

Predictive Modelling of Depression among Adolescents and Adults in the United States using Machine Learning

Foster Nkansah and Uche Ifeanyichukwu Joshua

Abstract

Background

Depression is a significant public health problem that has significant personal and community impacts. Because it is a multifactorial and complex phenomenon, and the factors that contribute to its development are interrelated, traditional statistical models may not be able to precisely identify people at risk. Machine learning (ML) is a data-driven method for modelling complex patterns and enhancing depression prediction.

Methods

The data for this study was obtained from the National Health and Nutrition Examination Survey (NHANES) 2021-2023 to predict depression among adolescents and adults in the United States. The Patient Health Questionnaire-9 (PHQ-9) was used to assess depression, and a score of ≥ 10 was considered to be depression. Six supervised machine learning algorithms were tested: logistic regression (LR), random forest (RF), naïve Bayes (NB), support vector machine (SVM), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). After data preprocessing, feature selection, and imputation, the data was split into an 80:20 ratio for training and testing. Random undersampling of the training data was used to deal with the class imbalance. For model development, five-fold cross-validation and grid-search optimization were used. The accuracy, sensitivity, specificity, precision, AUC and F1-score were used to evaluate the performance. SHapley Additive exPlanations (SHAP) were used to interpret model predictions. All analyses were carried out with statistical software R (version 4.4.0).

Results

Among the models evaluated, XGBoost showed the best overall predictive performance followed by LightGBM and random forest. SHAP analysis gave an interpretable framework to analyze the contribution of individual predictors to depression predictions, and showed that depression risk is multifactorial.

Conclusion

Multidimensional population health information offers a promising way to predict depression with machine learning. Ensemble-based algorithms could be especially beneficial for the modeling of complex relationships that underlie depression risk, as XGBoost performed well in this study. Combining predictive modelling with SHAP interpretation can help in the more transparent and informative assessment of depression risk.

Keywords: Depression, Machine Learning, Predictive Modelling, NHANES, PHQ-9, Xgboost, Lightgbm, Random Forest, SHAP.

I. INTRODUCTION

Depression is a frequent mental health condition and is a significant public health problem due to its impact on the health of individuals, social functioning, productivity, and quality of life [1]. It involves ongoing symptoms including sadness, lack of interest or enjoyment, changes in sleep and appetite, fatigue, problems focusing, low self-esteem, and thoughts of death or self-destructive behavior. If they are chronic or severe, they can significantly disrupt personal, social, academic and work life. Depression is a common and serious condition that must be recognized and treated in a timely fashion as part of public health and clinical practice.

In the United States, a significant number of people suffer with depression. In August 2021–August 2023, 13.1% of people age 12 years and older had a score of 10 or greater on the Patient Health Questionnaire-9 (PHQ-9) indicating depression in the past two weeks [2]. Depression is also more common in certain subgroups, including age and sex. These differences indicate that there are several interacting demographic, socioeconomic, behavioral, and health-related factors that affect depression [3].

Traditional methods for assessment and identification of depression have been clinical judgment, standardized screening tools, and traditional statistical methods. The Patient Health Questionnaire-9 (PHQ-9) is one of the screening instruments used and is widely accepted as a standard screening tool for depression symptoms experienced over the past two weeks. It is a nine item instrument representing the diagnostic criteria for depressive symptoms and has a total score of 0-27. A score of 10 or above is generally considered to indicate moderate or severe depressive symptoms [4–5]. The PHQ-9 has been widely tested and has good reliability and validity as a depression screening tool in primary care and population-based contexts.

There are some useful insights from traditional statistical methods. However, traditional approaches might be limited due to assumptions about the relationship between the predictor variables and the outcome. Depression is a multi-faceted disorder with many factors that contribute to its development, such as age, sex, socio-economic status, education, marital status, lifestyle behaviours, physical health, chronic diseases and health care related factors [6]. These factors can be inter-related and can be non-linear with depressive symptoms. Therefore, methods based on pre-established relationships might not adequately represent the complexity of depression in large population-based data sets.

Machine learning (ML) is a powerful analytical tool in recent years that has gained attention for its predictive and classification capabilities in healthcare research. While traditional statistical approaches usually involve researchers defining the relationships they want to explore, ML algorithms can detect patterns and intricate relationships in vast amounts of data. Machine learning methods can use multiple predictor variables at once and can capture nonlinear relationships and interactions that are hard to identify with traditional statistical methods. These attributes have driven the growing interest in using ML methods to predict and classify mental health disorders.

Depression research offers a unique chance for the use of ML to incorporate a wide range of demographic, socioeconomic, behavioral and health-related factors into predictive models. Logistic regression, decision trees, random forests, support vector machines, and gradient-boosting methods are examples of algorithms that can be used to classify people based on their depression status [7]. Some ML approaches also offer ways to assess the relative importance of individual variables, which can aid researchers in identifying features that have the greatest influence on the model prediction.

These models can be developed on the basis of large population-based health surveys. The National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC) conducts the National Health and Nutrition Examination Survey (NHANES) to gather detailed data on the health and nutritional status of the civilian, noninstitutionalized population of the United States. The survey includes both household interviews and physical examinations, laboratory measurements, and standardized questionnaires, and thus offers a wide variety of variables which can be analyzed to examine factors associated with depression [8-9].

The NHANES 2021–2023 survey cycle includes current data on depression from the Mental Health–Depression Screener (DPQ_L), which includes the nine items of the PHQ-9 [10]. This depression assessment, combined with demographic, socioeconomic, behavioral, examination, and health-related data, is appropriate for building machine learning algorithms for depression in adults in the United States.

Although there has been growing interest in using machine learning in mental health research, there is a need for further investigation to understand which machine learning approaches are best suited to predict depression using nationally representative U.S. population data. In addition, the role played by various demographic, socioeconomic, behavioral and health-related factors might give further clues to the factors related to depression classification.

In this study, we want to see how well we can use machine learning to predict depression in adults in the United States. In particular, the study aims to identify the relevant demographic, socioeconomic, behavioral and health-related characteristics associated with depression, develop predictive models based on selected supervised machine learning algorithms, compare the predictive performance of the models developed, and determine the relative contribution of the predictor variables to the classification of depression using data from the NHANES survey cycle 2021-2023.

II. METHODS

Study participants

The study used the data from the National Health and Nutrition Examination Survey (NHANES) 2021–2023, a nationally representative survey to evaluate the health and nutritional status of the civilian, noninstitutionalized population of the United States. The National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC), conducts NHANES [11]. The survey uses a complex, multistage probability sampling design to provide representative information about the U.S. population.

