ORIGINAL ARTICLE
https://doi.org/10.30773/pi.2024.0156

On-line ISSN 1976-3026
OPEN ACCESS

Comparison of Logistic Regression and Machine Learning Approaches in Predicting Depressive Symptoms: A National-Based Study

Xing-Xuan Dong[1], Jian-Hua Liu[1], Tian-Yang Zhang[1,2,3], Chen-Wei Pan[1], Chun-Hua Zhao[4] ✉, Yi-Bo Wu[5] ✉, and Dan-Dan Chen[6,7] ✉

[1]School of Public Health, Suzhou Medical College of Soochow University, Suzhou, China
[2]Research Center for Psychology and Behavioral Sciences, Soochow University, Suzhou, China
[3]Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Okayama, Japan
[4]Department of General Medicine, Medical Big Data Center, the Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Nanjing Medical University, Suzhou, China
[5]School of Public Health, Peking University, Beijing, China
[6]State Key Laboratory of Radiation Medicine and Protection, School of Radiation Medicine and Protection, Soochow University, Suzhou, China
[7]Collaborative Innovation Center of Radiological Medicine of Jiangsu Higher Education Institutions, Suzhou, China

Objective Machine learning (ML) has been reported to have better predictive capability than traditional statistical techniques. The aim of this study was to assess the efficacy of ML algorithms and logistic regression (LR) for predicting depressive symptoms during the COVID-19 pandemic.

Methods Analyses were carried out in a national cross-sectional study involving 21,916 participants. The ML algorithms in this study included random forest (RF), support vector machine (SVM), neural network (NN), and gradient boosting machine (GBM) methods. The performance indices were sensitivity, specificity, accuracy, precision, F1-score, and area under the receiver operating characteristic curve (AUC).

Results LR and NN had the best performance in terms of AUCs. The risk of overfitting was found to be negligible for most ML models except for RF, and GBM obtained the highest sensitivity, specificity, accuracy, precision, and F1-score. Therefore, LR, NN, and GBM models ranked among the best models.

Conclusion Compared with ML models, LR model performed comparably to ML models in predicting depressive symptoms and identifying potential risk factors while also exhibiting a lower risk of overfitting. **Psychiatry Investig 2025;22(3):267-278**

Keywords Logistic regression; Machine learning; Depressive symptoms; Risk factors.

INTRODUCTION

Machine learning (ML) encompasses statistical modeling techniques and automated algorithms for decision-making. These methods are better equipped to analyze large and complex datasets than traditional statistical methods.[1,2] Moreover, the predictive capability of ML has often been emphasized, especially in the modeling of rarely perfectly linear relationships,[1,3] and the predictive potential of ML for diseases has been demonstrated in previous studies.[1,3-5]

Received: May 7, 2024 **Revised:** September 1, 2024 **Accepted:** January 10, 2025

✉ **Correspondence:** Chun-Hua Zhao, MSc
Department of General Medicine, Medical Big Data Center, the Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Nanjing Medical University, Suzhou 215000, China
Tel: +86-18862243116, **E-mail:** chzhao@njmu.edu.cn

✉ **Correspondence:** Yi-Bo Wu, PhD, MPharm
School of Public Health, Peking University, Beijing 100191, China
Tel: +86-18810169630, **E-mail:** bjmuwuyibo@outlook.com

✉ **Correspondence:** Dan-Dan Chen, PhD
State Key Laboratory of Radiation Medicine and Protection, School of Radiation Medicine and Protection, Soochow University, 199 Renai Road, Suzhou 215123, China
Tel: +86-512-65880063, **E-mail:** dchen361@163.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

However, the use of ML models for disease prediction presents several challenges that warrant consideration. In imbalanced datasets, where the number of positive cases is far outweighed by the number of negative cases, this approach often poses a major challenge.[1,6,7] Overfitting poses another challenge, where a model that achieves high performance on the training set fails to generalize to new cases. While advanced sampling techniques and ensemble methods can help mitigate the risks of data imbalance and overfitting, large datasets are required for ML algorithms to reliably identify data patterns.[1,8]

While interest in using ML for disease prediction has increased rapidly, it remains unclear whether ML applications outperform traditional regression algorithms. Several studies have reported limited predictive abilities of ML models,[9] and others have shown that traditional models surpass ML models on certain statistical metrics, such as area under the receiver operating characteristic curve (AUC) and Akaike information criterion.[10-12] The predictive performance also varies among different ML models.[5,13,14] Therefore, it is crucial to systematically compare and understand the predictive capabilities of both traditional regression techniques and ML algorithms for disease prediction in large populations.

ML algorithms have been applied to identify predictors of psychological distress and depression in previous studies.[2] During the COVID-19 pandemic, higher scores of anxiety and depression were revealed than before the pandemic,[15] and the increasing tendency of this subclinical form of depression should not be ignored in public health.[16] Thus, this study evaluated the abilities of ML and logistic regression (LR) to predict the risks of depressive symptoms during the COVID-19 quarantine period in China. Our hypothesis is that ML and LR algorithms will exhibit similar predictive performance on depressive symptoms.

METHODS

Study population

We conducted an analysis of the psychology and behavior of Chinese residents, a national cross-sectional survey from 20 June to 31 August 2022 in mainland China that included 148 cities in all provinces. Approximately 30,000 people participated in the survey with or without the help of investigators (response rate 71.8%, eligibility rate 96.8%), and 21,916 participants (50% men) were included in this cross-sectional study. Ineligible participants included people who were less than 12 years old, who were not of any nationality in China, who resided abroad for more than one month, who were mentally disabled, and who were involved in other similar studies.

The data were also subjected to the following exclusion criteria. First, the questionnaires were excluded with a response time of less than 4 minutes. Second, the questionnaires were excluded with inconsistent LR outcomes (e.g., aged less than 18 years but with a marital status identified as married, divorced, or widowed). Third, questionnaires were excluded with incomplete information. Fourth, any repeated questionnaires or questionnaires completed in a regularly repeated manner were excluded. The final dataset was balanced, ensuring an equal number of man and woman respondents.

The survey was approved by the Medical Ethics Committee (No.JKWH-2022-02) and the Clinical Research Ethics Committee of the Second Xiangya Hospital of Central South University (No.2022-K050).

