

Deep Neural Network Prediction of Couple Relationship Quality Using Multidimensional Relational and Psychological Features

Nadine. Aeschlimann¹, Amanda. Leblanc^{2*}, Maria Lourdes. Villanueva³

¹ Department of Cognitive Psychology, University of Zurich, Zurich, Switzerland

² Department of Environmental Psychology, University of Guelph, Guelph, Canada

³ Department of Counseling Psychology, Ateneo de Manila University, Quezon City, Philippines

* Corresponding author email address: aleblanc@uoguelph.ca

E d i t o r	R e v i e w e r s
Shahram Vahedi  Professor, Department of Educational Psychology, Faculty of Educational Sciences and Psychology, Tabriz University, Tabriz, Iran vahedi117@yahoo.com	Reviewer 1: Mohammad Hassan Ghanifar  Assistant Professor, Department of Psychology, Birjand Branch, Islamic Azad University, Birjand, Iran. Email: ghanifar@iaubir.ac.ir Reviewer 2: Abotaleb Saadati Shamir  Assistant Professor, Department of Educational Sciences, University of Science and Research, Tehran, Iran. Email: psychology@iau.ac.ir

1. Round 1

1.1. Reviewer 1

Reviewer:

The missing-data procedure in the Data Analysis section is currently too vague. The manuscript states that “small amounts of missing continuous data were imputed using values estimated from the training data” and that categorical values were handled through an “appropriate training-set category-based procedure.” These descriptions are not reproducible. The authors should report the percentage of missingness for each feature, the overall proportion of missing data, the missingness mechanism considered plausible, the precise imputation algorithm used for continuous variables, and the procedure used for categorical variables. They should also specify whether imputation was fitted separately within cross-validation folds during hyperparameter tuning, because preprocessing performed before validation can introduce subtle information leakage.

The train-validation-test partition of 70%, 15%, and 15%, yielding 476, 102, and 102 participants, deserves stronger justification. With only 102 observations in the final test set, the reported $R^2 = .84$ may have considerable sampling uncertainty. The authors should provide uncertainty intervals for all test-set performance metrics using an appropriate resampling procedure, such as bootstrap confidence intervals. Reporting single-point estimates of MAE, RMSE, and R^2 can exaggerate apparent precision, particularly when differences between competing models are relatively modest. Confidence intervals would also

permit readers to assess whether the apparent superiority of the deep neural network over gradient boosting is statistically and practically robust.

The sentence “The final architecture contained four hidden layers with 128, 64, 32, and 16 neurons, respectively” is insufficient without reporting the dimensionality of the encoded input matrix and the resulting number of trainable parameters. Given the modest training sample, the parameter-to-observation ratio is a central concern. The authors should report the number of predictors before and after one-hot encoding, the number of parameters in each network layer, total trainable parameters, dropout probabilities, L2 penalty coefficient, batch-normalization placement, batch size, initial learning rate, learning-rate scheduling if used, maximum epochs, and early-stopping patience. Without these details, neither reproducibility nor the adequacy of regularization can be evaluated.

The hyperparameter-selection paragraph should specify the actual search procedure rather than stating that “candidate network configurations varied in depth, number of neurons per hidden layer, dropout probability, L2 regularization strength, batch size, and learning rate.” The authors should report whether tuning was performed by grid search, random search, Bayesian optimization, manual experimentation, or another strategy and provide the ranges or candidate values considered. More importantly, repeated tuning against a single validation set can overfit model selection to that validation sample. A nested cross-validation strategy or repeated cross-validation within the training data would provide a more rigorous estimate of the architecture and hyperparameters before final testing.

The robustness analysis reporting “test-set R^2 values remained between .81 and .84” across random initializations is useful but insufficiently specified. The authors should state exactly how many independent neural-network training runs were conducted, whether the same train-validation-test partition was retained, and whether random seeds affected only parameter initialization or also minibatch ordering and other stochastic components. Mean performance and standard deviation across repeated runs should be reported rather than only the range. If different data partitions were also tested, this should be distinguished explicitly from repeated weight initialization because these assess different sources of model instability.