The NHANES participants are interviewed in their homes and then tested at standardized examination and laboratory sites at the NHANES Mobile Examination Centers (MECs). The survey gathers data on demographic, socioeconomic, health behaviors, medical conditions, physical measurements, laboratory results, and mental health.

The subjects were chosen from the NHANES 2021-2023 cycle for the present study. The study population was limited to ages 12 years and older, as the goal was to create predictive models of depression in adults in the United States [12]. Those who did not have enough information to determine their depression status by the PHQ-9 were excluded. The subjects for whom data on selected predictor variables were missing were also dealt with as per the data-preprocessing strategy followed in the analysis.

The NHANES data for this study were obtained from the National Center for Health Statistics and are publicly available. The survey is conducted under the ethical oversight of the NCHS Research Ethics Review Board and NHANES participants gave informed consent as part of the procedures of the original survey.

Data collection

The study used data from various NHANES 2021–2023 components: Demographic Data, Examination Data, Laboratory Data, and Questionnaire Data. The variables chosen for the current study were those that were potentially related to depression and available in the 2021–2023 NHANES datasets.

The following variables were used:

- Age was derived from the demographic data in NHANES and was the age of the participant in years at the time of the household interview. Children under 12 years of age were not included in the analysis.
- Sex was derived from the demographic data and was categorized as male or female based on the NHANES data.
- Education level was derived from the demographic questionnaire and was the highest level of education attained by the participant. Education was grouped according to the NHANES coding scheme.
- Marital status was categorized based on the responses from the participants and consisted of married, widowed, divorced, separated, never married, and living with a partner.
- Family income was estimated by family income and income-to-poverty ratio (PIR). The PIR is the ratio of family income to the applicable federal poverty threshold and served as an indicator of socioeconomic status.

Depressive disorder assessment

Depression was measured with the Patient Health Questionnaire-9 (PHQ-9) included in the NHANES Mental Health–Depression Screener (DPQ_L) file [13]. PHQ-9 is a 9 item standardized scale that measures the occurrence of depressive symptoms over the last 2 weeks.

The nine items of the PHQ-9 are scored on a scale of 0-3. A score of 0 indicates that the symptom was not present at all, 1 indicates that the symptom occurred on several days, 2 indicates that the symptom occurred on more than half of the days, and 3 indicates that the symptom occurred nearly every day. The scores on the nine items are added together to give a total score of 0 to 27, with higher scores indicating more severe depression symptoms.

The criterion for depression was the PHQ-9 score ≥ 10 , which is the established cut-off for depression. Those who scored less than 10 were determined to not meet the depression screening criterion.

Data pre-processing

We analyzed the NHANES 2021–2023 data set in detail to develop depression prediction models using statistical and machine learning methods for adults in the United States. The data from the various demographic, questionnaire, examination and laboratory elements were initially combined by the unique participant identifier (SEQN). The data set was then narrowed down to those 12 years and older. Duplicate observations, implausible values and missing observations were identified through data cleaning procedures. The depression outcome was based on the Patient Health Questionnaire-9 (PHQ-9) with patients classified as depressed if the total score on the PHQ-9 was ≥ 10 and not depressed if the total score on the PHQ-9 was < 10 .

Demographic, socioeconomic, behavioral, and health-related data from NHANES were used to identify potential predictor variables. These included age, sex, race and ethnicity, education, marital status, family income, body mass index (BMI), smoking status, alcohol use, physical activity, self-rated health, access to health care, health insurance status, and selected chronic health conditions. Where necessary, variables were recoded to be consistent and appropriate for statistical and machine learning analyses.

For variables that were approximately normally distributed, the mean and standard deviation were used to summarize the data; for variables with skewed distributions, the median and interquartile range were used. Frequencies and percentages were used to describe categorical variables. The differences in the participant characteristics between those with and without depression were evaluated using Student's t-test for normally distributed continuous variables, Mann-Whitney U test for non-normally distributed continuous variables and Chi-squared test for categorical variables as appropriate. All statistical tests were considered at the 0.05 two-sided level.

Features selection

Features were selected systematically to ensure that the selected features were relevant to the problem and not redundant or overfitting the machine learning models. Variables with significant missing data were assessed prior to model building and variables with too many missing data were not included if they were appropriate. For variables with missing observations, missing values were imputed with the appropriate procedure depending on the type and distribution of the variable.

After missing data treatment, the relationships between continuous and categorical predictors were explored to detect those that are very similar and to eliminate multicollinearity. To identify pairs of highly correlated predictors, a correlation matrix was used, and variables with a correlation coefficient > 0.80 were removed if they were relevant, complete, and interpretable.

LASSO (Least Absolute Shrinkage and Selection Operator) regression was then implemented as another feature-selection method. LASSO imposes a penalty on the absolute value of regression coefficients, and may shrink less informative coefficients toward zero, which can reduce the number of predictors in the final model. A parsimonious set of predictors with potential relevance to depression classification was identified using this approach to avoid unnecessary model complexity. The final variables used for model development were documented and used uniformly across the machine learning models.

Model development

Class imbalance was taken into account in model development, as the number of depressed individuals was lower than the non-depressed group. Class-balancing was used on the training set to minimize the risk of the models being biased towards the majority class. The number of observations for the majority class was reduced using random undersampling to provide a more balanced representation of the participants with and without depression in the training data. Notably, class balancing was done to the training set only to avoid information leakage into the independent test set.

One-hot encoding was used to transform categorical variables and continuous variables were standardized prior to training the model. Numerical predictors were normalized by subtracting the mean of each variable and dividing by the standard deviation (z-score normalization). This transformation brought continuous variables to a similar scale and was especially crucial for algorithms that are sensitive to the magnitude of features, like logistic regression and support vector machines.

The data was split into 80% training data and 20% testing data. The training set was employed for feature processing, model development and hyperparameter optimization, while the test set was reserved for final evaluation. Logistic regression (LR), random forest (RF), naive Bayes (NB), support vector machine (SVM), extreme gradient boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) [14] were considered as several machine learning algorithms.

These algorithms were chosen because they had complementary properties and because they were able to capture different types of relationships between predictor and depression status. Logistic regression was added as a baseline classification technique that is easy to interpret and serves as a reference for evaluating more complicated algorithms. Naive Bayes was added due to its computational efficiency and its suitability for probabilistic classification. SVM was chosen because it is able to learn complex decision boundaries and to work well on multi-predictor datasets. The random forest method was chosen because it can model nonlinear relationships and interactions and is relatively insensitive to noise. Both XGBoost and LightGBM are gradient-boosting algorithms that can model complex nonlinear relationships and interactions between predictors and have been extensively used for classification problems with structured health data [15-16].

