Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine

Lorena Cecilia López Steinmetz[1,2]*, Margarita Sison[3†], Rustam Zhumagambetov[4†], Juan Carlos Godoy[2] and Stefan Haufe[1,3,4,5]*

[1]Inverse Modeling and Machine Learning, Chair of Uncertainty, Institute of Software Engineering and Theoretical Computer Science, Faculty IV Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany, [2]Instituto de Investigaciones Psicológicas (IIPsi), Facultad de Psicología, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad Nacional de Córdoba (UNC), Córdoba, Argentina, [3]Berlin Center for Advanced Neuroimaging (BCAN), Charité – Universitätsmedizin Berlin, Berlin, Germany, [4]Working Group 8.44 Machine Learning and Uncertainty, Mathematical Modelling and Data Analysis Department, Physikalisch-Technische Bundesanstalt Braunschweig und Berlin, Berlin, Germany, [5]Institute for Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin, Germany

Introduction: The COVID-19 pandemic has exacerbated mental health challenges, particularly depression among college students. Detecting at-risk students early is crucial but remains challenging, particularly in developing countries. Utilizing data-driven predictive models presents a viable solution to address this pressing need.

Aims: 1) To develop and compare machine learning (ML) models for predicting depression in Argentinean students during the pandemic. 2) To assess the performance of classification and regression models using appropriate metrics. 3) To identify key features driving depression prediction.

Methods: A longitudinal dataset (N = 1492 college students) captured T1 and T2 measurements during the Argentinean COVID-19 quarantine. ML models, including linear logistic regression classifiers/ridge regression (LogReg/RR), random forest classifiers/regressors, and support vector machines/regressors (SVM/SVR), are employed. Assessed features encompass depression and anxiety scores (at T1), mental disorder/suicidal behavior history, quarantine sub-period information, sex, and age. For classification, models' performance on test data is evaluated using Area Under the Precision-Recall Curve (AUPRC), Area Under the Receiver Operating Characteristic curve, Balanced Accuracy, F1 score, and Brier loss. For regression, R-squared (R2), Mean Absolute Error, and Mean Squared Error are assessed. Univariate analyses are conducted to assess the predictive strength of each individual feature with respect to the target variable. The performance of multi- vs univariate models is compared using the mean AUPRC score for classifiers and the R2 score for regressors.

Results: The highest performance is achieved by SVM and LogReg (e.g., AUPRC: 0.76, 95% CI: 0.69, 0.81) and SVR and RR models (e.g., R2 for SVR and RR: 0.56,

95% CI: 0.45, 0.64 and 0.45, 0.63, respectively). Univariate models, particularly LogReg and SVM using depression (AUPRC: 0.72, 95% CI: 0.64, 0.79) or anxiety scores (AUPRC: 0.71, 95% CI: 0.64, 0.78) and RR using depression scores ($R2$: 0.48, 95% CI: 0.39, 0.57) exhibit performance levels close to those of the multivariate models, which include all features.

Discussion: These findings highlight the relevance of pre-existing depression and anxiety conditions in predicting depression during quarantine, underscoring their comorbidity. ML models, particularly SVM/SVR and LogReg/RR, demonstrate potential in the timely detection of at-risk students. However, further studies are needed before clinical implementation.

Introduction

The COVID-19 pandemic has presented significant global challenges to mental health, particularly among college students. The sudden transition to remote learning, social isolation, and economic instability has exacerbated mental health issues, with depression becoming a prevalent concern for this population (1, 2). Depression can negatively impact academic performance, social relationships, and overall quality of life. Therefore, early identification of depression and its risk factors is crucial for timely interventions and prevention of further negative outcomes. However, identifying individuals who may be at risk of developing depression remains a significant challenge in this field.

The literature consistently identifies key psychological and demographic variables linked to or predicting depression. Anxiety often emerges as a predictor for depression (3, 4), notably among college students (5). Moreover, histories of diagnosed mental disorders (6) and suicidal behavior (7) closely correlate with college student depression. While sex (8) and age (9) differences in depression are well-documented, sex-based variations in depression may display age-specific trends (10). Amid COVID-19 lockdowns, mental health, particularly depression, was anticipated to deteriorate with extended measures, impacting well-being during and after implementation (11). Yet, achieving a comprehensive integration of these factors into accurate predictive models for depression remains incomplete and challenging.

Existing literature on depression has extensively analyzed variability between groups and within individuals (e.g., 12). Mixed-effects modeling (MEM), or hierarchical linear modeling, is a widely used statistical approach that allows for the estimation of both fixed and random effects, accounting for variability between groups and within individuals. MEM has a major strength in its ability to identify individual and group-level effects, as well as

interactions between them, which is particularly relevant in fields like psychology where individual differences are of interest.

However, MEM is primarily designed for inference rather than prediction, and offers limited flexibility for the handling of high-dimensional data, compromising their predictive accuracy compared to machine learning (ML) models. ML models are well-suited for prediction tasks, handling complex relationships between variables, and relying less on strong data assumptions (13–16). MEM, in particular, relies on assumptions about predictor-response relationships (i.e., linearity), normality, and independence of residuals, which can impact their predictive accuracy for new data. Conversely, ML is a data-driven approach that uses algorithms to identify patterns in the data for prediction. However, ML models require extensive data for training and can be prone to overfitting with too complex models or small datasets. Therefore, while ML has emerged as an alternative or complementary approach to traditional statistics in mental health research, it is an open challenge to leverage ML for predicting individuals at risk for depression.

In the field of mental health assessment and ML applications, the challenge of diagnosis prediction can be tackled using a dual methodology, encompassing both classification and regression approaches. In the context of binary classification, the goal is to categorize individuals into two groups: those with depression and those without. This approach yields a straightforward determination of depression presence, aiding in identifying individuals requiring additional evaluation. Conversely, in the context of regression, the aim is to estimate or forecast numerical depression scores. This approach offers a continuous prediction of depression severity, affording a deeper insight into the condition and potentially enabling personalized treatment strategies.

In a prior study, we utilized MEM to investigate the aforementioned key psychological and demographic factors linked to or predictive of depression in college students during Argentina's

COVID-19 pandemic quarantine, which was one of the most rigorous and extended lockdowns globally (17). In the current study, we seek to reassess the same dataset using ML algorithms to evaluate their potential as an alternative or complementary technique to MEM for predicting depression diagnosis. Through this, we hope to enhance predictive precision and effectively identify individuals with depression, while gaining deeper insights into the intricate interplay of factors contributing to depression during pandemics. Furthermore, we expect to determine whether the key factors commonly acknowledged as depression predictors remain as significant input features when utilizing ML algorithms for data analysis, in contrast to other widely used statistical methodologies in psychology. The objectives of this study are: 1) To develop and compare various ML models, including linear logistic regression classifiers/ridge regression, random forest classifiers/regressors, and support vector machines/support vector regression models, to predict depression in college students during the pandemic based on psychological inventory scores, basic clinical information, quarantine sub-period information, and demographics as features. 2) To assess the performance of classification and regression models using appropriate metrics. 3) To identify the key features that drive the prediction of depression by applying univariate methods.

A recent review following the PRISMA guidelines identified only 33 peer-reviewed studies in the domain of utilizing ML algorithms for predicting mental health diagnoses, with only 31% specifically focusing on depression/anxiety disorders. Additionally, most studies have been conducted on clinical samples, primarily consisting of adults or older patients (18). It is worth noting that many of these studies rely on unavailable facilities and resources for diagnosis, particularly in developing countries, such as MRI or blood samples. Therefore, one novel aspect of this study is its focus on exploring the potential of ML algorithms to identify risk factors for depression, during a pandemic, in a large longitudinal sample of quarantined and apparently healthy college students from a developing country. By doing so, it aims to provide further insights into this important area of research by enhancing the understanding of key determinants in depression detection and may have significant implications for the development of more effective screening tools and interventions for depression in college students during pandemics.

Materials and methods

Research design and dataset

This study employs a longitudinal dataset featuring two-repeated measurements in college students during the Argentinean COVID-19 quarantine period. The first measurement (T1) was taken across various quarantine sub-periods, each characterized by varying levels of restrictions, and spanning up to 106 days. The follow-up measurement (T2) occurred one month later (as depicted in Supplementary Figure S1).

The choice of this longitudinal research design arises from its inherent capability to capture dynamic changes over time, a crucial aspect when investigating the effects of a rapidly evolving event like the COVID-19 pandemic. This approach transcends mere cross-sectional snapshots, providing a more thorough understanding of the intricate interplay between psychological states and external factors amid a quarantine. By incorporating two measurements at distinct time points, it becomes possible to discern short-term fluctuations and potential enduring effects. The T1 measurement establishes a baseline and assesses the immediate effects of quarantine sub-periods lasting up to 106 days. The T2 measurement, conducted a month later, offers insights into the persistence or evolution of mental health patterns, shedding light on potential longer-term consequences. This temporal depth captures nuances that a single-time assessment might overlook, providing a more nuanced portrayal of the challenges college students faced during this unprecedented period.

