

Original Paper

# Discovery of Depression-Associated Factors From a Nationwide Population-Based Survey: Epidemiological Study Using Machine Learning and Network Analysis

Sang Min Nam<sup>1\*</sup>, MD, PhD; Thomas A Peterson<sup>2\*</sup>, PhD; Kyoung Yul Seo<sup>3</sup>, MD, PhD; Hyun Wook Han<sup>4</sup>, MD, PhD; Jee In Kang<sup>5</sup>, MD, PhD

<sup>1</sup>Department of Ophthalmology, CHA Bundang Medical Center, CHA University, Seongnam, Republic of Korea

<sup>2</sup>UCSF REACH Informatics Core, Department of Orthopaedic Surgery, Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, United States

<sup>3</sup>Department of Ophthalmology, Institute of Vision Research, Eye and Ear Hospital, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

<sup>4</sup>Department of Biomedical Informatics, CHA University School of Medicine, CHA University, Seongnam, Republic of Korea

<sup>5</sup>Department of Psychiatry, Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

\*these authors contributed equally

**Corresponding Author:**

Jee In Kang, MD, PhD

Department of Psychiatry

Institute of Behavioral Science in Medicine

Yonsei University College of Medicine

50-1 Yonsei-ro, Seodaemun-gu

Seoul, 03722

Republic of Korea

Phone: 82 2 2228 1620

Fax: 82 2 313 0891

Email: [jeeinkang@yuhs.ac](mailto:jeeinkang@yuhs.ac)

## Abstract

**Background:** In epidemiological studies, finding the best subset of factors is challenging when the number of explanatory variables is large.

**Objective:** Our study had two aims. First, we aimed to identify essential depression-associated factors using the extreme gradient boosting (XGBoost) machine learning algorithm from big survey data (the Korea National Health and Nutrition Examination Survey, 2012-2016). Second, we aimed to achieve a comprehensive understanding of multifactorial features in depression using network analysis.

**Methods:** An XGBoost model was trained and tested to classify “current depression” and “no lifetime depression” for a data set of 120 variables for 12,596 cases. The optimal XGBoost hyperparameters were set by an automated machine learning tool (TPOT), and a high-performance sparse model was obtained by feature selection using the feature importance value of XGBoost. We performed statistical tests on the model and nonmodel factors using survey-weighted multiple logistic regression and drew a correlation network among factors. We also adopted statistical tests for the confounder or interaction effect of selected risk factors when it was suspected on the network.

**Results:** The XGBoost-derived depression model consisted of 18 factors with an area under the weighted receiver operating characteristic curve of 0.86. Two nonmodel factors could be found using the model factors, and the factors were classified into direct ( $P < .05$ ) and indirect ( $P \geq .05$ ), according to the statistical significance of the association with depression. Perceived stress and asthma were the most remarkable risk factors, and urine specific gravity was a novel protective factor. The depression-factor network showed clusters of socioeconomic status and quality of life factors and suggested that educational level and sex might be predisposing factors. Indirect factors (eg, diabetes, hypercholesterolemia, and smoking) were involved in confounding or interaction effects of direct factors. Triglyceride level was a confounder of hypercholesterolemia and diabetes, smoking had a significant risk in females, and weight gain was associated with depression involving diabetes.

**Conclusions:** XGBoost and network analysis were useful to discover depression-related factors and their relationships and can be applied to epidemiological studies using big survey data.

(*J Med Internet Res* 2021;23(6):e27344) doi: [10.2196/27344](https://doi.org/10.2196/27344)

## KEYWORDS

depression; epidemiology; machine learning; network; prediction model; XGBoost

## Introduction

### Importance of Depression

Depression is a common debilitating psychiatric condition characterized by a low-spirited mood, loss of interest, and a range of emotional, cognitive, physical, and behavioral symptoms. It has a high global disease burden and had been projected to become the second most common cause of disability-adjusted life years worldwide by 2020 [1]. Thus, the determination of risk factors causing depression could have important implications in its prevention and intervention efforts by reducing modifiable risk factors. However, proper prevention and treatment of depression have been difficult owing to heterogeneity in the etiology and pathophysiology of depression [2].

### Conventional Modeling for Depression

Given the complex biological, psychological, and sociocultural factors underlying the pathogenesis of depression, an integrated model with confounder adjustment may provide a better understanding and multifaceted individualized approach for depression. In a typical study, survey-weighted logistic regression is used to identify depression-associated factors. A simple regression model for each candidate factor is built to adjust for age and sex. Then, a complex model is presented by adding more covariates to control potential confounders [3].

### Principal Factor Identification Problem in Conventional Modeling

A confounder has associations with both exposure and disease but is not in the causal pathway between exposure and outcome [4]. Confounding is a mixing of effects that obscures the real effect of exposure [4]. We might fit a regression model on all the measured potential confounders and minimize the confounding risk [5]. However, too many variables can cause multicollinearity or redundancy and make a regression model unstable with a high coefficient variance [5,6]. In contrast, variable selection may cause the exclusion of essential confounders [5]. Identification of the best subset of variables is challenging when the number of explanatory variables is large or when multicollinearity is present within the data [7].

### Benefits of Machine Learning–Based Modeling for Big Survey Data

For big survey data with numerous variables, the machine learning model can detect principle factors for a condition by attempting to maximize its predicting performance. During the training, dimensionality reduction (which reduces the number of features) is essential because many features may cause overfitting, and the generated model may not be generalized

appropriately [8]. Moreover, feature selection can resolve multicollinearity by removing redundant features from a group of highly correlated features [8]. Regularization is a more advanced machine learning algorithm that reduces model complexity and improves generalizability [8]. A popular regularization method is LASSO (L1 regularization; least absolute shrinkage and selection operator) [8], also called a shrinkage method, which can be applied to collinear confounders and has shown advantages over conventional methods [5].

### Advantages of the Extreme Gradient Boosting Algorithm and Strategies of This Study

We speculate that extreme gradient boosting (XGBoost) can help select an optimal subset of essential variables from a large volume of survey data. XGBoost is an advanced implementation of a gradient-boosting decision-tree algorithm. XGBoost has been used in a few studies to predict or screen depression [9-11]. XGBoost is advantageous because of its high speed and performance, making it dominant in applied machine learning for structured data. XGBoost also offers regularized gradient boosting and feature importance scores using a trained predictive model, which can be used for feature selection [12].

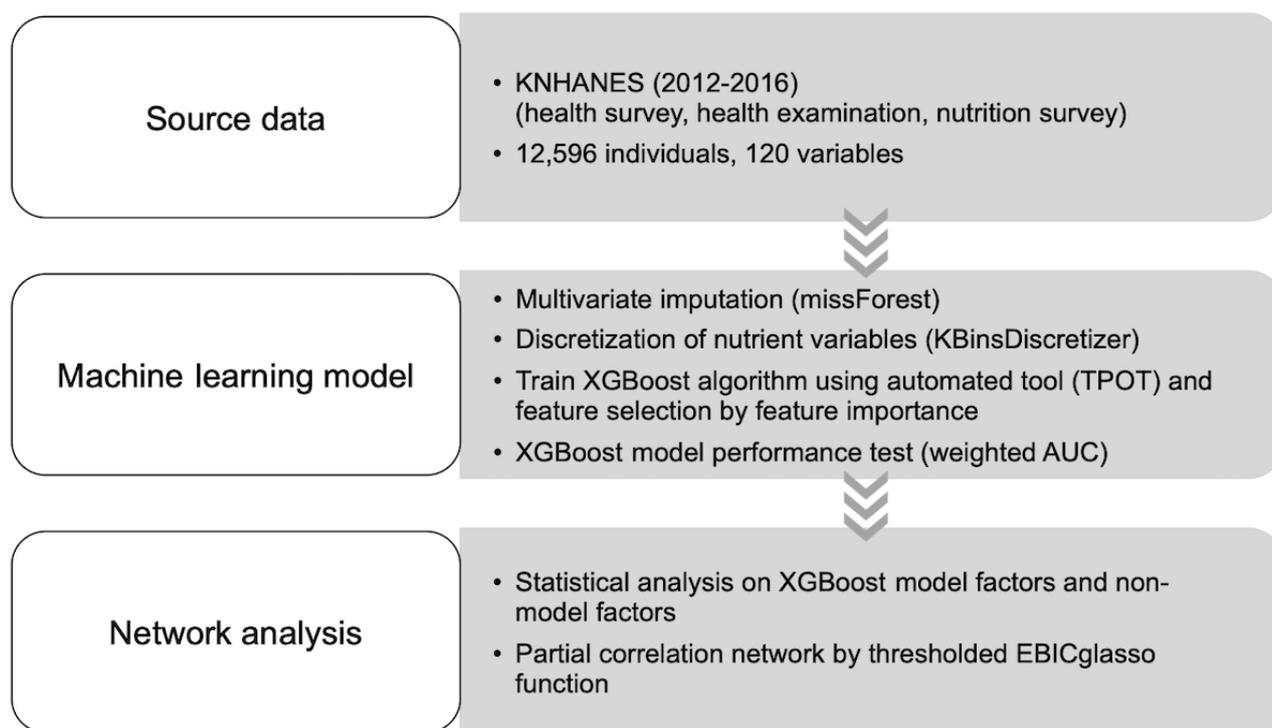
However, XGBoost has many hyperparameters that should be manually specified, and optimal parameter tuning can be challenging. To solve this problem, we have chosen an automated machine learning tool, the tree-based pipeline optimization tool (TPOT), along with genetic programming [13]. To the best of our knowledge, this study is the first attempt at an XGBoost-supported epidemiologic investigation regarding depression in the general population using national survey data.