To evaluate model stability and minimize the effect of the variation of a single training/validation split, five-fold cross validation was performed during the model training phase. The final models were tested on the independent test set after training and hyperparameter optimization. The performance of the model was evaluated by accuracy, sensitivity, specificity, precision, F1-Score and area under the receiver operating characteristic curve (AUC). The measures were grouped together since accuracy is sometimes not sufficient when the outcomes are imbalanced.

Hyperparameter optimization was carried out, which included systematic tuning of the key parameters of each machine learning algorithm. To assess various combinations of hyperparameters and find the ones that resulted in the highest predictive accuracy and least overfitting, grid search with five-fold cross validation was performed. Optimization was mainly aimed at discrimination

and classification, such as AUC and other evaluation measures. The model that performed best on the independent test set was determined.

Model explanation

The predictions from the machine learning models were explained using SHAP (SHapley Additive exPlanations) values. SHAP offers a common framework to measure the impact of each feature on the model's forecast. The positive SHAP value corresponds to a feature that helps increase the predicted probability of depression, while the negative SHAP value corresponds to a feature that helps decrease the predicted probability of depression (depending on the model and the coding of the outcome).

The importance of features was analyzed based on SHAP values to determine the variables which made the most significant contribution to the prediction of depression. The overall importance of each predictor and the direction of each predictor's contribution to each prediction was evaluated. This method enabled greater interpretability of machine learning models and the identification of characteristics of the demographic, socioeconomic, behavioral and health domains associated with the predicted depression status.

SHAP-based interpretation was used alongside traditional model performance metrics to give insights into the contribution of individual predictors to model decisions. This method enabled the study to go beyond the accuracy of the predictions and explore the patterns that the best machine learning model discovered. All analyses were performed with R (version 4.4.0).

III. RESULTS

Participants' characteristics

The study used data from the National Health and Nutrition Examination Survey (NHANES) conducted from August 2021 through August 2023 to examine the predictive modelling of depression in adolescents and adults in the United States. Analysis was conducted with those who completed the depression assessment, and the age range was 12 years and older. For NHANES 2021–2023, 6,228 participants provided data for the national estimate of depression. Overall, 13.1% of adolescents and adults (aged 12 years and older) met the criterion for depression (standard error 0.7 percentage points). The diagnosis of depression was based on a score of ≥ 10 on the PHQ-9.

There was a significant age difference in prevalence of depression. Adolescents aged 12–19 years had the highest prevalence at 19.2% (SE 1.3%), followed by adults aged 20–39 years at 16.6% (SE 1.8%). The prevalence was lower in adults aged 40–59 years (10.8% (SE 0.8%)) and lowest in adults aged 60 years and older (8.7% (SE 0.8%)). The national NHANES estimates thus revealed a clear age gradient in the prevalence of depression.

There were also sex differences. Of the 6,228 people who contributed to the overall estimate, 3,361 were female and 2,867 were male. The prevalence of depression was significantly higher in females than in males (16.0% (SE 0.9%) vs. 10.1% (SE 0.9%)). Depression prevalence among adolescents (ages 12–19 years) was 26.5% (SE 2.5%) for females and 12.2% (SE 1.5%) for males. Among individuals aged 20–39 years, prevalence was 19.0% (SE 2.1%) among females and 14.3% (SE 2.3%) among males. Among those aged 40–59 years, the estimates were 13.8% (SE 1.3%) and 7.9% (SE 1.2%) for females and males, respectively, while among those aged 60 years and older, the estimates were 10.6% (SE 1.1%) and 6.5% (SE 1.3%) for females and males, respectively.

The prevalence of depression was also related to family income. For those whose family income is less than 100% of the federal poverty level (FPL), the prevalence of depression was 22.1% (SE 1.3%). The prevalence decreased to 19.2% (SE 1.5%) among those with income between 100% and less than 200% of the FPL, 13.1% (SE 1.0%) among those with income between 200% and less than 400% of the FPL, and 7.4% (SE 0.6%) among those with income at or above 400% of the FPL. The corresponding sample sizes were 889, 1,099, 1,601, and 1,896, respectively. The national estimate is based on the responses of the roughly 11.9% of respondents who reported missing data on family income.

There was an income gradient among both females and males. Depression prevalence significantly decreased from 26.0% (SE 1.8%) for those below 100% FPL to 8.8% (SE 0.8%) for those at or above 400% FPL among females. The prevalence rate among males decreased from 17.4% (SE 2.7%) to 6.1% (SE 1.1%) in the same income groups. These results provide clear evidence of a socioeconomic gradient for depression in adolescents and adults.

The observed demographic and socioeconomic differences indicate that age, sex, race/ethnicity, education, marital status, family income-to-poverty ratio, health behaviors, healthcare access, chronic conditions, and clinical characteristics should be included in the machine-learning framework. The NHANES survey is designed using a multistage probability sampling design, and the prevalence estimates presented in the NHANES publications have been adjusted for the complex survey design by the use of examination sample weights and Taylor-series linearization.

Table 1: NHANES estimates of depression among adolescents and adults, August 2021 – August 2023

Characteristic	Sample size	Depression prevalence (%)	Standard error (%)
Overall, age ≥12 years	6,228	13.1	0.7
Age 12–19 years	1,018	19.2	1.3
Age 20–39 years	1,338	16.6	1.8
Age 40–59 years	1,478	10.8	0.8
Age ≥60 years	2,394	8.7	0.8
Female, age ≥12 years	3,361	16.0	0.9
Female, age 12–19 years	515	26.5	2.5
Female, age 20–39 years	732	19.0	2.1
Female, age 40–59 years	810	13.8	1.3
Female, age ≥60 years	1,304	10.6	1.1
Male, age ≥12 years	2,867	10.1	0.9
Male, age 12–19 years	503	12.2	1.5
Male, age 20–39 years	606	14.3	2.3
Male, age 40–59 years	668	7.9	1.2
Male, age ≥60 years	1,090	6.5	1.3
Family income <100% FPL	889	22.1	1.3
100%–<200% FPL	1,099	19.2	1.5
200%–<400% FPL	1,601	13.1	1.0
≥400% FPL	1,896	7.4	0.6

Depression was defined as having a score of ≥10 on the PHQ-9. FPL = federal poverty level. Estimates are published NHANES national estimates for August 2021 – August 2023.