Definitions of outcomes

The definition of depressive symptoms considered in this study was obtained from the 9-item Patient Health Questionnaire (PHQ-9) following Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, the diagnostic criteria for depression.[17] The PHQ-9 has been widely used in clinical practice and has good reliability, and its validity has been proven in previous studies.[16,18,19] Each item is rated from 0 (never) to 3 (nearly every day); thus, the PHQ-9 score ranges from 0 to 27, which corresponds to levels of depressive symptoms. The Cronbach's alpha of PHQ-9 was 0.921 in the present study. The definition of depressive symptoms was then dichotomized by a score threshold of 5, that is, the scores of 4 or less were classified as the absence of depressive symptoms.[17]

Predictors considered

Sociodemographic variables

The following demographic variables were collected: age, gender (man or woman), ethnicity (Han or non-Han nationality), residency (urban or rural), education level (graduate/doctoral, college, middle school, primary school, or illiterate/semiliterate), monthly income (categorized as ≤1,000 Chinese Yuan [CNY] per month, 1,001−3,000 CNY, 3,001−5,000 CNY, or >5,000 CNY), living arrangement (living alone or not), occupation (employed/student or not), obesity status (yes: body mass index [BMI] ≥25, or no: BMI<25 kg/m^2), and marital status (yes or no).

Social environment variables

The participants' perceptions of their neighborhood (assessed on a scale of 1 to 7), experiences of negative life events (binary response of yes or no), and self-assessed social status (rated on a scale of 1 to 7) were examined.

Subjective indicators

The following measures were used to assess subjective in-

dicators: stress (Perceived Stress Scale-4 items [PSS-4], and Cronbach's alpha=0.668),[20] self-efficacy (New General Self-Efficacy Scale [NGSES], and Cronbach's alpha=0.925),[21] social support (Multidimensional Scale of Perceived Social Support [MSPSS], Cronbach's alpha=0.888),[22] loneliness (Three-Item Loneliness Scale [T-ILS], Cronbach's alpha=0.862)[23] and subjective well-being (World Health Organization Well-Being Index [WHO-5], Cronbach's alpha=0.933).[24]

Individual behavior variables

The following variables were assessed in this study: Problematic internet use (Problematic Internet Use Questionnaire-Short Form-6 [PIUQ-SF6], Cronbach's alpha=0.932),[25-27] subjective sleep quality (Brief Version of the Pittsburgh Sleep Quality Index [B-PSQI], Cronbach's alpha=0.686),[28] late bedtime (frequency of late bedtime per week, rated from 1 to 7), screens before bedtime (≤0.5, 0.5−1, or >1 hour), physical activity (7-day recall of International Physical Activity Questionnaire [IPAQ-7]),[29,30] smoking (yes or no), alcohol drinking (yes or no), coffee drinking (≥3 cups per day, 1−2 or never), tea drinking (yes or no), sugar beverage (7, 4−6, 1−3 bottles per week, or never), takeaway food (3, 1−2, ≤1 time per week, or never), water intake (<1,200, 1,200−1,499, 1,500−1,699 mL per day, or 1,700−2,099 mL per day), chronic disease (yes or no), and presence of eye disease (yes or no).

COVID-19-related variables

The following variables were assessed in relation to the COVID-19 pandemic: home quarantine (yes or no), lockdown (yes or no), community closure (yes or no), and perceived influence on daily life (rated on a scale of 0−100 to indicate the extent of the pandemic's impact).

Modeling strategy

All analyses were conducted by using R software (version 3.5.1; R Core Team, 2019, https://www.R-project).

Data preprocessing

Several preprocessing steps were conducted to prepare data and select features. First, the data with missing values were removed. Second, categorical predictors were transformed into indicator variables. Continuous variables were standardized by subtracting the mean and dividing by the standard deviation. Third, for feature selection, the t-test was applied to determine if there was a significant mean difference in a predictor between the groups with and without depressive symptoms. The features (obesity and Han ethnicity) with the p-values larger than 0.05 were removed from further modeling process. Fourth, pair correlations were checked to avoid issues of collinearity. Based on the coefficient threshold 0.75, only the pair of community closure and city lockdown was identified as high correlation, and the predictor community closure was removed due to its lower correlation with the depression symptom. Therefore, 33 predictors were finally selected for our modeling.

Models considered

The following four models were considered: LR, random forest (RF), neural network (NN), support vector machine (SVM), and gradient boosting machine (GBM). The LR model was applied in two approaches. One was the traditional type with default configurations, and another was optimized by cross-validation with Lasso (L1) penalty. The LR model performance was also compared to that of ML models whose best hyperparameters were selected according to the highest AUC values (modeling details in Supplementary Material).

For GBM modeling, the hyperparameter tuning such as early stopping and regularization was concerned to control model complexity and preventing overfitting. Early stopping is a technique where model training stops automatically when the validation error stops improving for a certain number of iterations (trees). To incorporate the technique in our modeling, the key parameter is n.trees, which specifies the total number of boosting iterations (the number of trees) the model should build. It can be used to mimic early stopping by using cross-validation to identify the optimal number of trees.

In gradient boosting, L1 (Lasso) regularization encourages sparsity by driving some weights (or coefficients) to zero, effectively reducing the number of features. In our modeling, setting the parameter n.minobsinnode (minimum observations in terminal node) to a higher value is similar to L1 regularization. It means that each terminal node (leaf) in the tree must have at least the specified minimum number of observations, which avoids creating overly small leaves and then reduces model complexity. Another regularization L2 (Ridge) penalizes the sum of squared coefficients, which helps reduce the magnitude of the weights, making the model less sensitive to individual features. In our modeling, lower values of shrinkage (learning rate) force the model to learn slowly with smaller step sizes, which is similar to how L2 regularization limits large changes in weights. A smaller shrinkage value typically requires more boosting rounds (n.trees), but it helps smooth out the learning and preventing extreme responses to individual observations, similar to the coefficients smoothing in L2. In addition, the parameter interaction.depth (tree depth) limits the depth of each tree, and shallower trees are less likely to fit noise in the training data, making the model more regularized, acting similarly to L2 by reducing complexity.

Resampling strategy

The data were randomly divided into a training set (70%) and a testing set (30%). The optimal hyperparameters were specifically chosen by five-fold cross-validation using the training set. This resampling method split the training data into five folds equally, and then the model was trained on four of these folds and validated on the remaining one. This process was repeated five times, each time using a different fold as the validation set, and the performance metrics were averaged over five iterations. After the optimization, the model was run in the entire training sets, and then used for prediction on the testing set.