The dataset comprises responses from 1492 college students who completed an online survey during the mandatory restrictive quarantine. It encompasses assessments of depression, anxiety-trait, basic clinical information, demographics, and quarantine sub-periods. This dataset corresponds to the one utilized in our previous study (17). Further information on the research design and data collection can be found there and in López Steinmetz (19).

Input features

The prediction models were developed using the following input features: age, sex (female, male), history of diagnosed mental disorder (absent, present), history of suicidal attempt and/ or ideation (absent, present), depression scores from T1 using the Argentinean validation (20) of the Beck Depression Inventory (BDI) (21), anxiety-trait scores from T1 using the Spanish version of the State-Trait Anxiety Inventory (22), and three quarantine sub-periods to which participants' responses were chronologically assigned. These sub-periods were categorized according to decreasing levels of COVID-related restrictions over the 106-day duration, and participants were classified into one of three quarantine sub-periods based on the date of their response for measurement T1 (Supplementary Figure S1).

Target variable

The target variable, depression, was assessed using the Argentinean validation (20) of the BDI (21). For the classification task, depression scores from T2 were labeled as either according to the absence or presence of clinically relevant levels of depression, respectively. The standardized cut-off score of >20, indicative of depression presence in non-clinical populations (23), was employed. For the regression task, raw depression scores were used as a continuous variable.

Software

The analysis was performed using Python software (version 3.11.1) and relevant libraries, including scikit-learn (24), pandas (25), and numpy (26), among others.

Data analysis

Preprocessing

The dataset does not contain missing data. Before performing classification and regression tasks, the data were randomly divided into a training (75%) and a test (25%) set. In the classification task, the proportion of negative to positive labels was maintained in both sets. Categorical variables were encoded using a one-hot schema. Each variable was transformed into as many binary 0/1 variables as there were different values. Continuous features such as depression and anxiety scores from T1 and age were scaled using a quantile transformation, specifically, we used the QuantileTransformer method (24) setting the output distribution as normal, to reduce skewness and to enhance model performance. Principal component analysis (PCA) was applied for dimensionality reduction. PCA was set to retain the number of components essential to account for 95% of the variance present in the original data. Consequently, a total of 3 components were retained. Both feature scaling and dimensionality reduction were fitted on the training set and subsequently applied to both sets. The target variable in the regression task was scaled using the same quantile transform function.

Hyperparameter tuning in the classification and regression tasks

Both in the classification task and in the regression task hyperparameter tuning was conducted using the GridSearchCV function of the scikit-learn library (24) on the training set for each model. Stratified 10-fold cross-validation, which is typically used when dealing with imbalanced datasets, was applied for robust evaluation during hyperparameter tuning, that is, the training data were split into ten stratified subsets, where in each fold one of the subsets served as an inner validation set and nine as an inner training set. Three ML models for the classification task and three ML models for the regression task with different hyperparameter choices were optimized on each inner training set and tested on each inner validation set. In the classification task, the Average Precision (AP) score was used as the evaluation metric to select model hyperparameters within the inner cross-validation. For the regression task, the R-squared (R2) score was employed as the evaluation metric for hyperparameter selection within the inner cross-validation process. The state of the random number generator was set to 0 for reproducibility in all cases. The final model for each ML model was then trained on the entire training set using the selected hyperparameters. This comprehensive approach aimed to enhance the performance and robustness of the models in both classification and regression tasks.

Classification models

For the classification task, three ML algorithms are employed: Linear logistic regression classifier (LogReg) (13), random forest (RF) classifier (27), and support vector machine (SVM) (28). LogReg models are designed for predicting probabilities of K classes or categories in the classification problem via linear functions of input features x, ensuring they sum to one and stay within a valid probability range of [0, 1]. The model is expressed in terms of $K - 1$ log-odds or logit transformations and uses the last class as the denominator in the odds-ratios, with a special simplicity in binary classification scenarios (13). RF is an ensemble learning method that builds multiple decision trees by randomly sampling with replacement from the training data. Each tree is grown by recursively selecting random feature subsets and identifying optimal split points. For classification, the final output is determined by the majority vote of individual tree predictions during prediction. Thus, the model output is an ensemble of trees that collectively make predictions. Hyperparameters include the number of trees in the forest, the maximum depth of the trees, and the minimum number of samples required to split a node (13). The SVM algorithm classifies data by transforming each data point into a k-dimensional feature space, where k >> n and n is the number of features. It identifies a hyperplane that maximizes the margin between classes, minimizing classification errors. The margin is the distance between the decision hyperplane and the nearest instance of each class (29).

These algorithms were chosen due to their well-established competitive performance in classification tasks and adaptability to diverse scenarios. For instance, Uddin et al. (29) provide a comprehensive overview of the relative performance of different supervised ML algorithms for disease prediction, including LogReg, RF, and SVM. Their findings, despite variations in frequency and performance, underscore the potential of these algorithmic families in disease prediction.

Certain model hyperparameters were fixed for each algorithm while others were optimized using cross-validation. For the **LogReg** algorithm, an intercept term, which represents the log-odds of the baseline class, was included in the model by the default setting of scikit-learn. The parameter class_weight was set as balanced to adjust the weights assigned to classes during the training process. This indicates that the algorithm automatically adjusts the weights of the classes inversely proportional to their frequencies. The maximum number of iterations taken for the solver to converge was set to 500. Solver here refers to the optimization algorithm used to find the optimal values for the coefficients of the linear logistic regression model. L2-norm regularization was applied to prevent overfitting and improve the generalization of the model. Regularization involves adding a penalty term to the loss function, and in L2-norm regularization, the penalty is proportional to the square of the magnitudes of the coefficients. The regularization parameter C was set to values 0.0001, 0.001, 0.01, 0.1, 1, 10, and 1000. The optimal hyperparameter as identified using cross-validation is C = 0.1.

For the RF classifier, several model hyperparameters were tuned. The number of features to consider for the best split ('max_features')

was set to the square root of the total number of features. The function to measure the quality of a split ('criterion') was configured to use the Gini diversity index. The Gini index is employed by each decision tree in the RF ensemble as the measure for determining optimal split points during the training process. The number of trees ('n_estimators') was set to values 50, 100, 500, and 1000. The maximum depth of the tree ('max_depth') was set to values 5, 10, and 15. The minimum number of samples required to split an internal node ('min_samples_split') was set to values 2, 5, and 8. The minimum number of samples required to be at a leaf node ('min_samples_leaf') was set to values 1, 2, and 3. For tree building, a bootstrap method was used, involving sampling with replacement instead of using the whole training set to build trees. The optimal combination of hyperparameters as identified using cross-validation are n_estimators: 100, max_depth: 5, min_samples_split: 5, and min_samples_leaf: 2.

For the SVM algorithm, the parameter class_weight was set as balanced. Two variants of the SVM algorithm were tested. The first one used a non-linear radial basis function (rbf) kernel with kernel width gamma (kernel: rbf). Tested hyperparameter values included: C: 0.01, 0.1, 1, 10, 100, 500, 1000; gamma: 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10. The second variant used a parameterless linear kernel. Tested hyperparameter values for this choice included: C: 0.01, 0.1, 1, 10, 100, 500, 1000; kernel: linear. The optimal combination of hyperparameters as identified using cross-validation are C: 1000, gamma: 1e-05, kernel: rbf.

Results of the hyperparameter tuning and grid search for each classifier can be found in the Supplementary Material (Supplementary Tables S1–S3, Supplementary Figures S2, S3).

Benchmarking was conducted using dummy models representing baseline references. Baseline models serve as the null hypothesis, simulating scenarios where models possess minimal knowledge about individual sample labels, essentially reflecting the label distribution in the training data at best. These models are instrumental for assessing whether a sophisticated model's performance surpasses the "null" or "chance" level, providing a benchmark to evaluate the effectiveness of advanced algorithms.

Baseline models consisted of a uniform random baseline (randomly assigning class labels with equal probability, i.e., 50% for each class, without considering any input features), a most frequent baseline (it assigns the majority class label to all instances), and a stratified random baseline (which randomly assigns labels proportionally to their relative frequency in the training data) (14).

Regression models

For the regression task, three ML algorithms are utilized: Ridge regression (RR) (13), RF regressor (27), and support vector regressor (SVR) (28). RR, often referred to as L2-norm regularized least-squares regression, adds a regularization term to the standard linear regression objective function. In RR, the objective function to be minimized is the sum of the squared differences between the observed and predicted values (least squares term), and a penalty term that discourages overly complex models by adding the L2-norm of the coefficients multiplied by a regularization parameter (alpha).

The regularization term penalizes large coefficients, preventing overfitting and promoting a more stable and generalizable model. The strength of this penalty is controlled by the hyperparameter alpha (13). RF constructs multiple decision trees by bootstrapping the data and introducing randomness in variable selection at each node. The final prediction is an aggregation of predictions from all individual trees. For regression, the final output is determined by the average of individual tree predictions during prediction (13). SVR uses the principles of SVM for regression tasks. Similar to the latter, SVR introduces the concept of a margin. In SVR, the margin represents a range of values within which errors are tolerable. SVR aims to fit the function $f(x)$ in such a way that the differences between the predicted values and the actual values (errors) fall within the specified margin. SVR minimizes the errors while staying within the margin. The loss function penalizes deviations from the actual values but allows for a certain amount of error within the margin. To prevent overfitting, SVR includes a regularization term controlled by the parameter C, which is a user-defined hyperparameter. The regularization term controls the smoothness of the learned function (13).