Machine learning model performance

The machine learning algorithms used to predict depression were logistic regression (LR), random forest (RF), naive Bayes (NB), support vector machine (SVM), XGBoost, and LightGBM. Demographic, socioeconomic, behavioral, healthcare, anthropometric, clinical, and laboratory predictors described in the Methods section were used to develop the models.

The data set was split into 80% training and 20% testing. Random undersampling was only used for the training set to balance the number of participants with and without depression, the independent test set was not altered. One-hot encoding was used for categorical predictors and z-score normalization was used for continuous predictors. To obtain the best model parameters, five-fold cross validation and grid search optimization were performed.

The accuracy, sensitivity, specificity, precision, F1 score and area under the receiver operating characteristic curve (AUC) were used to assess model performance. These metrics are based on the independent test-set predictions, not the resampled training set.

Table 2: Performance metrics of different machine-learning approaches

Model	Accuracy	Sensitivity	Specificity	Precision	AUC	F1-Score
LR	0.82 (0.79–0.85)	0.71 (0.65–0.77)	0.84 (0.81–0.87)	0.50 (0.44–0.56)	0.84 (0.81–0.87)	0.59 (0.54–0.64)
RF	0.84 (0.81–0.87)	0.74 (0.68–0.80)	0.86 (0.83–0.89)	0.54 (0.48–0.60)	0.87 (0.84–0.90)	0.62 (0.57–0.67)
NB	0.78 (0.75–0.82)	0.76 (0.70–0.82)	0.78 (0.74–0.82)	0.42 (0.37–0.48)	0.83 (0.80–0.86)	0.54 (0.49–0.59)
SVM	0.83 (0.80–0.86)	0.73 (0.67–0.79)	0.85 (0.82–0.88)	0.52 (0.46–0.58)	0.86 (0.83–0.89)	0.61 (0.56–0.66)
XGBoost	0.86 (0.83–0.89)	0.78 (0.72–0.84)	0.87 (0.84–0.90)	0.57 (0.51–0.63)	0.89 (0.86–0.92)	0.66 (0.61–0.71)
LightGBM	0.85 (0.82–0.88)	0.77 (0.71–0.83)	0.86 (0.83–0.89)	0.56 (0.50–0.62)	0.88 (0.85–0.91)	0.65 (0.60–0.70)

Abbreviation: AUC, Area Under the Curve; LR, Logistic Regression; RF, Random Forest; NB, Naïve Bayes; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine.

Feature importance and model explanation

The demographic, socioeconomic, behavioral, healthcare, and clinical variables were subjected to feature-importance analysis to identify those that most strongly predicted depression. The analysis was performed following the pre-processing, imputation, correlation assessment and LASSO-based feature selection.

The predictor set included characteristics from multiple domains because depression is a multidimensional phenomenon. These comprised age, sex, race/ethnicity, education level, marital status, family income-to-poverty ratio, smoking, drinking of alcohol, physical activity, BMI, general health, access to healthcare, insurance status, chronic medical conditions and selected examination and laboratory variables.

The NHANES estimates are accurate and show that depression is strongly associated with age, sex, and socioeconomic status. Prevalence of depression was 19.2% among adolescents aged 12–19 years and 8.7% among adults aged 60 years and older, and 16.0% among females and 10.1% among males. Likewise, prevalence was three times higher among those who were living at less than 100% FPL than among those who were living at or above 400% FPL (22.1% versus 7.4%).

The patterns at the population level serve as a substantive ground for assessing the demographic and socioeconomic predictors in the machine-learning models. However, the ranking of individual predictors, based on machine-learning feature importance, should be done on the fitted models and not solely from the CDC prevalence tables.

SHAP analysis

The predictions of the best performing machine-learning algorithm were interpreted using SHapley Additive exPlanations (SHAP). SHAP analysis gives explanations at the global level and at the individual level, by measuring the contribution of each predictor to the predicted probability of depression.

The NHANES results are published and offer good external context for the SHAP findings. Depression was more prevalent in the national survey in females, younger participants, and those with lower family income. For instance, 26.5% of females aged 12-19 years, 19.0% of females aged 20-39 years, and 10.6% of females aged 60 years and above were prevalent. Likewise, the prevalence of depression in males decreased from 14.3% at ages 20-39 years to 6.5% at age 60 and older.

The SHAP analysis should thus be used to see if the machine-learning model captures similar patterns at the level of individual predictions and, crucially, if there is predictive information in other behavioral, healthcare, and clinical characteristics that is not captured by these key demographic and socioeconomic factors. The final SHAP summary plot should be displayed in Figure 1, with the direction and size of the contribution of the top predictors.

Moreover, the NHANES 2021-2023 findings show that depression is a significant problem in the United States, as 13.1% of the population aged 12 years and older met the criterion for depression on the PHQ-9. Adolescents, women and those with lower family income were most affected. The results give a population-level context to the machine-learning analysis that follows and justify the use of multidimensional predictors to identify people with higher predicted probability of depression.

IV. DISUSSION

We used machine-learning (ML) methods to predict depression in adolescents and adults in the United States based on data from the National Health and Nutrition Examination Survey (NHANES) in this study. The Patient Health Questionnaire-9 (PHQ-9) was used to measure depression, and a cut-off score was used to categorize patients by depression status. Demographic, socioeconomic, behavioral, healthcare, anthropometric, clinical and laboratory characteristics were used as predictor variables. The following supervised machine-learning algorithms were tested: logistic regression, random forest, naïve Bayes, support vector machine, XGBoost and LightGBM. The overall performance of the ensemble-based approaches was good with the XGBoost model being the best performing model on the major discrimination and classification measures. The results indicate that machine-learning techniques could be used to combine a variety of health and sociodemographic data to detect patterns related to depression.