## Methods

### Overview of Survey Data

The overall study flowchart is provided in [Figure 1](#). We used survey data obtained from 12,596 Koreans (aged 19-64 years) who participated in the Korea National Health and Nutrition Examination Survey (KNHANES, 2012-2016), which is an annual survey conducted by the Korea Centers for Disease Control and Prevention (KCDC) in the Republic of Korea, wherein population-wide health and nutritional statuses are assessed [14]. The maximum age was set to 64 years because the food frequency survey was conducted for those aged 19 to 64 years. The design and methods of the KNHANES and the data resource profile are available in previous reports [14-16]. KNHANES V-3 (2012), KNHANES VI (2013-2015), and KNHANES VII-1 (2016) were approved by the KCDC research ethics committee (2012-01EXP-01-2C, 2013-07CON-03-4C, 2013-12EXP-03-5C, and 2015-01-02-6C), and written informed consent was obtained from all the subjects.

**Figure 1.** Flow diagram of the study. AUC: area under the curve; EBICglasso: extended Bayesian information criterium graphical lasso; KNHANES: Korea National Health and Nutrition Examination Survey.



In the annual KNHANES (2012-2016) following a multistage clustered probability design, 192 primary sampling units (PSUs) were generated from approximately 200,000 geographically defined PSUs for the entire country; subsequently, 20 final target households were sampled for each PSU as secondary sampling units (ID\_fam) [16]. The sample weights were calculated using the inverse of selection probabilities and the inverse of response rates that were adjusted to gender- and age-specific Korean populations (poststratification). Therefore, the sampled individuals accurately represented the Korean population.

### Depression Case Definition

Positive depression diagnosed by a physician (DF2\_dg) and positive present depression (DF2\_pr) represented the positive case “current depression,” and these indicated that the individual had a diagnosis of depression and was currently experiencing

depression, respectively. In terms of the negative case, both DF2\_dg and DF2\_pr were negative and represented “no lifetime depression,” which indicated that the individual had never experienced depression. DF2\_dg and DF2\_pr were reported by the participants in this survey.

### XGBoost-Derived Model for Depression

We included as many variables as possible (Textbox 1) and imputed missing values using univariate imputation (filling with the last valid value for categorical or discrete variables) or multivariate imputation for missing continuous variables. The nutrient variables were discretized into three bins as follows: low, medium, and high. We converted categorical variables to numerical variables using one-hot encoding and dropped reference categories. We standardized variables, except for a binary dummy variable, by removing the mean and scaling them to unit variance before machine learning.

**Textbox 1.** Study variables of the Korea National Health and Nutrition Examination Survey data.

<b>Health Examination Data</b>
Physical examination
<ul style="list-style-type: none"> <li>Waist circumference, BMI, regularity of pulse, and periodontal disease</li> </ul>
Blood tests
<ul style="list-style-type: none"> <li>Anemia (hemoglobin &lt;13.0 g/dL for men; hemoglobin &lt;12.0 g/dL for nonpregnant women; hemoglobin &lt;11.0 g/dL for pregnant women), white blood cell count, platelet count, red blood cell count, aspartate aminotransferase, alanine aminotransferase, creatinine, urea nitrogen, hepatitis B surface antigen, and hepatitis C antibody</li> </ul>
Fasting (≥8 hours) blood tests
<ul style="list-style-type: none"> <li>Sugar level, high-density lipoprotein cholesterol (HDL), and triglyceride (TG)</li> </ul>
Urine strip test
<ul style="list-style-type: none"> <li>pH, nitrite, specific gravity, protein, glucose, ketones, bilirubin, blood, urobilinogen</li> </ul>
Pulmonary function test
Hypercholesterolemia: total cholesterol (TC) ≥240 mg/dL or lipid-lowering medication
Hypertension: systolic pressure ≥140 mmHg or diastolic pressure ≥90 mmHg or medication
Diabetes mellitus: fasting blood sugar level ≥126 mg/dL, diagnosis, medications, or insulin injections
Prediabetes: fasting blood sugar level ≥100 and <126 mg/dL
Low-density lipoprotein cholesterol (LDL) = $TC - (HDL + TG/5)$ , where $TG \leq 400$ mg/dL
<b>Health Survey Data</b>
<ul style="list-style-type: none"> <li>Age, marital status, educational level, occupational class, household income, weight changes in the past year, alcohol consumption frequency, sleep hours per day, perceived stress, current smoking status, walking for more than 10 minutes (days/week), muscle-strengthening activities (days/week), childbirth experience, and EuroQol five-dimension three-level (EQ-5D-3L) questionnaire (mobility, usual activities, self-care, and pain/discomfort; three levels for each) [17]</li> </ul>
Morbidity
<ul style="list-style-type: none"> <li>Chronic obstructive pulmonary disease, stroke, ischemic heart disease, osteoarthritis, rheumatoid arthritis, pulmonary tuberculosis, asthma, thyroid disease, gastric cancer, hepatoma, colon cancer, breast cancer, cervix cancer, lung cancer, thyroid cancer, atopic dermatitis, renal failure, hepatitis B, hepatitis C, and liver cirrhosis</li> </ul>
Oral health
<ul style="list-style-type: none"> <li>Difficulty chewing, caries treatment within the last 1 year, root canal treatment within the last 1 year, and prosthetic treatment within the recent 1 year</li> </ul>
<b>Food Frequency Survey Data (daily intake)</b>
<ul style="list-style-type: none"> <li>Carbohydrates, proteins, saturated fatty acids, monounsaturated fatty acids, polyunsaturated fatty acids, n3 fatty acids, n6 fatty acids, n6:n3 fatty acid ratio, cholesterol, fiber, vitamin A, vitamin B1, vitamin B2, vitamin C, niacin, iron, calcium, potassium, phosphorus, and sodium</li> </ul>

The training set included 80% “current depression” and 80% “no lifetime depression” (10,076 cases). The remaining cases were assigned to the test set (2520 cases). We used TPOT to choose the best XGBoost hyperparameters from the training set to predict “current depression” from “no lifetime depression.” Then, the model features were selected further using the feature importance of XGBoost. The XGBoost model was explained using shapley additive explanation (SHAP) values that showed each feature’s impact on the model prediction [18].

The weighted area under the receiver operating characteristic curve (AUC) of the final model was calculated using the sample weight [19]. The optimal threshold was computed as the prediction probability at which Youden’s index (sensitivity + specificity – 1) was the maximum on the test set. We also

estimated the weighted AUC of the model for the Patient Health Questionnaire-9 (PHQ-9) score (depression when score ≥10) instead of the reported depression.

### Statistical Test for Model Factors and Exploring Nonmodel Factors

Survey-weighted multiple logistic regression was performed using the model features from all samples (training and test sets). We defined the “direct” factor of depression as the model feature whose multiple logistic regression coefficient significantly differs from zero ( $P < .05$ ); the other model features were designated as “indirect” factors. Subsequently, we performed a statistical test on factors excluded by the model to identify significant “nonmodel” factors that were not chosen

by XGBoost as optimal for predicting depression. Nonmodel factors whose coefficients of multiple logistic regression with the model features were significantly different from zero ( $P < .05$  with Bonferroni correction) were chosen.

### Network Analysis for Depression-Related Factors

A correlation matrix of model features and nonmodel factors was generated using all samples. Based on the correlation matrix, a thresholded EBICglasso (extended Bayesian information criterion graphical lasso) network was plotted, in which the network model was estimated using graphical LASSO regularization with EBIC model selection. Each edge weight was the correlation coefficient between two nodes after controlling all other network correlations [20,21].

Three centrality indices (strength, closeness, and betweenness) were computed. Strength centrality is the absolute sum of the edge weights connected to the node. Closeness centrality is the sum of the shortest distances from the node to all other network nodes. Betweenness centrality is the number of times when the node lies on the shortest path between two other nodes [21,22].

### Confounding or Interaction Effects on Indirect Risk Factors

If an indirect risk factor was positively connected to direct risk factors, the direct factor confounding effects were tested. If the indirect factor's coefficient on survey-weighted multiple logistic regression became significant without the direct factors, we assumed they were the indirect factor's confounders.