Certain model hyperparameters were fixed for each algorithm while others were optimized using cross-validation. For the **RR** algorithm the regularization strength parameter, denoted as alpha, was set to values 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, and 1000. The optimal model's alpha as identified using cross-validation is 100.

For the RF regressor, several model hyperparameters were tuned. The number of features to consider for the best split ('max_features') was set to 1, as empirically justified in Geurts et al. (30). The function to measure the quality of a split ('criterion') was configured to use the mean squared error. The number of trees ('n_estimators') was set to values 50, 100, 500, and 1000. The maximum depth of the tree ('max_depth') was set to values 5, 10, and 15. The minimum number of samples required to split an internal node ('min_samples_split') was set to values 2, 5, and 8. The minimum number of samples required to be at a leaf node ('min_samples_leaf') was set to values 1, 2, and 3. For tree building, a bootstrap method was used, involving sampling with replacement instead of using the whole training set to build trees. The optimal combination of hyperparameters as identified using cross-validation are n_estimators: 50, max_depth: 5, min_samples_split: 8, and min_samples_leaf: 3.

For the SVR training, the shrinking trick described by (31) was employed. Shrinking attempts to reduce the optimization problem by removing the elements that have already been bound. Two variants of the SVR algorithm were tested. The first one used a rbf kernel and the hyperparameters C, epsilon, and gamma are varied. The tested values included: C: 0.01, 0.1, 1, 10, 100, 500, 1000; epsilon: 0.001, 0.01, 0.1, 1; gamma: 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10; kernel: rbf. The second variant used a linear kernel and only the hyperparameters C and epsilon are varied. The hyperparameter values for this choice included: C: 0.01, 0.1, 1, 10, 100, 500, 1000; epsilon: 0.001, 0.01, 0.1, 1; kernel: linear. The optimal hyperparameters combination as identified using cross-validation are C: 500, epsilon: 0.1, gamma: 0.00001, kernel: rbf.

Results of the hyperparameter tuning and grid search for each regressor can be found in the Supplementary Materials (Supplementary Tables S4, S5, Supplementary Figures S4, S5).

Similar to the classification models, dummy models were included as benchmark references for regression tasks. These models encompassed a randomly shuffled baseline (predicting the target variable by randomly shuffling actual target values), a mean and a median baseline (predicting the arithmetic average and the median value of the target variable, respectively, for all instances, without considering any input features or patterns) (14).

Evaluation of classification model performance

The performance of the classification models on test data was assessed using five metrics: Area Under the Precision-Recall Curve (AUPRC), Area Under the Receiver Operating Characteristic curve (AUROC), Balanced Accuracy, F1 score, and Brier loss. These metrics all range from 0 to 1, where 0 indicates worst and 1 indicates best performance. An exception is Brier loss, for which 0 indicates perfect model performance and uncertainty calibration, while a loss of 1 indicates the worst performance.

Precision measures a classifier's ability to avoid incorrectly labeling negative examples as positives, while Recall (also known as "sensitivity" or "true positive rate") measures its ability to identify all positive examples correctly as positives (32). The AUPRC represents the tradeoff between Precision and Recall over all possible classifier thresholds as quantified by the area under the Precision curve integrated over all Recall values. The AUROC represents the tradeoff between Recall and the true negative rate (also known as "specificity", 33, 34). The Balanced Accuracy score calculates the average of sensitivity (true positive rate) and specificity and it is used to ensure that the model's performance on imbalanced datasets is not overestimated. The F1 score is the harmonic mean of Precision and Recall scores (32). The Brier loss is calculated as the mean squared difference between the predicted probability and the true binary label (35). It measures whether the probabilities provided by the model are well-calibrated. These predicted probabilities are derived from the model's internal mechanisms, which assess the input features and generate probability estimates for each instance in the dataset.

Evaluation of regression model performance

To assess the performance of the regression models on test data, three key metrics were employed: R2, Mean Absolute Error (MAE), and Mean Squared Error (MSE). The R2 score measures the proportion of variance in the target variable explained by the model and is calculated as 1 minus the ratio of the model's MSE to that of a mean baseline model. The R2 score is a coefficient of determination and, theoretically, it can range from -∞ to 1, where 1 indicates perfect prediction, 0 indicates no improvement over the mean model, and values less than 0 indicate poorer performance than the baseline (16, 36). MSE and MAE represent the average squared and absolute differences between predicted and actual scores, respectively, and are used to assess the model's prediction accuracy (16, 36). For both MSE and MAE, lower values signify better model performance.

Bootstrapping with replacement

To enhance the reliability of performance evaluation, bootstrapping with replacement was utilized. Specifically, we resampled the test set 100 times with replacement to create distributions of performance estimates. Each resampled sample had the same dimensions as the original sample, and mean performance estimates with 95% confidence intervals (CI) were derived using this method (37).

Feature importance analysis: comparing multi- vs. univariate models for depression prediction

Feature importance analysis was conducted to assess the predictive strength of each individual feature with respect to the target variable, depression. For this analysis, all ML algorithms previously employed for the classification (LogReg, RF, and SVM classifiers) and regression (RR, RF regressor, and SVR) tasks were applied to single features. Using different algorithms can provide a more comprehensive view of feature importance. The purpose of this analysis was to determine the extent to which individual features are predictive of depression.

Here, the performance of multivariate models, which incorporate all features simultaneously, is compared with that of univariate models, each including a single feature. The metrics used for this comparative analysis are the mean AUPRC score (with 95% CI) for classifiers and the mean R2 score (with 95% CI) for regressors. In this assessment, each multivariate model, whether for classification or regression tasks, serves as a benchmark. This enables an evaluation of whether the collective impact of features enhances prediction accuracy.

Ethical considerations

This study was approved by the Ethics Committee of the Psychological Research Institute, Universidad Nacional de Córdoba (CEIIPsi-UNC-CONICET; comite.etica.iipsi@psicologia.unc.edu.ar), 14/02/20–23/03/20. During data collection, participants' names were not requested. However, email addresses were obtained solely for survey submission during the follow-up measurement, after which they were promptly deleted from the database. Only the principal researcher had access to the survey system and the raw dataset. The privacy and confidentiality of participants' data were ensured, and informed consent was obtained from all participants before their participation. Guidelines and regulations regarding the use of human subjects in research were also followed. Additionally, the current study utilized the dataset that is available in open access.

Results

### Classification models: identifying depression presence

The performance of ML classifiers for predicting depression in college students during the COVID-19 pandemic is evaluated across five metrics and compared to the metrics of dummy

baseline models. The baseline models serve as sanity checks for evaluating the performance of the ML classifiers.

Figure 1 summarizes the performance metrics for both the trained and tested ML models and the tested baseline models. As expected and across all metrics, the ML models during the test phase achieve a superior predictive ability using the provided input features compared to random chance.

The SVM classifier is on par with or outperforms both alternative classifiers, attaining the highest AUPRC score of 0.76 (95% CI: 0.69, 0.81) and the lowest Brier score loss of 0.15 (0.14, 0.17). The LogReg classifier closely follows, displaying an identical AUPRC score of 0.76 (0.69, 0.81) and a slightly higher Brier score loss of 0.16 (0.14, 0.18).