Predictive performance

Six indices of predictive performance were considered: AUC, accuracy, precision, sensitivity, specificity, and F1-score. The AUCs with the 95% confidence interval (CI) were calculated for both training and testing sets, and the AUC differences were considered to indicate the potential risk of model overfitting. The AUC values of different models were compared by paired Delong's tests. The F1-score is the harmonic mean of precision and recall, which symmetrically represents both precision and recall in one metric. The model with the highest training AUC was selected to rank the variable importance for the risks of depressive symptoms.

RESULTS

Characteristics of the data

Table 1 presents the characteristics of 21,916 participants in this cross-sectional study, including sociodemographic variables, social environmental indicators, subjective indicators, individual behavior, and COVID-19-related variables. The man population represented 50% of the sample, with a mean age of 39.4±18.9 years. Additionally, 9.9% of the participants were adolescents (aged 12−18 years), and 18.8% were older than 60 years. More than half of the participants were urban (69.3%), and only a small number lived alone (14.3%). The education levels included primary (39.8%), college (40.8%), and postgraduate/doctoral (3.8%) education; 5.4% were illiterate/semiliterate. The monthly household income per capita (in CNY) was considered to be at four different levels: >5,000 CNY (36.6%), 3001−5,000 (30.4%), 1,001−3,000 (26.7%), and <1,000 (6.3%). The prevalence of obesity was 16.8%, and 56.7% of participants were married.

The risk of depressive symptoms was 57.6% among the participants in this study. The univariate analysis in Table 1 revealed that the young participants tended to have more depressive symptoms than the older participants ($p<0.001$), and the women were more likely to develop depressive symptoms than were the men ($p<0.001$). It was also more common in urban residents and solitary or unmarried individuals (all $p<0.001$). Its relationship with the social environment was found in poorer neighborhoods, with negative life-event experiences and with lower social status (all $p<0.001$). The subjective indicators of the participants suggested a strong correlation with depressive symptoms (all $p<0.001$), such as higher stress, lower self-efficacy, lower social support, greater loneliness, and lower subjective well-being. Furthermore, greater impacts of COVID-19 were reported among participants with depressive symptoms than among those without depressive symptoms ($p<0.001$). In addition, significant correlations were observed between certain individual behaviors and depressive symptoms. Participants were more likely to experience depressive symptoms ($p<0.001$) if they went to bed late, had poor sleep quality, or had long screen time before sleep. Depressive symptoms were more common with longer problematic internet use and less physical activity. Participants with depressive symptoms had the following dietary habits: smoking, alcohol intake, more coffee or tea intake, more sugar beverage, more takeaway food consumption, and less water intake (all $p<0.001$). Additionally, the presence of chronic diseases and eye diseases was associated with a greater likelihood of depressive symptoms (all $p<0.001$).

Predictive performance

Table 2 presented the predictive performances in training and testing datasets of all models. There were small AUC differences between the training and testing cohorts among ML algorithms except for RF, which may imply that the issue of overfitting cannot be neglected in RF. Among the non-overfitting models, the LR, SVM and NN algorithms had similar AUC values around 0.86 (LR=0.862, 95% CI=0.853−0.871; SVM=0.860, 95% CI=0.851−0.869; NN=0.866, 95% CI=0.857−0.875), while the GBM had slightly better performance with the highest AUC (0.874, CI=0.865−0.883), specificity (0.818), accuracy (0.798), precision (0.769), and F1-score (0.771). According to AUC differences between training and testing, LR had the smallest difference (0.69%) and GBM had a bit larger one (2.3%). Furthermore, it seemed that the performances between the two LR models were almost the same either with or without optimization, which implied that the traditional LR was good enough for our data analysis.

The training and testing AUC values of the different models were compared by the paired Delong's test (Supplementary Table 1). Although there were significant differences among those models in the training set ($p<0.01$), the LR was found to correlate with other ML models except RF in the testing set. Overall, the performances of LR and the other models seemed systematically similar except that of RF in our study.

Table 1. Data characteristics of this study

Characteristics	Total (N=21,916)	Dep (N=12,618)	No Dep (N=9,298)	p
Sociodemographic variable				
Age (yr)	39.4±18.9	38.3±18.7	41.0±19.0	<0.001
12–17	2,178 (9.9)	1,203 (9.5)	975 (10.5)	
18–39	9,318 (42.5)	5,852 (46.4)	3,466 (37.3)	
40–59	6,297 (28.7)	3,342 (26.5)	2,955 (31.8)	
≥60	4,123 (18.8)	2,221 (17.6)	1,902 (20.5)	
Man	10,958 (50.0)	6,138 (48.6)	4,820 (51.8)	<0.001
Han ethnicity	19,970 (91.1)	11,504 (91.2)	8,466 (91.05)	0.759
Urban residency	15,188 (69.3)	8,622 (68.3)	6,566 (70.6)	<0.001
Education				<0.001
Graduate/doctoral	830 (3.8)	494 (3.9)	336 (3.6)	
College	8,943 (40.8)	5,489 (43.5)	3,454 (37.1)	
Middle school	8,731 (39.8)	4,828 (38.3)	3,903 (42.0)	
Primary school	2,234 (10.2)	1,153 (9.1)	1,081 (11.6)	
Illiterate/semi-literate	1,178 (5.4)	654 (5.2)	524 (5.6)	
Monthly income (CNY)				<0.001
≤1,000	1,375 (6.3)	902 (7.1)	473 (5.1)	
1,001–3,000	5,854 (26.7)	3,316 (26.3)	2,538 (27.3)	
3,001–5,000	6,655 (30.4)	3,674 (29.1)	2,981 (32.1)	
>5,000	8,032 (36.6)	4,726 (37.5)	3,306 (35.6)	
Living alone	3,141 (14.3)	2,029 (16.1)	1,112 (12.0)	<0.001
Employed/student	16,790 (76.6)	9,803 (77.7)	6,987 (75.1)	<0.001
Obesity	3,690 (16.8)	2,061 (16.3)	1,629 (17.5)	0.02
Married	12,437 (56.7)	6,664 (52.8)	5,773 (62.1)	<0.001
Social environment				
Neighborhood	5.8±1.3	5.6±1.3	6.0±1.1	<0.001
Negative life event	9,160 (41.8)	6,513 (51.6)	2,647 (28.5)	<0.001
Social status	4.4±1.3	4.2±1.3	4.5±1.3	<0.001
Subjective indicator				
Stress	6.1±3.1	6.7±2.9	5.3±3.2	<0.001
Self-efficacy	7.8±2.4	7.3±2.3	8.5±2.4	<0.001
Social support	12.0±3.8	11.1±3.5	13.3±3.8	<0.001
Loneliness	4.6±1.6	5.3±1.6	3.6±1.0	<0.001
Subjective well-being	14.4±6.0	12.6±5.4	16.8±5.9	<0.001
Individual behavior				
Problematic internet use	11.7±5.5	13.5±5.6	9.3±4.3	<0.001
Subjective sleep quality	4.4±2.95	5.1±3.0	3.3±2.5	<0.001
Sleep latency (days per wk)	1.5±1.8	1.9±1.9	1.0±1.4	<0.001
Screens before bedtime (h)				<0.001
≤0.5	10,947 (49.9)	5,370 (42.6)	5,577 (60.0)	
0.5–1	5,831 (26.6)	3,602 (28.5)	2,229 (24.0)	
>1	5,138 (23.4)	3,646 (28.9)	1,492 (16.0)	
Physical activity	2.0±1.1	1.9±1.2	2.1±1.0	<0.001
Smoking	4,015 (18.3)	2,546 (20.2)	1,469 (15.8)	<0.001

