

Deep Neural Network Prediction of Depression Risk Among Young Adults Based on Psychological, Sleep, Lifestyle, and Social Variables

Daniela. Cevallos¹, Samantha. Clarke^{2*}, Victor. Rusu³

¹ Department of Clinical Psychology, Pontifical Catholic University of Ecuador, Quito, Ecuador

² Department of Behavioral Psychology, University of Lethbridge, Lethbridge, Canada

³ Department of Experimental Psychology, Moldova State University, Chişinău, Moldova

* Corresponding author email address: samantha.clarke@uleth.ca

E d i t o r	R e v i e w e r s
Ahmad Amani  Associate Professor, Counseling Department, University of Kurdistan, Sanandaj, Iran a.amani@uok.ac.ir	Reviewer 1: Zahra Yousefi  Assistant Professor, Department of Psychology, Khorasgan Branch, Islamic Azad University, Isfahan, Iran. Email: yousefi1393@khuisf.ac.ir Reviewer 2: Mehdi Rostami  Department of Psychology and Counseling, KMAN Research Institute, Richmond Hill, Ontario, Canada. Email: dr.mrostami@kmanresce.ca

1. Round 1

1.1. Reviewer 1

Reviewer:

The statement that “the study was conducted in accordance with the ethical principles of the Declaration of Helsinki” is insufficient as the sole ethics information for a study involving human participants and mental health screening. Please report the name of the approving Canadian research ethics board, institution, approval/reference number, and date of approval. This is especially important because the PHQ-9 includes an item assessing thoughts of death or self-harm. The manuscript should also specify what safety procedure was activated when participants endorsed this item, including whether crisis resources, emergency contact information, or referral guidance were automatically displayed. The handling of potentially acute suicide-related responses is an important ethical and methodological component of this type of study.

In the Data Collection Tools paragraph, the manuscript states that “participants with a PHQ-9 score of 10 or higher were categorized as being at elevated risk of clinically significant depression.” This threshold is reasonable for screening, but the terminology throughout the article should be carefully standardized so that the model is not interpreted as predicting a diagnosis of major depressive disorder. The outcome is a questionnaire-defined elevated depressive-symptom category, not a clinician-confirmed depressive disorder. I recommend consistently using terminology such as “elevated depressive symptom risk” or

“PHQ-9-defined elevated depression risk” and explicitly distinguishing this endpoint from diagnostic status in the Introduction, Methods, Discussion, Conclusion, abstract, and figure captions.

The Data Collection Tools section lists numerous scales, including the Perceived Stress Scale, GAD-7, Rosenberg Self-Esteem Scale, UCLA Loneliness Scale, resilience measure, Pittsburgh Sleep Quality Index, and Multidimensional Scale of Perceived Social Support, but the manuscript does not provide sufficient psychometric information for the present sample. Please identify the exact version and number of items used for every instrument, the scoring range, interpretation of higher scores, and internal consistency estimates calculated in this dataset. Cronbach’s alpha and, preferably, McDonald’s omega should be reported for multi-item scales. This information is particularly important because measurement error in input variables can affect both neural-network performance and SHAP-based importance rankings.

Table 2 presents extremely large standardized group differences, including Cohen’s $d = 1.97$ for generalized anxiety, $d = 1.76$ for perceived stress, and $d = 1.54$ for social support. These effects are unusually strong and deserve careful verification because the depression outcome and several psychological predictors may share considerable conceptual and item-level content. Please examine whether the GAD-7, perceived stress, and PHQ-9 scores contain overlapping constructs that create predictor-outcome proximity. Reporting zero-order correlations among all continuous predictors and PHQ-9 scores, along with variance inflation factors or a correlation heat map in supplementary material, would help readers assess redundancy. The manuscript should also discuss whether the excellent predictive performance partly reflects concurrent measurement of closely related symptoms rather than prediction from distal risk factors.