The RF classifier generally achieves lower performance than SVM and LogReg, attaining an AUPRC score of 0.73 (0.66, 0.80)
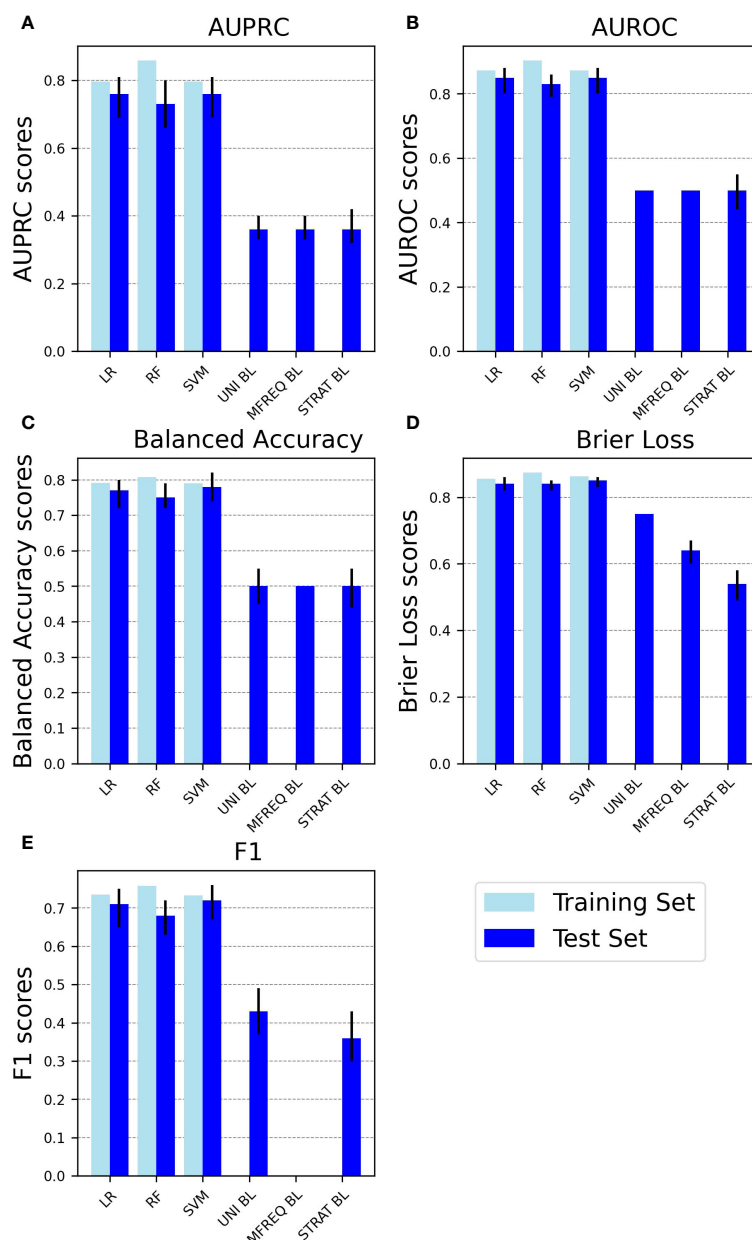
FIGURE 1
Performance of machine learning classifiers on the training and test sets as well as of baseline classifiers on the test set. The bar charts depict the performance of machine learning classifiers (linear logistic regression, random forest classifier, and support vector machine) on the training set and the test set and of baseline classifiers (uniform random, most frequent, and stratified random) on the test set using five performance metrics: **(A)** AUPRC, **(B)** AUROC, **(C)** Balanced accuracy, **(D)** Brier loss, and **(E)** F1. Light blue bars represent performance scores on the training set, while blue bars with error bars represent mean test set scores with 95% confidence intervals. To ensure that higher values consistently indicate better performance across all methods, the Brier loss is plotted as 1-Brier. AUPRC, Area Under the Precision-Recall Curve; AUROC, Area Under the Receiver Operating Characteristic Curve; LR, Linear logistic regression; RF, Random forest classifier; SVM, Support vector machine; UNI BL, Uniform random baseline; MFREQ BL, Most frequent baseline; STRAT BL, Stratified random baseline.

and a Brier score loss of 0.16 (0.15, 0.18). RF also shows some evidence of overfitting, as indicated by a higher drop in AUPRC performance (from 0.86 on the training set to 0.73 on the test set) compared to both alternative classifiers. Refer to Figure 1 and Table S6 for a visualization of these and the remaining performance metrics (i.e., AUROC, Balanced accuracy, and F1 scores) employed in this study for each ML classifier.

Refer to Figure 2 for the Precision-Recall Curves and Receiver-Operating Characteristic depicting depression prediction. This plot aids in comparing models' ability to distinguish between positive and negative classes (Depressed and Non-Depressed, respectively) at different clinically relevant operating points such as high sensitivity and high specificity). Confusion matrices for

each ML classifier are displayed in the Supplementary File (Supplementary Figure S6).

Regression models: quantifying depression severity

The performance of ML regressors in predicting depression among college students during the COVID-19 pandemic is evaluated across three key metrics and compared to the results of these metrics on dummy baseline models. The baseline models serve as benchmarks for evaluating the performance of the ML regressors.

FIGURE 2
Classifier performance comparison: Precision–Recall Curves and Receiver-Operating Characteristic Curves for depression prediction. These curves characterize the predictive performance of trained models (linear logistic regression, random forest classifier, and support vector machine) and dummy classifiers (uniform random, most frequent, and stratified random baselines) for depression prediction for varying classifier thresholds, where the positive class indicates depression presence. **(A)** Precision-Recall Curve (PRC) displays how precision changes with different recall levels for various models/baselines. **(B)** Receiver-Operating Characteristic (ROC) curves illustrate the balance between true positive rate (sensitivity) and false positive rate (1-specificity) as probability thresholds change. Each curve depicts how sensitivity changes with different specificity levels for different models/baselines. SVM, Support vector machine. AP: Precision-Recall curve values (the Precision-Recall curve display in scikit-learn uses the term "average precision" (AP) to refer to the area under the precision-recall curve). AUC, Area Under the Receiver Operating Characteristic values.
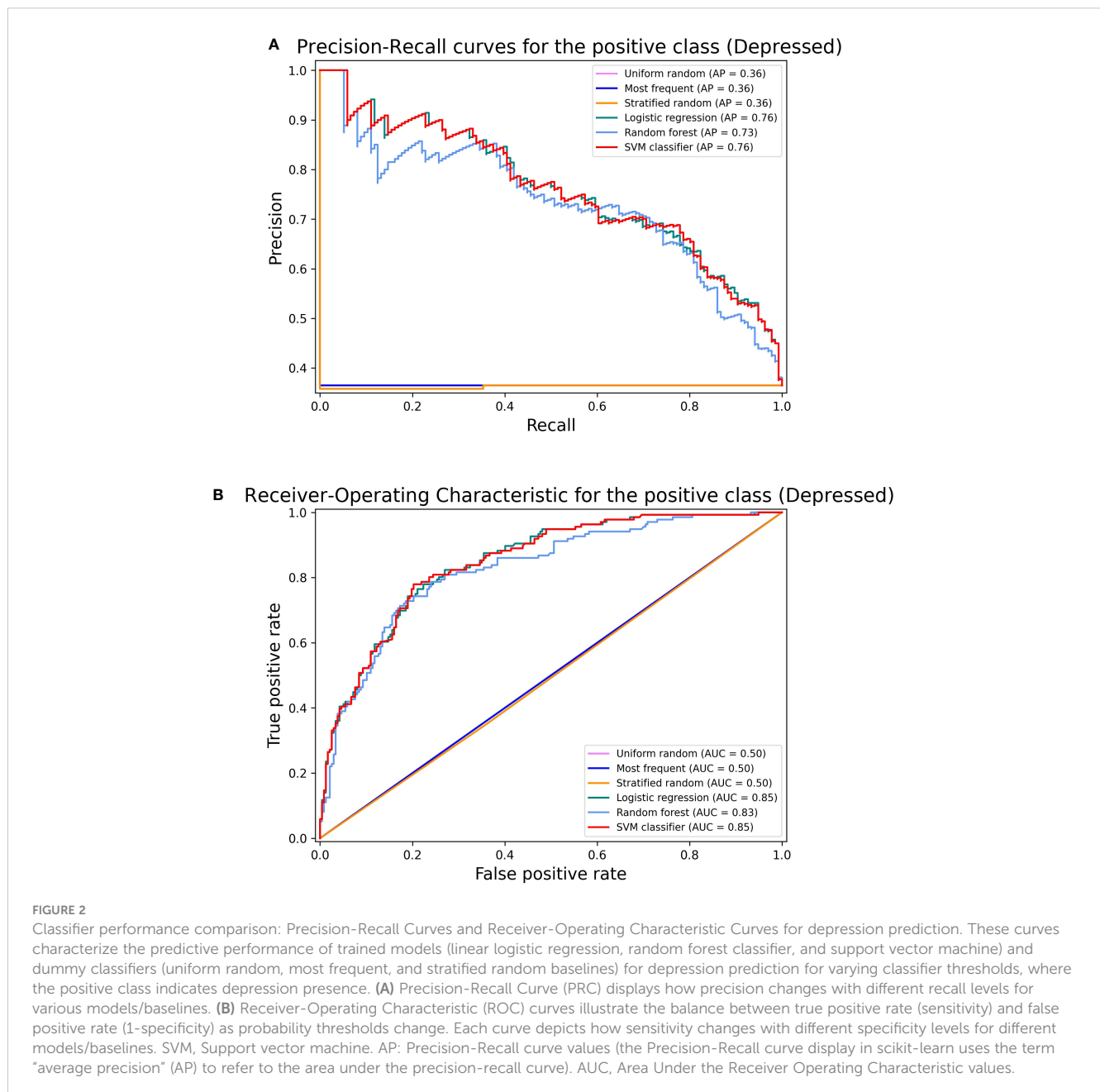
Figure 3 provides an overview of the performance metrics for both the trained and tested ML models, along with the tested baseline models. As expected, the tested ML models consistently demonstrate superior predictive capabilities using the provided input features compared to random chance, as evidenced by all metrics.