Table 1. Data characteristics of this study (continued)

Characteristics	Total (N=21,916)	Dep (N=12,618)	No Dep (N=9,298)	p
Alcohol	6,702 (30.6)	4,301 (34.1)	2,401 (25.8)	<0.001
Coffee (cups per day)				<0.001
Never	16,994 (77.5)	9,236 (73.2)	7,758 (83.4)	
1–2	4,438 (20.3)	3,039 (24.1)	1,399 (15.0)	
≥3	484 (2.2)	343 (2.7)	141 (1.5)	
Tea	11,005 (50.2)	6,810 (54.0)	4,195 (45.1)	<0.001
Sugar beverage (bottles per wk)				<0.001
Never	8,403 (38.3)	4,021 (31.9)	4,382 (47.1)	
1–3	9,595 (43.8)	5,857 (46.4)	3,738 (40.2)	
4–6	2,768 (12.6)	1,959 (15.5)	809 (8.7)	
≥7	1,150 (5.25)	781 (6.2)	369 (4.0)	
Takeaway food (times per wk)				<0.001
Never	4,648 (21.2)	2,193 (17.4)	2,455 (26.4)	
<1	8,011 (36.6)	4,594 (36.4)	3,417 (36.7)	
1–2	6,105 (27.9)	3,775 (29.9)	2,330 (25.1)	
≥3	3,152 (14.4)	2,056 (16.3)	1,096 (11.8)	
Water intake (mL)				<0.001
<1,200	7,453 (34.0)	4,806 (38.1)	2,647 (28.5)	
1,200–1,499	8,037 (36.7)	4,430 (35.1)	3,607 (38.8)	
1,500–1,699	4,805 (21.9)	2,552 (20.2)	2,253 (24.2)	
1,700–2,099	1,621 (7.4)	830 (6.6)	791 (8.5)	
Chronic disease	5,684 (25.9)	3,697 (29.3)	1,987 (21.4)	<0.001
Eye disease	10,211 (46.6)	6,674 (52.9)	3,537 (38.0)	<0.001
COVID-19 related variables				
Home quarantine	828 (3.8)	543 (4.3)	285 (3.1)	<0.001
Lockdown	1,245 (5.7)	792 (6.3)	453 (4.9)	<0.001
Community closure	932 (4.3)	603 (4.8)	329 (3.5)	<0.001
Influence on daily life (0–100)	60.5 (26.8)	61.1 (26.2)	59.6 (27.4)	<0.001

Values are presented as number (%) or mean±standard deviation. Dep, depressive symptom; CNY, Chinese Yuan

Figure 1 is the AUC learning curves for different models, where they were trained on increasing subsets of the training set. The modeling performances were evaluated not only by the training set (dash-dot-dot green), but also by the testing set (dash red) and the averaged results of cross-validation during the model training process (solid blue). There was a large gap between the training and the validation AUCs in the RF model, which suggested a variance problem which probably led to overfitting issues. The three curves of LR and SVM models almost overlapped with high values of AUCs, which implied the models were well-generalized. In terms of GBM model, its curves seemed to be a bit higher than those of LR, but its curve gap was also larger though. All of these observations agreed with the results of Table 2 above.

Variable importance

The variable importance of 36 predictors from all models were ranked in Table 3. The rank in LR was based on the absolute value of the z-statistic, and the ranks of ML models were obtained by the function "varImp" in R. In RF, the variable importance was approached by the mean decrease in Gini impurity; in GBM, it was based on the relative influence of each feature, considering both frequency and gain across all trees; and in SVM, it was determined by the absolute values of the model coefficients with the linear kernel we used.

Although the top 10 variables of different models were not exactly the same, the five were always listed: loneliness, stress, problematic internet use, late bedtime, and subjective well-being. It implied a statistical correlation of subjective indicators (loneliness, stress) and individual behaviors (late bedtime,

Table 2. AUCs of the models for training and testing datasets

Model	Training						Testing						
	AUC [95% CI]	Sens	Spec	Accu	Prec	F1	AUC [95%CI]	Sens	Spec	Accu	Prec	F1	ΔAUC (%)
LR	0.868 [0.862, 0.874]	0.767	0.805	0.788	0.751	0.759	0.862 [0.853, 0.871]	0.770	0.803	0.789	0.754	0.762	0.69
LR2	0.868 [0.862, 0.873]	0.766	0.805	0.788	0.752	0.759	0.862 [0.853, 0.872]	0.771	0.803	0.789	0.754	0.762	0.69
RF	1.000 [1.000, 1.000]	1.000	1.000	1.000	1.000	1.000	0.877 [0.868, 0.885]	0.767	0.829	0.802	0.778	0.773	23.3
SVM	0.867 [0.861, 0.873]	0.781	0.798	0.791	0.749	0.765	0.860 [0.851, 0.869]	0.777	0.792	0.785	0.745	0.761	0.81
GBM	0.896 [0.891, 0.901]	0.788	0.834	0.814	0.785	0.787	0.874 [0.865, 0.883]	0.773	0.818	0.798	0.769	0.771	2.5
NN	0.880 [0.875, 0.886]	0.778	0.816	0.799	0.764	0.771	0.866 [0.857, 0.875]	0.768	0.804	0.788	0.754	0.761	1.6