If an indirect risk factor was positively connected to a direct preventive factor or negatively related to a direct risk factor, the interaction effect was tested using the interaction term on survey-weighted multiple logistic regression.

### Statistics and Software

We used the following two programming languages: Python (version 3.8.5, Python Software Foundation) and R (version 3.6.3, R Foundation for Statistical Computing).

#### Scikit-Learn Tools

The following scikit-learn tools [23] were applied: for imputing missing continuous values, *IterativeImputer* class with the *ExtraTreesRegressor* estimator; for discretization, *KBinsDiscretizer* with a k-means strategy based on a k-means clustering procedure; for standardization, *StandardScaler*; and for weighted AUC, sensitivity, specificity, and threshold, *roc\_auc\_score* and *roc\_curve*.

#### TPOT

TPOT starts from a collection of random models (first generation). Subsequently, those with higher performance are chosen and copied into the next generation's population. The offspring crossover with other offspring or are randomly changed by mutation. The algorithm repeats this evaluate-select-crossover-mutate process for multiple generations. Finally, the best model is selected from the run [13].

We used TPOT to tune the hyperparameters of the XGBClassifier, such as the number of boosting rounds

("n\_estimators"), boosting learning rate ("learning\_rate"), maximum tree depth for base learners ("max\_depth"), minimum sum of instance weight needed in a child ("min\_child\_weight"), minimum loss reduction required to make a further partition on a leaf node of the tree ("gamma"), subsample ratio of the training instance ("subsample"), subsample ratio of columns when constructing each tree ("colsample\_bytree"), L1 regularization term on weights ("reg\_alpha"), and L2 regularization term on weights ("reg\_lambda"). To control the balance of positive and negative weights, we set "scale\_pos\_weight" as the number of negative cases/the number of positive cases. The "objective" option was set to a default value, "binary:logistic" (ie, logistic regression for binary classification).

Two parameters of TPOTClassifier, namely, number of iterations to run the XGBoost optimization process ("generation") and number of individuals to retain in the genetic programming population every generation ("pupulation\_size"), were set to 100 (default) each. Finally, TPOT evaluated models using AUC ("roc\_auc").

### Extreme Gradient Boosting (XGBoost) and Feature Selection

XGBoost incorporates an ensemble technique (boosting) of adding new models to correct errors made by previous models. When adding new models, gradient boosting uses a gradient descent algorithm to minimize the loss. More details on XGBoost's working principles can be found in published articles [24]. For feature selection, we sorted the model features according to their importance (importance\_type = "gain," the average gain of splits that use the feature) and selected a subset of features by omitting the least important feature. If the AUC of the model with the subset of features was not compromised or became higher, we omitted the next least important feature. Subsequently, we retested it until the model performance reached the best AUC.

#### R Packages

The *Survey* package allows to specify a multistage sampling survey design and provides functions to estimate total population counts, means, and variances for a survey sample [25]. In this study, the survey design of the KNHANES was applied using the variables *psu* for PSU, *ID\_fam* (family ID) for the secondary sampling unit, *kstrata* for strata, and sample weight (*wt\_tot/5*) for weight. Additionally, we used the "svyglm" function from the *survey* package for survey-weighted multiple logistic regression. The odds ratio was calculated by exponentiating the coefficient of multiple logistic regression.

Receiver operating characteristic (ROC) curves were generated using the *ggplot2* package [26], and the weighted AUC was calculated by using the "WeightedAUC" function with the sample weight (*WeightedROC* package) [19].

#### Network Graph

For a correlation matrix in network analysis, a correlation between two variables was calculated using the *wtd.cor* function with the sample weight for "weight" in R's *weights* package, which produced weighted Pearson correlation coefficients for

survey data [27]. The network graph was generated using the `qgraph` function in the `qgraph` package of R by setting the graph argument to “glasso” and layout to “spring” with threshold [20]. For node sizes in the network, the effective size of the variable’s odds ratio was calculated by exponentiating the absolute coefficient of multiple logistic regression. For the calculation of centrality indices, the `qgraph` object and the `centralityTable` function in the `qgraph` package of R were used [20].

## Results

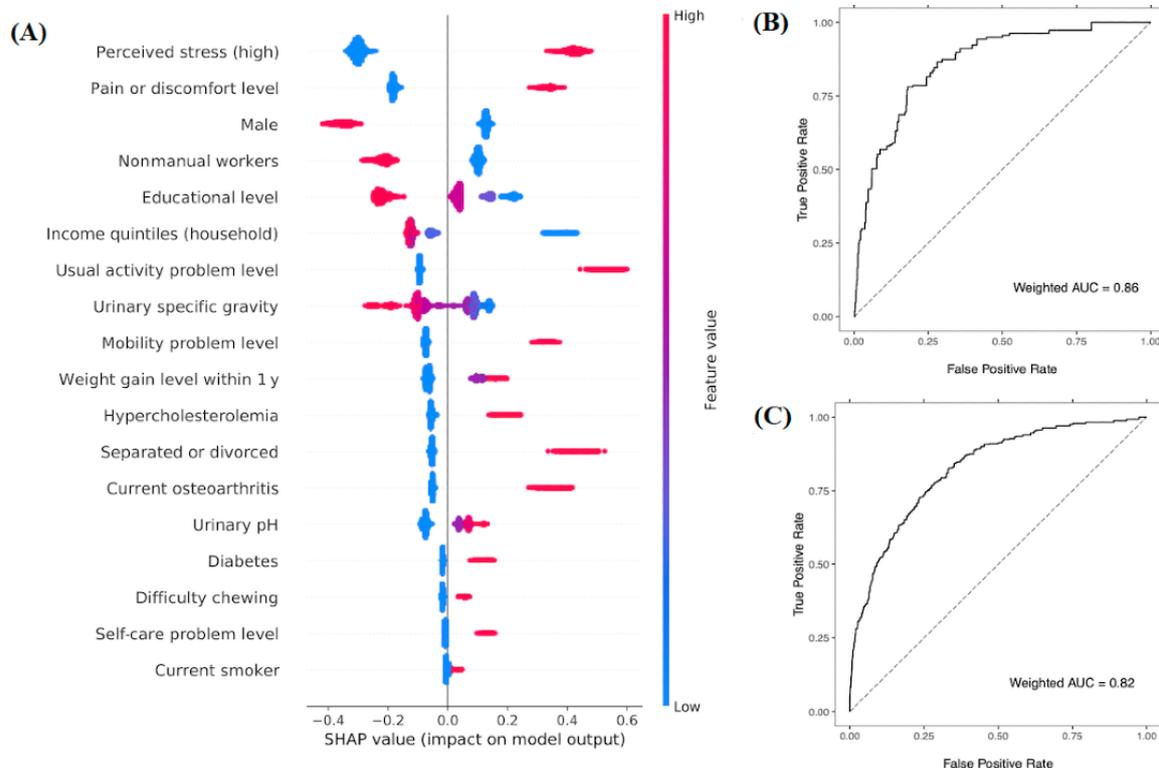
### XGBoost-Derived Model and the Performance

The population was estimated to be 22,262,880 (SE 355,977) (97.3%) for no depression reported during lifetime and 616,082

(SE 40,273) (2.7%) for current depression (complex-survey-design-based estimation). After TPOT training, the number of XGBoost model features was reduced from 120 to 81 (33% reduction) by the L1 regularization term of XGBoost. Furthermore, we decreased the feature number to 18 (78% reduction) by applying feature selection without performance loss.

The final model features included mental health, quality of life (QoL), socioeconomic factors, morbidity, sex, marital status, urinalysis, and health behavior (Figure 2). Nutrition factors failed to be included in the model. For the test set, the weighted AUC of the model was 0.86 for reported depression and 0.82 for PHQ-9 score depression (Figure 2). The model’s accuracy was 0.82 at the best threshold of the test data (Table 1).

**Figure 2.** XGBoost model for the depression and performance test. Red dots represent positive for a binary factor or high feature values for a continuous factor on the beeswarm plot of shapley additive explanation (SHAP) values (log odds of the current depression) for the training data (A). On the weighted receiver operating characteristic curve (ROC) (B), sensitivity is 0.78 and specificity is 0.82 at the best threshold of probability of 0.461 for test samples. The model is also tested using the Patient Health Questionnaire-9 (depression when score  $\geq 10$ ) for 6098 samples (C). AUC: area under the curve.



**Table 1.** Confusion matrix for the test data set.

Actual depression	Predicted depression <sup>a</sup>	
	Current	No lifetime
Current	103,182	28,933
No lifetime	804,195	3,657,604

<sup>a</sup>Case numbers are estimated using complex-survey-design weights at the best threshold.

### Statistical Significance of Depression-Associated Factors

Perceived stress and current asthma were top-ranked statistically significant risk factors (Table 2). Triglycerides, current asthma,

and farm workers were nonmodel factors having a significant association with depression and added on the network with model factors. Eight features out of 18 model factors were not statistically significant and classified as indirect (Table 2).

