

Early Identification of Families at Risk of Child Maltreatment: A Fairness-Audited Machine-Learning Model Integrating Caregiver Stress, Social Isolation, Economic Hardship, Family History, and Neighborhood Disadvantage

Petar. Ivanov¹, Salma. Abdelnour^{2*}, Mariana. Coutinho³

¹ Department of Developmental Psychology, Sofia University, Sofia, Bulgaria

² Department of Social Psychology, Helwan University, Helwan, Egypt

³ Department of Clinical Psychology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

* Corresponding author email address: salma.abdelnour@helwan.edu.eg

E d i t o r	R e v i e w e r s
Salami Mutiu Olagoke  Associate Profesor, Head of Counseling Psychology and Guidance Department, Ibn Haldun University, Turkey salami@fpm.upsi.edu.my	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 economic-hardship measure is described as a “structured household economic strain questionnaire,” and the family-history measure is described as a form “developed for the study.” These instruments require substantially more methodological detail. The authors should provide the complete item domains, response options, scoring procedure, theoretical basis, expert-review process, pilot testing, and psychometric properties. Because both measures contributed strongly to the final model, insufficiently validated measurement could materially affect model performance and interpretability. The full instruments or a detailed item appendix should be supplied as supplementary material.

Table 2 reports a test AUROC of .852 for extreme gradient boosting and values between .812 and .845 for competing models, yet no statistical comparison between models is presented. The authors should report confidence intervals for all test-set

performance estimates and formally compare the AUROCs using an appropriate paired method, while also comparing AUPRC and Brier scores through bootstrap resampling. With only 28 positive outcomes in the test set, the apparent superiority of extreme gradient boosting may be uncertain. The manuscript should avoid claiming definitive superiority unless the observed differences are shown to be robust.

The Findings section states that the classification threshold of .18 yielded a sensitivity of 82.1% and specificity of 79.2%, but the threshold-selection criterion is insufficiently explicit. The authors should describe the exact objective function or clinical constraint used to select this value in the validation set. For example, it should be stated whether the threshold maximized the F1 score, balanced accuracy, net benefit, or sensitivity subject to a minimum specificity. The consequences of alternative thresholds should be reported so readers can understand how referral burden and missed-case rates would change under different operational priorities.

Response: Revised and uploaded the new document.

1.2. Reviewer 2

Reviewer:

In the neighborhood-disadvantage paragraph, the manuscript states that an index was created from unemployment, education, housing, poverty, service access, transportation, and environmental conditions. The authors should identify the exact administrative data sources, reference years, geographical unit of linkage, and weighting procedure used to construct this index. It should also be clarified whether area-level indicators were standardized across the entire sample or within governorates. Because families were nested within neighborhoods and recruitment sites, the analysis should quantify clustering and assess whether model performance was inflated by geographic overlap between the development and test datasets.

The primary outcome is defined as either a substantiated child-protection report or a “high-probability maltreatment concern” determined by a safeguarding panel. This composite outcome requires stronger justification because substantiation and panel adjudication may represent different evidentiary thresholds. The authors should report how many of the 186 positive outcomes arose from formal substantiation and how many arose solely from panel determination. Inter-rater agreement, the proportion of cases requiring discussion, and the procedures for resolving disagreement should be presented. A sensitivity analysis restricted to formally substantiated cases should be reported in the main Findings section rather than mentioned only as an analytical possibility.

The manuscript states that outcome adjudicators were blinded to baseline predictor scores “except for information required to understand the safeguarding event.” This exception introduces possible incorporation bias. The authors should specify which baseline variables were visible to adjudicators and whether information such as prior child-protection involvement, intimate partner violence, or caregiver mental-health problems was included in the case summaries. If adjudicators had access to variables later used as predictors, the observed association between these variables and the outcome may partly reflect the outcome-definition process rather than independent prospective prediction.

In the Data Analysis section, the manuscript lists penalized logistic regression, elastic-net regression, random forest, gradient-boosted trees, and extreme gradient boosting, but it does not report the final hyperparameters or feature-engineering decisions. The authors should provide a reproducible description of the tuning grid, optimization metric, number of repeated cross-validation cycles, tree depth, learning rate, subsampling parameters, regularization settings, and early-stopping criteria. The final predictor list should also be reported explicitly, including whether total scale scores and their subscales were entered simultaneously. Entering highly overlapping total and domain scores could create redundancy and unstable feature-importance estimates.

The missing-data paragraph states that median, mode, model-based, and multiple-imputation procedures were considered, but the manuscript does not report the actual amount or pattern of missingness. The authors should provide variable-level missing-data percentages, explain which imputation strategy was used for the primary model, and confirm that imputation was performed separately within each cross-validation training fold. The phrase “an explicitly coded missing category when the

absence of information was considered potentially informative” requires justification because missingness indicators can encode site, socioeconomic, or cultural differences and may therefore affect fairness.

In Table 1, all comparisons between outcome-positive and outcome-negative families are reported as $p < .001$. The authors should supplement significance tests with effect sizes and confidence intervals, such as standardized mean differences for continuous variables and risk ratios or odds ratios for categorical variables. Given the large sample size and the exploratory nature of these univariable comparisons, statistical significance alone does not convey clinical importance. The table should also state whether continuous variables met parametric assumptions and whether any correction for multiple testing was considered.

Response: Revised and uploaded the new document.

2. Revised

Editor’s decision: Accepted.

Editor in Chief’s decision: Accepted.