ΔAUC is the percentage of AUC differences between training and testing. AUC, area under the receiver operating characteristic curve; LR, logistic regression; LR2, optimized logistic regression with Lasso (L1) penalty; RF, random forest; SVM, support vector machine; GBM, gradient boosting machine; NN, neural network; CI, confidence interval; Sen, sensitivity; Spec, specificity; Accu, accuracy; Prec, precision

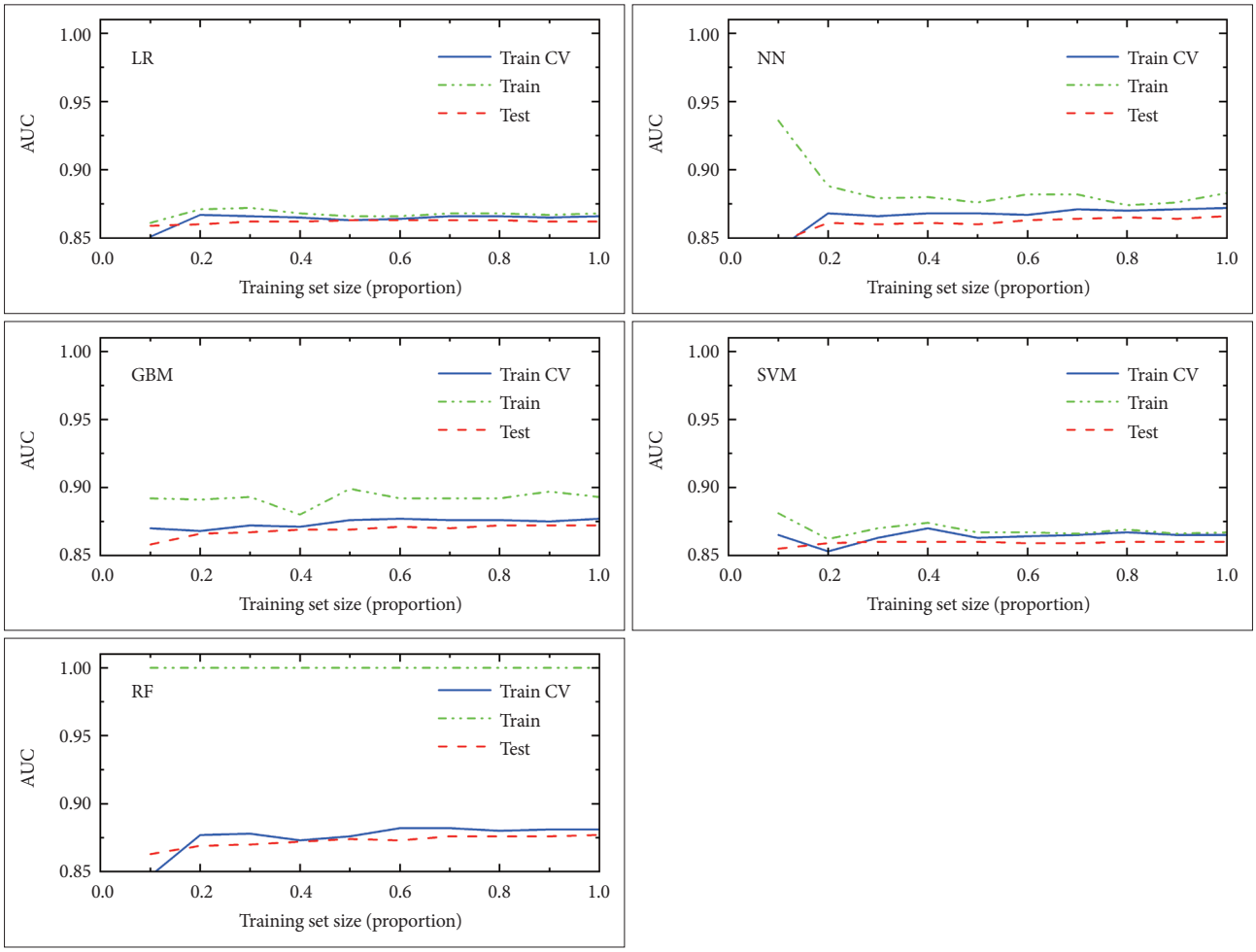

Figure 1. The AUC learning curves of the five models (LR, RF, NN, SVM, and GBM) which were optimized by CV. The x-axis was the proportion of training subsets. The blue solid line was the mean AUC from the CV, the green dash-dot-dot was from the prediction on training subsets, and the red dash was from the prediction on the testing set. AUC, area under the receiver operating characteristic curve; LR, logistic regression; RF, random forest; NN, neural network; SVM, support vector machine; GBM, gradient boosting machine; CV, cross-validation.

problematic internet use) with depressive symptoms. Meanwhile, the COVID-19-related variables (home quarantine, lockdown, community closure, and influence on daily life) were ranked relatively low as the less important predictors, indicating that their influence on depressive symptoms was limited.

Interactions among predictors

Figure 2 shows the GBM performance by the relationship between the mean training AUC and the maximum interaction depth with the tree numbers from 50 to 300. The AUC for the low tree number 50 was the lowest at interaction depth 1,

Table 3. Predictor importance for the five models

	LR	RF	NN	GBM	SVM
1	Loneliness	Loneliness	Loneliness	Loneliness	Loneliness
2	Stress	SubWB	Stress	SubWB	SubWB
3	ProbNet	Stress	SubWB	Stress	Stress
4	NegEvent	ProbNet	ProbNet	ProbNet	ProbNet
5	ChroDis	SocSupp	ChroDis	LatBed	socialsuppor
6	LatBed	Age	NegEvent	Age	Self-efficacy
7	SubWB	InfDL	LatBed	SocSupp	LatBed
8	EyeDis	LatBed	Gender	InfDL	NegEvent
9	Gender	Self-efficacy	Married	NegEvent	ScrB
10	Smoking	SubSQ	SocSupp	ChroDis	SugBev
11	Married	SocSta	Smoking	SocSta	EyeDis
12	Alcohol	Neighborhood	EyeDis	Self-efficacy	Neighborhood
13	HQua	TakFood	HQua	SugBev	TakFood
14	SocSupp	Water intake	Alcohol	Neighborhood	SocSta
15	Residency	MInc	Lockdown	SubSQ	Water intake
16	Self-efficacy	Education	SocSta	EyeDis	Coffee
17	Age	NegEvent	ScrB	ScrB	Married
18	SugBev	SugBev	SugBev	PA	ChroDis
19	Neighborhood	ScrB	Self-efficacy	TakFood	Age
20	SocSta	PA	Residency	Gender	Alcohol
21	ScrB	EyeDis	SubSQ	Water intake	Tea
22	SubSQ	Tea	LivAlone	Education	Education
23	Water intake	Gender	InfDL	MInc	PA
24	Emp/Stu	ChroDis	Neighborhood	Smoking	Gender
25	MInc	Residency	Water intake	Alcohol	Smoking
26	TakFood	Alcohol	Coffee	Tea	LivAlone
27	LivAlone	Coffee	Age	Coffee	Emp/Stu
28	Education	Married	Emp/Stu	LivAlone	InfDL
29	Tea	Smoking	Education	Residency	Residency
30	Lockdown	Emp/Stu	MInc	Emp/Stu	Lockdown
31	PA	LivAlone	Tea	Married	SubSQ
32	InfDL	Lockdown	PA	Lockdown	HQua
33	Coffee	HQua	TakFood	HQua	MInc