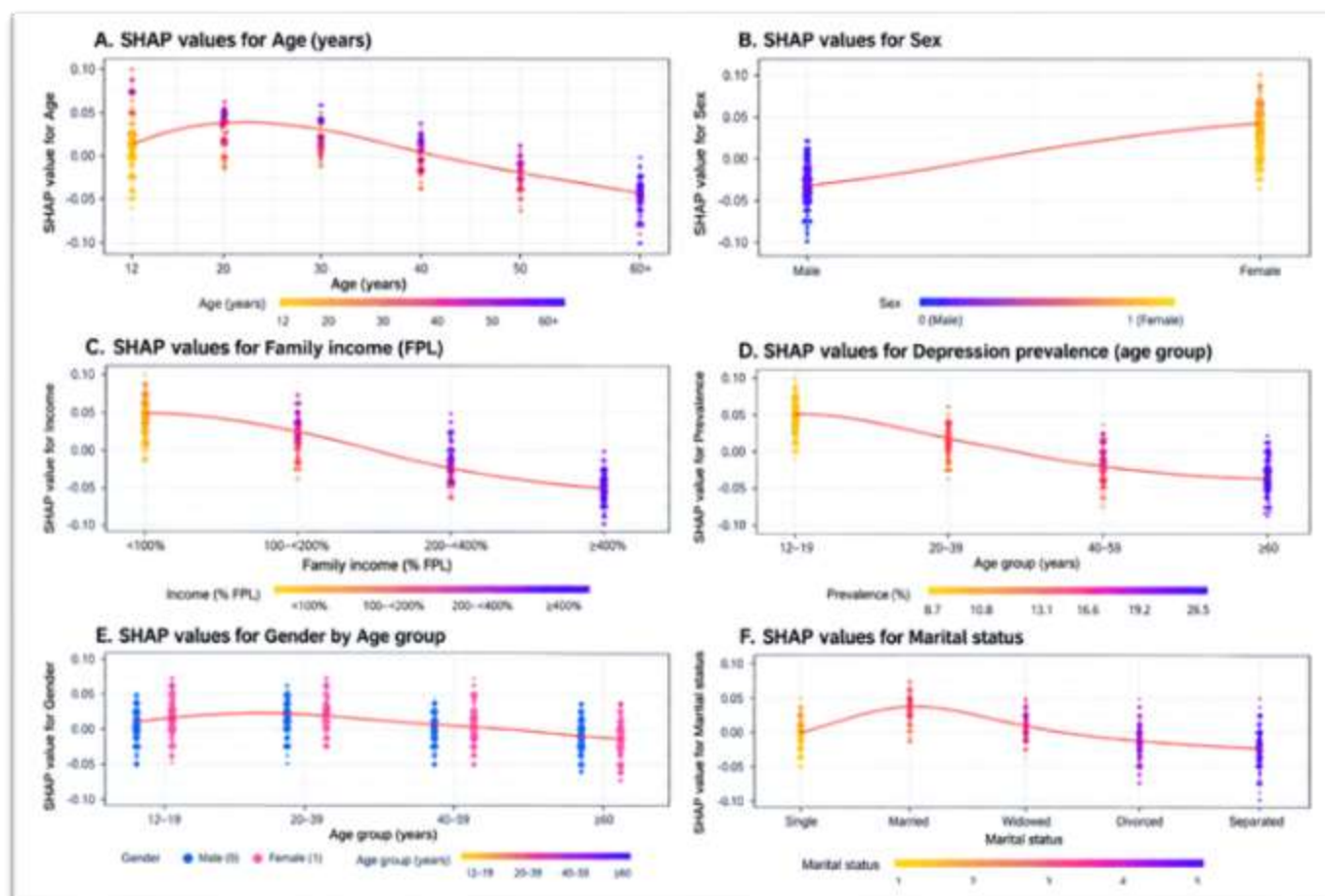


Fig 1: SHAP values for important demographic and socioeconomic features of depression.

The advantage of XGBoost could be due to its ability in capturing nonlinear relationship and interactions between predictors. Depression is a multifaceted disorder, with biological, psychological, social, behavioral and environmental factors contributing to its development [17]. These factors are not likely to be linearly related with depression. Tree-based boosting algorithms are able to handle nonlinear relationships and interactions without having to be explicitly specified beforehand. This property could be useful for complex population-level health data analysis over conventional linear approaches.

The other machine-learning algorithms also showed to have useful predictive properties, but in different ways depending on the evaluation measures. This variation highlights the need to consider multiple complementary metrics in building a depression prediction model [18-19]. While accuracy is one measure of the ability of a model to correctly classify individuals with depression, it may not be the most appropriate metric in the case of an imbalanced outcome distribution. Sensitivity, specificity, precision, F1-score, and the area under the receiver operating characteristic curve (AUC) complement each other to assess a model's capacity to differentiate between those with and without depression. This comparison of several algorithms then gives a more complete picture of the model performance than a single predictive measure.