**Table 2.** Survey-weighted multiple logistic regression analysis of depression-related factors.

Variable	No lifetime depression <sup>a</sup> (N=22,262,880)	Current depression <sup>a</sup> (N=616,082)	Odds ratio (95% CI) <sup>b</sup>	P value
<b>High perceived stress (<math>\geq 3/4</math> points), n (%)</b>				
Yes	5,999,970 (26.95%)	372,138 (60.40%)	3.3 (2.5-4.3)	<.001
No (reference)	16,262,910 (73.05%)	243,944 (39.60%)	N/A <sup>c</sup>	N/A
<b>Sex/gender, n (%)</b>				
Male	11,072,884 (49.74%)	165,770 (26.91%)	0.5 (0.3-0.7)	<.001
Female (reference)	11,189,996 (50.26%)	450,312 (73.09%)	N/A	N/A
<b>Marital status, n (%)</b>				
Separated or divorced	852,170 (3.83%)	105,567 (17.14%)	2.2 (1.4-3.5)	<.001
Single	6,153,788 (27.64%)	133,952 (21.74%)	1.4 (0.9-2.2)	.11
Widowed	397,387 (1.78%)	36,923 (5.99%)	1.4 (0.8-2.6)	.25
Married (reference)	14,859,535 (66.75%)	339,640 (55.13%)	N/A	N/A
<b>Occupation, n (%)</b>				
Nonmanual	10,304,597 (46.29%)	161,758 (26.26%)	0.6 (0.4-0.8)	<.001
Manual	4,394,739 (19.74%)	102,028 (16.56%)	0.7 (0.4-1.0)	.06
Farm <sup>d</sup>	498,239 (2.24%)	10,848 (1.76%)	0.4 (0.2-0.8)	.02
Unemployed (reference)	7,065,305 (31.73%)	341,448 (55.42%)	N/A	N/A
<b>Asthma<sup>d</sup>, n (%)</b>				
Current	205,771 (0.92%)	34,769 (5.64%)	3.1 (1.5-6.5)	.002
Past	303,213 (1.36%)	10,387 (1.69%)	1.3 (0.5-3.2)	.57
No lifetime (reference)	21,753,896 (97.72%)	570,926 (92.67%)	N/A	N/A
<b>Hypercholesterolemia, n (%)</b>				
Yes	2,963,312 (13.31%)	146,742 (23.82%)	1.4 (0.9-1.9)	.10
No (reference)	19,299,568 (86.69%)	469,340 (76.18%)	N/A	N/A
<b>Osteoarthritis, n (%)</b>				
Current	825,130 (3.71%)	100,307 (16.28%)	1.4 (0.9-2.1)	.17
Past	155,266 (0.70%)	8,385 (1.36%)	0.7 (0.1-4.6)	.71
No lifetime (reference)	21,282,484 (95.59%)	507,390 (82.36%)	N/A	N/A
<b>Diabetes, n (%)</b>				
Yes	1,466,133 (6.59%)	84,663 (13.74%)	1.4 (0.9-2.2)	.11
Prediabetes	4,504,230 (20.23%)	125,783 (20.42%)	1.0 (0.7-1.4)	.81
No (reference)	16,292,517 (73.18%)	405,636 (65.84%)	N/A	N/A
<b>Difficulty chewing, n (%)</b>				
Yes	3,398,898 (15.27%)	173,323 (28.13%)	0.9 (0.7-1.3)	.65
No (reference)	18,863,982 (84.73%)	442,759 (71.87%)	N/A	N/A
<b>Current smoker, n (%)</b>				
Yes	5,178,518 (23.26%)	139,892 (22.71%)	1.2 (0.8-1.8)	.47
No	17,084,362 (76.74%)	476,190 (77.29%)	N/A	N/A
Pain or discomfort level (1 to 3 points), mean (SE)	1.2 (0.004)	1.6 (0.034)	1.3 (1.1-1.5)	<.001
Fasting serum triglyceride (mg/dL) <sup>d</sup> , mean (SE)	134.8 (1.3)	161.8 (12.0)	1.2 (1.1-1.3)	.002
Weight gain level within 1 year (0 to 3 points), mean (SE)	0.4 (0.009)	0.6 (0.047)	1.2 (1.1-1.3)	.003

Variable	No lifetime depression <sup>a</sup> (N=22,262,880)	Current depression <sup>a</sup> (N=616,082)	Odds ratio (95% CI) <sup>b</sup>	P value
Urine specific gravity, mean (SE)	1.020 (0.00007)	1.018 (0.0004)	0.8 (0.7-0.9)	.003
Usual activity problem level (1 to 3 points), mean (SE)	1.0 (0.002)	1.3 (0.029)	1.1 (1.0-1.3)	.006
Income quintiles (household) (1 to 5 points), mean (SE)	3.4 (0.02)	2.6 (0.09)	0.8 (0.7-0.9)	.007
Educational level (1 to 4 points), mean (SE)	3.2 (0.01)	2.6 (0.06)	0.8 (0.7-1.0)	.01
Urine pH, mean (SE)	5.7 (0.009)	5.8 (0.050)	1.1(1.0-1.3)	.07
Mobility problem level (1 to 3 points), mean (SE)	1.1 (0.002)	1.3 (0.032)	1.1 (1.0-1.2)	.22
Self-care problem level (1 to 3 points), mean (SE)	1.0 (0.001)	1.1 (0.017)	1.0 (0.9-1.1)	.98
Age (years) <sup>c</sup> , mean (SE)	40.7 (0.2)	45.1 (0.8)	1.1 (0.8-1.3)	.61

<sup>a</sup>Population counts (n), means, and standard errors are estimated using complex-survey-design weights.

<sup>b</sup>For continuous factors, the value was standardized by removing the mean and scaling to unit variance.

<sup>c</sup>N/A: not applicable.

<sup>d</sup>Nonmodel factors were found by controlling for model features and age.

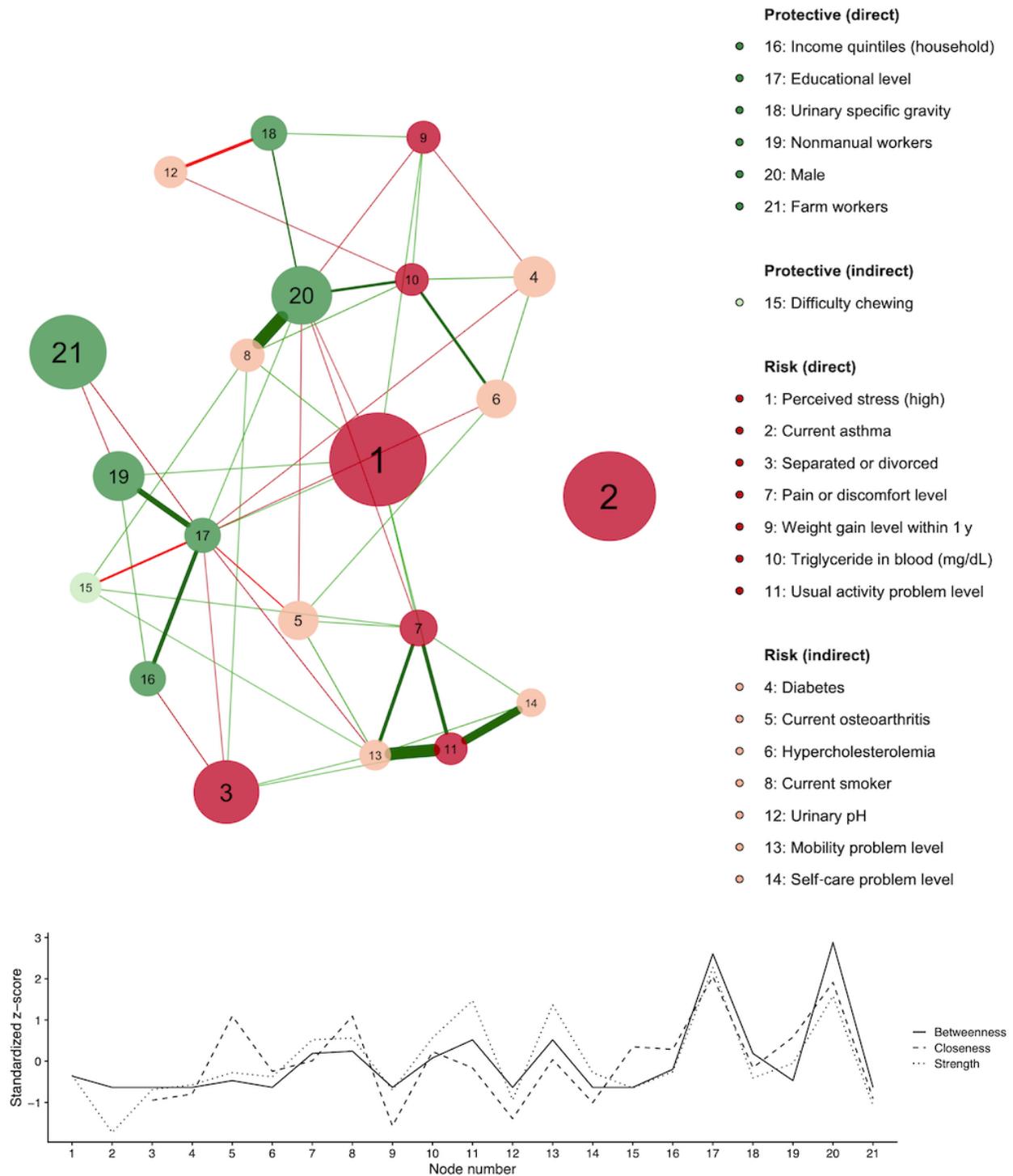
<sup>e</sup>Not a depression-related factor, but included to control confounding effects of age.