LR, logistic regression; RF, random forest; NN, neural network; GBM, gradient boosting machine; SVM, support vector machine; ProbNet, problematic internet use; SubWB, subjective well-being; NegEve, negative life event; ChroDis, chronic disease; SocSupp, socialsupport; LatBed, late bedtime; MInc, monthly income; EyeDis, eye disease; ScrB, screens before bedtime; InfDL, influence on daily life; SocSta, social status; HQua, home quarantine; PA, physical activity; SubSQ, subjective sleep quality; Emp/Stu, employed/student; SugBev, sugar beverage; LivAlone, living alone; TakFood, takeaway food

but it made a large jump at tree depth 2, and then the increased steps became smaller with deeper interaction. The AUC values of tree number 100 were higher at the low interaction depth than tree number 50, but the gap between these two curves was shrinking as the interaction depth increased. Similar tendency was found among other large tree numbers, where their curves were closer between each other as the interaction depth increased and reached a plateau around 0.875 at the tree depth 6. After the optimization of cross-validation, the hyperparameters of the best GBM model were found to be tree number 150 and max tree depth 6.

Nonlinearity

Figure 3 showed the nonlinear effect of age on the predic-

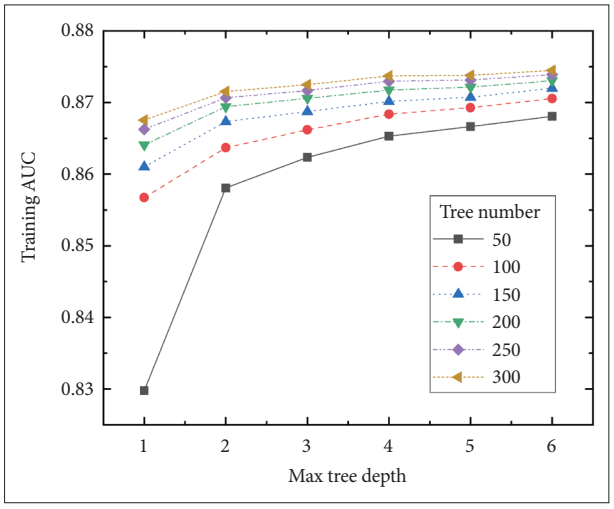

Figure 2. AUC performances of GBM models for the training data as the function of maximum interaction depth. The symbol points from black square to yellow triangle correspond to the range of tree numbers from 50 to 300. AUC, area under the receiver operating characteristic curve.

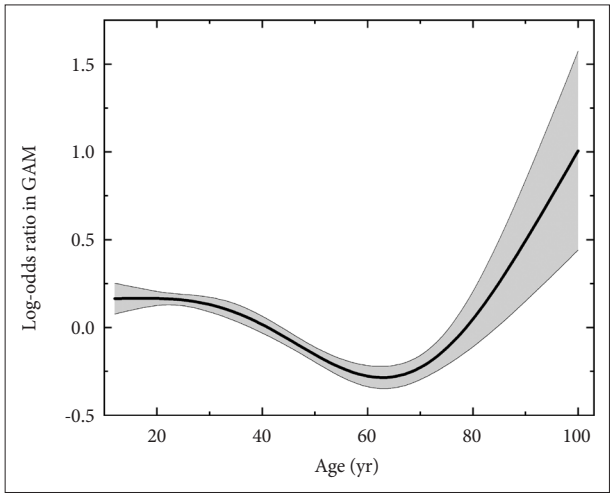

Figure 3. The nonlinear effect of age on the prediction of depressive symptoms using a GAM. GAM, generalized additive model.

tion of depressive symptoms using a generalized additive model (GAM). The smooth curve had a minimum at approximately 60 years of age according to visual inspection; the degrees of freedom were greater than 1 (2.95), and the p-value associated with each smoothing term was less than 0.001. Thus, the visual examination and the statistical tests indicated the presence and significance of nonlinearity. Considering the nonlinearity effect of age, we recalculated the variable importance for the age subgroups 18–39, 40–59, and 60–100 years old (Supplementary Table 2).

DISCUSSION

This large-scale epidemiological study demonstrated that the traditional LR model is comparable to the ML models for predicting depressive symptoms. The AUC values on the testing set estimated the similar performances of LR, SVM, and NN models, while those of GBM and RF were found larger by a small amount. The likelihood of overfitting was investigated by comparing the differences between the AUCs on training and testing data, which turned out to be the smallest in LR (0.69%) but the largest in RF (23.3%). Each pair of models were also compared by the pair Delong's test on these AUCs. There were no significant differences between LR and other machine models except the RF model which had an obvious overfitting issue. Furthermore, the traditional LR was much simpler without optimization, and easier to interpret comparing with those complex ML models like NN and GBM. Thus, our study suggested that ML models may not be warranted because of their better performance than traditional LR models in predicting depressive symptoms in large populations.