SVR and RR achieve the highest R2 score of 0.56 (95% CI: SVR: 0.45, 0.64 and RR: 0.45, 0.63). This R2 value indicates that the models can account for approximately 56% of the variability in the target variable. The RF regressor achieves a lower R2 value of 0.50 (0.40, 0.59).
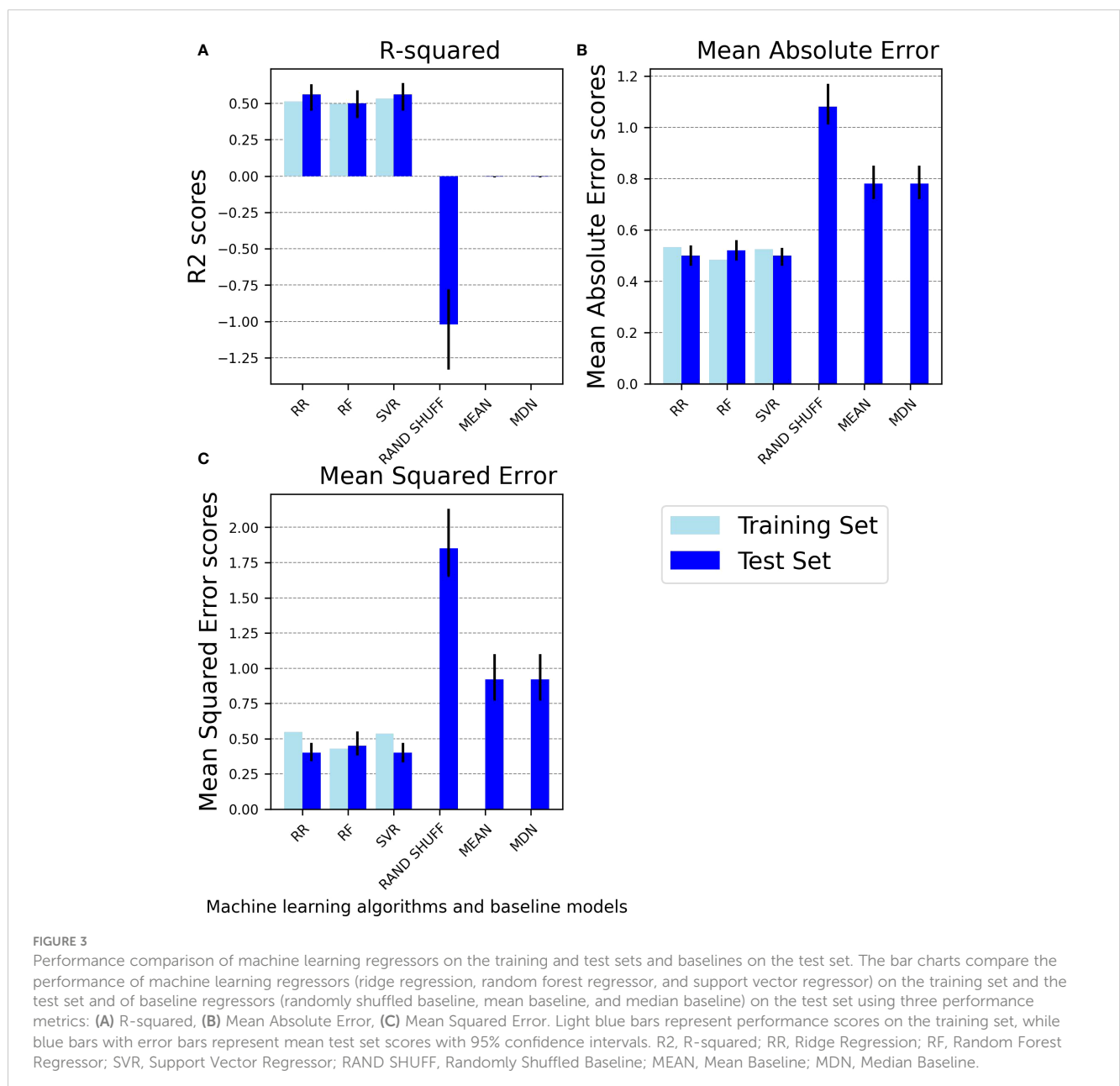
The RF regressor also exhibits inferior MAE and MSE performance on the test set compared to the SVR and RR models. Both the MAE and MSE metrics for the RR algorithm mirror the SVR's performance, underscoring their comparable predictive

capabilities (Figure 3, Supplementary Table S7). The correlations between actual and predicted values for each ML algorithm in the test set are depicted in the Supplementary File (Supplementary Figure S7).

Evaluation of feature importance: comparison of multivariate to univariate models

Comparison of multivariate to univariate models in the classification task

In the classification task for predicting depression presence, the multivariate model generally outperforms univariate models, except for a small difference in favor of the RF classifier using depression

FIGURE 3
Performance comparison of machine learning regressors on the training and test sets and baselines on the test set. The bar charts compare the performance of machine learning regressors (ridge regression, random forest regressor, and support vector regressor) on the training set and the test set and of baseline regressors (randomly shuffled baseline, mean baseline, and median baseline) on the test set using three performance metrics: **(A)** R-squared, **(B)** Mean Absolute Error, **(C)** Mean Squared Error. Light blue bars represent performance scores on the training set, while blue bars with error bars represent mean test set scores with 95% confidence intervals. R2, R-squared; RR, Ridge Regression; RF, Random Forest Regressor; SVR, Support Vector Regressor; RAND SHUFF, Randomly Shuffled Baseline; MEAN, Mean Baseline; MDN, Median Baseline.

scores (at T1) as a single feature (AUPRC multivariate RF classifier: 0.73, 95% CI: 0.66, 0.80; AUPRC univariate RF classifier: 0.74, 95% CI: 0.68, 0.79) (Figure 4, Supplementary Table S8). However, it's crucial to note that the multivariate model for the RF classifier shows some signs of overfitting compared to both multivariate alternative classifiers (Figure 1).

Univariate classifiers, in particular LogReg and SVM using depression (at T1) or anxiety scores as single features, exhibit comparative performance levels (in terms of mean test AUPRC) that are close to those of the corresponding multivariate model that include all features. On the other hand, all other univariate models perform significantly worse, with mean AUPRC scores slightly above chance-level performance. Importantly, this trend is consistent across the three ML algorithms tested (Figure 4, Supplementary Table S8).

Comparison of multivariate to univariate models in the regression task

In the regression task for predicting depression scores, the multivariate model consistently outperforms all univariate models. Notably, when considering all univariate models, depression (at T1) and anxiety scores emerge as the most predictive single features, achieving the highest performance levels (mean R^2 scores), particularly when using the RR algorithm. It is noteworthy that univariate models excluding both depression (at T1) and anxiety scores exhibit performance levels slightly above (e.g., suicidal behavior history) or comparable (e.g., quarantine sub-period) to

chance-level. Importantly, this pattern is consistent across the three ML algorithms. Contrary to the results on the comparison between multivariate classifiers' performance, when comparing univariate models' performance, RF models perform better than SVR models (Figure 5, Supplementary Table S9).

Discussion

This study leveraged a longitudinal dataset for the prediction of depression among college students amidst the challenging backdrop of the COVID-19 pandemic, allowing for the examination of evolving mental health patterns over time. The results indicate a gain in predictive efficacy of multivariate ML algorithms as compared to their univariate counterparts.

As previously mentioned, the dataset analyzed using ML in this study underwent scrutiny through MEM in a prior study (17). In that study, variables such as sex (female), age (younger), mental disorder history (presence), and suicidal behavior history (presence) were identified as having significant main effects on depression. However, following Cohen's conventions for effect sizes (ES; 38), these effects were deemed small for sex (ES = 0.09), age (ES = 0.09), and mental disorder history (ES = 0.17), while being of medium magnitude for suicidal behavior history (ES = 0.34) (17). Notwithstanding, the outcomes of that prior MEM study align with existing literature, underscoring that variables associated with adverse mental health effects after and during the COVID-19
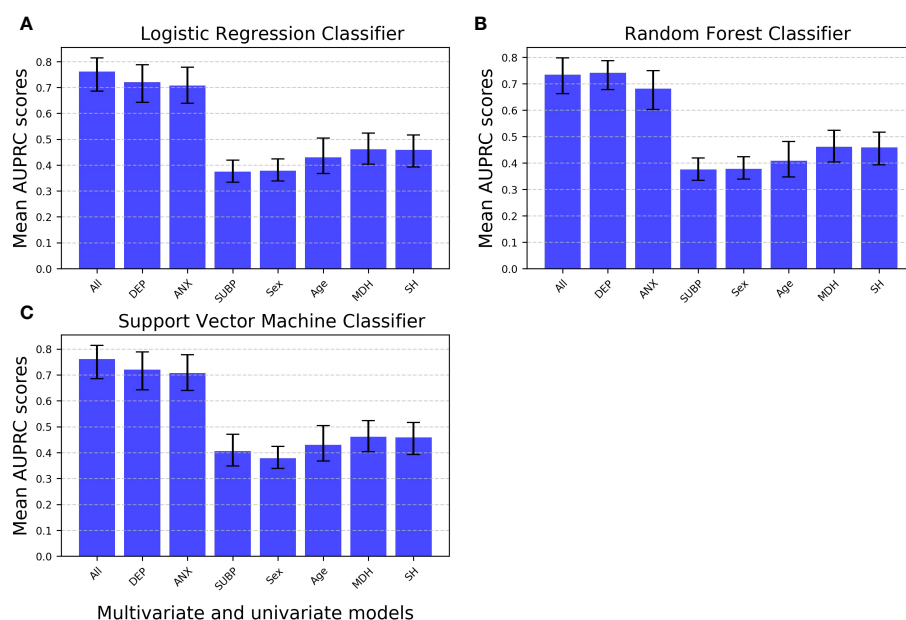