### High-Centrality Nodes and Clusters in the Depression Network

Based on the centrality indexes, the network had two high-centrality nodes, namely, “educational level” and “male,” which meant both nodes were highly connected with various factors (Figure 3). Some factors were strongly connected to

form a cluster (eg, the socioeconomic status [SES] cluster consisted of nonmanual workers, farm workers, educational level, and household income) (Figure 3). The QoL cluster comprised four EuroQol five-dimension three-level (EQ-5D-3L) domains (“pain or discomfort level,” “usual activity problem level,” “mobility problem level,” and “self-care problem level”) (Figure 3).

**Figure 3.** Partial correlation network graph and centrality indices for depression-associated factors. The factors can be positively (risk) or negatively (protective) related to current depression. Node size is proportional to the effective size of the odds ratio. Green and red edges represent positive and negative correlations, respectively. The edge with the highest absolute weight has full-color saturation and the widest width.



**Confounders and Interaction Effects of Indirect Factors**

Two indirect factors, namely, hypercholesterolemia and diabetes, were positively connected with triglyceride on the network, and

triglyceride was a confounder of hypercholesterolemia and diabetes (Table 3). Another indirect factor, current osteoarthritis, became statistically significant after excluding its positive associates, namely, “pain or discomfort level” and “mobility problem level” (Table 3).

**Table 3.** Possible confounders on indirect factors in depression.

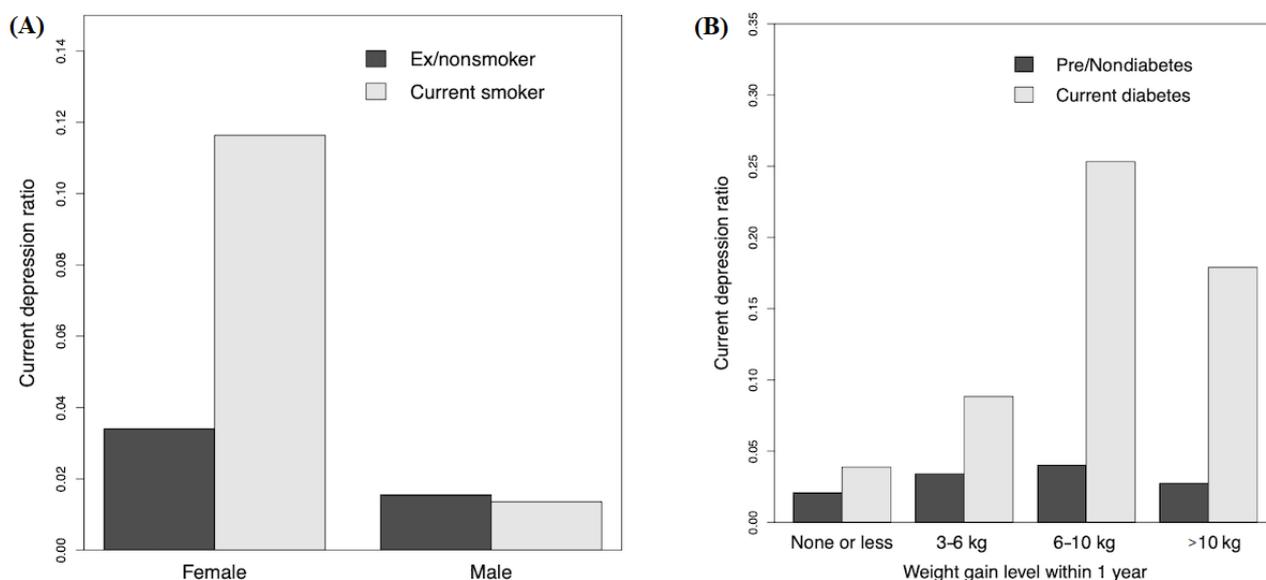
Indirect factors	Confounders <sup>a</sup>	Odds ratio (95% CI)	
		With confounders	Without confounders
Current osteoarthritis	Pain or discomfort level and mobility problem level	1.4 (0.9-2.1)	1.6 (1.1-2.5)
Hypercholesterolemia	Triglyceride	1.4 (1.0-1.9)	1.5 (1.0-2.1)
Diabetes	Triglyceride and hypercholesterolemia	1.4 (0.9-2.2)	1.7 (1.1-2.6)

<sup>a</sup>A group of confounders to make a coefficient of the indirect factor statistically insignificant.

Another indirect factor, “current smoker,” interacted with sex and was statistically significant in females (odds ratio 2.2, 95% CI 1.3-3.7;  $P=.003$ ) (Figure 4). Current diabetes interacted with

the weight gain level that was not statistically significant for prediabetes or nondiabetes ( $P=.05$ ) (Figure 4).

**Figure 4.** The proportion of current depression according to sex and smoking (A) and according to weight gain and diabetes (B). The interaction effects between them are significant ( $P=.001$  and  $.01$ , respectively; survey-weighted logistic regression).