Table 3 reports that the deep neural network outperformed logistic regression, random forest, and gradient boosting, but no statistical comparison between models is presented. Statements such as “the deep neural network achieved the strongest overall predictive performance” should be supported by uncertainty estimates and inferential comparison. Please report 95% confidence intervals for ROC-AUC, sensitivity, specificity, F1 score, and preferably PR-AUC. A DeLong test or bootstrap comparison could be used to compare ROC-AUCs, while bootstrap differences could quantify uncertainty for other performance metrics. Without these analyses, differences such as ROC-AUC 0.93 versus 0.91 may not represent meaningful superiority.

Authors uploaded the revised manuscript.

1.2. Reviewer 2

Reviewer:

The predictor construction requires clarification in the paragraph beginning “Sleep-related information included average nightly sleep duration.” Several sleep variables appear potentially redundant with components of the Pittsburgh Sleep Quality Index, including sleep duration, sleep-onset latency, and subjective sleep quality. If both PSQI total score and individual variables derived from overlapping items were simultaneously entered into the neural network, information duplication and multicollinearity may inflate the apparent contribution of the sleep domain. Please specify precisely which variables were derived independently of the PSQI and which were components of it. A sensitivity analysis excluding overlapping sleep indicators would help establish whether the model’s performance and predictor rankings remain stable.

The manuscript states that “categorical variables were converted into machine-readable indicator variables, while continuous variables were retained at their original measurement level and subsequently standardized.” This preprocessing description should be expanded sufficiently to ensure computational reproducibility. Please specify the exact categories used for each categorical variable, reference categories where relevant, treatment of rare categories, whether ordinal variables were one-hot encoded or treated numerically, and whether any variables were winsorized, transformed, or truncated. The final number of model input features after encoding should also be reported. A reader should be able to reconstruct the complete input matrix from the Methods section.

The missing-data strategy described in the sentence “Continuous variables with missing values were imputed using the median calculated exclusively from the training data, whereas categorical variables were imputed using the modal category” appropriately attempts to prevent leakage, but the extent and structure of missingness are not reported. Please provide the

percentage of missing observations for each major predictor, the proportion of participants with any missing predictor data, and the rationale for the >20% missing-data exclusion threshold. It would also strengthen the analysis to compare the current simple imputation procedure with a more robust alternative, such as multiple imputation or model-based imputation, particularly if missingness is nontrivial. At minimum, a sensitivity analysis should demonstrate that predictive performance is not materially dependent on the imputation strategy.

The train-validation-test allocation requires correction or additional explanation. The manuscript reports a 70%/15%/15% split yielding “729 participants, 156 participants, and 157 participants,” which sums correctly to 1,042, but a single random split can produce unstable performance estimates, particularly when the independent test set contains only 157 observations and approximately 56 elevated-risk cases given the reported prevalence. Please report the random seed and exact stratification procedure and consider repeated stratified cross-validation, nested cross-validation, or bootstrap confidence intervals in addition to the holdout test evaluation. Point estimates such as ROC-AUC = 0.93 should be accompanied by 95% confidence intervals to communicate the statistical uncertainty associated with the relatively small test set.

The architecture description, “three fully connected hidden layers containing 128, 64, and 32 neurons,” is not sufficiently justified. Please explain how this architecture was selected and report the hyperparameter search space for number of layers, units, learning rate, dropout, L2 regularization, batch size, and training epochs. The Methods states that hyperparameters were selected according to validation performance, but it does not specify whether this involved manual tuning, grid search, random search, Bayesian optimization, or another procedure. Please also provide the final learning rate, dropout probabilities, L2 coefficient, batch size, maximum epochs, early-stopping patience, and optimizer parameters. These are essential details for reproducibility and for determining whether the neural network comparison with benchmark models was fair.

The manuscript should address potential overparameterization of the neural network relative to the sample size. With only 729 observations in the training set but three dense layers containing 128, 64, and 32 neurons plus numerous encoded predictors, the model may contain a substantial number of trainable parameters. Please report the exact number of trainable parameters and provide training and validation learning curves. Evidence that validation loss converged without substantial divergence from training loss would strengthen the claim that dropout, L2 regularization, and early stopping adequately controlled overfitting. It would also be informative to compare the chosen architecture with a smaller network to determine whether similar test performance can be achieved with materially lower model complexity.

Authors uploaded the revised manuscript.

2. Revised

Editor’s decision after revisions: Accepted.

Editor in Chief’s decision: Accepted.