang H, Tian M. Predicting Coronary Heart Disease using an Improved LightGBM Model: performance analysis and comparison. *IEEE Access*. 2023;11:23366–80. <https://doi.org/10.1109/ACCESS.2023.3253885>.
24. Zhang Z, Jackson SL, Gillespie C, Merritt R, Yang Q. Depressive symptoms and mortality among US adults. *JAMA Netw Open*. Oct. 2023;6(10):e2337011. <https://doi.org/10.1001/jamanetworkopen.2023.37011>.
25. Yi H, et al. Nonlinear associations between the ratio of family income to poverty and all-cause mortality among adults in NHANES study. *Sci Rep*. May 2024;14(1):12018. <https://doi.org/10.1038/s41598-024-63058-z>.
26. Albert PR. Why is depression more prevalent in women? *J Psychiatry Neurosci*. Jul. 2015;40(4):219–221. <https://doi.org/10.1503/jpn.150205>
27. Zare H, Meyerson NS, Nwankwo CA, Thorpe RJ. How Income and Income Inequality Drive depressive symptoms in U.S. adults, does sex matter: 2005–2016. *Int J Environ Res Public Health*. May 2022;19(10):6227. <https://doi.org/10.3390/ijerph19106227>.
28. Freeman EW. Treatment of depression associated with the menstrual cycle: premenstrual dysphoria, postpartum depression, and the perimenopause. *Dialogues Clin Neurosci*. Jun. 2002;4(2):177–91. <https://doi.org/10.31887/DCN.5.2002.4.2/efreeman>.
29. Thayer JF, Yamamoto SS, Brosschot JF. The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors. *Int J Cardiol*. May 2010;141(2):122–31. <https://doi.org/10.1016/j.ijcard.2009.09.543>.
30. Golbidi S, Frisbee JC, Laher I. Chronic stress impacts the cardiovascular system: animal models and clinical outcomes. *Am J Physiol Heart Circ Physiol*. Jun. 2015;308(12):H1476–H1498. <https://doi.org/10.1152/ajpheart.00859.2014>
31. Meng L, Chen D, Yang Y, Zheng Y, Hui R. Depression increases the risk of hypertension incidence. *J Hypertens*. May 2012;30(5):842–51. <https://doi.org/10.1097/HJH.0b013e32835080b7>.
32. Hamer M, Endrighi R, Venuraju SM, Lahiri A, Steptoe A. Cortisol responses to Mental stress and the progression of coronary artery calcification in healthy men and women. *PLoS ONE*. Feb. 2012;7(2):e31356. <https://doi.org/10.1371/journal.pone.0031356>.
33. Li C, Li X, Li Y, Niu X. The Nonlinear Relationship Between Body Mass Index (BMI) and Perceived Depression in the Chinese Population. *Psychol Res Behav Manag*. Jun. 2023;16:2103–2124. <https://doi.org/10.2147/PRBM.S411112>
34. Badillo N, Khatib M, Kahar P, Khanna D. Correlation between body Mass Index and Depression/Depression-Like symptoms among different genders and races. *Cureus*. Feb. 2022. <https://doi.org/10.7759/cureus.21841>.
35. Patsalos O, Keeler J, Schmidt U, Penninx BWJH, Young AH, Himmerich H. Diet, obesity, and Depression: a systematic review. *J Pers Med*. Mar. 2021;11(3):176. <https://doi.org/10.3390/jpm11030176>.
36. Luppino FS, et al. Overweight, obesity, and Depression. *Arch Gen Psychiatry*. Mar. 2010;67(3):220. <https://doi.org/10.1001/archgenpsychiatry.2010.2>.
37. Dalle Grave R, Sartirana M, Calugi S. Personalized cognitive-behavioural therapy for obesity (CBT-OB): theory, strategies and procedures. *Biopsychosoc Med*. Dec. 2020;14(1). <https://doi.org/10.1186/s13030-020-00177-9>.
38. Taple BJ, Chapman R, Schalet BD, Brower R, Griffith JW. The Impact of Education on Depression Assessment: Differential Item Functioning Analysis. *Assessment*. Mar. 2022;29(2):272–284. <https://doi.org/10.1177/1073191120971357>
39. Patria B. The longitudinal effects of education on depression: finding from the Indonesian national survey. *Front Public Health*. Oct. 2022;10. <https://doi.org/10.3389/fpubh.2022.1017995>.
40. Berk M, et al. So depression is an inflammatory disease, but where does the inflammation come from? *BMC Med*. Dec. 2013;11(1):200. <https://doi.org/10.1186/1741-7015-11-200>.
41. Hassamal S. Chronic stress, neuroinflammation, and depression: an overview of pathophysiological mechanisms and emerging anti-inflammatories. *Front Psychiatry*. May 2023;14. <https://doi.org/10.3389/fpsy.2023.1130989>.
42. Wadood MA, Karim MR, Md. A. S, Alim HM, Rana MM, Hossain MG. Factors affecting depression among married adults: a gender-based household cross-sectional study. *BMC Public Health*. Oct. 2023;23(1):2077. <https://doi.org/10.1186/s12889-023-16979-9>
43. Zhao L, Zhang K, Gao Y, Jia Z, Han S. The relationship between gender, marital status and depression among Chinese middle-aged and older people: mediation by subjective well-being and moderation by degree of digitization. *Front Psychol*. Oct. 2022;13. <https://doi.org/10.3389/fpsyg.2022.923597>.
44. Bullock AGM, Williams JVA, Lavorato DH, Patten SB. The depression and marital status relationship is modified by both age and gender. *J Affect Disord*. Dec. 2017;223:65–68. <https://doi.org/10.1016/j.jad.2017.06.007>
45. Wang W-L, et al. The prevalence of depression and the association between depression and kidney function and health-related quality of life in elderly patients with chronic kidney disease: a multicenter cross-sectional study. *Clin Interv Aging*. May 2019;14:905–13. <https://doi.org/10.2147/CIAS.203186>.
46. Gupta S, Patil N, Karishetti M, Tekkalaki B. Prevalence and clinical correlates of depression in chronic kidney disease patients in a tertiary care hospital. *Indian J Psychiatry*. 2018;60(4):485. https://doi.org/10.4103/psychiatry.IndianJPsychiatry_272_18.
47. El-Sherbiny N, Elsary A. Smoking and nicotine dependence in relation to depression, anxiety, and stress in Egyptian adults: a cross-sectional study. *J Family Community Med*. 2022;29(1):8. https://doi.org/10.4103/jfcm.jfcm_290_21.
48. Albarak DA et al. Dec., The Association Between Nicotine Dependence and Mental Health in the General Population of Saudi Arabia: A Cross-Sectional Analytical Study. *Int J Gen Med*. 2023;16:5801–5815. <https://doi.org/10.2147/IJGM.S429609>.
49. Bainter T, Selya AS, Oancea SC. A key indicator of nicotine dependence is associated with greater depression symptoms, after accounting for smoking behavior. *PLoS ONE*. May 2020;15(5):e0233656. <https://doi.org/10.1371/journal.pone.0233656>.
50. Pawlina MMC, Rondina RDC, Espinosa MM, Botelho C. Nicotine dependence and levels of depression and anxiety in smokers in the process of smoking cessation. *Rev Psiqiatr Clin*. Aug. 2014;41(4):101–5. <https://doi.org/10.1590/0101-60830000000020>.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.