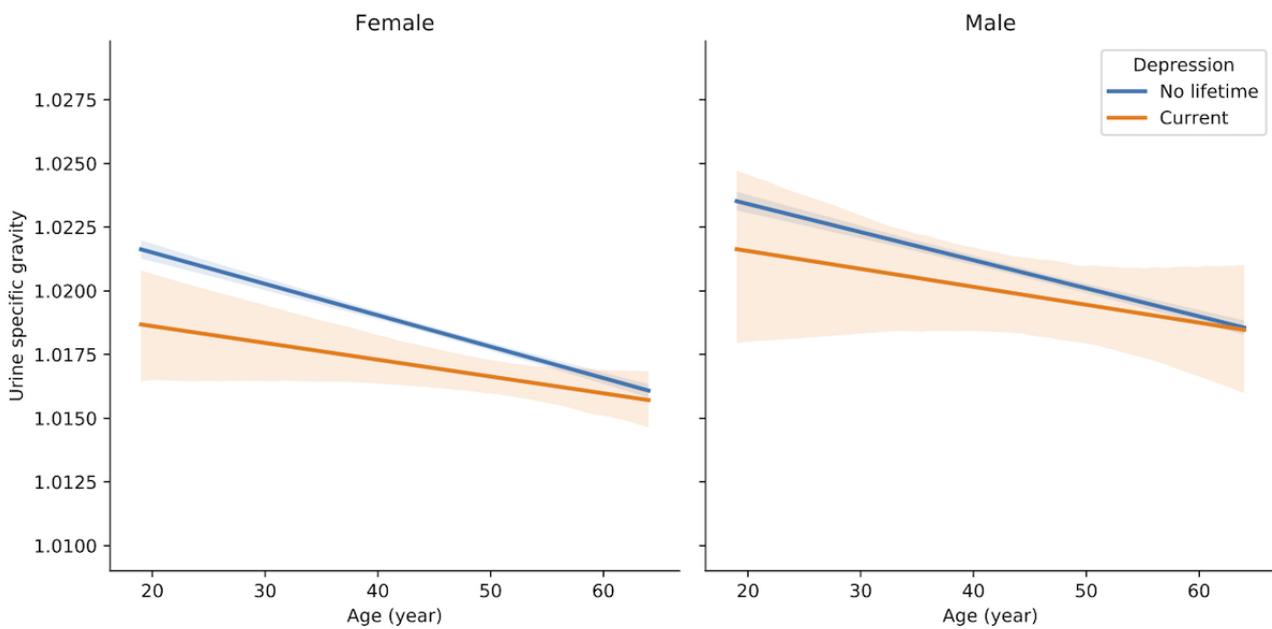


### Urine Specific Gravity and Depression

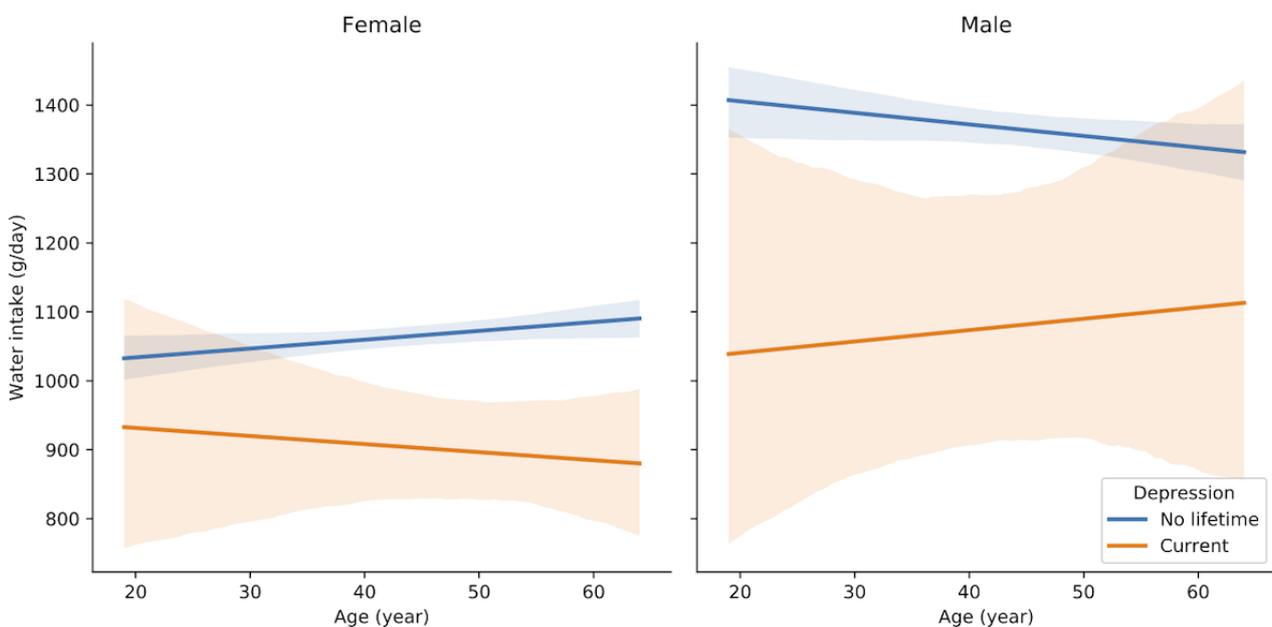
We found an interesting relationship between urine specific gravity (USG) and depression. Because USG was associated with the “male” node on the network (Figure 3), we stratified it into males and females and plotted it according to age. USG

was lower in “current depression” than in “no lifetime depression,” especially in females who were in their early 50s or younger (Figure 5). However, the daily water intake was not higher in “current depression” than in “no lifetime depression” (Figure 6).

**Figure 5.** Linear regression plots between age and urine specific gravity in females and males. In the images, 95% confidence intervals for the regression estimate are drawn using translucent bands around the regression line.



**Figure 6.** Linear regression plots between age and daily water intake in females and males. In the images, 95% confidence intervals for the regression estimate are drawn using translucent bands around the regression line.



## Discussion

### Addressing Class Imbalance and Generalization Problems

Our data set was highly imbalanced, with a 2.7% prevalence of depression. Therefore, it may be difficult for a machine learning algorithm to predict depression because this diagnosis is not sufficiently represented [28]. We used the “scale\_pos\_weight” parameter of XGBoost that enables class-weighted or cost-sensitive learning to make the model more sensitive to misclassification in the minority class [12]. The performance of our XGBoost model showed that the TPOT-trained

class-weighted XGBoost model was working well for our imbalanced classification.

There would have been a generalization problem with this survey data because we did not train the model using the sample weight. However, the weighted AUC for the test set (0.86) was not lower than the mean unweighted AUC for the training set (0.84, SD 0.02; seven repeats of five-fold cross-validation). Therefore, our model was not overfitted, and we expect that our model can be applied to the general population.

### Reliability of Self-Reported Depression

A model may be less reliable when it is based on self-reported depression than on clinically diagnosed depression. However,

for PHQ-9 scores, a clinical module, the model performance was still satisfactory when the model was retested with available cases (Figure 2). PHQ-9 is a validated screening tool for depression and is known to exhibit an acceptable diagnostic accuracy for cutoff scores between 8 and 11 [29,30]. It might be possible because we strictly defined depression by requiring both conditions (ie, current symptoms and a physician's diagnosis). In addition, we restricted "negative depression" only to the cases when depression was never diagnosed, and respondents with a past diagnosis of depression were excluded.

### Combination of the XGBoost Model and Statistical Analysis: Feature Reduction, Indirect Factors, and Nonmodel Factors

XGBoost L1 regularization did not ultimately reduce the feature number; however, feature selection using XGBoost feature importance could decrease the feature number more at a higher reduction rate. Therefore, we suggest feature reduction using XGBoost feature importance to obtain a sparser model that contains the most principle factors of the disease.

Because of the relatively small number of predictors, a sparse model is interpretable without irrelevant features, which could be shown by the impact on the depression of each feature (Figure 2). Furthermore, we performed statistical tests on the sparse model features using multiple logistic regression at a lower risk of multicollinearity or redundancy problems. Overall, 44% of the model features were not statistically significant, and we classified them as indirect factors because they affected depression without a direct statistical association with depression.

Additionally, nonmodel factors (triglyceride, current asthma, and farm worker) exhibited statistical significance by multiple regression with depression model factors. By controlling principle depression factor effects (eg, confounding effects), nonmodel factors' significance was reliable. Therefore, we might use a sparse disease model as a testing tool for candidate factors.

### Elements of the Depression Network: Node Size and Edge

Our network was composed of differently sized nodes (statistical strength of the association) and edges (net correlation between two factors). "Perceived stress" and "current asthma" were prioritized in depression risk factor control because of their large nodes (Figure 3). In addition to genetic factors, stressful events affect the onset of depression and cause depression through psychological stress responses, such as activating the hypothalamic-pituitary-adrenal axis [31]. Additionally, the network indicated that "perceived stress" could accompany other risk factors, such as "pain and discomfort" and "weight gain," and we might consider them together for risk factor control (Figure 3). In contrast, current asthma was an independent node, which suggests asthma may directly link with depression (eg, asthma medications or dysregulation of specific stress-sensitive biological processes) and be an individual control target [32].

### Depression Network Centrality and Clusters: Indication of Predisposing Factors and Factor Groups

In terms of the network's centrality indexes, "male" and "educational level" showed the highest values. Scale-free networks are characterized by growth and preferential attachment in which earlier nodes in the network increase their connectivity at a higher rate [33]. Therefore, high-centrality nodes might be preceding factors. For example, the high centrality of "educational level" might explain that a higher educational level has accumulated protective effect throughout life [34]. Therefore, the network centrality suggested that gender and the education level would be predisposed to biological and socioeconomic depression factors. Statistically, gender and the education level might be must-have covariates because of their multiple relations to other factors.

Clusters in the disease network can be potent risk intervention targets because factors in a cluster are connected and controlled together [35]. The SES cluster on our depression network is reported as the most prominent risk among psychosocial risks [35], and low SES is common in depression because of poor care, low treatment compliance, and job strain [36]. For the QoL cluster, a study reported that the usual activity problem and the pain level of EQ-5D are mainly affected by depression, and the treatment of depression improves the EQ-5D index score [37]. Depression symptoms were negatively connected to health-related QoL in the network in a previous study [38].

### Confounding and Interaction Effects of Indirect Risk Factors

If an indirect risk factor is connected to a direct risk factor, the direct factor could be the confounder. Furthermore, if the direct factor is on the causal pathway, it can be a mediator of the indirect factor action. We found potential confounders of current osteoarthritis, hypercholesterolemia, and diabetes using the network (Table 3). Osteoarthritis causes pain and physical disability, which can reduce QoL and lead to the development of depression [39]. This can be seen in our network based on the relationships among current osteoarthritis, pain, and mobility problems in the QoL cluster (Figure 3).

Obesity and metabolic syndrome may mediate the relationship between diabetes and depression [40]. This fact could be captured in our network because diabetes was associated with weight gain, higher triglycerides, and hypercholesterolemia (Figure 3). Additionally, the elevated triglyceride level was a confounder of hypercholesterolemia (Table 3), which may explain a recent meta-analysis reporting that the first episode of major depressive disorder is associated with elevated triglyceride levels, not low-density lipoprotein cholesterol or total cholesterol levels [41]. Previous studies reported that an elevated triglyceride level is associated with depression and suicidality in men [3] or women [42]. However, there was no interaction effect between "male" and triglyceride level in our study ( $P=.82$ ).

If an indirect risk factor is connected to a direct protective factor or negatively associated with a direct risk factor, its role could be revealed by the interaction effect. "Current smoker," which is an indirect factor strongly connected to "male," was a

significant risk factor in females, but not in males (Figures 3 and 4). Smoking cessation might be considered, especially for female depression patients [43]. Weight gain occurred in depression in current diabetes, but not clearly in prediabetes or nondiabetes (Figure 4).