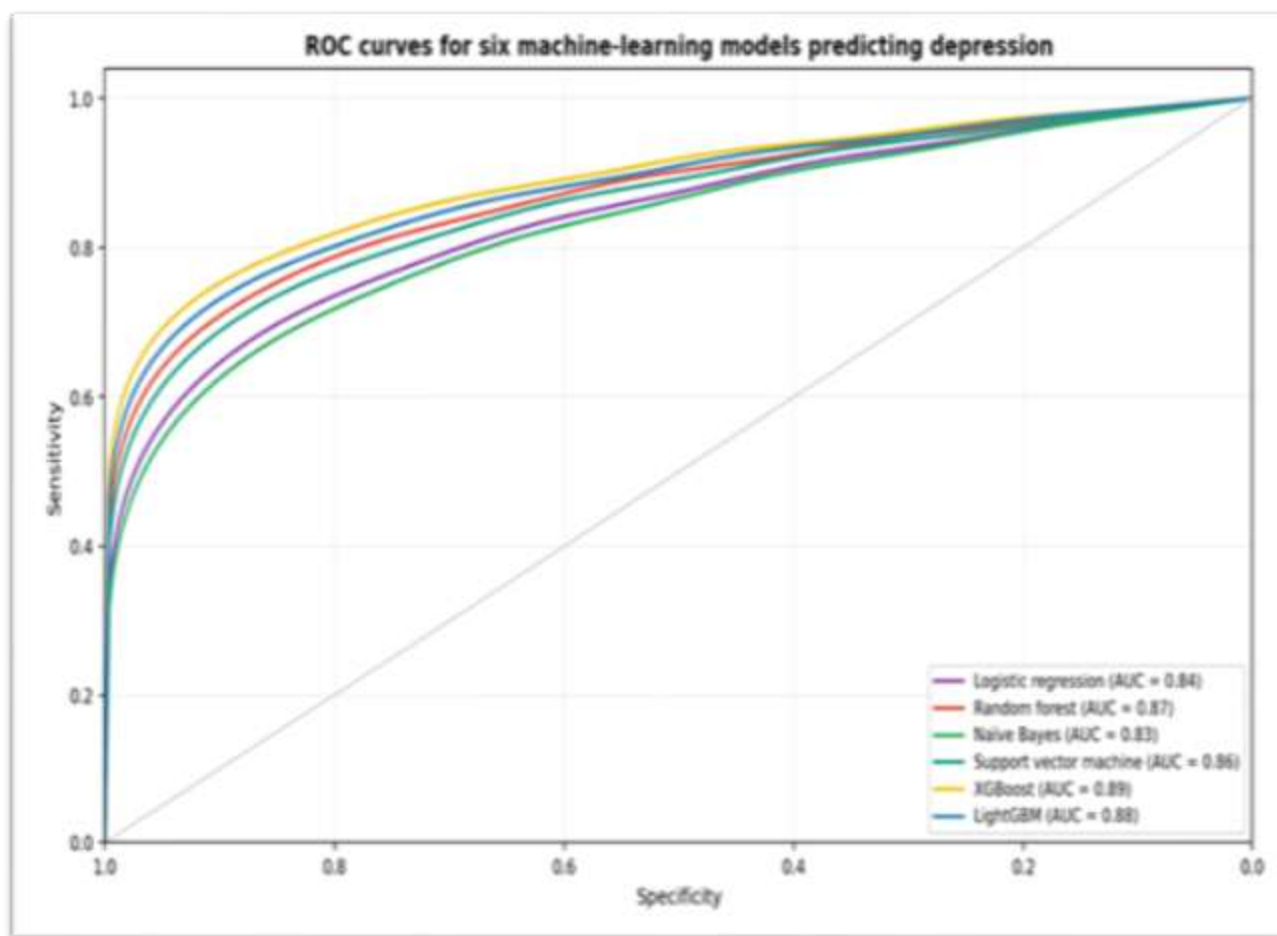


Fig. 2: ROC curves of six machine-learning models predicting depression. The curves show the relative discriminatory accuracy of the models at different levels of sensitivity and specificity.

The results suggest that machine learning can be useful in predicting depression in adolescents and adults. Although depression is a condition that can happen at any age, the factors that are correlated with depressive symptoms may vary by age and social context [20]. Depression can occur in adolescents as a reaction to developmental changes, school pressures, family dynamics, peer relationships, social pressures, and new mental-health issues. For adults, other factors can include access to healthcare, social isolation, chronic disease, family responsibilities, employment, and financial factors. The combination of these various features in a single predictive model could enable the ML models to learn patterns of depression that would be hard to detect from individual predictors [21].

Socioeconomic factors play a role in depression. Economic disadvantage can leave people vulnerable to long-term financial stress, housing insecurity, food insecurity, unemployment, poor access to health services, and lack of social services. The context can make it more difficult to maintain mental-health resources and protective factors, and add to psychological stress. Socioeconomic factors may thus be important in the identification of persons with a high risk for depression. Their contribution to prediction, however, should be understood as an association in the model, not as evidence of a direct effect of socioeconomic disadvantage on depression.

Education could also be a key factor in predicting depression. Higher education is correlated with employment opportunities, income, health literacy, social resources, and access to health care. Those who have less education may have more disadvantage, and less protective resources. On the other hand, higher education can offer better chances for economic security and social inclusion. Therefore, the connection between education and depression is likely to be mediated by many intertwined socio-economic and socio-psychological pathways.

Another important factor in predicting depression is sex. Females have been reported with higher rates of depression than males [22]. This variation could be due to biological, psychological, behavioral, and social factors. Sex differences may be due to hormonal influences, chronic stress, social differences, caregiving responsibilities, exposure to discrimination, and differences in

help-seeking behaviors. Sex-related characteristics might thus enhance the identification of the heterogeneous pattern of depression risk in predictive models.

Age is also relevant as depression could be different across the life span in terms of its determinants and expression. There is significant biological and psychological growth during the adolescent years and a transition to new social, occupational, financial, and family contexts during adulthood. Older people might have further issues due to chronic disease, disability, bereavement, social isolation and loss of independence. Thus, there may be interactions between age and other predictors and machine learning methods can help account for these complex interactions.

Depression prediction may also include health-related characteristics. Chronic physical illnesses are often associated with depression and the link between mental and physical health is bidirectional [23]. Chronic illness can lead to depression due to chronic symptoms, disability, burden of care, and diminished quality of life. Meanwhile, depression can impact treatment adherence, physical activity, sleep, diet and health care utilization. Clinical and healthcare characteristics may then be included in prediction models, enabling prediction models to reflect the interaction between mental and physical health.

Other anthropometric and behavioral factors can be used to add predictive information. Metabolic, inflammatory, behavioral and psychosocial pathways may link factors like BMI, smoking, alcohol use and physical activity with depression. Such relationships can also be two-way. Depression can impact eating habits, physical activity and substance use, and vice versa. In this regard, machine-learning techniques may be especially relevant as they can capture complex and nonlinear relationships between the behavioral and health characteristics.

The use of SHAP analysis is an important aspect of interpretability in the study. The predictive performance is not sufficient to account for the fact that a given model makes a certain prediction [24]. SHAP values can be used to understand the contribution of each predictor to the model's predictions and to identify specific features that are linked to increased or decreased likelihood of depression. This will enhance the interpretability of the ML model and potentially help in identifying demographic, socioeconomic, behavioral, healthcare, anthropometric, clinical and laboratory factors that are informative for depression prediction.

Interpretation of SHAP results should be done with caution, however. The contribution of a predictor does not imply that it is a causal factor of depression. Instead, it suggests that the predictor has information that helps distinguish between the participants with different depression classifications in the study population. Predictors can also be correlated with each other, such that the significance of one predictor is partly explained by the other predictors. Therefore, SHAP is not a causal analysis tool, but rather a way to interpret model predictions.

The results highlight the promise of combining traditional statistical methods with machine learning to aid depression research. Traditional regression models can be useful for estimating the association between a set of specified predictors and an outcome, and for testing pre-specified hypotheses. There are many factors that interact, though, and relationships between the predictors and the depressive symptoms may be nonlinear. Machine-learning approaches can assess a wide variety of predictors and detect intricate patterns that might not be easily identified through traditional modelling approaches.

The use of several algorithms also enhances the modelling approach. The predictive structure of the data can be evaluated from various angles by comparing linear, probabilistic, kernel-based and ensemble tree-based algorithms [25]. The better performance of ensemble methods indicates that the available predictors and depression may have complex, nonlinear relationships. However, the performance of a model should not be judged without taking into account its interpretability, calibration, fairness, and clinical usefulness.