Our results were consistent with those of previous studies on depression, and the AUC was in line with that of previous studies.[2] The results also agreed with those of an algorithm study, which revealed that the simple LR method yielded as good a performance as did more sophisticated models according to twenty-two decision trees, nine statistical algorithms, and two NN algorithms compared.[2] A systematic review also indicated that the lack of evidence supporting the clinical prediction of ML was greater than LR after investigating 71 studies out of 927.[12] In addition, previous studies revealed that the LR model could have good performance even for low-risk diseases among large-size investigations since class imbalance and overfitting challenge ML applications.[9-11]

The properties of our data may facilitate the good performance of LR, with weak interactions between predictors based on the pair correlation analysis and no complex nonlinearity relationships between depressive symptoms and predictors. Approximately half of the population in our study had depressive symptoms (57.6%), and the possibility of data imbalance could be ignored for the ML models like SVM and RF, which are sensitive to this issue according to previous studies.[31] A nonlinear effect of age was found by GAM in the whole data, and thus the data processing was conducted repeatedly in the sub age groups to eliminate the nonlinearity (Supplementary Table 2). Although the PHQ-9 in this study has been considered as the gold standard for screening depressive symptoms, other potential approaches on depression were also investigated in previous studies to compare different models, such as natural language processing (NLP) and electronic health records (HERs). ML models seemed to perform better in feature extraction from NLP of Twitter datasets,[32] but the LR was proven to be as effective as ML techniques for HERs.[33]

In our study, loneliness, stress, and problematic internet use

were the top risk factors associated with an increase in depression during the pandemic, and these findings are consistent with previous studies.[16,34-36] Loneliness associated with enforced social distancing leads to an increase in depression and anxiety, particularly in children and adolescents.[37] A study involving United Kingdom (UK) adults aged 50 years and older revealed that living alone was among the key risk factors.[34] Stress from lower incomes during the pandemic has also been investigated in various studies in the United States (US),[38,39] Malaysia,[35] and the UK.[34] Consistent with those studies, our research showed that income had a more negative impact on adults younger than 40 years than on those older than 60 years. One possible explanation is the greater financial status (higher position and greater stability) of older adults could contribute to better resilience to economic stress and depression.[38,39] A European study provided an explanation on the association between high use of social media and greater depressive symptoms.[36] It indicated that the negative information (e.g., bulletins about new deaths) from media increases fear and anxiety, and then long durations of problematic internet use could lead to sleep issues and subsequently mental health.

For the general population, lockdowns and COVID-19-related restrictions were associated with worse psychological outcomes.[37] In our study, city lockdown was among the most important predictors for the elderly over 60 years, while it was less important for younger adults. One of the possible explanations was that the social connections among younger adults were more dependent on social media and internet usage.[36] Furthermore, older people may need more medical support, which was quite limited during the lockdown. It was also noted that the lockdown seemed to have a protective effect against depression in women aged older than 65 years due to hormonal influence.[36] The difference between age groups led to a nonlinear age effect, presented by the GAM with a significant slope change at approximately 60 years of age (Figure 2).

Our study revealed that late bedtime was a sensitive predictor of depressive symptoms, as was sleep quality quantified by the B-PSQI[28,40] (including difficulty falling asleep, staying asleep, and experiencing restless and unsatisfying sleep[41,42]). One study demonstrated that normal sleepers were at risk for insomnia during the COVID-19 pandemic, with depression being a significant contributing factor to worsening insomnia. These findings suggested that certain neural markers and functional connectivity between certain brain regions mediate the association between depressive symptoms and insomnia symptoms during the COVID-19 pandemic.[43] It was also noted that different courses of insomnia development (such as lessening, slightly worsening, or developing mild insomnia) could be predicted based on pre-COVID-19 depressive symptoms and brain functional connectivity.[43]

The strengths of this study are as follows. First, the study involved more than 20,000 participants in the nationwide population, which satisfied the large amount of data required for ML algorithms. Second, the nonlinear effect of ages was investigated, and further studies on the sub-age groups provided more impact details of those predictors. Third, the algorithms in our study covered the typical model classes, that is, NNs, tree algorithms, SVMs, and LRs. Their performances were compared and discussed in terms of AUCs on testing sets, overfitting issues and the details of hyperparameter optimization. The training codes are available in the Supplementary Material with tunning and griding settings for hyperparameter optimization.

This study has several limitations. First, depressive symptoms were identified by the PHQ-9 index instead of through clinical diagnosis, and the predictors were limited to individual behaviors and social environments without any clinical predictors, such as metabolites. This may imply that our conclusions cannot be generalized to a broader range of clinical predictors on depression. Second, the possibility of depressive symptoms was more than 50%, and so similar conclusions may not be drawn for disease prediction with low risks. Third, limitations of the cross-sectional design prevented the determination of relationships among the predictors, and longitudinal studies are needed for further investigations.

In summary, this study examined the prediction of depressive symptoms in a large national population during the COVID-19 pandemic. Our findings illustrated that the traditional LR performs comparably to ML models, such as NN, RF, SVM, and GBM models. Several risk factors were shown to be associated with depressive symptoms according to the LR model (age, income, stress, loneliness, late bedtime, and problematic internet use). This study provides a foundation for further investigations into depressive symptoms, and future research is needed to validate these results in diverse populations with varying characteristics.

Supplementary Materials

The Supplement is available with this article at https://doi.org/10.30773/pi.2024.0156.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Data curation: Yi-Bo Wu. Formal analysis: Chun-Hua Zhao, Dan-Dan Chen. Funding acquisition: Chen-Wei Pan. Investigation: Jian-Hua Liu, Tian-Yang Zhang, Yi-Bo Wu. Writing—original draft: Xing-Xuan Dong. Writing—review & editing: Dan-Dan Chen.

ORCID iDs

Xing-Xuan Dong https://orcid.org/0000-0001-6791-4833
Jian-Hua Liu https://orcid.org/0009-0007-8946-2322
Tian-Yang Zhang https://orcid.org/0000-0003-4519-0939
Chen-Wei Pan https://orcid.org/0000-0003-3362-4613
Chun-Hua Zhao https://orcid.org/0009-0007-2751-430X
Yi-Bo Wu https://orcid.org/0000-0001-9607-313X
Dan-Dan Chen https://orcid.org/0000-0002-8064-488X

Funding Statement

A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Acknowledgments

We appreciate the contributions of Bing Shi and Yueqing Huang on our manuscript revisions.