### USG as a Novel Depression-Related Factor

To the best of our knowledge, USG was a novel factor of depression, and the level was low in female depressive subjects (Figure 5). USG correlates with urine osmolality and reflects the subject's hydration status or the kidney's concentrating ability [44]. However, our study showed that low USG in depression might not be caused by hydration because daily water intake was not low in depression (Figure 6). One study suggested that patients with depression concentrate the urine less well and excrete less solute in the urine [45].

### Limitations

This cross-sectional study involved people aged 19 to 64 years and could analyze only associations, not causalities. Therefore, the associations established in this study, such as weight gain in diabetes, elevated triglyceride, low USG, and smoking in females, require future clinical research to prove their efficacy in depression control. We used Bonferroni correction for

nonmodel factors, and this conservative method might miss other possibly significant factors. Several cancers (hepatoma, gastric cancer, colon cancer, breast cancer, and lung cancer), renal failure, pulmonary tuberculosis, liver cirrhosis, and hepatitis C might not have enough positive cases to be tested. Finally, psychiatrists' structured interviews to diagnose depression would be more valid than the self-reported approach to identify depression used in this study.

### Conclusions

We successfully created a sparse model for depression using TPOT-assisted XGBoost training and feature selection based on the feature importance of XGBoost from a large number of variables in KNHANES data. Because of the data-driven approach, we could discover a novel factor. We constructed a network of the depression-associated factors using association strength and interfactor correlations. The model factors were classified into direct and indirect according to their statistical significance, and the role of indirect factors was explained by confounding or interaction effects. The network also indicated predisposing factors by high centrality and cluster factors by a closely connected edge. Therefore, XGBoost and network analysis can be useful for discovering and understanding disease-associated factors in epidemiological studies.

### Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2019R1C1C1007663, 2020R1F1A106842312, and 2019R1A2C1084611).

### Authors' Contributions

HWH and JIK contributed equally as corresponding authors.

### Conflicts of Interest

None declared.

### References

1. National Collaborating Centre for Mental Health. Depression: The NICE Guideline on the Treatment and Management of Depression in Adults (Updated Edition) (Nice Guidelines). Leicester, UK: British Psychological Society; 2010.
2. Hasler G. Pathophysiology of depression: do we have any solid evidence of interest to clinicians? *World Psychiatry* 2010 Oct;9(3):155-161 [FREE Full text] [doi: [10.1002/j.2051-5545.2010.tb00298.x](https://doi.org/10.1002/j.2051-5545.2010.tb00298.x)] [Medline: [20975857](https://pubmed.ncbi.nlm.nih.gov/20975857/)]
3. Kim B, Park EY. The combined effect of socioeconomic status and metabolic syndrome on depression: the Korean National Health and Nutrition Examination Survey (KNHANES). *BMC Public Health* 2020 May 04;20(1):617 [FREE Full text] [doi: [10.1186/s12889-020-08778-3](https://doi.org/10.1186/s12889-020-08778-3)] [Medline: [32366283](https://pubmed.ncbi.nlm.nih.gov/32366283/)]
4. Jager KJ, Zoccali C, Macleod A, Dekker FW. Confounding: what it is and how to deal with it. *Kidney Int* 2008 Feb;73(3):256-260 [FREE Full text] [doi: [10.1038/sj.ki.5002650](https://doi.org/10.1038/sj.ki.5002650)] [Medline: [17978811](https://pubmed.ncbi.nlm.nih.gov/17978811/)]
5. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol* 2008 Mar 01;167(5):523-9; discussion 530. [doi: [10.1093/aje/kwm355](https://doi.org/10.1093/aje/kwm355)] [Medline: [18227100](https://pubmed.ncbi.nlm.nih.gov/18227100/)]
6. Stoltzfus JC. Logistic regression: a brief primer. *Acad Emerg Med* 2011 Oct;18(10):1099-1104 [FREE Full text] [doi: [10.1111/j.1553-2712.2011.01185.x](https://doi.org/10.1111/j.1553-2712.2011.01185.x)] [Medline: [21996075](https://pubmed.ncbi.nlm.nih.gov/21996075/)]
7. Lima E, Davies P, Kaler J, Lovatt F, Green M. Variable selection for inferential models with relatively high-dimensional data: Between method heterogeneity and covariate stability as adjuncts to robust selection. *Sci Rep* 2020 May 14;10(1):8002 [FREE Full text] [doi: [10.1038/s41598-020-64829-0](https://doi.org/10.1038/s41598-020-64829-0)] [Medline: [32409668](https://pubmed.ncbi.nlm.nih.gov/32409668/)]
8. Dankers FJWM, Traverso A, Wee L, van Kuijk SMJ. Prediction Modeling Methodology. In: Kubben P, Dumontier M, Dekker A, editors. *Fundamentals of Clinical Data Science*. Cham: Springer; 2019:101-120.
9. Tlachac ML, Rundensteiner EA. Depression Screening from Text Message Reply Latency. *Annu Int Conf IEEE Eng Med Biol Soc* 2020 Jul;2020:5490-5493. [doi: [10.1109/EMBC44109.2020.9175690](https://doi.org/10.1109/EMBC44109.2020.9175690)] [Medline: [33019222](https://pubmed.ncbi.nlm.nih.gov/33019222/)]
10. Wang S, Pathak J, Zhang Y. Using Electronic Health Records and Machine Learning to Predict Postpartum Depression. *Stud Health Technol Inform* 2019 Aug 21;264:888-892. [doi: [10.3233/SHTI190351](https://doi.org/10.3233/SHTI190351)] [Medline: [31438052](https://pubmed.ncbi.nlm.nih.gov/31438052/)]

