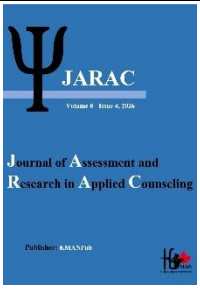
PUBLISHING INSTITUTE
Journal Website

Article history:
Received 18 July 2026
Revised 11 September 2026
Accepted 20 September 2026
Published online 01 October 2026

Journal of Assessment and Research in Applied Counseling

Open-Peer Review



E-ISSN: 3041-8518

Machine-Learning Prediction of Major Depressive Symptoms from Rumination, Sleep Quality, Perceived Stress, and Social Support Using Random Forest Classification

Parsa. Nikoei Rad[1], Pedram. Bahram[1*]

[1] Department of Psychology, University of New Brunswick, Fredericton, NB E3B 5A3, Canada

*** Corresponding author email address**: Pedram.bahram@unb.ca

Editor	Reviewers
José Aparecido Da Silva Full Professor, Department of Psychology, University of Sao Paulo, Ribeirao Preto, Sao Paulo, Brazil jadsilva@ffclrp.usp.br	**Reviewer 1:** Jacinto dos Santos Lino Department of Applied Social Sciences, State University of Southwest Bahia, State of Bahia, Brazil. Email: jacintolino@ufba.br **Reviewer 2:** Cheng-Xiong Lu Mac Anxiety Research Centre, McMaster University, Suite L02, Hamilton, ON, Canada. Email: cheng-xiong-lu@mcmaster.ca

1. Round 1

1.1. Reviewer 1

Reviewer:

The sentence “After removal of respondents who did not provide informed consent, failed to complete the principal study measures, submitted substantially incomplete questionnaires, or provided responses indicative of careless or duplicate participation, the final analytic sample consisted of exactly 684 participants” requires a transparent accounting of exclusions. At present, the exclusion process cannot be reproduced because the number of participants removed for each criterion is not reported and terms such as “substantially incomplete” and “careless participation” are not operationally defined. Please state how many respondents were excluded for lack of consent, incomplete questionnaires, duplicate participation, failed attention checks, implausibly short completion time, patterned responding, or any other quality-control criterion. In addition, define the threshold used for “careless” responding and clarify whether exclusion rules were prespecified before examining outcome status.

The manuscript should justify the sample size of 684 participants in relation to the intended machine-learning analysis. Although Random Forest models can be estimated with this number of observations, the Methods contain no a priori sample-size rationale, learning-curve analysis, events-per-feature consideration, or simulation-based justification. This is especially

relevant because only 261 participants belonged to the positive PHQ-9 class, and the independent test set contained only 52 positive cases. Please explain why the selected sample was considered adequate for model training, hyperparameter tuning, estimation of ROC-AUC, and stable assessment of sensitivity and specificity. A learning curve showing how predictive performance changed as training sample size increased would further strengthen the argument that the model had reached reasonably stable performance.

The sentence "Participants with a total score of 10 or higher were classified as presenting clinically significant major depressive symptoms" should be revised to distinguish screening status from psychiatric diagnosis more consistently throughout the manuscript. A PHQ-9 score ≥10 identifies elevated or clinically significant depressive symptomatology but does not establish major depressive disorder in the absence of a diagnostic interview. The current title, "Machine-Learning Prediction of Major Depressive Symptoms," is relatively appropriate, but several passages refer to the "major depressive symptom group" in ways that could be interpreted as diagnostic classification. Please standardize terminology throughout the Methods, Findings, Discussion, Abstract, tables, and figure so that the positive outcome is consistently described as "clinically significant depressive symptoms based on the PHQ-9 cutoff" rather than as confirmed major depression.

The very large univariate differences in Table 1 raise an important question about the incremental value of the Random Forest algorithm. Perceived stress alone shows a Cohen's d of 1.43, rumination d = 1.17, and poor sleep quality d = 1.30. Under these circumstances, the reported ROC-AUC of .901 may primarily reflect strong linear separation rather than complex nonlinear interactions that specifically require Random Forest modeling. The manuscript should therefore include benchmark models, at minimum logistic regression and preferably penalized logistic regression, and compare their cross-validated and test-set ROC-AUC, accuracy, sensitivity, specificity, F1 score, and calibration against the Random Forest. Without such benchmarking, the claim that machine learning provides added predictive value remains insufficiently demonstrated.