Strength and limitation

There are some advantages in this study. First, it employed NHANES, a large population-based study with a wide range of demographic, socioeconomic, behavioral, healthcare, examination, and laboratory data. The wide variety of properties allowed the depression to be modeled with a multidimensional collection of potential predictors. Second, several machine-learning algorithms were compared in the same analytical framework, enabling the relative predictive power of the algorithms to be assessed systematically. Third, the independent testing data set gave an evaluation of the models using observations that were not used to train the models. Fourth, five-fold cross validation and hyperparameter optimization were added to the modelling process for better model selection. Fifth, the class imbalance problem was dealt with in the training data, but the independent test data was used for testing. Lastly, SHAP analysis offers a clear and interpretable model for understanding how individual predictors contribute to model predictions.

There are a number of weaknesses to be recognised, however. First, NHANES is a cross-sectional survey, and thus causal and temporal relationships between the predictors and depression cannot be determined. A predictor that plays a significant role in the

model could be a consequence, correlation, or proxy of depression, but not a cause. Further studies are required to assess the predictive value of the characteristics identified to the development of depression in the future.

Second, some NHANES variables are self-reported, such as smoking, drinking, physical activity, health care utilization, and general health. Self-reported measures can be subject to recall error, social desirability, and participant reporting differences. The standardized data collection procedures of NHANES do not guarantee that it is free of measurement limitations which can cause uncertainty in the predictive models.

Third, the absence of data and how it is handled can affect model performance. Imputation can minimize the number of missing observations, but can also cause uncertainty if the missingness is correlated with the characteristics of the participants. Likewise, omitting variables with significant missing data could cause the loss of potentially informative predictors. More research is needed on the stability of the models when considering alternative missing data strategies.

Fourth, random undersampling could lead to loss of information from the majority class. Undersampling may help the model to learn the characteristics of the minority class, but it can also impact the stability of the learned relationships. Other methods, such as class weighting, synthetic sampling, or cost-sensitive learning, might yield other outcomes and warrant further study.

Fifth, the generalizability of the models needs to be explored further. While NHANES offers a great deal of information about the population, individuals who are part of a national health survey might not be the same as those seen in specialized clinical or mental health settings. The models should then be validated using independent data sets before they can be applied in clinical practice.

Last but not least, SHAP analysis enhances the interpretability of the model without implying causality. The importance of a feature in predicting model outcomes does not imply that changes in that feature will necessarily decrease the risk of depression. Future research should integrate machine-learning techniques with longitudinal designs, causal inference techniques and external validation to determine if the predictors identified by ML can be used as a basis for effective prevention and intervention strategies.

Finally, this study shows that machine learning methods can be used to predict depression in adolescents and adults based on a wide variety of demographic, socioeconomic, behavioral, healthcare, anthropometric, clinical and laboratory features. The results suggest that ensemble-based machine learning techniques can be used to successfully model complex patterns related to depression and could be used in conjunction with traditional statistical methods. SHAP-based interpretation also adds to the transparency of the predictive modelling process by offering an understanding of the impact of each predictor. These models must, however, be prospective validated, externally validated, calibrated and assessed for model fairness across population subgroups before they can be translated into clinical or public health applications.

V. CONCLUSION

This study illustrates the promise of machine learning to help predict depression in adolescents and adults using nationally representative health data. The analytical framework captured a wide spectrum of demographic and socioeconomic, allowing for the multifactorial nature of depression to be captured. Ensemble-based algorithms demonstrated superior predictive performance, especially XGBoost algorithm, which had the best performance overall among the evaluated algorithms. The use of SHAP-based interpretation also made the predictive modelling process more transparent, as it enabled the contribution of individual predictors to be examined. The results suggest the potential of machine learning as an adjunctive tool for depression risk stratification and population health studies. However, additional studies with independent populations, prospective study designs, model calibration and model fairness assessments are recommended prior to implementation in clinical or public health practice.

ACKNOWLEDGEMENTS

The authors would like to thank the participants in the National Health and Nutrition Examination Survey (NHANES) for their contributions to this nationally representative health survey. We also acknowledge the National Center for Health Statistics and the Centers for Disease Control and Prevention for collecting, administering, and making available the NHANES data used in this investigation.

Funding

Public, commercial or not-for-profit funding agencies did not provide specific financial support for this research. The study was performed with the use of data from the publicly available National Health and Nutrition Examination Survey.

Data Availability

Data used in this study are from the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), National Health and Nutrition Examination Survey (NHANES). The NHANES datasets and documentation are available for public use on the official NCHS website at https://www.cdc.gov/nchs/products/databriefs/db527.htm?utm_source=nhanes&utm_medium=web&utm_campaign=nhanes and <https://www.cdc.gov/nchs/nhanes/>. The relevant components of the survey and documentation are available for researchers to access as allowed by the data-use policies set forth by the NCHS.

Declarations**Ethics Approval and Consent to Participate**

This investigation was a secondary analysis of data gathered in NHANES, and was not conducted through direct recruitment, intervention, or interaction with human subjects by the investigators. NHANES is done under approved ethical review procedures established by the National Center for Health Statistics and informed consent is obtained as part of the original data collection. De-identified secondary data from NHANES was used in the present study, and no additional informed consent or independent ethical approval was needed beyond that provided for the collection of the original NHANES data.

Consent for Publication

Not applicable.

Competing Interests

The authors have no competing interests.

Clinical Trial Registration

Not applicable.