The discussion paragraph stating that “The higher accuracy of the neural network compared with both linear and tree-based benchmark models is consistent with the broader theoretical literature showing that relationship functioning reflects interdependent systems...” should be revised to avoid overinterpreting algorithmic superiority as substantive evidence of psychological nonlinearity. Better predictive performance by a neural network can arise from several factors, including flexible functional approximation, regularization differences, optimization choices, or benchmark tuning, and does not by itself prove that the underlying psychological process is nonlinear. The discussion should distinguish between evidence that nonlinear models predict better and evidence concerning the actual form of the underlying relational mechanisms.

Response: Revised and uploaded the manuscript.

1.2. Reviewer 2

Reviewer:

The benchmark comparison in Table 2 requires greater methodological parity across algorithms. The manuscript reports that the neural network outperformed multiple linear regression, ridge regression, random forest regression, and gradient-boosted regression, but it does not describe how the hyperparameters of the tree-based benchmarks were tuned. A heavily tuned neural network compared against default or minimally optimized benchmark models would not constitute a fair comparison. The authors should report the parameter grids and validation procedures for random forest and gradient boosting, including tree depth, number of trees, minimum node size, feature subsampling, learning rate, and other relevant parameters, and confirm that model selection for every algorithm used precisely the same training information without access to the test sample.

The claim in the Findings section that the deep neural network “reduced RMSE by approximately 15.8% relative to the gradient-boosted model” should be supplemented with an inferential comparison of predictive errors. Since all models generated predictions for the same 102 test observations, paired prediction-error comparisons are possible. The authors could use bootstrap resampling of paired absolute or squared errors or another appropriate paired model-comparison procedure to

estimate uncertainty in the performance difference. Without such analysis, the manuscript should avoid interpreting the numerical superiority as conclusively demonstrating that the deep neural network is genuinely better rather than reflecting sampling variation in one test split.

Table 3 and the associated interpretation of SHAP values require more methodological detail. The manuscript states that “SHAP-based model interpretation was used,” but different SHAP implementations can yield different computational approximations for deep networks. The authors should specify the exact SHAP explainer used, the background/reference dataset, the number of background samples, whether values were computed for the entire test set or another subset, and how categorical dummy variables were handled when calculating aggregate feature importance. The statement that certain predictors showed a “predominant direction of association” should also be qualified because SHAP values describe model-specific contributions conditional on the fitted feature distribution and should not be interpreted as causal effects.

The interaction interpretations are scientifically interesting but presently underdocumented. The Findings section states that “the negative effect of destructive conflict was particularly strong when perceived partner responsiveness was low” and that “perceived stress also interacted with emotional regulation,” yet no quantitative interaction statistics, SHAP interaction values, or corresponding figures are provided. If these interaction effects constitute substantive findings, the authors should present formal SHAP interaction estimates or dependence plots and report the magnitude of these interactions. Otherwise, the language should be softened to indicate exploratory model patterns rather than established interaction effects.

Figure 1 should be strengthened methodologically. The observed-versus-predicted plot appropriately includes the identity line and test-set metrics, but the manuscript should additionally consider showing a fitted calibration line with its slope and intercept or reporting these values in the text. A model can exhibit a high R^2 while being systematically miscalibrated, particularly at distributional extremes. The current narrative acknowledges larger error at the lower end of relationship quality, but this observation should be quantified. Reporting calibration intercept, calibration slope, mean signed error, and perhaps performance within outcome quantiles would provide a more rigorous assessment of whether the network systematically compresses extreme predicted values toward the mean.

Response: Revised and uploaded the manuscript.

2. Revised

Editor’s decision after revisions: Accepted.

Editor in Chief’s decision: Accepted.