

Explainable Machine Learning Prediction of Adolescent Depression From Family Functioning, Parenting Practices, Peer Relationships, Cultural Values, and Digital Media Behaviors

Mathieu. Schmit¹, Emily. Chen^{2*}, Tanja. Kralj³

¹ Department of Cognitive Psychology, University of Luxembourg, Esch-sur-Alzette, Luxembourg

² Department of Educational Psychology, University of Victoria, Victoria, Canada

³ Department of Social Psychology, University of Ljubljana, Ljubljana, Slovenia

* Corresponding author email address: echen@uvic.ca

E d i t o r

Salami Mutiu Olagoke
Associate Profesor, Head of
Counseling Psychology and
Guidance Department, Ibn Haldun
University, Turkey
salami@fpm.upsu.edu.my

R e v i e w e r s

Reviewer 1: Sara Nejatifar
Department of Psychology and Education of People with Special Needs, Faculty of
Educational Sciences and Psychology, University of Isfahan, Isfahan, Iran.
Reviewer 2: Zahra Yousefi
Assistant Professor, Department of Psychology, Khorasgan Branch, Islamic Azad
University, Isfahan, Iran.
Email: yousefi1393@khuif.ac.ir

1. Round 1

1.1. Reviewer 1

Reviewer:

The participant paragraph reports an exact analytical sample of 1,000 adolescents but does not provide a sample-size justification. The authors should explain whether the target sample was determined through a formal power analysis, an events-per-variable criterion, simulation-based learning-curve analysis, or practical recruitment considerations. For machine-learning classification, the relevant information includes the number of outcome-positive cases, the number of candidate predictors, anticipated model complexity, and expected precision of the ROC-AUC. A post hoc statement that 286 participants met the depression criterion is insufficient to demonstrate that the sample adequately supported hyperparameter tuning, interaction detection, subgroup analyses, and independent testing.

The eligibility criterion requiring participants to have “resided in Canada for at least one year” needs a stronger conceptual and methodological justification. This restriction may systematically exclude newly arrived immigrant and refugee adolescents, who may experience substantial acculturative stress and may be particularly relevant to the cultural-value component of the

study. The authors should explain why one year was selected, indicate how many potential participants were excluded under this criterion, and discuss how the decision affects the representativeness of the Canadian adolescent population.

The ethical-procedures paragraph states that adolescents reporting severe symptoms or self-harm thoughts were shown “an immediate standardized support message.” Given that the study involved minors and included a suicide-related PHQ-A item, an automated message alone may not constitute an adequate risk-management protocol. The manuscript should specify whether responses were anonymous or confidential, whether immediate human follow-up was possible, what threshold activated intervention, who was responsible for risk assessment, and how parental notification was handled. The authors should also provide the ethics committee name, approval number, approval date, and consent procedures for participants who had reached the locally applicable age of independent consent.

The authors should address the possibility of school-level data leakage. Because adolescents were recruited through clustered school sampling, randomly assigning individual students from the same school to both development and holdout sets may inflate predictive performance by allowing the model to learn school-specific patterns that recur in the test sample. The sentence “The Canadian dataset was divided through stratified random sampling” suggests that the split occurred at the individual level. The authors should repeat the primary analysis using group-stratified splitting by school or district and report whether discrimination and calibration remain stable.

Table 1 presents conventional Pearson correlations among variables with ordinal scales, bounded symptom counts, and potentially skewed distributions, including problematic social-media-use scores ranging from 0 to 9 and nighttime media duration. The authors should state whether distributional assumptions were examined and whether Pearson, Spearman, or polychoric correlations were used. The paragraph reporting that “all principal psychosocial and digital-media variables were significantly associated” should also avoid overemphasizing statistical significance in a sample of 1,000 participants and instead interpret effect sizes, confidence intervals, and practical importance.

Response: Revised and uploaded the new document.

1.2. Reviewer 2

Reviewer:

The Data Collection Tools section identifies the PHQ-A as the primary depression measure but does not provide evidence for its reliability and validity in the specific linguistic and cultural groups represented in the sample. The statement “English and French versions of all instruments were administered according to the participant’s language preference” should be supplemented with citations to validated Canadian English and French versions, evidence of measurement invariance where available, and internal-consistency estimates by language group. Because the model was evaluated across English- and French-speaking adolescents, differences in measurement properties could affect both the outcome classification and subgroup performance.

The decision to classify elevated depression using “a total score of 10 or higher” should be justified in relation to adolescent validation studies and the intended screening context. The authors should report the sensitivity and specificity of this threshold in relevant adolescent populations and clarify whether functional impairment criteria were considered. Since the PHQ-A is a screening measure rather than a diagnostic interview, the manuscript should consistently use terms such as “elevated depressive symptoms” or “screen-positive status” rather than implying that the outcome represents confirmed depressive disorder.

Several measures described in the Methods are insufficiently specified for reproducibility. For example, the manuscript refers to “an adolescent-appropriate version of the Portrait Values Questionnaire,” “standardized questions regarding social exclusion,” and “brief questions concerning family cultural orientation,” but does not provide exact instrument versions, item counts, scoring procedures, source references, or psychometric properties. The authors should distinguish established validated instruments from study-specific items and provide enough detail to permit replication. Study-specific measures should be included in supplementary material and evaluated for content validity, factor structure, and reliability.

The manuscript states that “item-level responses were retained during exploratory model development” for family functioning, while several subscale scores and individual behavioral indicators were also entered as predictors. This approach

creates a risk of excessive dimensionality and redundancy because total scores, subscale scores, and their constituent items may encode overlapping information. The authors should provide a complete predictor inventory, identify whether item-level and aggregate variables were entered simultaneously, and describe how correlated features were handled. A sensitivity analysis comparing item-level and scale-level models would help determine whether predictive performance resulted from substantive information or from redundant measurement representations.

The Data Analysis section states that missing values were imputed using “a multivariate procedure fitted exclusively within the training data,” but the exact algorithm is not named. The authors should specify whether multiple imputation, iterative chained equations, k-nearest-neighbor imputation, missForest, or another method was used, including all parameter settings. They should also report the amount and pattern of missingness by variable, examine whether missingness was associated with depression status, and clarify whether imputation uncertainty was incorporated into performance estimates. The current description is too general for a reproducible machine-learning study.

The manuscript reports that the dataset was divided into an 80% development set and a 20% holdout set and that “repeated stratified cross-validation” was used within the development sample. The authors should specify the number of folds, number of repeats, random seed, optimization criterion, and whether nested cross-validation was used for unbiased hyperparameter selection. Without nested resampling, performance estimates used during model comparison may be optimistic because the same cross-validation process can influence both tuning and model selection. A complete analytical workflow diagram would substantially improve transparency.

The hyperparameter-optimization description is not sufficiently precise. The sentence “Hyperparameters were optimized within the training folds using randomized or Bayesian search procedures” leaves unclear which method was applied to each algorithm, which hyperparameters were tuned, what ranges were searched, how many iterations were performed, and what objective function determined the optimal configuration. These details should be reported in the manuscript or supplementary material. The final hyperparameter values for the selected extreme gradient-boosting model should also be presented so that the analysis can be independently reproduced.

Response: Revised and uploaded the new document.

2. Revised

Editor’s decision: Accepted.

Editor in Chief’s decision: Accepted.