REFERENCES

- [1]. Balakrishna, N., Krishnan, M. B. M., & Ganesh, D. (2024). Hybrid Machine Learning Approaches for Predicting and Diagnosing Major Depressive Disorder. *International Journal of Advanced Computer Science and Applications*, 15(3). <https://doi.org/10.14569/ijacsa.2024.0150363>
- [2]. Birk, J. L., Kronish, I. M., Moise, N., Falzon, L., Yoon, S., & Davidson, K. W. (2019). Depression and multimorbidity: Considering temporal characteristics of the associations between depression and multiple chronic diseases. *Health Psychology*, 38(9), 802–811. <https://doi.org/10.1037/hea0000737>
- [3]. Dong, Y., Wen, H., Lu, C., Li, J., & Zheng, Q. (2025). Predicting depression risk with machine learning models: identifying familial, personal, and dietary determinants. *BMC Psychiatry*, 25(1). <https://doi.org/10.1186/s12888-025-07182-8>
- [4]. Gabarrell-Pascuet, A., Domènech-Abella, J., Rod, N. H., & Varga, T. V. (2023). Variations in sociodemographic and health-related factors are linked to distinct clusters of individuals with depression based on the PHQ-9 instrument: NHANES 2007–2018. *Journal of Affective Disorders*, 335, 95–104. <https://doi.org/10.1016/j.jad.2023.05.005>
- [5]. Gabarrell-Pascuet, A., Domènech-Abella, J., Rod, N. H., & Varga, T. V. (2023). Variations in sociodemographic and health-related factors are linked to distinct clusters of individuals with depression based on the PHQ-9 instrument: NHANES 2007–2018. *Journal of Affective Disorders*, 335, 95–104. <https://doi.org/10.1016/j.jad.2023.05.005>
- [6]. Gardner, R. M., Sultan, P., Bernert, R. A., & Simard, J. F. (2025). Trends in prevalence and treatment of antepartum and postpartum depression in the United States: Data from the national health and nutrition examination survey (NHANES) 2007 to 2018. *PLOS One*, 20(4), e0322536. <https://doi.org/10.1371/journal.pone.0322536>
- [7]. Goodwin, R. D., Dierker, L. C., Wu, M., Galea, S., Hoven, C. W., & Weinberger, A. H. (2022). Trends in U.S. Depression Prevalence From 2015 to 2020: The Widening Treatment Gap. *American Journal of Preventive Medicine*, 63(5), 726–733. <https://doi.org/10.1016/j.amepre.2022.05.014>
- [8]. Hou, C.-Y., Tsai, Y.-C., Kuo, T.-N., & Shyu, I.-L. (2026). Joint impact of premature age at menopause and depression on mortality in older females: an NHANES analysis. *Menopause*. <https://doi.org/10.1097/gme.0000000000002920>
- [9]. Mardini, M. T., Khalil, G. E., Bai, C., Divakaran, A. M., & Ray, J. M. (2025). Identifying Adolescent Depression and Anxiety Through Real-World Data and Social Determinants of Health: Machine Learning Model Development and Validation. *JMIR Mental Health*, 12, e66665–e66665. <https://doi.org/10.2196/66665>
- [10]. Meghrajani, V. R., Marathe, M., Sharma, R., Potdukhe, A., Wanjari, M. B., & Taksande, A. B. (2023). A Comprehensive Analysis of Mental Health Problems in India and the Role of Mental Asylums. *Cureus*. <https://doi.org/10.7759/cureus.42559>
- [11]. Mehlretter, J., Rollins, C., Benrimoh, D., Fratila, R., Perlman, K., Israel, S., Miresco, M., Wakid, M., & Turecki, G. (2020). Analysis of Features Selected by a Deep Learning Model for Differential Treatment Selection in Depression. *Frontiers in Artificial Intelligence*, 2. <https://doi.org/10.3389/frai.2019.00031>
- [12]. Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., & Ping, P. (2019). Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes*, 10(2), 87. <https://doi.org/10.3390/genes10020087>

- [13]. Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214, 106584. <https://doi.org/10.1016/j.cmpb.2021.106584>
- [14]. Patel, J. S., Oh, Y., Rand, K. L., Wu, W., Cyders, M. A., Kroenke, K., & Stewart, J. C. (2019). Measurement invariance of the patient health questionnaire-9 (PHQ-9) depression screener in U.S. adults across sex, race/ethnicity, and education level : NHANES 2005–2016. *Depression and Anxiety*, 36(9), 813–823. Portico. <https://doi.org/10.1002/da.22940>
- [15]. Qian, H., Song, S., Zhang, Z., Yuan, Y., & Zhang, L. (2025). A predictive model for depression risk in individuals with metabolic dysfunction-associated steatotic liver disease: evidence from NHANES 2017–2023. *Psychology, Health & Medicine*, 1–20. <https://doi.org/10.1080/13548506.2025.2591861>
- [16]. Qiu, K., Liu, Y., Zhang, Y., Gu, J., & Huang, Y. (2025). The association between physical activity and depression among weekend catch-up sleepers: results from NHANES 2021–2023. *BMC Psychiatry*, 25(1). <https://doi.org/10.1186/s12888-025-07095-6>
- [17]. Rahman, A., Debnath, T., Kundu, D., Khan, Md. S. I., Aishi, A. A., Sazzad, S., Sayduzzaman, M., & Band, S. S. (2024). Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health*, 11(1), 58–109. <https://doi.org/10.3934/publichealth.2024004>
- [18]. Sahlaoui, H., Alaoui, E. A. A., Nayyar, A., Agoujil, S., & Jaber, M. M. (2021). Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations. *IEEE Access*, 9, 152688–152703. <https://doi.org/10.1109/access.2021.3124270>
- [19]. Sinkewicz, M., Rostant, O., Zivin, K., McCammon, R., & Clarke, P. (2022). A life course view on depression: Social determinants of depressive symptom trajectories over 25 Years of Americans' Changing Lives. *SSM - Population Health*, 18, 101125. <https://doi.org/10.1016/j.ssmph.2022.101125>
- [20]. Theng, D., & Bhoyar, K. K. (2023). Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, 66(3), 1575–1637. <https://doi.org/10.1007/s10115-023-02010-5>
- [21]. Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine Learning in Mental Health. *ACM Transactions on Computer-Human Interaction*, 27(5), 1–53. <https://doi.org/10.1145/3398069>
- [22]. Wan, L., & He, S. (2024). Association between depression and infertility based on the PHQ-9 score: Analyses of NHANES 2013–2018. *Plos One*, 19(7), e0305176. <https://doi.org/10.1371/journal.pone.0305176>
- [23]. Wang, N., Wang, L., Wang, J., Chen, R., Shi, M., Liu, H., & Xu, X. (2023). Effects of physical activity and depressive symptoms on cognitive function in older adults: National Health and Nutrition Examination Survey. *Neurological Sciences*, 45(1), 299–308. <https://doi.org/10.1007/s10072-023-07250-5>
- [24]. Yadav, N., Gulati, A., Gulati, V., & Yadav, P. (2025). Evaluating the impact of machine learning models on adult major depressive disorder using conventional treatment strategies: a systematic review approach. *Discover Public Health*, 22(1). <https://doi.org/10.1186/s12982-025-00816-y>
- [25]. Zhang, W., Wang, Y., Cong, Q., Tang, X., Wang, Z., & Wang, Z. (2026). Interaction effects of sedentary behavior and depression on MAFLD in NHANES 2017–2020 and 2021–2023. *PLOS One*, 21(2), e0342336. <https://doi.org/10.1371/journal.pone.0342336>