FIGURE 4

Mean AUPRC scores for multivariate versus univariate machine learning models predicting depression in the classification task. Linear logistic regression, random forest, and support vector machine classifiers were trained either as multivariate models (encompassing all features) or univariate models (each incorporating just a single feature). Error bars represent 95% confidence intervals. AUPRC, Average Precision-Recall Curve; All, Multivariate model with all features included; Univariate models: DEP, Depression scores measured at time 1 as the single feature; ANX, Anxiety scores as the single feature; SUBP, Quarantine sub-periods as the single feature; Sex, Biological sex as the single feature; Age, Age as the single feature; MDH, Mental disorder history as the single feature; SH, Suicidal behavior history as the single feature.
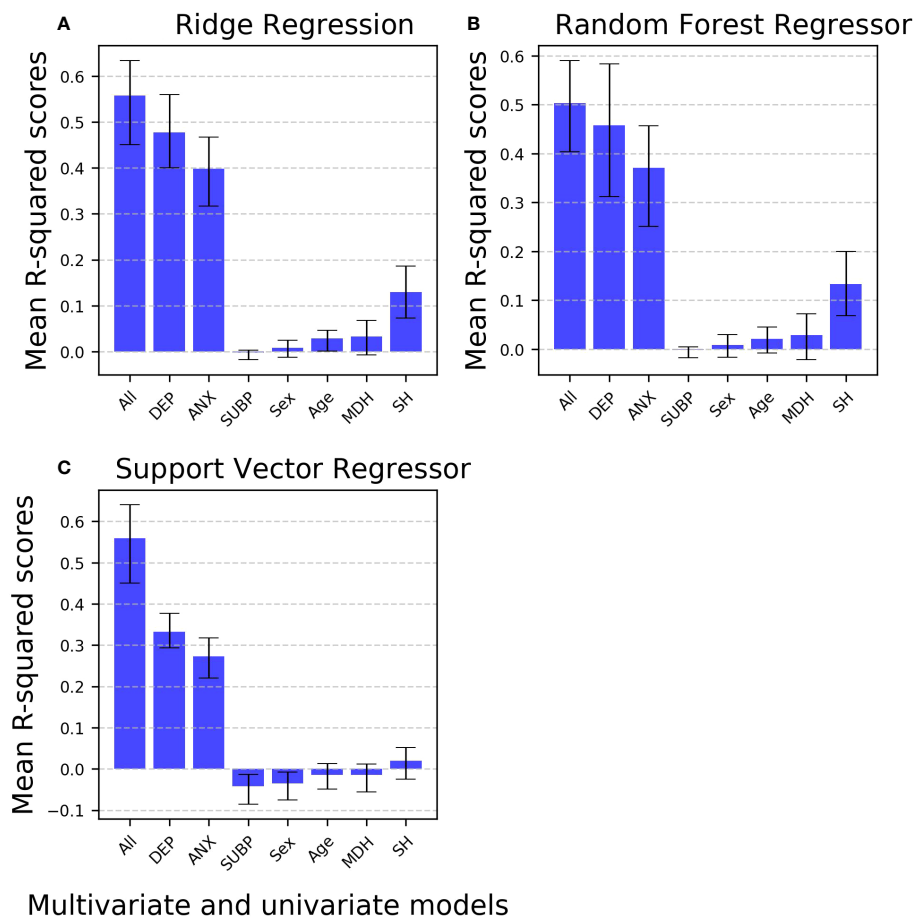
FIGURE 5
Comparative analysis of mean R-squared scores for multivariate versus univariate machine learning models predicting depression in the regression task. Bar plot illustrating the comparison of mean R2 scores for various machine learning models (ordinary least-squares regression, random forest regressor, and support vector regressor) in the regression task of predicting depression. The models evaluated include multivariate models (encompassing all features) and univariate models (each incorporating a single feature). The bar plots include error bars representing 95% confidence intervals and using markers to highlight data points. All: Multivariate model with all features included; Univariate models: DEP, Depression scores measured at time 1 as the single feature; ANX, Anxiety scores as the single feature; SUBP, Quarantine sub-periods as the single feature; Sex, Biological sex as the single feature; Age, Age as the single feature; MDH, Mental disorder history as the single feature; SH, Suicidal behavior history as the single feature.

pandemic encompass baseline depression, female sex, younger age, the presence of mental disorders, and student status, among others (39, 40). In the current study, the application of ML models enabled the exploration of various features' predictive potential, aligning with those examined in the previous study (17). Notably, anxiety, a feature included in this ML-based approach but absent in the previous study, was revealed as one of the most relevant features, alongside depression at T1, for identifying depression presence and quantifying depression severity. The significance of anxiety as a predictor of depression underscores the interconnectedness of these mental health conditions. It aligns with existing literature that emphasizes the comorbidity and shared features between anxiety and depression (41–43) particularly in college students during COVID-19 (44, 45). As for depression at T1, these findings suggest that depressive symptoms persist over time during COVID-19. This aligns with research conducted in college students (46) and general populations in developed countries (47)

and Argentina (48), demonstrating the enduring nature of depression symptoms amid the pandemic.

As for the ML algorithms' performance in this study, both SVM/SVR and LogReg/RR demonstrated the best results in both classification and regression tasks. SVM has consistently been identified as a high-performing algorithm in predicting depression among students (see, e.g., 49–51). However, RF also emerged as the top-performing algorithm in some studies (see, e.g., 50–52), which contrasts with the present findings where it displayed some signs of overfitting and exhibited comparative lower performance. The observed overfitting in the RF algorithm in both classification and regression tasks within this study can be attributed to various factors. Although RFs are renowned for their adaptability and capacity to discern intricate data relationships, this characteristic may lead to overfitting, particularly if the model's complexity surpasses the dataset's requirements. The profusion of decision trees and their interactions may contribute to an overfitted model. Additionally, the

hyperparameters of the RF play pivotal roles. In this study, a set of hyperparameters was explored, including the number of trees, the maximum depth of the tree, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node. For tree building, a bootstrap method involving sampling with replacement was used. Likewise, RFs automatically perform feature selection by considering a subset of features at each split. While advantageous, this process can lead to overfitting if the model emphasizes noise or irrelevant features. However, as demonstrated in this study, an examination of multi-versus univariate models revealed two dominant features (depression at T1 and anxiety), with the remaining features exhibiting limited significance. Just like SVM/SVR and LogReg/RR models, RF multi- and univariate models also captured these key features. As for the discrepancies in algorithm performance between different studies, these may be ascribed to several factors. To mention some of them, the diverse array of features integrated into the models can significantly impact performance when comparing studies. Additionally, variations in study designs, predominantly the widespread use of a cross-sectional design in studies analyzing questionnaire-collected data, further contribute to these differences. Depression assessment methods, relying on self-reported and clinician-administered questionnaires, exhibit inherent limitations (53). However, it is crucial to emphasize that, particularly in developing countries, accessing potent features capable of identifying biomarkers, such as those related to neuroimaging (54), is often impeded by their high cost. Consequently, there is a pressing need to identify reliable and cost-effective predictors to develop risk prediction models capable of enhancing clinical decision-making in depression diagnosis and care. This study represents a first exploratory step toward addressing this goal.

In addition, addressing the challenge of interpretability in ML models is crucial for their effective deployment. This challenge is particularly pronounced in the transparency of decision-making processes within advanced models, often referred to as "black-boxes" (55). Notably, complex models like RF, which leverage numerous decision trees in a non-linear interaction, exacerbate the difficulty in comprehending underlying mechanisms (27). In contrast, linear models such as logistic regression and RR are frequently considered to be easier to interpret due to their transparent mathematical formulations (56). Nevertheless, even the interpretation of multivariate linear models and decision trees can be highly misleading (57–59). Moreover, additional challenges persist in high-dimensional datasets (60). For these reasons, we here resort to univariate analyses to unanimously clarify the aptitude of individual features for prediction (57). In this study, while multivariate ML models encompassing all features demonstrate robust performance in both classification and regression tasks, particularly with LogReg/RR and SVM/SVR, a closer examination reveals a reliance on two key features: depression (at T1) and anxiety. The relevance of these variables in identifying depression aligns with existing literature (3–5, 39).