Authors revised and uploaded the document.

1.2. Reviewer 2

Reviewer:

In the Data Collection Tools section, the manuscript describes the PHQ-9, RRS, PSQI, PSS-10, and MSPSS, but reliability estimates for the present Canadian sample are not reported. Instrument validity established in prior studies does not eliminate the need to report internal consistency or other relevant psychometric characteristics in the analyzed sample. Please provide Cronbach's alpha and preferably McDonald's omega for each scale and, where appropriate, for relevant subscales. Because the predictors are subsequently treated as quantitative machine-learning features, measurement reliability directly affects the quality of model inputs. The manuscript should also identify the precise version of the RRS used, because different versions and scoring conventions exist.

The paragraph stating that "Missing item-level responses within otherwise sufficiently completed questionnaires were handled according to the scoring recommendations of the relevant instruments when permissible" is too vague for reproducible analysis. Please provide the exact proportion of missing data for every predictor and the PHQ-9, the permissible number or percentage of missing items for each scale, the procedure used to calculate scores in partially completed measures, and whether any imputation was performed after scale construction. If complete-case analysis was ultimately used for model development, this should be explicitly stated. Importantly, any data-driven imputation procedure must be fitted exclusively within training folds and subsequently applied to validation data to avoid information leakage.

The manuscript states that "the complete analytic dataset of 684 participants was divided using stratified random sampling" into 547 training cases and 137 test cases. Please report the exact random seed and clarify whether all data-cleaning decisions were completed before the split. More importantly, the authors should explain why an 80/20 single holdout strategy was selected rather than nested cross-validation or repeated stratified cross-validation, particularly given that the test sample contains only 137 participants. A single random partition may produce unstable estimates depending on which participants

JARAC Assessment and Research in Applied Counseling
E-ISSN: 3041-8518

enter the test set. Sensitivity analyses using repeated stratified resampling or nested cross-validation would provide a more reliable estimate of expected generalization performance.

The paragraph describing hyperparameter optimization states that "candidate values were evaluated for the number of trees, maximum tree depth, minimum number of observations required for node splitting, minimum observations permitted in terminal nodes, and number of predictors considered at each split," but the actual search space is absent. Reporting only the final configuration of 500 trees, maximum depth 8, minimum split size 10, minimum leaf size 4, and square-root feature sampling is insufficient for reproducibility. Please report every candidate value or distribution considered, specify whether grid search, randomized search, or another optimization strategy was used, identify the software package and version, and clearly state the optimization objective. It should also be clarified whether the decision to prioritize ROC-AUC while simultaneously considering sensitivity and F1 score involved a prespecified decision rule or post hoc judgment.

The sentence "To reduce possible bias resulting from unequal frequencies of the two outcome classes, class weighting was incorporated during model training" needs further methodological justification. The class distribution of 38.2% positive versus 61.8% negative is moderately imbalanced but not extreme, and class weighting can alter probability calibration as well as sensitivity-specificity tradeoffs. Please specify the exact class weights used and how they were calculated. The authors should ideally compare the weighted Random Forest against an otherwise identical unweighted model to demonstrate that weighting materially improved clinically relevant performance rather than merely shifting the classification threshold.

Table 1 reports independent-samples t tests and Cohen's d values for all four predictors, including very large standardized differences, yet the Methods do not describe the assumptions, inferential purpose, or multiple-testing strategy for these comparisons. For example, perceived stress yielded "$t = 18.37$, $p < .001$, $d = 1.43$," which represents extremely strong univariate separation. Please clarify whether Welch's t tests or equal-variance t tests were applied, report confidence intervals for the mean differences or effect sizes, and state whether any adjustment for multiple testing was considered. More importantly, because the principal objective is prediction rather than group-comparison inference, the authors should explain the role of these tests and avoid presenting them as independent evidence that the machine-learning model is valid.

Authors revised and uploaded the document.

2. Revised

Editor's decision after revisions: Accepted.
Editor in Chief's decision: Accepted.

JARAC Assessment and Research in Applied Counseling
E-ISSN: 3041-8518