11. Gu S, Zhou J, Yuan C, Ye Q. Personalized prediction of depression in patients with newly diagnosed Parkinson's disease: A prospective cohort study. *J Affect Disord* 2020 May 01;268:118-126. [doi: [10.1016/j.jad.2020.02.046](https://doi.org/10.1016/j.jad.2020.02.046)] [Medline: [32158001](https://pubmed.ncbi.nlm.nih.gov/32158001/)]
12. Brownlee J. *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn*. San Juan, PR: Machine Learning Mastery; 2018.
13. Olson RS, Moore JH. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In: Hutter F, Kotthoff L, Vanschoren J, editors. *Automated Machine Learning*. The Springer Series on Challenges in Machine Learning. Cham: Springer; 2019:151-160.
14. Kweon S, Kim Y, Jang M, Kim Y, Kim K, Choi S, et al. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). *Int J Epidemiol* 2014 Feb;43(1):69-77 [FREE Full text] [doi: [10.1093/ije/dyt228](https://doi.org/10.1093/ije/dyt228)] [Medline: [24585853](https://pubmed.ncbi.nlm.nih.gov/24585853/)]
15. Yoon KC, Choi W, Lee HS, Kim S, Kim S, Kim CY, et al. An Overview of Ophthalmologic Survey Methodology in the 2008-2015 Korean National Health and Nutrition Examination Surveys. *Korean J Ophthalmol* 2015 Dec;29(6):359-367 [FREE Full text] [doi: [10.3341/kjo.2015.29.6.359](https://doi.org/10.3341/kjo.2015.29.6.359)] [Medline: [26635451](https://pubmed.ncbi.nlm.nih.gov/26635451/)]
16. Korea National Health and Nutrition Examination Survey. Korea Centers for Disease Control and Prevention. URL: <https://knhanes.kdca.go.kr/knhanes/eng/index.do> [accessed 2021-06-05]
17. EuroQol. EQ-5D-3L. URL: <https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/> [accessed 2021-06-05]
18. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020 Jan;2(1):56-67 [FREE Full text] [doi: [10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9)] [Medline: [32607472](https://pubmed.ncbi.nlm.nih.gov/32607472/)]
19. WeightedROC: Fast, Weighted ROC Curves. r-project. URL: <https://CRAN.R-project.org/package=WeightedROC> [accessed 2021-06-05]
20. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: Network Visualizations of Relationships in Psychometric Data. *J. Stat. Soft* 2012 Sep 1;48(4):1-18. [doi: [10.18637/jss.v048.i04](https://doi.org/10.18637/jss.v048.i04)]
21. Pereira-Morales AJ, Adan A, Forero DA. Network analysis of multiple risk factors for mental health in young Colombian adults. *J Ment Health* 2019 Apr;28(2):153-160. [doi: [10.1080/09638237.2017.1417568](https://doi.org/10.1080/09638237.2017.1417568)] [Medline: [29265896](https://pubmed.ncbi.nlm.nih.gov/29265896/)]
22. Robinaugh DJ, Millner AJ, McNally RJ. Identifying highly influential nodes in the complicated grief network. *J Abnorm Psychol* 2016 Aug;125(6):747-757 [FREE Full text] [doi: [10.1037/abn0000181](https://doi.org/10.1037/abn0000181)] [Medline: [27505622](https://pubmed.ncbi.nlm.nih.gov/27505622/)]
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011;12:2825-2830 [FREE Full text]
24. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August, 13, 2016; San Francisco, CA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
25. Lumley T. *Complex Surveys: A Guide to Analysis Using R*. Hoboken, NJ: John Wiley & Sons, Inc; 2010.
26. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer; 2016.
27. Pasek J. weights: Weighting and Weighted Statistics. r-project. URL: <https://CRAN.R-project.org/package=weights> [accessed 2021-06-05]
28. Brownlee J. *Imbalanced Classification with Python*. San Juan, PR: Machine Learning Mastery; 2020.
29. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire*. *JAMA* 1999 Nov 10;282(18):1737-1744. [doi: [10.1001/jama.282.18.1737](https://doi.org/10.1001/jama.282.18.1737)] [Medline: [10568646](https://pubmed.ncbi.nlm.nih.gov/10568646/)]
30. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ* 2012 Feb 21;184(3):E191-E196 [FREE Full text] [doi: [10.1503/cmaj.110829](https://doi.org/10.1503/cmaj.110829)] [Medline: [22184363](https://pubmed.ncbi.nlm.nih.gov/22184363/)]
31. Yang L, Zhao Y, Wang Y, Liu L, Zhang X, Li B, et al. The Effects of Psychological Stress on Depression. *Curr Neuropharmacol* 2015;13(4):494-504 [FREE Full text] [doi: [10.2174/1570159x1304150831150507](https://doi.org/10.2174/1570159x1304150831150507)] [Medline: [26412069](https://pubmed.ncbi.nlm.nih.gov/26412069/)]
32. Van Lieshout RJ, Bienenstock J, MacQueen GM. A review of candidate pathways underlying the association between asthma and major depressive disorder. *Psychosom Med* 2009 Feb;71(2):187-195. [doi: [10.1097/PSY.0b013e3181907012](https://doi.org/10.1097/PSY.0b013e3181907012)] [Medline: [19073754](https://pubmed.ncbi.nlm.nih.gov/19073754/)]
33. Barabasi, Albert R. Emergence of scaling in random networks. *Science* 1999 Oct 15;286(5439):509-512 [FREE Full text] [doi: [10.1126/science.286.5439.509](https://doi.org/10.1126/science.286.5439.509)] [Medline: [10521342](https://pubmed.ncbi.nlm.nih.gov/10521342/)]
34. Bjelland I, Krokstad S, Mykletun A, Dahl AA, Tell GS, Tambs K. Does a higher educational level protect against anxiety and depression? The HUNT study. *Soc Sci Med* 2008 Mar;66(6):1334-1345. [doi: [10.1016/j.socscimed.2007.12.019](https://doi.org/10.1016/j.socscimed.2007.12.019)] [Medline: [18234406](https://pubmed.ncbi.nlm.nih.gov/18234406/)]
35. Elovainio M, Hakulinen C, Pulkki-Råback L, Juonala M, Raitakari OT. A network approach to the analysis of psychosocial risk factors and their association with health. *J Health Psychol* 2020 Sep;25(10-11):1587-1600. [doi: [10.1177/1359105318765624](https://doi.org/10.1177/1359105318765624)] [Medline: [29607696](https://pubmed.ncbi.nlm.nih.gov/29607696/)]

36. Ervasti J, Vahtera J, Pentti J, Oksanen T, Ahola K, Kivimäki M, et al. Depression-related work disability: socioeconomic inequalities in onset, duration and recurrence. *PLoS One* 2013;8(11):e79855 [FREE Full text] [doi: [10.1371/journal.pone.0079855](https://doi.org/10.1371/journal.pone.0079855)] [Medline: [24278194](https://pubmed.ncbi.nlm.nih.gov/24278194/)]
37. Sobocki P, Ekman M, Agren H, Krakau I, Runeson B, Mårtensson B, et al. Health-related quality of life measured with EQ-5D in patients treated for depression in primary care. *Value Health* 2007;10(2):153-160 [FREE Full text] [doi: [10.1111/j.1524-4733.2006.00162.x](https://doi.org/10.1111/j.1524-4733.2006.00162.x)] [Medline: [17391424](https://pubmed.ncbi.nlm.nih.gov/17391424/)]
38. Solmi M, Koyanagi A, Thompson T, Fornaro M, Correll CU, Veronese N. Network analysis of the relationship between depressive symptoms, demographics, nutrition, quality of life and medical condition factors in the Osteoarthritis Initiative database cohort of elderly North-American adults with or at risk for osteoarthritis – CORRIGENDUM. *Epidemiol Psychiatr Sci* 2019 Mar 08;29. [doi: [10.1017/s2045796019000064](https://doi.org/10.1017/s2045796019000064)]
39. Jung JH, Seok H, Kim J, Song GG, Choi SJ. Association between osteoarthritis and mental health in a Korean population: a nationwide study. *Int J Rheum Dis* 2018 Mar;21(3):611-619. [doi: [10.1111/1756-185X.13098](https://doi.org/10.1111/1756-185X.13098)] [Medline: [28544532](https://pubmed.ncbi.nlm.nih.gov/28544532/)]
40. Semenkovich K, Brown ME, Svrakic DM, Lustman PJ. Depression in type 2 diabetes mellitus: prevalence, impact, and treatment. *Drugs* 2015 Apr;75(6):577-587. [doi: [10.1007/s40265-015-0347-4](https://doi.org/10.1007/s40265-015-0347-4)] [Medline: [25851098](https://pubmed.ncbi.nlm.nih.gov/25851098/)]
41. Wei Y, Cai D, Liu J, Liu R, Wang S, Tang Y, et al. Cholesterol and triglyceride levels in first-episode patients with major depressive disorder: A meta-analysis of case-control studies. *J Affect Disord* 2020 Apr 01;266:465-472. [doi: [10.1016/j.jad.2020.01.114](https://doi.org/10.1016/j.jad.2020.01.114)] [Medline: [32056914](https://pubmed.ncbi.nlm.nih.gov/32056914/)]
42. Oh J, Kim T. Serum lipid levels in depression and suicidality: The Korea National Health and Nutrition Examination Survey (KNHANES) 2014. *J Affect Disord* 2017 Apr 15;213:51-58. [doi: [10.1016/j.jad.2017.02.002](https://doi.org/10.1016/j.jad.2017.02.002)] [Medline: [28189965](https://pubmed.ncbi.nlm.nih.gov/28189965/)]
43. Sarris J, O'Neil A, Coulson CE, Schweitzer I, Berk M. Lifestyle medicine for depression. *BMC Psychiatry* 2014 Apr 10;14:107 [FREE Full text] [doi: [10.1186/1471-244X-14-107](https://doi.org/10.1186/1471-244X-14-107)] [Medline: [24721040](https://pubmed.ncbi.nlm.nih.gov/24721040/)]
44. Simerville JA, Maxted WC, Pahira JJ. Urinalysis: a comprehensive review. *Am Fam Physician* 2005 Mar 15;71(6):1153-1162 [FREE Full text] [Medline: [15791892](https://pubmed.ncbi.nlm.nih.gov/15791892/)]
45. Ellis GG, Copen A, Glen AI. Urine concentration in depressive illness. *J Neurol Neurosurg Psychiatry* 1971 Feb;34(1):30-31 [FREE Full text] [doi: [10.1136/jnnp.34.1.30](https://doi.org/10.1136/jnnp.34.1.30)] [Medline: [5551692](https://pubmed.ncbi.nlm.nih.gov/5551692/)]

## Abbreviations

**AUC:** area under the receiver operating characteristic curve  
**EBICglasso:** extended Bayesian information criterium graphical lasso  
**EQ-5D-3L:** EuroQol five-dimension three-level  
**KCDC:** Korea Centers for Disease Control and Prevention  
**KNHANES:** Korea National Health and Nutrition Examination Survey  
**LASSO:** least absolute shrinkage and selection operator  
**PHQ-9:** Patient Health Questionnaire-9  
**PSU:** primary sampling unit  
**QoL:** quality of life  
**ROC:** receiver operating characteristic  
**SES:** socioeconomic status  
**TPOT:** tree-based pipeline optimization tool  
**USG:** urine specific gravity  
**XGBoost:** extreme gradient boosting

*Edited by R Kukafka; submitted 22.01.21; peer-reviewed by X Cheng, M Pradhan; comments to author 26.02.21; revised version received 06.03.21; accepted 06.05.21; published 24.06.21*

### *Please cite as:*

Nam SM, Peterson TA, Seo KY, Han HW, Kang JI

*Discovery of Depression-Associated Factors From a Nationwide Population-Based Survey: Epidemiological Study Using Machine Learning and Network Analysis*

*J Med Internet Res* 2021;23(6):e27344

URL: <https://www.jmir.org/2021/6/e27344/>

doi: [10.2196/27344](https://doi.org/10.2196/27344)

PMID:

Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.