However, this study's outcomes also emphasize the limited significance of certain variables, encompassing quarantine duration and restrictiveness (referred to as quarantine sub-periods), sex, age, mental disorder history, and suicidal behavior history, not only in identifying depression presence (i.e.,

classification task) but particularly in quantifying depression severity (i.e., regression task). This departure from the established psychological literature, which traditionally links these variables to depression prediction and diagnosis both before and during pandemics (e.g., 6–9, 11, 61), is noteworthy. Nevertheless, it is also worth noting that some recent research, such as a systematic review and meta-analysis, found no significant association between sex and depression among undergraduates (62). Likewise, it is important to consider that much of the established psychological literature is derived from studies analyzing samples in developed countries, often leading to underrepresentation from developing southern countries. Consequently, the role and impact of these variables in predicting depression can vary across different contexts and populations. Overall, the presented findings suggest that, when considered in isolation, these variables may not reliably predict depression within this population. This underscores the need for a more holistic approach to depression prediction, emphasizing the intricate interplay of multiple factors. Additionally, there is a call for further studies to explore additional factors that may be linked with depression prediction, particularly in samples from developing southern countries.

The observed disparities among these current ML-based results, prior MEM analysis, and existing literature bear implications for translational research, as well as for health decision-makers and policymakers striving to improve mental health, particularly to offer scalable and effective evidence-based interventions addressing depression (63). Reliable predictors and prediction models are crucial for effective mental health interventions, particularly those targeting depression prevention (64). While internal validation provides strong support for the performance of the models tested in this study, it is imperative to underscore the necessity for additional external validation and replication studies to solidify the implications of these findings. The prevailing landscape of developing clinical prediction models related to mental disorders calls for more rigorous validation procedures. The current studies in this field often exhibit a deficiency in both external and internal validation (63), emphasizing a critical weakness that needs to be addressed in future research.

Especially in the post-COVID-19 era, there is a growing imperative to leverage advanced technologies like ML algorithms and artificial intelligence to enhance the precision of depression screening among college students. This, in turn, can enable proactive prevention efforts and contribute significantly to improving mental health outcomes (65). However, the applicability of ML algorithms to depression prediction and the development of evidence-based tools also entails ethical challenges. These include optimizing predictions across diverse datasets, ensuring data protection and privacy, assessing feasibility in clinical practice, addressing issues of fairness, accountability, and transparency, as well as identifying and mitigating potential biases of ML models. Many of these challenges have not yet been fully addressed or solved by research studies (66–68). Likewise, as discussed above, the non-trivial interpretability and explainability of ML algorithms is a well-recognized problem. Furthermore, potential future application of these findings, such as widespread

depression screening using ML and automatic diagnosis of disorders, raise additional questions. These include whether clinicians need to confirm diagnoses and eventually communicate them to individuals. Thus, the involvement of these algorithms in diagnosis decision-making has to be accompanied by ethical reflection (67).

This study comprises a valuable exploration of using ML models to predict depression among college students during the COVID-19 quarantine. However, the interpretation of these results is subject to certain limitations. The focus on a specific population and cultural context limits the generalizability and applicability of the developed ML models. Future studies should endeavor to incorporate data from diverse demographics and cultural backgrounds. Additionally, potential biases introduced by the data collection process, which are discussed elsewhere (17), should also be considered. Moreover, the limitations of the ML models, as acknowledged above, should be considered.

In conclusion, this study not only advances our understanding of depression prediction in college students from a developing country using ML algorithms but also highlights the need for further exploration and validation of predictive models in diverse populations. Future research should aim to refine models, consider additional psychological and socio-environmental factors, and enhance external validation to ensure the broad applicability of these findings. For instance, further studies should include larger and more diverse datasets, incorporating the features analyzed in this study as well as other potentially relevant ones. Specifically, future studies may involve additional psychological measurements, such as capturing somatic symptoms of coronavirus-related anxiety, and analyze specific mental disorder diagnoses instead of examining a general category of mental disorder background, as done in this study. Additionally, a comparison with another similar cohort should be conducted as part of future work. Moreover, future research should focus on improving the interpretability of multivariate models. As suggested in this study, comparing multivariate and univariate models may help elucidate the interpretability of the former. This way, the implementation of scalable and effective interventions for diagnosing depression may become more attainable. This initial research study serves to explore the viability of data-driven algorithms in detecting depression among college students. While this study reveals interesting correlations, these findings do not necessarily imply immediate practical consequences. To thoroughly understand and establish causal relationships between anxiety and depression, or to develop evidence-based tools, further investigation is needed. In the future, models similar to those presented in this study may be employed for online pre-screening of students for depression at home.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: The dataset is available in the Open Science Framework (OSF) repository, https://doi.org/10.17605/OSF.IO/2V84N. The reproducible Python code is available in the Open Science Framework (OSF) repository, https://doi.org/10.17605/ OSF.IO/QT2GU, and in the GitHub repository, https://github. com/cecilost/machine-learning-to-predict-depression-in-college-students-during-the-COVID-19-pandemic.

Ethics statement

Author contributions

LL: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. MS: Writing – original draft, Formal analysis. RZ: Writing – original draft, Formal analysis. JG: Writing – review & editing, Supervision, Resources. SH: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition.

Funding

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2024.1376784/full#supplementary-material

References

1. Li Y, Wang A, Wu Y, Han N, Huang H. Impact of the COVID-19 pandemic on the mental health of college students: a systematic review and meta-analysis. *Front Psychol.* (2021) 12:669119. doi: 10.3389/fpsyg.2021.669119

2. Wang C, Wen W, Zhang H, Ni J, Jiang J, Cheng Y, et al. Anxiety, depression, and stress prevalence among college students during the COVID-19 pandemic: a systematic review and meta-analysis. *J Am Coll Health.* (2021) 1:1–8. doi: 10.1080/07448481.2021.1960849

3. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders. 4th ed.* Washington, DC: American Psychiatric Association (2000).

4. Horn PJ, Wuyek LA. Anxiety disorders as a risk factor for subsequent depression. *Int J Psychiatry Clin Pract.* (2010) 14:244–7. doi: 10.3109/13651501.2010.487979

5. Cassady JC, Pierson EE, Starling JM. Predicting student depression with measures of general and academic anxieties. *Front Educ.* (2019) 4:11. doi: 10.3389/feduc.2019.00011

6. Sheldon E, Simmonds-Buckley M, Bone C, Mascarenhas T, Chan N, Wincott M, et al. Prevalence and risk factors for mental health problems in university undergraduate students: a systematic review with meta-analysis. *J Affect Disord.* (2021) 287:282–92. doi: 10.1016/j.jad.2021.03.054

7. Kisch J, Leino EV, Silverman MM. Aspects of suicidal behavior, depression, and treatment in college students: results from the spring 2000 national college health assessment survey. *Suicide Life Threat Behav.* (2005) 35:3–13. doi: 10.1521/suli.35.1.3.59263

8. Piccinelli M, Wilkinson G. Gender differences in depression. Critical review. *Br J Psychiatry.* (2000) 177:486–92. doi: 10.1192/bjp.177.6.486

9. Nolen-Hoeksema S, Aldao A. Gender and age differences in emotion regulation strategies and their relationship to depressive symptoms. *Pers Individ Differ.* (2011) 51:704–8. doi: 10.1016/j.paid.2011.06.012

10. Jorm AF. Sex and age differences in depression: a quantitative synthesis of published research. *Aust N Z J Psychiatry.* (1987) 21:46–53. doi: 10.3109/00048678709160898

11. Brooks S, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *Lancet.* (2020) 395:912–20. doi: 10.1016/S0140-6736(20)30460-8

12. Jacobson NC, Newman MG. Anxiety and depression as bidirectional risk factors for one another: a meta-analysis of longitudinal studies. *Psychol Bull.* (2017) 143:1155–200. doi: 10.1037/bul0000111

13. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction. 2nd ed.* New York, US: Springer Science & Business Media (2009).

14. Murphy KP. *Machine learning: a probabilistic perspective.* Cambridge, MA: MIT press (2012).

15. Kuhn M, Johnson K. *Applied predictive modeling.* New York: Springer (2013).

16. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning: with applications in R.* New York: Springer (2017).

17. López Steinmetz LC, Godoy JC, Fong SB. A longitudinal study on depression and anxiety in college students during the first 106-days of the lengthy Argentinean quarantine for the COVID-19 pandemic. *J Ment Health.* (2023) 32:1030–9. doi: 10.1080/09638237.2021.1952952

18. Iyortsuun NK, Kim SH, Jhon M, Yang HJ, Pant S. A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare (Basel).* (2023) 11:285. doi: 10.3390/healthcare11030285

19. López Steinmetz LC. Dataset and R Code for: A longitudinal study on depression and anxiety in college students during the first 106-days of the lengthy Argentinean quarantine for the COVID-19 pandemic. *OSF Repository.* (2023). doi: 10.17605/OSF.IO/2V84N

20. Brenlla ME, Rodríguez CM. Adaptacion Argentina del Inventario de Depresion de Beck (BDI-II) [Argentinean adaptation of the Beck Depression Inventory (BDI-II)]. In: Beck AT, Steer RA, Brown GK (editors) *Inventario de Depresion de Beck, BDIII [Beck Depression Inventory, BDI-II]* (2006). p. 11–37. Buenos Aires, Paidós.