REFERENCES

1. Deo RC. Machine learning in medicine. Circulation 2015;132:1920-1930.
2. Lin S, Wu Y, Fang Y. Comparison of regression and machine learning methods in depression forecasting among home-based elderly Chinese: a community based study. Front Psychiatry 2022;12:764806.
3. Yland JJ, Wang T, Zad Z, Willis SK, Wang TR, Wesselink AK, et al. Predictive models of pregnancy based on data from a preconception cohort study. Hum Reprod 2022;37:565-576.
4. Belsti Y, Moran L, Du L, Mousa A, De Silva K, Enticott J, et al. Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. Int J Med Inform 2023;179:105228.
5. Hu P, Li Y, Liu Y, Guo G, Gao X, Su Z, et al. Comparison of conventional logistic regression and machine learning methods for predicting delayed cerebral ischemia after aneurysmal subarachnoid hemorrhage: a multicentric observational cohort study. Front Aging Neurosci 2022;14:857521.
6. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6:27.
7. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 2016;5:221-232.
8. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079-2107.
9. Taylor RA, Moore CL, Cheung KH, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. PLoS One 2018;13:e0194085.
10. Nusinovici S, Tham YC, Chak Yan MY, Wei Ting DS, Li J, Sabanayagam C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol 2020;122:56-69.
11. Song YX, Yang XD, Luo YG, Ouyang CL, Yu Y, Ma YL, et al. Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: a retrospective study. CNS Neurosci Ther 2023;29:158-167.
12. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22.
13. Iranitalab A, Khattak A. Comparison of four statistical and machine learning methods for crash severity prediction. Accid Anal Prev 2017;108:27-36.
14. Arora A, Lin JJ, Gasperian A, Maldjian J, Stein J, Kahana M, et al. Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. J Neural Eng 2018;15:066028.
15. Bareeqa SB, Ahmed SI, Samar SS, Yasin W, Zehra S, Monese GM, et al. Prevalence of depression, anxiety and stress in china during COVID-19 pandemic: a systematic review with meta-analysis. Int J Psychiatry Med 2021;56:210-227.
16. Dong XX, Li DL, Miao YF, Zhang T, Wu Y, Pan CW. Prevalence of depressive symptoms and associated factors during the COVID-19 pandemic: a national-based study. J Affect Disord 2023;333:1-9.
17. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606-613.
18. Wang W, Bian Q, Zhao Y, Li X, Wang W, Du J, et al. Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry 2014;36:539-544.
19. Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. Gen Hosp Psychiatry 2015;37:567-576.
20. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24:385-396.
21. Chen G, Gully SM, Eden D. Validation of a new general self-efficacy scale. Organ Res Methods 2001;4:62-83.
22. Zimet GD, Powell SS, Farley GK, Werkman S, Berkoff KA. Psychometric characteristics of the multidimensional scale of perceived social support. J Pers Assess 1990;55:610-617.
23. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res Aging 2004;26:655-672.
24. Bech P, Gudex C, Johansen KS. The WHO (Ten) Well-Being Index: validation in diabetes. Psychother Psychosom 1996;65:183-190.
25. Demetrovics Z, Király O, Koronczai B, Griffiths MD, Nagygyörgy K, Elekes Z, et al. Psychometric properties of the problematic internet use questionnaire short-form (PIUQ-SF-6) in a nationally representative sample of adolescents. PLoS One 2016;11:e0159409.
26. Demetrovics Z, Szeredi B, Rózsa S. The three-factor model of internet addiction: the development of the problematic internet use questionnaire. Behav Res Methods 2008;40:563-574.
27. Göktaş S, Aygar H, Akbulut Zencirci S, Önsüz MF, Alaiye M, Metintaş S. Problematic internet use questionairre-short form-6 (PIUQ-SF 6): a validity and reliability study in Turkey. Int J Res Med Sci 2018;6:2354-2360.
28. Buysse DJ, Reynolds CF 3rd, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28:193-213.
29. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003;35:1381-1395.
30. Sjostrom M, Ainsworth BE, Bauman A, Bull FC, Hamilton-Craig, CR, Sallis JF. Guidelines for data processing analysis of the International Physical Activity Questionnaire (IPAQ) - Short and long forms (2005) [Internet]. Available at: https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/ipaq_analysis.pdf. Accessed May 7, 2024.
31. Batuwita R, Palade V. Class imbalance learning methods for support vector machines. In: He H, Ma Y, editors. Imbalanced Learning. New Jersey: Wiley, 2013, p.83-99.
32. Sudhishna KS, Kumar CS, Kishore BS, Arun A. Comparative analysis of different machine learning algorithms to predict depression. International Conference on Sustainable Computing and Smart Systems (ICSCSS); 2023 Jun 14-16; Coimbatore. Piscataway(NJ): IEEE; 2023. p.148-153.
33. Nickson D, Meyer C, Walasek L, Toro C. Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. BMC Med Inform Decis Mak 2023;23:271.
34. Rutland-Lawes J, Wallinheimo AS, Evans SL. Risk factors for depression during the COVID-19 pandemic: a longitudinal study in middle-aged and older adults. BJPsych Open 2021;7:e161.

35. Razali S, Tukhvatullina D, Hashim NA, Raduan NJN, Anne SJ, Ismail Z, et al. Sociodemographic factors of depression during the COVID-19 pandemic in Malaysia: the COVID-19 mental health international study. East Asian Arch Psychiatry 2022;32:82-88.
36. Tesarova S, Pekacek O, Porrovecchio A. Predictors of depression: lifestyle choices during the pandemic. Front Psychol 2023;14:1194270.
37. Kupcova I, Danisovic L, Klein M, Harsanyi S. Effects of the COVID-19 pandemic on mental health, anxiety, and depression. BMC Psychol 2023;11:108.
38. Ettman CK, Abdalla SM, Cohen GH, Sampson L, Vivier PM, Galea S. Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic. JAMA Netw Open 2020;3:e2019686.
39. Ettman CK, Cohen GH, Abdalla SM, Sampson L, Trinquart L, Castrucci BC, et al. Persistent depressive symptoms during COVID-19: a national, population-representative, longitudinal study of U.S. adults. Lancet Reg Health Am 2022;5:100091.
40. Sancho-Domingo C, Carballo JL, Coloma-Carmona A, Buysse DJ. Brief version of the Pittsburgh Sleep Quality Index (B-PSQI) and measurement invariance across gender and age in a population-based sample. Psychol Assess 2021;33:111-121.
41. Plante DT. The evolving nexus of sleep and depression. Am J Psychiatry 2021;178:896-902.
42. Riemann D, Krone LB, Wulff K, Nissen C. Sleep, insomnia, and depression. Neuropsychopharmacology 2020;45:74-89.
43. Tian Y, Vitiello MV, Wang H, Wang Y, Dong D, Xu H, et al. Risk of insomnia during COVID-19: effects of depression and brain functional connectivity. Cereb Cortex 2023;33:7015-7025.