21. Beck AT, Steer RA, Brown GK. *Manual for the Beck depression inventory II.* (San Antonio: Psychol Corporation). (1996).

22. Spielberger C, Gorsuch RL, Lushene RE. *Manual for the state-trait anxiety inventory. STAI (Form Y), self-evaluation questionnaire.* (Palo Alto, California: Consulting Psychologists Press) (1983).

23. Kendall PC, Hollon SD, Beck AT, Hammen CL, Ingram RE. Issues and recommendations regarding use of the Beck Depression Inventory. *Cogn Ther Res.* (1987) 11:289–99. doi: 10.1007/BF01186280

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* (2011) 12, 2825–30.

25. McKinney W. (2010). Data structures for statistical computing in Python, in: *Proceedings of the 9th Python in Science Conference.* pp. 56–61. doi: 10.25080/Majora-92bf1922-00a

26. van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng.* (2011) 13:22–30. doi: 10.1109/MCSE.2011.37

27. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

28. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi:10.1007/BF00994018

29. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak.* (2019) 19:281. doi: 10.1186/s12911-019-1004-8

30. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn.* (2006) 63:3–42. doi: 10.1007/s10994-006-6226-1

31. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol.*. (2011) 2(3):1–27. doi: 10.1145/1961189.1961199

32. Powers DM. Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol.* (2011) 2:37–63.

33. Davis J, Goadrich M. (2006). The relationship between Precision-Recall and ROC curves, in: *Proceedings of the 23rd international conference on Machine learning.* New York: ACM (Association for Computing Machinery). pp. 233–40.

34. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* (2006) 27:861–74. doi: 10.1016/j.patrec.2005.10.010

35. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* (1950) 78:1–3. doi: 10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

36. Bishop CM. *Pattern recognition and machine learning* Vol. 4. (New York: Springer) (2006).

37. Davison AC, Hinkley DV. *Bootstrap methods and their application.* New York: Cambridge University Press (1997).

38. Cohen J. *Statistical power analysis for the behavioral sciences. 2nd Edition.* New York: Lawrence Erlbaum Associates (1988).

39. Liu Y, Zhang N, Bao G, Huang Y, Ji B, Wu Y, et al. Predictors of depressive symptoms in college students: a systematic review and meta-analysis of cohort studies. *J Affect Disord.* (2019) 244:196–208. doi: 10.1016/j.jad.2018.10.084

40. Xiong J, Lipsitz O, Nasri F, Lui LMW, Gill H, Phan L, et al. Impact of COVID-19 pandemic on mental health in the general population: a systematic review. *J Affect Disord.* (2020) 277:55–64. doi: 10.1016/j.jad.2020.08.001

41. Brown TA, Campbell LA, Lehman CL, Grisham JR, Mancill RB. Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *J Abnorm Psychol.* (2001) 110:585–99. doi: 10.1037//0021-843x.110.4.585

42. Lamers F, van Oppen P, Comijs HC, Smit JH, Spinhoven P, van Balkom AJ, et al. Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the Netherlands Study of Depression and Anxiety (NESDA). *J Clin Psychiatry.* (2011) 72:341–8. doi: 10.4088/JCP.10m06176blu

43. Kalin NH. The critical relationship between anxiety and depression. *Am J Psychiatry.* (2020) 177:365–7. doi: 10.1176/appi.ajp.2020.20030305

44. Bai W, Cai H, Liu S, Chen X, Sha S, Cheung T, et al. Anxiety and depressive symptoms in college students during the late stage of the COVID-19 outbreak: a network approach. *Transl Psychiatry.* (2021) 11:638. doi: 10.1038/s41398-021-01738-4

45. Chang JJ, Ji Y, Li YH, Pan HF, Su PY. Prevalence of anxiety symptom and depressive symptom among college students during COVID-19 pandemic: a meta-analysis. *J Affect Disord.* (2021) 292:242–54. doi: 10.1016/j.jad.2021.05.109

46. Zhao Y, Liang K, Qu D, He Y, Wei X, Chi X. The longitudinal features of depressive symptoms during the COVID-19 pandemic among Chinese college students: a network perspective. *J Youth Adolesc.* (2023) 52:2031–44. doi: 10.1007/s10964-023-01802-w

47. 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. doi: 10.1016/j.lana.2021.100091

48. Canet-Juric L, Andrés ML, Del Valle M, López-Morales H, Poó F, Galli JI, et al. A longitudinal study on the emotional impact cause by the COVID-19 pandemic quarantine on general population. *Front Psychol.* (2020) 11:565688. doi: 10.3389/fpsyg.2020.565688

49. Choudhury AA, Khan MRH, Nahim NZ, Tulon SR, Islam S, Chakrabarty A. (2019). Predicting depression in Bangladeshi undergraduates using machine learning, in: *2019 IEEE Region 10 Symposium (TENSYMP).* New York: IEEE (Institute of Electrical and Electronics Engineers) pp. 789–94. doi: 10.1109/TENSYMP46218.2019.8971369

50. Gil M, Kim SS, Min EJ. Machine learning models for predicting risk of depression in Korean college students: identifying family and individual factors. *Front Public Health.* (2022) 10:1023010. doi: 10.3389/fpubh.2022.1023010

51. Qasrawi R, Vicuna Polo SP, Abu Al-Halawa D, Hallaq S, Abdeen Z. assessment and prediction of depression and anxiety risk factors in schoolchildren: machine

learning techniques performance analysis. *JMIR Form Res*. (2022) 6:e32736. doi: 10.2196/32736

52. Rois R, Ray M, Rahman A, Roy SK. Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms. *J Health Popul Nutr*. (2021) 40:50. doi: 10.1186/s41043-021-00276-5

53. Fried EI, Nesse RM. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *J Affect Disord*. (2015) 172:96–102. doi: 10.1016/j.jad.2014.10.010

54. Wise T, Cleare AJ, Herane A, Young AH, Arnone D. Diagnostic and therapeutic utility of neuroimaging in depression: an overview. *Neuropsychiatr Dis Treat*. (2014) 10:1509–22. doi: 10.2147/NDT.S50156

55. Ribeiro MT, Singh S, Guestrin C. (2016). "Why should I trust you?" Explaining the predictions of any classifier, in: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM pp. 1135–44. doi: 10.1145/2939672.2939778

56. Hosmer Jr. DW, Lemeshow S. *Applied logistic regression. 2nd Ed*. New York: John Wiley & Sons (2013).

57. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*. (2014) 87:96–110. doi: 10.1016/j.neuroimage.2013.10.067

58. Wilming R, Budding C, Müller KR, Haufe S. Scrutinizing XAI using linear ground-truth data with suppressor variables. *Mach Learn*. (2022) 111:1903–23. doi: 10.1007/s10994-022-06167-y

59. Wilming R, Kieslich L, Clark B, Haufe S. (2023). Theoretical behavior of XAI methods in the presence of suppressor variables, in: *Proceedings of the 40th International Conference on Machine Learning*. ML Research Press, PMLR Vol. 202. pp. 37091–107.

60. Belloni A, Chernozhukov V, Hansen C. High-dimensional methods and inference on structural and treatment effects. *J Econ Perspect*. (2014) 28:29–50. doi: 10.1257/jep.28.2.29

61. Solomou I, Constantinidou F. Prevalence and predictors of anxiety and depression symptoms during the COVID-19 pandemic and compliance with precautionary measures: age and sex matter. *Int J Environ Res Public Health*. (2020) 17:4924. doi: 10.3390/ijerph17144924

62. Yang L, Yuan J, Sun H, Zhao Y, Yu J, Li Y. Influencing factors of depressive symptoms among undergraduates: a systematic review and meta-analysis. *PloS One*. (2023) 18:e0279050. doi: 10.1371/journal.pone.0279050

63. Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol Psychiatry*. (2022) 27:2700–8. doi: 10.1038/s41380-022-01528-4

64. Bernardini F, Attademo L, Cleary SD, Luther C, Shim RS, Quartesan R, et al. Risk prediction models in psychiatry: toward a new frontier for the prevention of mental illnesses. *J Clin Psychiatry*. (2017) 78:572–83. doi: 10.4088/JCP.15r10003

65. Liu XQ, Guo YX, Zhang WJ, Gao WJ. Influencing factors, prediction and prevention of depression in college students: a literature review. *World J Psychiatry*. (2022) 12:860–73. doi: 10.5498/wjp.v12.i7.860

66. Lawrie SM, Fletcher-Watson S, Whalley HC, McIntosh AM. Predicting major mental illness: ethical and practical considerations. *BJPsych Open*. (2019) 5:e30. doi: 10.1192/bjo.2019.11

67. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. (2020) 46:205–11. doi: 10.1136/medethics-2019-105586

68. Fusar-Poli P, Manchia M, Koutsouleris N, Leslie D, Woopen C, Calkins ME, et al. Ethical considerations for precision psychiatry: A roadmap for research and clinical practice. *Eur Neuropsychopharmacol*. (2022) 63:17–34. doi: 10.1016/j.euroneuro.2022.08.001